ADVANCES IN ANGER DETECTION WITH REAL LIFE DATA

Felix Burkhardt, Richard Huber**, Joachim Stegmann**

**T-Systems Enterprise Services GmbH, **Sympalog Voice Solutions GmbH*

Felix.Burkhardt@t-systems.com

Abstract: We report on the progress with respect to an emotion-aware voice portal concerning several directions. A comprehensive new data collection has been carried out and gives new insight on the nature of real life data. The labeling process and the structure of the data will be discussed. Experiments with the anger detector on that data indicate that the acoustic features based on voicing don't play an important role for this data. It shows once again, that classification problems are highly data specific as this data contains a lot of noise. Also using Support Vector Machines instead of Gaussian Mixture Models as classification algorithm turned out to give better results with respect to anger recall. As a consequence for these findings we plan for a new classifier that's more flexible with respect to choice of features and algorithm in order to adapt for different data.

1 Introduction

The automation of business processes in call centers based on Interactive-Voice-Response (IVR) systems has been introduced in many companies for cost reduction purposes. In state-of-the-art IVR systems automation based on automatic speech recognition (ASR) is often used for customer self services like e.g. checking of account balances, product information, or tariff changes. In this context, it can be helpful to detect potential problems that arise from an unsatisfactory course of interaction with the system to help the customer by either offering the assistance of a human operator or trying to react with appropriate dialog strategies. An important decision criterion for such changes in the call flow is the automatic detection of anger from the caller's voice that can be monitored during the entire dialog. A respective technology module can be introduced in the IVR system running in parallel to the ASR component.

In principal, most classification algorithms for the detection of anger are based on a three-step approach [11]: First, a set of acoustic, prosodic, or phonotactic features are calculated from the input speech signal. In a second step, classification algorithms like e.g. Gaussian Mixture Models (GMMs), Artificial Neural Networks (ANNs) or Support Vector Machines (SVMs) 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. The labeling process is cumbersome because the perception of anger is a subjective impression of the listener within the specific application context. Therefore, it is important to work with multiple labelers and find out where their ratings agree.

This paper is organized as follows: After a short overview of the related work we explain the process of data acquisition and labeling for our experiments. Then, we describe the classifiers under evaluation and discuss the obtained results.

2 Related Work

Although emotional speech is in the research focus since many years, often results can not be applied with telephone services in mind. The experiences with the life data acquisition discussed in the next section reveal some important points.

- 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. This makes it quite difficult to draw the line between angry and not angry speech.
- People tend to over-pronounce, speak slow and loud or even in a "robot-like" manner due to the erroneous belief to 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 "per se" irrespective of problems that may result from the interaction.

There are quite a few studies that deal with telephone data, but most of the data differs from the above mentioned points in some ways. Yacoub et al [4] use data that was performed by actors, Ang et al [5] is based on the analysis of a simulated customer portal, in Walker et al [10], Lee et al [3] and Liscombe et al [8] some of the features are based on manual annotation. Nonetheless we can learn a lot from these studies regarding e.g. the expectation we have on the performance of different classifiers.

Shafran et al [6] studied, beneath 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.

Ang et al [5] 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. Liscombe et al [8] operate also on the HMIHY database and reached an accuracy of about 80%. Beneath 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 [3] studied data of an automated flight reservation application and introduced the concept of "emotionally salient words" which we might explore if the underlying speech recognition allows for a less restricted recognition. Walker et al [10] also operated on a subset of the HMIHY database, but focused on the classification of whole dialogs after the interaction took place, which might be interesting to be used for quality measurement.

3 Data Acquisition

The client of this study delivered about 120 hours of tuning data from a voice portal where customers can report problems with their phone connection and get pre-selected by an automated voice dialog before being connected to an agent. The recordings were done during

10 working days distributed widely in 2007. Of these, we chose randomly 21 hours of voice data.

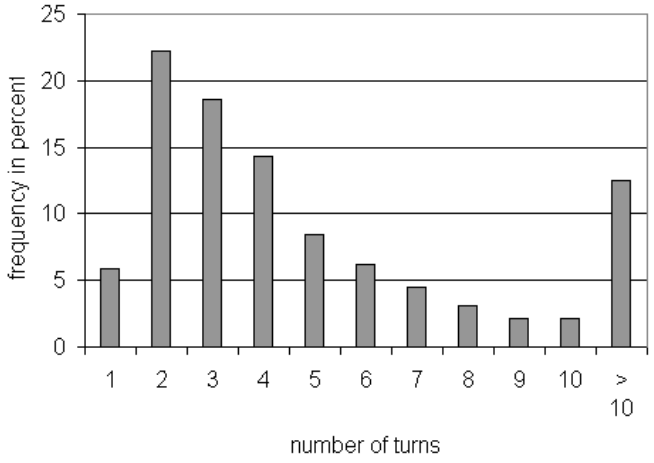

Figure 1 – Distribution of turn number per session

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, as can be seen from the distribution displayed in Figure 1.

All turns were listened to by three humans, two women and one man, and got annotated an anger value with the aid of a graphical labeling tool that was especially implemented for this task.

In order to achieve a consistent rating behaviour, the labelers got written label instructions and took part in a common session where some examples were discussed.

For each turn, the labelers had 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). Garbage turns included a multitude of turns that could not be classified for some reason, e.g. DTMF tunes, coughing, baby crying or lorries passing by.

We unified the ratings and mapped them to four classes "not angry", "unsure", "angry" and "garbage" for further processing conducting the following algorithm; in order to calculate a mean value for the three judgments, we assigned the value 0 to the "garbage" labels. All turns reaching a mean value below 0.5 were than assigned as "garbage", below 1.5 as "not angry", below 2.5 as "unsure" and all turns above that as "angry".

One special rule was used beforehand: if at least half of the labelers voted for 0, we decided for "garbage". Otherwise a turn with the labels 0, 0 and 4, mean value 1.3, would have been counted as "not angry" which is surely not correct.

Labeler	**Agreement in percent**	**Cohen's Kappa**
L1/L2	72,2 %	0,63
L1/L3	66%	0,55
L2/L3	64,8%	0,53
L1/L2/L3	55,4%	0,52

Table 1 – Inter-labeler agreement

The pairwise agreement between the three labelers is shown in Table 1. It is given in the first column as percentage of agreement, and in the second 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.

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 [9]. The table reveals that two of the labelers agreed much better than the third one (almost excellent), but still even between all three labelers the agreement is fair. Once more the point is underlined that emotion rating is by far an exact science but depends a lot on the subjective view of the labeler.

The distribution of all four classes is shown in Figure 2. It reveals that about 10% of the turns doesn't even contain analyzable speech. Of 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.

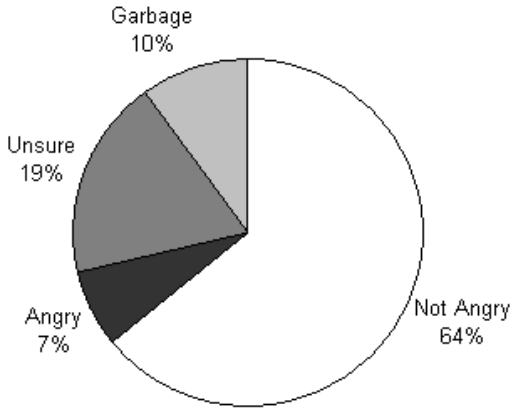

Figure 2 – Distribution of anger and garbage in the data

4 Features and Classifiers

The evaluation was on the one hand done with a classifier based on Gaussian Mixture models (GMM) which we have already used in prior projects [1], [2], and on the other hand with a new classifier based on Support Vector Machines (SVM), which is a fundamental different technology.

Also the set of features has been varied: on the one hand in the original classifier a set of prosodic as well as phonemic features has been used, on the other experimentally a reduced set that did not contain the phonemic features. In both settings the same prosodic features, calculated over the pitch and energy contour were used. For example we used the maximum and minimum of the pitch and the energy respectively. Altogether 31 prosodic features were computed.

In the original feature set, with the additional phonemic features, we used a phoneme recognizer to extract the position and length of vocals and non-vocals and out of it computed seven further features, for example the total length of all vocals and of all non-vocals respectively. Hence the prosodic/phonemic feature vector contained 38 features.

We computed the features over complete utterances thus for every utterance we got one feature vector for training and classification.

In order to model the strong speaker dependency of emotional expression, we used the first utterance of every dialog as a reference for a non-angry utterance. Based on this "reference vector" we additionally calculated the difference of every prosodic/phonemic feature to the value of the feature from the reference vector, the so-called "delta features". Therefore we used altogether 76 (38 prosodic/phonemic and 38 delta) and 62 (31 only prosodic and 31 delta) features respectively.

One of the characteristics of the SVM classifier is that the result is not a probability but a binary decision between one of the two classes "anger" and "non anger", while the GMM classifier delivers a probability value for each class. In order to compare the performance of both classifiers more easily, we adjusted a threshold on the Gaussian mixture models so that the recall value for the non-angry turns nearly equaled that of the SVM classifier.

5 Results

There were 647 dialogs that contained at least one angry utterance. Of these, 90% were randomly selected for the training set and the other 64 dialogs as test. In order not to have only dialogs containing anger in the test set, we extended it by 36 dialogs randomly selected from the remaining dialogs that don't contain any anger. The training set contains 1761 angry and 2502 non angry turns, the test set 190 angry turns and 302 non-angry.

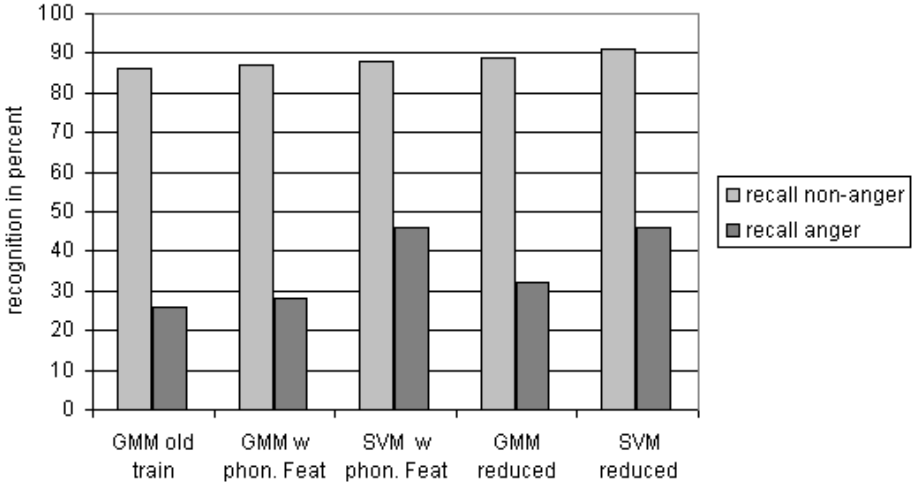

Figure 3 – Recall for anger and non anger with different configurations

Our tuning efforts can be comprised by the following three points.

- We trained the classifier with speech data from the same voice portal that the test data came from.
- We tried out a new classifier algorithm (SVM).
- We tested a reduced feature set by not regarding for phoneme based features.

The results of the evaluation are depicted in Figure 3.

For comparison we used a GMM based classifier with the complete feature set which was trained on a completely different data set and classified the test data set. As can be seen this classifier performed worse than all other new trained classifiers, underlining once again the data dependence of statistically based classification approaches.

As can be seen, the SVM based classifier performed better than the GMM approach in any case. Furthermore, the results with the reduced feature set are better for the GMM as well as for the SVM. The difference for the recall of anger between GMM and SVM for the feature vector with phonemic features (76 features) amounts to 18 percent (28% versus 46%). On the other hand, the difference for the reduced feature set between GMM and SVM amounts to 14 percent (32% versus 46%). The difference for the recall of anger between the two feature sets for the GMM approach amounts to 4% (28% versus 32% for the reduced set).

In the case of the SVM approach the recall of anger remains at 46% but the recall of non-anger raised from 89% for the complete feature set to 91% for the reduced feature set. Thus, despite our findings reported in [1], where the phoneme based features resulted in a higher accuracy, for this dataset better results were obtained without them.

6 Discussion and Outlook

As expected the recall for anger is worse than the recall for non-anger. One reason for this is that the acoustic variation (training and test) with anger stimuli is generally higher than with the non angry ones. Another reason is that the expression of emotion is highly speaker dependent. Although we tried to model the speaker dependency with the adding of the delta features we did not train an own classifier for every speaker. The reason for this is that the emotion classifier should be used in a voice portal and therefore it has to be a speaker independent classifier.

The following paragraphs list some ideas for further development.

Unlike the classifier with the GMM approach the SVM approach is at this stage only an offline classifier, hence not integrated in a voice portal (we used the SVM classifier inside the Weka system [11]). Because the SVM approach performs far better and we can not use speaker dependent classifiers, the next step to improve our system will be to implement an own SVM classifier and to integrate it into our system.

The data analysis of the current voice portal revealed a high percentage of garbage turns that can not be classified. The modeling of this garbage turns would enhance the accuracy of the classifier with respect to the given reality in the voice portal. However, this is a difficult task as the diversity of origin of these garbage turns is very high.

In order to provide for a linguistic detection of anger, the training of “emotion salient” words as described in [3] seems promising. As a pre-condition, the application of a large vocabulary grammar in the voice portal would be needed.

Given a multitude of classifiers, e.g. GMM and SVM with several feature sets (prosodic and linguistic), the adoption of a meta classifier would be an issue.

Finally, the detection of problematic dialogs for statistical reasons differs fundamentally from the detection of angry turns with real time dialog adaption in mind. An approach that models a dialog as a turn wise vector of anger probabilities might be better suited for the statistical application. Unaffected by this is the additional modeling of dialog stages, e.g. a turn that was uttered after repeatedly haven’t been understood by the system should have a raised anger probability.

7 Acknowledgements

The authors would like to thank Deutsche Telekom Laboratories for funding of the work described in this paper.

References

[1] Burkhardt, F., van Ballegooy, M., Englert, R. and Huber, R.: An Emotion-Aware Voice Portal, ESSP 2005

[2] Burkhardt, F., Ajmera, J., Englert, R., Stegmann, J. and Burleson, W.: Detecting Anger in Automated Voice Portal Dialogs, Interspeech (ICSLP) 2006,

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

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

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

[6] Shafran, I., Riley, M. and Mohri, M. "Voice signatures," in Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2003.

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

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

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

[11] Ian H. Witten and Eibe Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, 2005.