Modeling A-Priori Likelihoods for Angry User Turns with Hidden Markov Models

Alexander Schmitt[1] *and Tim Polzehl*[2] *and Wolfgang Minker*[1]

[1]Institute for Information Technology, University of Ulm, Germany
[2]Quality and Usability Lab TU Berlin / Deutsche Telekom Laboratories, 10587 Berlin, Germany
alexander.schmitt@uni-ulm.de, tim.polzehl@telekom.de, wolfgang.minker@uni-ulm.de

Abstract

Current studies dealing with the detection of angry users in automated telephone-based speech applications take acoustic and sometimes linguistic information into account in order to classify the emotional state of the caller in single user turns. Angry user turns, however, don't appear from nowhere and the likelihood of observing another angry turn rises substantially when anger has already been observed previously in the discourse. In this contribution we examine the context of angry user turns in two different telephone corpora. We then introduce Hidden Markov Models (HMM) as classifiers modeling the temporal aspects of anger across single turns. As additional information source, the HMMs improve our acoustic classifier serving as baseline substantially. Performance gains of 1-4 % can be reported by performing late fusion of the acoustic classifier and the HMMs.

1. Introduction

With a further increase in complexity of telephone-based automated speech services, more advanced techniques for the robust detection of problematic dialogue situations becomes necessary. A typical procedure in modern telephone-based customer care is to start with automation, i.e. with an Interactive Voice Response (IVR) system, and escalate to a human operator when problems occur. In the vast majority of today's IVR systems, a problematic situation is considered to occur when frequent *no-input* and *no-match* events are observed. The former means that the user did not reply to a question within a certain time slot, the latter accounts for the fact that the Automatic Speech Recognition (ASR) did not get what the user meant because the user utterance did not match the active grammars. Several subsequent no-match and no-input events are then the reason for escalation.

Annoyance is considered as another symptom for a call gone bad and both carriers and the research community are striving for a robust solution to detect angry users online. Offline, anger detection can also be of valuable help for system designers: flaws in dialogue design can be detected in logged calls allowing for an adaption and improvement of the dialogue system at critical points.

Speech-based emotion research regarding telephone applications has been increasingly discussed in the speech community. While in early studies acted corpora were used, such as in [1], training and testing data in later studies has been more and more based on real-life data [2]. Most studies are limited to acoustic/prosodic features that have been extracted from the audio data. Linguistic information was additionally exploited in [3] resulting in a 45.7% accuracy improvement compared to using only acoustic features.

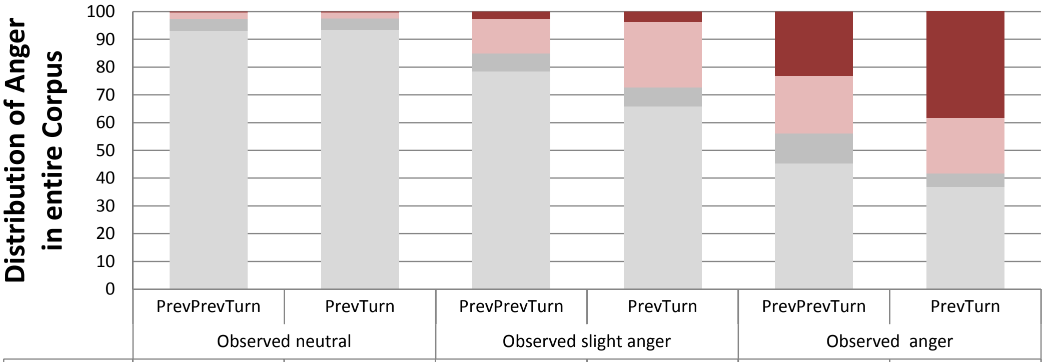
In previous work we have shown, that anger detection can be enhanced with contextual features, i.e. non-acoustic information, delivering a performance gain of approximately 2% [4]. Additionally to acoustic features such as pitch, loudness or jitter, the classifier is enhanced with features originating from the dialogue context. In [5] the lexical and prosodic features were additionally enriched with dialogue act features leading to an increase in accuracy of 2.3%.

The remainder of this paper is organized as follows: first, we present an analysis of angry user turns within the dialogue context in Section 2. Then we provide details about the two employed corpora serving as test and training material for our evaluation in Section 3. In Section 4 we present the acoustic classifier serving as baseline followed by an architecture description of the HMMs in Section 5. The performance of the models is evaluated in Section 6 and the overall system performance after fusion with the baseline classifier is described in Sections 7 and 8.

2. The Previous Emotional State

An important factor has been neglected so far when dealing with IVR corpora. Users don't get angry out of the blue sky. A certain "history of anger" can be observed as depicted in Figure 1. The data originates from a US-American IVR application being able to resolve Internet-related problems jointly with the caller. It can be seen in the first two bars of the chart that it is very unlikely that a user who is non-angry in the current dialogue turn was slightly angry or strong angry in the two previous turns. By contrast it is interesting to analyze the anger history of turns where the caller showed slight or even strong anger. For example, if we observe that the user is slightly angry in the 5th dialogue turn, the likelihoods that the user has already been slightly angry in the two previous turns, i.e. the 4th and the 3rd turn are 23.69% and 12.5% respectively. In other words, when a user is angry in the current turn we have a very high probability that he will be also angry in the next turn.

The role of garbage turns is also striking. In this context, garbage turns are turns where the caller did cross- and off-talk or when other background noise and non-speech events such as coughing or sneezing have been recorded. When comparing non-angry and angry turns it is interesting to see that angry turns were more frequently preceded by garbage turns. The likelihood of having had a garbage turn two steps prior to a hot anger turn is even 10.77% (second last bar). Such a garbage turn inevitably leads to an ASR error which then causes the system to re-prompt the question. Our assumption is that users are provoked by frequent ASR errors and become angry. A closer



	Observed neutral		Observed slight anger		Observed anger	
	PrevPrevTurn	PrevTurn	PrevPrevTurn	PrevTurn	PrevPrevTurn	PrevTurn
■ angry	0,19	0,19	2,56	3,69	23,08	38,30
■ slightly angry	2,51	2,31	12,50	23,69	20,77	19,86
■ garbage	4,26	4,01	6,48	6,81	10,77	4,96
■ non-angry	93,04	93,49	78,46	65,82	45,38	36,88

Figure 1: Emotional state of the caller in the two previous dialogue turns of the currently considered turn. E.g. when we observe anger in the current turn (see last bar), the likelihood that the caller has already been angry in the previous turn is 38%.

look to the relationship between no-match events and anger confirms this estimation. We considered a window of 3 previous turns prior to a current turn. In 5.21% of all non-angry turns we counted at least one no-match event within this window, whereas 11.8% of all slight or strong anger turns were preceded by at least one no-match event within this window.

3. Corpora

Obviously, there exists a relationship between a currently observed angry user turn and the emotional state of the caller in previous turns which, to our best knowledge, has not been exploited in previous work. In this study we are modeling the likelihood that a current turn is an angry one with information about the emotional state of the caller in the preceding turns.

To provide a more reliable statement, we employ training material from two different Interactive Voice Response corpora containing real-life emotions serving also as training data for our acoustic anger detection system as described in [6]. The first corpus, which has also been employed for the introducing analysis, is originating from a US-American Internet troubleshooter and contains English narrow-band short utterances. It contains 1,911 calls comprising 22,724 utterances. Three labelers divided the corpus into *angry*, *annoyed* and *non-angry* utterances (Cohen's $\kappa = 0.70$ [7]). The final label was defined based on majority voting resulting in 90.2% non-angry, 5.1% garbage, 3.4% annoyed and 0.7% angry utterances. Merely 0.6% of the samples in the corpus were sorted out since all three raters had different opinions. The raters were asked to label "garbage" when the utterance is incomprehensible or consists of non-speech events. While the number of angry and annoyed utterances seems very low, 429 calls (i.e. 22.4%) contained annoyed or angry utterances.

The German database comprises 683 dialogues with 4263 user turns from an Interactive Voice Response (IVR) portal. For each turn, 3 labelers assigned one of the following labels: "not angry", "not sure", "slightly angry", "clear anger", "clear rage" or marked the turns as "non applicable" when encountering garbage. The labels were mapped in two classes, anger and non-anger, by clustering according to a threshold over the average of all voters' labels as described in [2]. Following Cohen's Kappa (κ) we obtain a value of 0.52 which corresponds to fair inter labeler agreement. Finally, our setup contained 1761 angry turns and 2502 non-angry turns which roughly corresponds to a 40/60 split of anger/non-anger distribution.

In order to create similar conditions in both corpora, we limit the English corpus to a 40/60 split by removing dialogues where users are not angry and collapsed the classes "slight anger" and "hot anger" to "anger" resulting in 850 angry and 1233 non-angry turns. By that both corpora contain binary classes (angry vs. non-angry) and the same distribution of 40/60 of angry and non-angry turns.

4. Baseline Classifier

As baseline for our experiments we use our current acoustic anger classification system for Interactive Voice Response systems being able to predict angry user turns [6].

It consists of a prosodic and an acoustic feature definition unit calculating a broad variety of information about vocal expression patterns such as pitch, loudness, intensity, Mel Frequency Cepstral Coefficients (MFCC), formants, harmonics-to-noise ratio etc. A statistical unit derives means, moments of first to fourth order, extrema and ranges from the respective contours in the first place. Special statistics are then applied to certain descriptors. Pitch, loudness and intensity are further processed by a Discrete Cosine Transform (DCT) in order to model the behavior over time. In order to exploit the temporal behavior at a certain point in time we append first and second order derivatives to the pitch, loudness and intensity contours and calculate statistics on them alike. The complete feature space comprises 1450 features per user utterance which are then subject to a feature selection unit based on Information Gain Ratio ranking. The final anger detection unit is based on a discriminative classifier, a Support Vector Machine (SVM) with linear kernel. When performing 10-fold cross validation with the data described in Section 3 it yields f1 scores of 73.2% on the train set of the German DB and 77.3% on the English DB.

5. HMMs for Modeling Anger History

Our anger detection system is based on information from a single user utterance not taking into account the context a user utterance originates from. Based on the findings in Section 2 we enrich our system with a subunit that models the probability of observing an angry turn when having previously seen angry turns. Therefore, we consider the immediate context of the utterance we want to classify. Figure 2 outlines the probabilities regarding the emotional state of the previous two user turns. It depicts both, the English and the German corpus including the final distribution of anger and non-anger in our evaluation material.

For each of the two emotional states anger and non-anger we trained a Hidden Markov Model (HMM) containing 2 hidden states and discrete observation symbols with training data from the two corpora. The first HMM was trained with observation sequences leading to an angry turn, in other words it contained the hand labels of the 3 preceding turns, for example:

$$ANA$$

$$NAN$$

$$NAA$$

etc.

The second HMM has been trained with observation sequences leading to a non-angry turn.

$$NNN$$

$$NAN$$

$$NNN$$

etc.

Turns where only two or one previous turn was available have been considered separately. In this case the HMMs have been trained with observation sequences of length 2 and 1 respectively. Figure 3 shows the architecture of the HMMs.

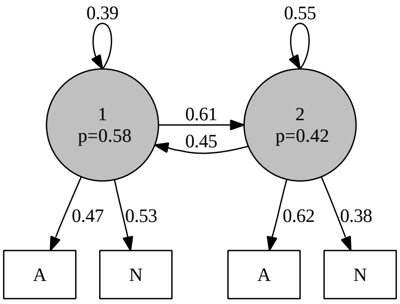
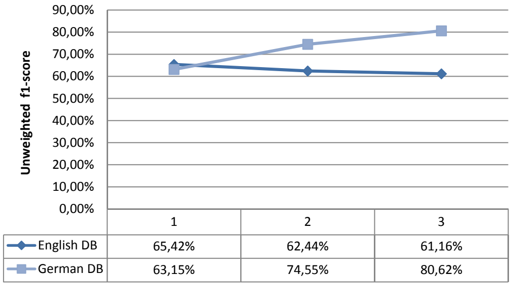


Figure 3: HMM architecture including priors, transition and emission probabilities after training with data from the English DB. The depicted HMM contains the probabilities for observing an angry turn.

6. Decoding

For estimating the likelihood that the *currently* observed turn is angry or non-angry, we let both HMMs decode the observation sequence of the *previous* 3 turns, where available. In turns where only 2 or 1 previous turns are available we proceed analogously. We are aware of the fact that we use hand labels, and by that assume that the emotion recognizer would deliver 100% correct results. A slight decrease in performance is to be expected when using classifier predictions instead of manually annotated labels. Both HMMs generate a likelihood score indicating the probability that the observed sequence has been generated by this specific HMM. Evaluation of the models is performed with 10-fold cross validation respectively with the English and German corpus. In order to deal with the unbalanced class distribution in both databases we calculate f1 measures and use it as evaluation criterion. The f1 measurement is defined as the arithmetic mean of F-measures from all classes. The F-measure accounts for the harmonic mean of both precision and recall of a given class. Note that an accuracy measurement would allow for false bias since it follows the majority class to a greater extent than other classes.

Results are depicted in Figure 4. The chart visualizes the f1-score of the models when fed with 1,2 and 3 previously observed turns.



	1	2	3
English DB	65,42%	62,44%	61,16%
German DB	63,15%	74,55%	80,62%

Figure 4: Avg. unweighted f1-score performance of the HMMs when tested with observation sequences that are available so far in the dialogue.

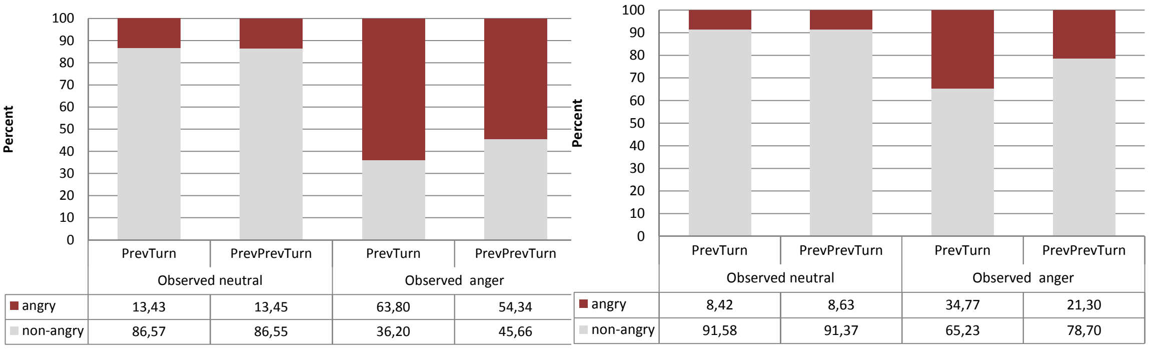
With increasing number of available turns the performance is increased considerably in the German DB while there is a slight decrease visible in the English DB. It is not surprising that the models perform better on the German DB: the likelihood of having another angry turn after having seen an angry turn is much higher in this corpus.

In order to evaluate the performance of the model in a realistic scenario, we use the full English corpus as test set after training with the 40/60 split corpus. The resulting avg. f1 values are 62.4%, 68.0% and 57.8% for 1, 2 and 3 turns respectively. A possible explanation of the comparatively bad performance of the HMM tested with 3 turns can be due to the fact that the window of 3 contains already irrelevant data.

7. Fusion

The initial idea of modeling anger history was to improve the classification of a single turn. Our new system combines the information from the acoustic classifier with information from the HMMs. As meta-classifier we apply again a SVM with linear kernel that is trained with feature vectors containing 3 attributes: the prediction of the acoustic subsystem, 0 for non-angry or 1 for angry, as well as the two probabilities of the two HMMs each generating a likelihood score that the turn under consideration is an angry or non-angry one given the 1-3 previous turns in the dialogue.

Our current acoustic setup reaches a performance of 77.3% f1 (English DB) and 73.4% (German DB) when 10-fold speaker independent cross validation is applied. When incorporating the meta-classifier including anger history the average f1-scores are increased to 78.5 % and 77.5% respectively.



(a) German Corpus

(b) English Corpus

Figure 2: Anger state in the two previous user turns of the currently considered one: when observing an angry user turn the likelihood of observing another one is 63.8% in the German DB (a) and 34.7% in the English (b) DB.

8. Creating Deployment Conditions

Until now, we trained the HMMs with manually annotated labels which are not available in real-life conditions. During testing, decoding the HMMs takes places with prediction labels rather than manual labels. In a first step, we thus apply the acoustic subsystem on the training material in 10-fold cross validation. The HMMs remain trained on the hand labels; however, testing is now performed with the prediction labels of the acoustic subsystem.

Although we are using prediction labels a final combination, again with a SVM meta-classifying both subsystems, yields a considerable performance (see Figure 5). Only a slight decrease in performance on the English DB can be observed when using the labels from the acoustic subsystem instead of the hand labels. The German DB, however, even performs slightly better.

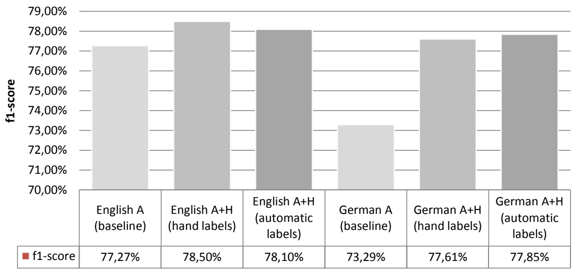


Figure 5: f1 performance of the acoustic baseline system (A), the performance of the enhanced systems (A+H) with hand labels and with automatic prediction labels

9. Conclusion and Discussion

Taking the previous emotional state of the caller into account when classifying angry user turns in IVRs can improve the performance in anger detection tasks. In the presented approach that was applied on two different IVR corpora to show the generalizability we modeled the emotional behavior of the user with Hidden Markov Models and trained them with up to 3 prediction results of our acoustic subsystem. The HMMs generate likelihood scores that the next turn will be an angry one based on the previous emotional state which are then used as additional information source for a meta-classifier. The final system consists of a Support Vector Machine being trained on the prediction of the acoustic subsystem and the anger history subsystem (HMMs). The fact that it makes virtually no difference whether decoding of the HMMs is carried out with hand labels of human annotators or predictions of the acoustic subsystem can be attributed to the robustness of HMMs. Single previous misclassifications of the acoustic subsystem are smoothed by the HMM. The late-fusion of both subsystems yields performance gains of 0.83% (English DB) and 4.56% (German DB). In future work we will incorporate further knowledge sources such as contextual and linguistic information that will be subject to late fusion.

10. Acknowledgements

This research has been carried out in the context of the Transregional Collaborative Research Centre SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” funded by the German Research Foundation (DFG).

11. References

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

[2] F. Burkhardt, T. Polzehl, J. Stegmann, F. Metze, and R. Huber, “Detecting real life anger,” in *Proc. of ICASSP*, Taipeh, Apr. 2009.

[3] C. M. Lee, S. S. Narayanan, and R. Pieraccini, “Combining acoustic and language information for emotion,” *Proc. of Eurospeech, Denver*, 2002.

[4] A. Schmitt, T. Heinroth, and J. Liscombe, “On nomatchs, noinputs and bargeins: Do non-acoustic features support anger detection?” in *Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue, SigDial Conference 2009*. London (UK): Association for Computational Linguistics, Sep. 2009.

[5] J. Liscombe, G. Riccardi, and D. Hakkani-Tur, “Using context to improve emotion detection in spoken dialog systems,” in *Proc. of Interspeech, Lisboa*, 2005.

[6] T. Polzehl, A. Schmitt, and F. Metze, “Comparing features for acoustic anger classification in german and english ivr portals,” in *First International Workshop on Spoken Dialogue Systems (IWSDS)*, Kloster Irsee (Germany), Dec. 2009.

[7] 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 *Proceedings of ICASSP 2005, Philadelphia*, vol. 1, 2005, pp. 317–320.