# Detecting Anger in Automated Voice Portal Dialogs

*F. Burkhardt*, J. Ajmera**, R. Englert**, J. Stegmann*, W. Burleson***

T-Systems Enterprise Services GmbH*, Deutsche Telekom Laboratories**
Berlin, Germany

{felix.burkhardt|joachim.stegmann}@t-systems.com,
{jitendra.ajmera|roman.englert|winslow.burleson}@telekom.de

## Abstract

Anger detection is a topic that is gaining more and more attention with voice portal carriers, as it can be useful for quality measurement and emotion-aware dialog strategies. In the context of a prototype voice portal we describe methods to search for training data, report on the performance of the prosodic classifier under real world conditions and explore the use of dialog information for anger prediction. The results show that, although significantly worse than under laboratory conditions, anger detection still works well above chance level and can be used to enhance real world voice-portal usability.

**Index Terms**: Emotion Recognition, Voice Portal, Speech Classification, Dialogue System.

## 1. Introduction

Anger detection is a topic that is gaining more and more attention with voice portal carriers, as it can be useful for quality measurement and empathic dialog strategies [1, 2]. In the context of customer care voice portals it can be helpful to detect potential problems that arise from a unsatisfactory course of interaction in order to help the customer by either offering the assistance of human operators or trying to react with appropriate dialog strategies. In an industrial real world deployment a set of requirements are to be considered:

- The anger-detection module must integrate into the existing architecture.
- The delay caused by the processing must not obstruct the dialog flow.
- The classification must be based solely on data gained automatically.
- The algorithm has to work on short one-word commands and poor audio condition.
- The procedure must attend to economic issues, e.g. algorithms that are IPR protected must be avoided and manual labor should be restricted.

We developed a concept of an emotion-aware VoiceXML-based voice-portal and described the architecture, the dialog strategies and the underlying prosodic classifier as well as first results based on laboratory data in [3]. This paper now reports on findings that resulted from the integration of the emotion-module in a pilot-phase and our experiments with *real-life data*. The features we analyze are mainly prosodic, i.e. pitch, energy and duration. A linguistic analysis consists simply of swear-word spotting. Although more sophisticated approaches based on machine learning are reported in the literature ([4, 5, 6]), they are limited in their applicability, as ASR systems are generally based on a rule-based grammar allowing only for the recognition of a limited set of words.

The article is structured as follows. We survey related work in Section 2. Section 3 describes the data on which the reported findings are based on. It consists of voice-portal dialogs that date from a pilot portal in the customer care domain. Because the classification of speech into emotional categories is a non-trivial, ambiguous task, Section 4 describes the process of annotating user-turns with emotional labels. The subsequent Section deals with the an exploration of the use of dialog features and reports on outcomes of the acoustic classifier. We conclude with a combined summary and outlook in Section 6.

## 2. Related Work

Although emotional speech has been a research focus for many years, the results often can not be applied to telephone services for several reasons:

- The speech signal is of low bandwidth, often coded by a GSM codec and disturbed by noisy environments so feature extraction is error-prone.
- The dialog-turns are typically short, often consisting of only a very limited set of command words.
- People don't need to follow the politeness rules that apply for human-human dialogs but address machines in an inherently unfriendly "bossy" undertone.
- People tend to over-pronounce, speak slow and loud or even in a "robot-like" manner due to the erroneous belief that this will ease the automatic speech recognition.
- In many applications customers call with some kind of complaint in mind and tend to speak with quite a negative undertone irrespective of problems that may result from the interaction [7].

There are quite a few studies that deal with telephone data, but most of the data differs from the above mentioned points or the challenges that arise from the industrial integration are not fully addressed. Devillers et al [4] e.g. don't use human-computer dialogs but human-human interaction, Yacoub et al [1] use data that was performed by actors, Ang et al [6] is based on the analysis of a simulated customer portal, in Walker et al [8], Lee et al [5] and Liscombe et al [9] some of the features are based on manual annotation. Nonetheless we can learn a lot from these studies to inform our expectation of the performance of different classifiers. Shafran et al [2] studied, beyond gender, age and dialect, the automatic classification of emotional expression on a subset of AT&T's HMIHY (How-may-I-help-you) database. After collapsing originally seven discreet emotion labels to two (negative vs. positive/neutral), a HMM-based classifier resulted

in an error rate of about 31 % based on cepstral features, additional pitch information did not result in a significant increase.
Devillers et al [4] investigated call-center dialogs. As their automatic classifier is based mainly on word analysis, their results are not directly comparable to our work, a fact that is also true for Walker et al [8].
Ang et al [6] analyzed voice portal dialogs, although from an application that was specially designed for research purpose, i.e. the callers did not use the service as part of a real task. A CART-based classifier resulted in a 30% error rate for a ternary decision (annoyed, frustrated, else) based solely on automatically extracted acoustic features. [9] also reported experiments on the HMIHY database and reached an accuracy of about 80%. Beyond prosodic, lexical and dialog act features they modeled the dialog history as a set of *contextual features*. However the recognition rate is enhanced by only 3% if dialog acts and contextual features are taken into account.

Lee et al [5] studied data of an automated flight reservation application and introduced the concept of "*emotionally salient words*" which we could explore if the underlying speech recognition allows for a less restricted recognition. Walker et al [8] also operated on a subset of the HMIHY database, but focused on the classification of whole dialogs after the interaction took place, which would also be interesting to use quality measurement. Petrushin [10] achieved about 77% accuracy with a neural net classifier on voice mails containing faked emotional expression.

## 3. Anybody Angry?

The pilot voice-portal provided during the evaluation time span for 18500 turns in 2300 dialogs, about 22 hours of data. As we didn't have the resources to manually label this amount, we classified the data based on a training set of "faked anger" data gained in an earlier phase of the project. The problem with this approach was that this still resulted in a data-set that was too large, because the provisional anger-detector tended to misclassify the non-angry turns. The recall value of the non-angry turns (see section 5.2) was under 50%.

This is probably caused by the fact that the faked data was performed under good audio conditions and contained clearly distinguishable emotional expression, while the real data is highly distorted and differences between anger and non-anger are often very tiny. It shows once more that training sets from laboratory data are not easily applicable to real world problems.

Thus we used a threshold on the non-anger value (see section 5.2) and selected 2232 turns in 167 dialogs based on a threshold of 0.8. Although there were still many misclassifications, later experiments with higher thresholds showed that this subset indeed contained most of the angry turns. In addition we looked at those turns that were recognized by the speech recognizer containing swear-words, but this didn't bring much gain, as most of them were actually misrecognitions (see section 5.1).

## 4. Decide What's Anger

In order to annotate the data, we formed a labeler group of three listeners and instructed them as follows. "*If you listen to the speech turn, do you think the dialog-manager should interact because the speaker is angry?*". This question could be answered on a five point scale: *1: no, 2: not sure, 3: yes, slightly angry, 4: yes, clear anger, 5: clear rage*. A sixth label ("*NA*") could be used on turns not containing speech (like DTMF tones).

As a way to measure inter-labeler agreement the kappa-statistics $K$ has often been used (e.g. [5]),which sets the percentage of agreement in relation to the agreement expected by chance: $K = \frac{P(A)-P(E)}{P(E)}$, where $P(A)$ is the average time the labelers agreed and $P(E)$ the time they'd have agreed on chance level. A 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. The kappa values computed for the three labelers can be seen in table 1. We compared the differences between three labelers ($K$, $F$ and $W$) and the outcomes of the automatic classification ($M$, see section 5). As

Table 1: *Kappa values comparing three labelers and automatic classification.*

| | $F/W$ | $F/K$ | $W/K$ | $M/F$ | $M/K$ | $M/W$ |
|---|---|---|---|---|---|---|
| Kappa | -0.31 | 0.79 | -0.33 | 0.41 | 0.38 | -0.1 |

can be seen, labeler $W$ differs strongly from labelers $K$ and $F$, which agree to a much higher degree then frequently reported in experiments dealing with emotional speech [5, 11] (about 0.45). The automatic classification results in a similarity comparable to the literature and to the human labelers, an outcome that was also reported in [11]. We conclude that for further labeling the instructions for the group of labelers should be more thorough, e.g. by conducting a collective training session. Although we simply continued by disregarding $W$'s labels (in a majority voting he'd be outnumbered in most cases anyway), we feel that developing a group of "trained labelers" with a large enough number to enable a representative voting is quite essential for further evaluations and enhancements by new training sets.

## 5. Recognition Results

This Section elaborates on the insights we gained by looking at the classification results. It is divided into two parts; the first one discusses classification based on the non-acoustic features, the second one reports on results from the prosodic classifier in dependence of thresholds.

### 5.1. Non-acoustic features

As reported in [8, 5, 9, 7], anger detection can benefit from the analysis of lexical or dialog features.

The only non-prosodic feature that we used on turn-by-turn basis was the detection of swear words by the speech recognizer. The swear-word detection though did not work very well due to the difficulty in finding an adequate swear-word grammar. The grammar that was used as a basis caused many misrecognitions from the speech recognizer (about 99% of the detected swear words were false alarm), while only a fraction of the (rarely used) swear words were actually detected. As outlined in section 6, we will look into other strategies to detect anger on a semantic level in the future.

On a turn-by-turn basis following were investigated to predict Current anger, the current turn state (angry/non-angry):

1. Number of no-matches (NNMs), so far,
2. Number of anger turns, so far,
3. Number of turns, so far,
4. Last turn anger, and
5. Last turn NM (no match)

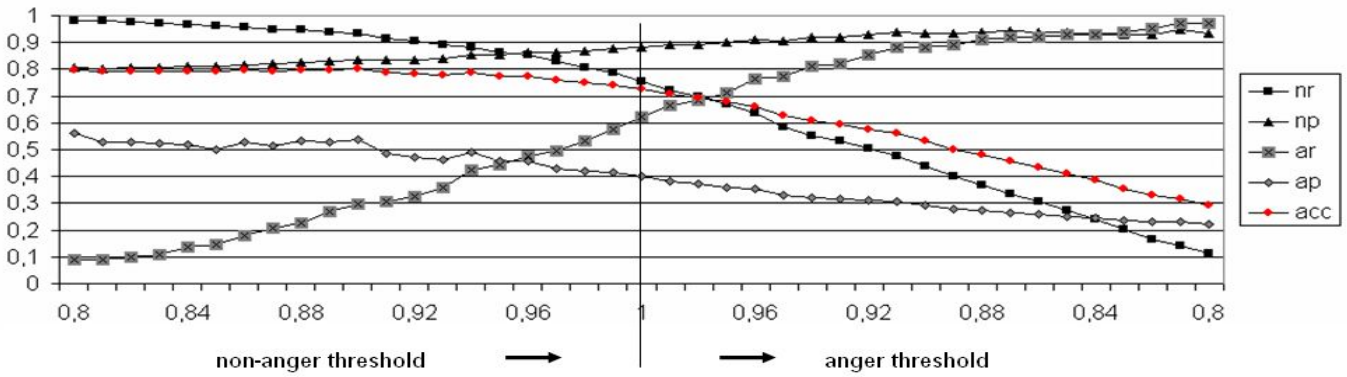


Figure 1: Recall and precision values in dependence of thresholds (see text). n: non-angry, a: angry, r: recall, p: precision, acc: accuracy

Regression analysis showed that each of these measures significantly predicted the *Current anger* (see table 2). These variables taken together also explained a significant proportion of variance in the *current anger*, $R^2 = .106, F(5, 2794) = 65.94, p < .001$. Furthermore, the analysis shows that after accounting for the *Last turn anger* and the *Number of angry turns* the *Last turn NM* has an important impact on the current state of anger, while the *NNMs* and the *Number of turns* do not have as important a role ($\beta = -.05$ and $-.06$, respectively) in the user's *Current anger* state.

It is interesting to note that NNMs inversely predict anger, i.e. there are more NNMs for non-angry dialogs as compared to angry dialogs. We tried to study this in the light of two kind of NMs.

1. The user says something and the recognizer cannot make sense of it, or
2. The recognizer reacts to background noise.

The anger is likely to be caused by the first situation only. Therefore, we tried to use other parameters (e.g. durational) to discriminate between these two kinds of NMs. Our analysis showed that very long and very short turns are very likely to be NMs. However, subsequent analysis of NNMs and their effect on anger, taking into account durational parameters was not conclusive.

The durational parameters (maximum. minimum and average turn lengths) themselves did not correlate well with angry/non-angry state. This can be attributed to the fact that the final state achieved by the user and even more importantly the intended final state of the user differ from dialog to dialog.

Which leaves us with the conclusion that the most influential factor for angry/non-angry states are "misrecognitions". Confidence scores (a measure provided by the recognizer indicating correctness of recognition) combined with durational parameters should provide significant cues about correct and incorrect recognitions and also the two kind of NMs mentioned above. It was difficult to verify this hypothesis in absence of confidence scores in the dialog logs used for this study and we will look into this in future.

### 5.2. Acoustic features

In order to evaluate the performance of our acoustic classifier, we performed several experiments with different training and test sets that differed by the way a unified label was achieved from the different labelers and whether anger was labeled for step 2 ("not sure") onward or step 3 onward (see section 4). The following results are based on a disjunct test and training set based on the decisions of one labeler alone, containing (randomly selected) 10 minutes anger out of 48 in the training and 6.5 minutes anger out of 28 in the test set.

Table 2: *Regression model for predicting Current anger.*

| | | | |
|---|---|---|---|
| Last turn anger | $\beta = -.23$ | $t(2789) = 12.35$ | $p < .001$ |
| Number of turns | $\beta = -.06$ | $t(2789) = -2.87$ | $p < .005$ |
| Last turn NM | $\beta = .12$ | $t(2789) = 6.07$ | $p < .001$ |
| Number of angry turns | $\beta = .15$ | $t(2789) = 7.49$ | $p < .001$ |
| Number of NM | $\beta = -.05$ | $t(2789) = -2.36$ | $p < .02$ |

We report in this section our results from the prosodic classifier which uses pitch, energy and phoneme durations as described in [3]. This classifier gives as results two probability values, one for non-anger ($N$) and one for anger ($A$). These values, coming from the Gaussian mixture models, origin from negative logarithms but are already normalized to each other, i.e. they are not independent but one is always 1 while the other is between 0 and 1.

As reported in [3], we control the trade-off between false acceptance and false rejection by means of thresholds, i.e. if we want to avoid situations where users are accused of being angry although they were not, we disregard $A$ and decide only for anger, if $N$ is lower than a threshold $T_N$.

Classification results are often given in the literature as recall and precision values, the recall of a class meaning the percentage of correctly *identified* cases and the precision the percentage of correctly *predicted* cases for each class. The recall $Rec_A$ of a class $A$ is given by the relation between the correctly identified cases ($C_A$) and the total of existing cases $T_A$:
$Rec_A = \frac{C_A}{T_A}$
In contrast, the precision $Prec_A$ corresponds to the fraction of $C_A$ and the total of predicted cases $P_A$:
$Prec_A = \frac{C_A}{P_A}$

In figure 1 we display recall and precision values for non-anger and anger detection as well as the overall accuracy (the total percentage of correctly identified cases) as a function of the threshold for non-anger (left hand side) and anger (right hand

side). Note that because the values are normalized it makes no sense to display results for both thresholds at the same time, as one value will always be 1.

We can see that the anger recall rises with the increase of the non-angry threshold, as less and less samples get classified as non-angry. If the non-anger threshold reaches its limit and we start to lower the anger threshold, the anger recall keeps on rising until it will reach its maximum of 1, the case where we always decide on anger, irrespective of the classifier's outcome. The rise is monotonous, because the less turns get classified as non-angry the more they get classified as angry. At the same time the recall value for the non-angry turns drops, because more and more of them are misclassified as anger. The non-angry precision rises with the neutral-threshold because the less turns get classified as non-angry the higher the percentage of the correctly identified ones. The angry precision in contrast does not depend on the neutral threshold and therefore the curve is not monotonous in the left hand side.

All these statements get reversed on the right hand side of the figure, that displays recall and precision as a function of the anger-threshold. The fact that the overall accuracy is falling is a result of the far greater number of non-angry turns, i.e. the accuracy is influenced mostly by the non-angry recall.

The optimal threshold to be used depends of course on the application. If one is primarily interested in identifying all the angry turns, he/she might opt for a lower anger-threshold, while a lower threshold for no-anger decision will be advised in order to avoid false anger detection.

We realize that the reported results stand inferior compared to those reported in the literature e.g. [1, 9], but one has to take into account that they were gained on "real world" data and are based exclusively on automatically gained features without manual processing.

## 6. Summary and Outlook

We reported on the first findings resulting from the operation of our emotion-aware voice-portal concept in a pilot portal. Although the acoustic classifier performed significantly worse under real conditions than with "laboratory" data, it still gives results well above chance level. As anger detection from short command-style utterance under low audio quality conditions will always be a problem and the occurrence of false alarms can not be excluded, the resulting dialog strategies will have to be conservative in nature. It was shown once more, that clear anger expression appears rarely in real world data and we still have to work on the data collection and refine the labeling process.

The analysis of lexical features, which consisted of simple swear word spotting in our case, proved to be unsuccessful due to the problem of finding an adequate finite grammar. Future portals based on statistical grammars and statistical learning methods will probably show better results.

Each voice portal design incorporates different strategies to deal with user's disaffection (e.g. often the transfer to a human agent is not possible due to financial limitations). Emotional expression that is to be detected depends on the conciliation strategy; the anger concepts have to be defined in cooperation with the dialog designers. Thus, we have to work on a theoretic framework that serves as a basis to label the data in a way that can be utilized for different strategies to calm down the user.

The huge difference in the inter-labeler agreement showed that the concepts were not clear and we have to work on a consistent way of labeling.

Reusing data from different voice-portal applications and working with a set of standardized dialog tasks as well as a standard way of emotional labeling would be desirable and will be conducted as part of our work in standard bodies and EU-projects.

Another issue concerns the work on more advanced methods to detect anger semantically, e.g. using machine-learning methods on data like reported in [5]. Inspired by a similar direction is the idea to combine acoustic and (enhanced) semantic detection with dialog features (like number of no-match events or number of turns) and even application specific features (e.g. current task). A concept that deals with the fusion of the different recognizers still has to be developed.

## 7. Acknowledgments

This work was partially funded by the EU NoE HUMAINE.

## 8. References

[1] S. Yacoub, S. Simske, X. Lin, and J. Burns, "Recognition of emotions in interactive voice response systems," in *Eurospeech 2003 Proc*, 2003.

[2] M. R. u. M. M. I. Shafran, "Voice signatures," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2003.

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

[4] L. Devillers, L. Lamel, and I. Vasilescu, "Annotation and detection of emotion in a task-oriented human-human dialog corpus," in *Proc. ISLE Workshop on dialogue tagging*, 2002.

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

[6] J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke, "Prosody-based automatic detection of annoyance and frustration in human-computer dialog," in *Proc. ICSLP*, 2002.

[7] V. Botherel and V. Maffiolo, "Regulation of emotional attitudes for a better interaction: Field study in call centres," in *Proc. 20th International Symposium on Human Factors in Telecommunication*, 2006.

[8] M. Walker, I. Langkilde-Geary, H. Wright, J. Wright, and A. Gorin, "Automatically training a problematic dialogue predictor for a spoken dialogue system," *Journal of Artificial Intelligence Research, 16: 293-319.*, 2002.

[9] J. Liscombe, G. Riccardi, and D. Hakkani-Tür, "Using context to improve emotion detection in spoken dialog systems," *Proc. of Interspeech 05, Lisbon*, 2005.

[10] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *Artificial. Neu. Net. In Engr. (ANNIE)*, 1999.

[11] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, "Of all things the measure is man - classification of emotions and inter-labeler consistency," in *Proc. of ICASSP 2005*, 2005.