Emotion Detection in Dialog Systems: Applications, Strategies and Challenges

Felix Burkhardt, Markus van Ballegooy, Klaus-Peter Engelbrecht, Tim Polzehl, Joachim Stegmann
Deutsche Telekom Laboratories
Berlin, Germany
Felix.Burkhardt@telekom.de

Abstract

Emotion plays an important role in human communication and therefore also human machine dialog systems can benefit from affective processing. We present in this paper an overview of our work from the past few years and discuss general considerations, potential applications and experiments that we did with the emotional classification of human machine dialogs. Anger in voice portals as well as problematic dialog situations can be detected to some degree, but the noise in real life data and the issue of unambiguous emotion definition are still challenging. Also, a dialog system reacting emotionally might raise expectations with respect to its intellectual abilities that it can not fulfill.

1. Introduction

No humans are ever non emotional. We speak emotionally, perceive others emotions and communicate emotionally. Despite this, contemporary human machine dialog systems always speak with the same unmoved voice and ignore customer's irony, anger or elation. This is partly due to insufficient technological performance with respect to recognition and simulation, and partly to a gap with respect to the necessary artificial intelligence to support emotional behavior.

In Figure 1 we display some possibilities of emotional processing in human machine interaction [1]. Emotional awareness can be included in several places of an information-processing system [19]:

a) Broadcast: In telecommunication it might be desirable to provide a special channel for emotional communication. A popular example are the so-called '*emoticons*' used in e-mail communication.

b) Recognition: The human emotional expression can be analyzed in different modalities, and this knowledge is used to alter the system reaction.

c) Simulation: Emotional expression can be mimicked by the system in order to enhance a natural interface or to access further channels of communication, like uttering urgent messages in an agitated speech style.

d) Modeling: Internal models of emotional representations can be used to represent user or system states or as models for artificial intelligence, e.g. influence decision making.

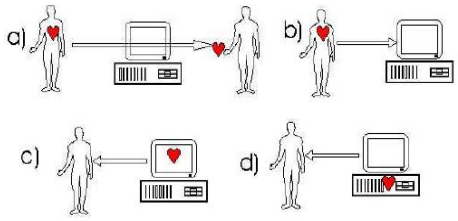

Figure 1. Different uses of emotional processing in computer systems.

In cases a), b) and d), emotional speech analysis can be used to recognize and react on emotional states. Thinking of scenarios, the following lists some ideas:

- Fun applications, e.g. "how enthusiastic do I sound"
- Problematic dialog detection
- Alert systems, i.e. analysis of urgency in speaker's voice
- Adapted dialog and/or persona design
- Believable agents, artificial humans

This list is ordered in an ascending time line when these applications can be expected. Since a technology has to be developed for a long time before it is stable and able to work under pressure, first applications will be about less serious topics like gaming and entertainment or will be adopted by users that have a strong motivation like elderly people being

able to live independently while being monitored by stress detection systems.

The applications further down the list are closely related to the development of artificial intelligence. Because emotions and intelligence are closely mingled [9], great care is needed when computer systems appear to react emotional without the intelligence to meet the user's expectations with respect to intellectual abilities.

We have been working for several years on the development of anger detection in customer care voice portals with mainly the following applications in focus:

1. enable the dialog manager to react on the user's mood, e.g. by transfer to a human agent before he hangs up.
2. evaluate system performance (Faulty systems make the callers angry) and detect problematic dialog situations.
3. measure customer satisfaction (Satisfied customers are less angry).

The progress in this work was reported in [7, 4, 17, 5, 6]. During this period, we experimented with different acoustic feature sets and different classifier algorithms. We investigated the differences between simulated and real life data and conducted user experience studies with respect to emotion aware dialog systems. This paper comprises our experiences so far.

2. How to React?

Asking for the purpose of uttering negative emotions in every day communication, three main functions can be identified: Uttering negative emotions may serve to

1. inform your communication partner about your own emotional status in order to give him a complete comprehension of the information you want to express (your emotional appraisal of the information given).
2. inform your communication partner about the perceived, respectively the desired quality of relation between the communication partner and yourself (e.g. denial or distrust)
3. induce a certain action or behavior of your communication partner that you want him to show (e.g. yielding or flinching).

In functional and professional communication contexts, the occurrence of negative emotions (especially anger) is often not regarded as goal leading and therefore mostly unwanted. Firstly, strong negative emotions may distract the communication from the actual communication topic and therefore lead to a loss of efficiency in communication, secondly the exposure to strong negative emotions itself poses a strong emotional strain on the communicators. For this reason so called conciliation or deescalation strategies are frequently used by professional communicators in order disarm an emotionally charged situation quickly and switch back to the actual communication topic again.

Table 1 shows examples of effective conciliation strategies for negative emotions.

Table 1. *Overview on conciliation strategies.*

Distraction/ change of topic without showing an interest for negative emotions.
Providing information which rebuts the possibly wrong assumptions that led to a negative emotional state.
Feedback / mirroring the perceived emotions in order to show awareness for the speaker's emotional state.
Empathic utterances showing one's sympathy for the speaker's emotions.
Further encouragement of the speaker to express his emotional state.
Pointing out alternatives for being angry or frustrated.
Humor / joking / teasing Reason (appeal to the speaker's senses).

While transferring conciliation strategies from the inter-human communication into dialog behavior of an emotional-aware voice portal we face two major constraints:

1. We might detect the customer's anger but don't know the reason. To be credible, only those strategies can be transferred into man-machine dialogs that are sufficiently generic so that it is not required to make references to the content subject of the negative emotion.
2. As soon as communication between man and machine leaves the immediate task the user actually intends to handle with the dialog system and switches over to abstract topics (e.g. to the reason why the user is angry), the result is a complexity of the dialog being neither calculable nor manageable in a speech dialog system.

Therefore, conciliation strategies in a dialog system must be designed very straightforward and narrow in order for the user not to be encouraged to stray from the task. Despite these constraints we tried to transfer two conciliation strategies into the speech dialog of our emotion aware voice portal. It was set up as a self service application offering information to customers about their telephone bills and mobile phone tariffs.

1. When slight anger was detected we used "conciliation by mirroring". The goal of this strategy is to show

the user that his emotions are recognized but that it is better to continue the task.

2. In dialog situations where strong anger was detected in combination with hints that the dialog will not be completed successfully we used "conciliation by empathy and delegation", i.e. we offered the possibility to be transferred to a human agent. Before the connection was put through, the system verbally showed empathy for the users situation.

We evaluated the user acceptance of affective processing with a user evaluation study. 200 paid test subjects called a voice portal that reacted on user anger with feedback strategies, doing five representative tasks. The users rated the system afterwards in questionnaires. 20% noticed an emotional system reaction and 70% of these judged the system reaction as helpful. After interaction with the prototype system, 70% of the test users judged the use of emotion detection in voice portal systems as reasonable.

3. Experiments

In this section we describe some experiments we conducted with respect to two applications of emotion detection in human machine communication: firstly to detect anger in voice portal services and secondly to predict user satisfaction in dialog recordings.

3.1. Finding Anger in Voice Portals

3.1.1 Data Acquisition

Our main database, originating from service evaluation data of a voice portal service, consists of 21 hours recordings in which customers report problems with their phone connection. The data amounted to 26970 turns in 4683 dialogs, i.e. about 5.8 turns per dialog. Most of the dialogs are very short: more than 50 % contain at most three turns. Most of the turns contain only 2-3 words as is typical for voice command applications, the average audio duration is 2.8 seconds whereas the standard deviation is quite big (2.2) due to the fact that the data contains, besides longer turns, i.e. spelled telephone numbers, "garbage" turns which are not directed to the voice service. The distribution of angry, non angry and garbage turns as well as turns were the labelers were unsure (see following section) is shown in Figure 2. The number of turns which do not contain any analyzable speech is about 10 %. With almost 20% of the turns the listeners were unsure whether anger is revealed in the turn. That leaves about 70 % of utilizable turns, 7 % were classified as angry, which amounts to about 1.8 hours of angry speech.

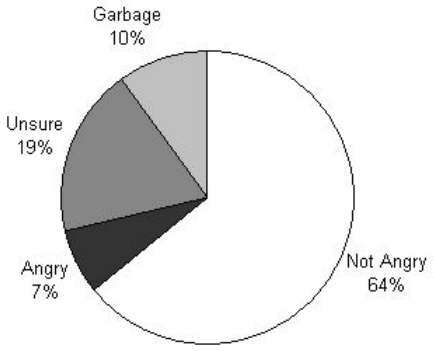

Figure 2. Distribution of anger and garbage turns in the data.

3.1.2 Data Labeling

For simulated anger data, the assignment of the data to the emotional classes is part of the generation process. For real life data though, the question of the data labeling becomes difficult.

Since is no objective measure for anger and the decision of a single person seems to be not stable enough, we label the voice data with several labelers (at least three) of mixed gender. In order to achieve a consistent rating behavior, the labelers got written label instructions and took part in a common session where some examples were discussed. For each turn, the labelers have the choice to assign an anger value between 1 and 5 (1: not angry, 2: not sure, 3: slightly angry, 4: clear anger, 5: clear rage), or mark the turn as "non applicable" (garbage) using a self developed GUI based labeling tool. Garbage turns included a multitude of turns that could not be classified for some reason, e.g. DTMF tones, coughing, baby crying or trucks passing by.

The pairwise agreement between labelers is usually expressed as Cohen's Kappa, which sets the agreement in relation with the chance level, in order to allow for the fact that agreement is less probable with a higher set of choices: $K = \frac{P(A)-P(E)}{P(E)}$, where $P(A)$ is the average time the labelers agreed and $P(E)$ the time they agree by chance level. A Kappa value of 0 means no agreement, values between 0.4 and 0.7 are usually regarded as fair agreement and values above denote excellent agreement. Our labelers reached Kappa values between 0.79 for an early data collection [4] and 0.55 for the data set described here [6]. This indicates (not surprisingly) that agreement depends strongly on the data as well as the labelers. It would probably help to use more than three labelers but that is an economic issue as well.

3.1.3 Results

We reported the results for automatic classification of this data in [6]. To summarize, we reached with our Gaus-

sian Mixture Models and prosodic features based classifier which is integrated in the pilot system mentioned in section 2, a f1 value of 0.412. F1 is the harmonic mean of precision and recall and computes as $\frac{2r_s p_a}{r_a + p_a}$ with r_a and p_a being recall and precision for "anger". When experimenting with Support Vector Machines and additional spectral-related features, we could improve this results up to a f1 value of 0.564.

3.2. Predicting User Dissatisfaction in Voice Dialogs

This section deals with a different application of emotion detection in human machine communication: the automatic measurement of user dissatisfaction from dialog data. With respect to the acoustic classification, we only did a first pilot investigation.

3.2.1 Data

Sometimes data must be simulated, e.g. with specific research questions demanding the control of system or dialog characteristics, such as recognition errors. In such cases, a wizard-of-oz test is conducted in a laboratory. A related feature of anger, when it comes to dialog system evaluation, is user satisfaction, which can usually only be measured by asking the user [10].

In order to investigate the users' judgments of problematic situations in the dialog, we conducted such an experiment. Previous research has shown that users differ in how they judge dialogs (e.g. [11] or [16]). Because of that, all users should be confronted with the same dialogs, such that the distribution of judgments could be measured and groups of users (e.g. low and high affinity to dialog systems) could be compared. We designed five dialog scripts incorporating a number of known dialog problems. Each of the 25 users performed and rated each dialog during the experiment. In addition, in order to track down the user judgments to specific situations in the dialog, we gathered a rating after each dialog turn using a number pad with a rating scale attached. The resulting data set consists of 1027 recorded turns and the corresponding user judgments of the dialog. The distribution of judgments is: 43 "bad", 143 "poor", 229 "fair", 389 "good" and 223 "excellent".

3.2.2 Classification

In a first step, the development of user satisfaction over the dialogs was modeled in dependence of dialog features, such as different types of recognition errors or confirmation strategy. We use a Hidden Markov Model, in which the dialog features are the emissions and the probabilities for the users to change their judgments are represented in the state transitions. The distribution of judgments at each turn can then be predicted using forward recursion. Results are documented in [10].

In contrast to the anger labeled data described in section 3.1, in this case the emotional annotation was done by the user herself (assumed that user satisfaction is related to emotional states). In fact, no one listened to the data and said: "*yes, here the user sounds satisfied*" or: "*based on how the voice sounds here, the user seems to have a problem*".

Whether the satisfaction can be predicted by acoustical measurements is a fascinating question, given that no evidence by manual inspection exists beforehand. Based on the Praat toolkit [3], we extracted the following features from the audio files:

- Pitch: Maximum, minimum, mean and standard deviation.
- Voicing: Relation of voice and unvoiced frames.
- Formants: Mean of first five formants.
- Spectrum: Mean of 12 MFCCs.

For a first preliminary experiment, we merged the user satisfaction annotations into a binary decision: judgments from 1 to 3 were taken as unsatisfied and 4 to 5 as satisfied. We also tried to take labels 3 to 5 as satisfied but this resulted in a much worse automatic classification. Utilizing the WEKA classification toolkit [24], we tried out all available classifiers (with default parametrization) in the standard distribution (version 3.6). About 15 of the (approx. 40) classifiers had an accuracy rate above the trivial classifier that always decides for the majority class (ZeroR). The algorithms that performed best were "classification by regression" [12], a decision tree-like approach[1] with linear regression functions at its leaves, and the logistic classifier [14], which minimizes a matrix distance also based on regression functions. It must be noted for this preliminary investigation, that we used all classifiers in the standard configuration, so a real comparison between the different classifiers is not possible.

3.2.3 Results

Usually results in literature are compared with the "baseline" or "ground truth" meaning the accuracy of a trivial classifier that always decides for the majority class. In the available data, about 64% of the turns were annotated as "satisfied". To test equal distributed data, we excluded 14% of the satisfied turns from our training and test corpus in order to achieve a set of equal distribution. In table 2 we present the results for accuracy and f1 (for dissatisfaction).

[1] We used the default parametrization with an M5P tree algorithm as classifier.

Table 2. *f1 and accuracy for different classifier configurations.*

data set	classifier	f1	accuracy
all data	ZeroR	0	0.645
	Logistic	0.45	0.693
	Class. via regr.	0.457	0.694
equal distr.	ZeroR	0	0.49
	Logistic	0.626	0.638
	Class. via regr.	0.6	0.614

While the overall accuracy gain for all data is only 0.049, the gain in f1 is considerably higher (the f1 for ZeroR is 0 because the dissatisfied class is never predicted) and even higher when using an equally distributed train and test set. It is probably not a good idea to rely on results based on only a subset of the application data because the real situation is not modeled adequately than, but we wanted to raise that issue anyway.

It is quite remarkable, given that the classification of this data is not based on listener impressions but on speaker introspection, the fact that acoustical analysis can help to predict the data. As a next step, the modeling of the time-dynamic of the dialog progression as well as speaker adaption can be investigated.

4. Related Literature

The recognition of emotional states from speech is a research topic with a long history as it is connected with the general research on the acoustical correlates of affective speech. In the following short review we concentrate on studies dealing with telephone data.

Most classification algorithms for the detection of anger are based on a three-step approach: First, a set of acoustic, prosodic, or phonotactic features are calculated from the input speech signal. In a second step, different classification algorithms, e.g. Gaussian Mixture Models, [7] or [15], Artificial Neural Networks, e.g. [8], Support Vector Machines, [21], other vector clustering algorithms like k-nearest neighbor, [15] or linear discriminant analysis, [2], are applied to derive a decision whether the current dialog turn is angry or not angry. Finally, post-processing technologies can be utilized for consideration of time dependencies of subsequent turns or for combination of the results of different classifiers. All these algorithms heavily depend on the availability of suitable acoustic training data that should be derived from the target application.

With respect to the features that are used to classify the speech data, mainly prosodic features, often in conjunction with lexical based and/or dialog related features, were investigated [7], [15], [21], while newer studies also include spectral features derived from Mel Frequency Cepstral Coefficients [2]. Several studies have shown that the inclusion of dialog features can help to enhance the classification accuracy [8] [15]). As we had no dialog context information for our data (other than the order of turns), the inclusion of these features will be future work. The same goes for linguistic features, i.e. the classification based on the words that were spoken.

Regarding the prediction of user dissatisfaction as a variable related to anger, previous approaches mainly considered features derived from the interaction history. Walker et al. [23] measure dissatisfaction with a questionnaire and train a prediction model for the user ratings by applying linear regression, using interaction parameters (e.g. dialog length) as predictors. In other studies [22], unsuccessful dialogs (e.g. because of user hang-up) are predicted from interaction parameters describing the first N dialog turns. Attempts were made to incorporate emotion recognition in such predictions [13]. Direct relations between audio features and user ratings are analyzed in [18], however, unlike in our study, features are averaged across entire dialogs, as judgments were collected only once after the interaction.

5. Conclusion and Outlook

We presented our work with respect to the analysis of emotion related states in spoken dialog contexts. Several applications are envisaged and automatic classification/prediction based on dialog and acoustic features is possible, although complicated by real world constraints like highly noisy data. We found that the comparison of different classifier approaches is worthwhile and a systematic analysis of the reasons for performance differences between classifiers will be done.

The next steps for our research are:

- The design of a unified framework for speaker classification problems for research, development and industrial deployment.
- Explore better strategies to deal with real world noisy data.
- Model dialog dynamics and speaker characteristics.
- Speech-to-text techniques are showing progress now and this opens up interesting possibilities to incorporate ideolectal and semantic features.
- Investigate the possible enhancements related to the next generation wide-band telephony audio data.

Statistical classification is based on data and therefore data acquisition and labeling is vital for emotional analysis. One of our activities with respect to interchangeable formats for emotional data collections is the W3C Emotion Incubator Working Group [20]. Within this focus, we work on an XML based formalism to annotate data emotionally for labeling, classification, simulation and modeling.

6. Acknowledgments

This work was partially funded by the EU FP6-IST project HUMAINE (Human-Machine Interaction Network on Emotion). We further are grateful to the development team of the WEKA toolkit, which greatly simplifies experiments with different classification strategies. The same thanks go to the developers of the Praat tool for acoustical analysis and annotation.

References

[1] A. Batliner, F. Burkhardt, M. van Ballegooy, and E. Nöth. A taxonomy of applications that utilize emotional awareness. In *Proc. of the Fifth Slovenian and First International Language Technologies Conference Ljubljana*, pages 246–250, 2006.

[2] C. Blouin and V. Maffiolo. A study on the automatic detection and characterization of emotion in a voice service context. *Proc. Interspeech, Lisbon*, 2005.

[3] P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 5.1.04), April 2009.

[4] F. Burkhardt, J. Ajmera, R. Englert, J. Stegmann, and W. Burleson. Detecting anger in automated voice portal dialogs. *Proc. ICSLP, Pittsburgh*, 2006.

[5] F. Burkhardt, R. Huber, and J. Stegmann. Advances in anger detection with real life data. In *Proceedings of Elektronische Sprachsignal Verarbeitung (ESSV) 2008*, september 2008.

[6] F. Burkhardt, T. Polzehl, J. Stegmann, F. Metze, and R. Huber. Detecting real life anger. In *Proceedings ICASSP, Taipei; Taiwan*, 4 2009.

[7] F. Burkhardt, M. van Ballegooy, R. Englert, and R. Huber. An emotion-aware voice portal. In *Proc. Electronic Speech Signal Processing ESSP, Prague*, 2005.

[8] Z. Callejas and R. López-Cózar. Influence of contextual information in emotion annotation for spoken dialogue systems. *Speech Commun.*, 50(5):416–433, 2008.

[9] A. R. Damasio. *Descartes' error: emotion, reason, and the human brain*. Avon Books, 1994.

[10] K.-P. Engelbrecht, F. Gödde, F. Hartard, H. Ketabar, and S. Möller. Modeling user satisfaction with hidden markov models. *Proc. SigDial 09*, 2009.

[11] K.-P. Engelbrecht, S. Möller, R. Schleicher, and I. Wechsung. Analysis of paradise models for individual users of a spoken dialog system. In *Proc. of ESSV*, pages 86–93, 2008.

[12] E. Frank, Y. Wang, S. Inglis, G. Holmes, and I. H. Witten. Using model trees for classification. *Machine Learning*, 32(1):63–76, 1998.

[13] O. Herm, A. Schmitt, and J. Liscombe. When calls go wrong: How to detect problematic calls based on log-files and emotions? In *Proc. of the International Conference on Speech and Language Processing (ICSLP)*, Sept. 2008.

[14] S. le Cessie and J. van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201, 1992.

[15] C. M. Lee and S. S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing*, 2005.

[16] R. M. W. M. A. Okun. Toward a judgment model for college satisfaction. *Educational Psychological Review*, 1990.

[17] F. Metze, R. Englert, U. Bub, F. Burkhardt, B. Kaspar, and J. Stegmann. Getting closer: Tailored human-computer speech dialog. *UAIS journal, special issue on Vocal Interaction: Beyond Traditional Automatic Speech Recognition*, 8(2), 2008.

[18] S. Möller, K. P. Engelbrecht, M. Pucher, P. Fröhlich, L. Huo, U. Heute, and F. Oberle. Tide: A testbed for interactive spoken dialogue system evaluation. In *Proc. of the XII International conference Speech and Computer (SPECOM 2007), Moscow, Russia*, 2007.

[19] R. Picard. *Affective computing*. MIT Press, 1997.

[20] M. Schröder, E. Zovato, H. Pirker, C. Peter, and F. Burkhardt. W3c emotion incubator group report. http://www.w3.org/2005/Incubator/emotion/XGR-emotion/, 2008.

[21] I. Shafran and M. Mohri. A comparison of classifiers for detecting emotion from speech. In *Proc. ICASSP, Philadelphia*, 2005.

[22] M. Walker, I. Langkilde, J. Wright, A. Gorin, and D. Litman. Learning to predict problematic situations in a spoken dialogue system: experiments with how may i help you? In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 210–217, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[23] M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. Paradise: a framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, pages 271–280, Morristown, NJ, USA, 1997. Association for Computational Linguistics.

[24] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition edition, 2005.