DETECTING REAL LIFE ANGER

*Felix Burkhardt, *Tim Polzehl, Joachim Stegmann, *Florian Metze* — *Richard Huber*

Deutsche Telekom Laboratories / *Technische Universität Berlin
Berlin, Germany

Sympalog Voice Solutions
Erlangen, Germany

ABSTRACT

Acoustic anger detection in voice portals can help to enhance human computer interaction. A comprehensive voice portal data collection has been carried out and gives new insight on the nature of real life data. Manual labeling revealed a high percentage of non-classifiable data. Experiments with a statistical classifier indicate that, in contrast to pitch and energy related features, duration measures do not play an important role for this data while cepstral information does. Also in a direct comparison between Gaussian Mixture Models and Support Vector Machines the latter gave better results.

Index Terms— emotion, detection, speech, classification

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. 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 [1]: 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 (GMMs, e.g. [2], [3], [4]), Support Vector Machines (SVMs, e.g. [5], [6]) or other vector clustering algorithms like k-nearest neighbor (KNN, e.g. [7], [4]) or linear discriminant analysis (LDA, e.g. [8]) 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 (e.g. [3], [4], [5]), while newer studies also include spectral features derived from Mel Frequency Cepstral Coefficients (MFCCs), e.g. [6], [8], [2] or [7].

There is quite a difference between telephone data as investigated by [4], [9] [5] or [8] and speech recorded with high quality microphones, as noted e.g. by [2] in a direct comparison. The difference between real life data and acted speech is so big that a direct comparison does not seem to make sense, e.g. [7] report recognition results for acted emotions far better than those reported on voice portal data. Most of the studies are based on data, even if it stems from customer voice portals, that was selected for laboratory investigation. Until today, the problem of how to deal with non-classifiable turns has not mentioned.

2. DATA ACQUISITION

The database consists of 21 hours recordings from a German voice portal where customers report problems with their phone connection and get preselected by an automated voice dialog before being connected to an agent. The recordings were done during 10 working days distributed widely in 2007. 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 shown in Figure 1. Most of the turns contain only 2-3 words as is typical for voice command applications, the average audio duration is 2.8 seconds while 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.

As there is no objective measure for anger, we labeled the data with three labelers, two women and one man. 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

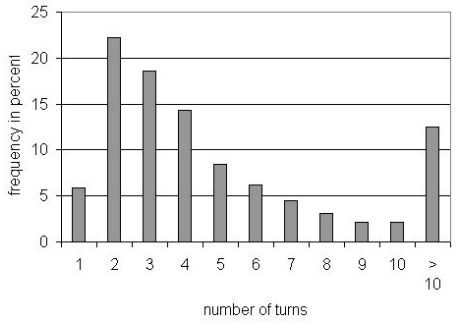

Fig. 1. Turn frequency per dialog.

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 tones, coughing, baby crying or lorries passing by.

We unified the ratings by mapping them to four classes ("not angry", "unsure", "angry" and "garbage") to further process as follows: 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 then assigned as "garbage", below 1.5 as "not angry", below 2.5 as "unsure" and all turns above that as "angry".

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 [10], 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. The table reveals that two of the labelers agreed much better than the third one, but still even between all three labelers the agreement is fair.

Table 1. *Agreement and Kappa values comparing three labelers.*

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

The distribution of all four classes is shown in Figure 2. The number of turns which does 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.

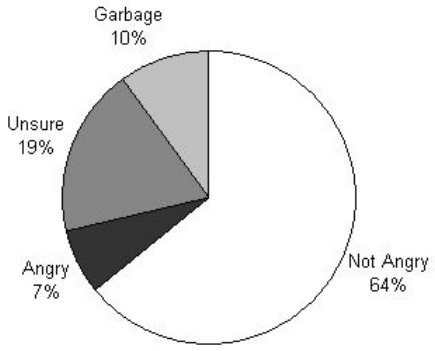

Fig. 2. Distribution of anger and garbage turns in the data.

A closer analysis of the turns marked as "garbage" revealed that 642(23,7 %) appear as first turn in a dialog, and 550 of these are followed by a non-garbage turn in the dialog. This is relevant because our classifier uses the first turn of each dialog to adapt the classifier to the speaker, i.e. the recognition of at least 2 % of the turns could be enhanced if garbage turns would be detected.

3. CLASSIFIERS AND FEATURES

The experiments reported here are an extension of our work described in [3] and [9]. Our original classifier is based on Gaussian Mixture Models (GMM) and prosodic features. As a first step a voiced/unvoiced decision is used as a starting point for a frame-based pitch detection algorithm based on dynamic programming. The pitch-values are then transformed to semitones in order for the later comparisons to operate on relative intervals rather than absolute pitch values. The duration related values are computed with respect to vowel vs. non-vowel phases in the speech. From these pitch, intensity and duration values, prosodic features are extracted like e.g. mean, minimum, standard deviation, regression coefficients etc. which are listed in detail in [3]. The feature vector is then classified into one of two classes using an algorithm based on Gaussian Mixture Models (GMM). A likelihood for every class is calculated, which is the minimum of all negative logarithms from the evaluation of the corresponding densities.

In order to model the strong speaker dependency of emotional expression, we use the first utterance of every dialog as a reference for a non-angry utterance. Based on this "reference vector" we additionally calculate the difference of every prosodic/phonemic feature to the value of the feature from the reference vector, the so-called "delta features".

3.1. Configuration 1

In a first experiment, we tested Support Vector Machines (SVM), which is a fundamental different technology. For a first evaluation we used the SVM classifier inside the Weka system [1]. 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 likelihood threshold on the Gaussian Mixture Models so that the recall value for the non-angry turns nearly equaled that of the SVM classifier.

Also the set of features has been varied: on the one hand in the original classifier a set of pitch and energy related features as well as duration features based on the length of voiced speech parts were used, on the other a reduced set that did not contain the duration features.

3.2. Configuration 2

A second experiment explored mainly the possible benefit of spectral related features in addition to prosodic ones. We used a different GMM-based classifier, stemming from the experiments described in [11], using pitch- and energy-related as well as MFCC features. We did feature reduction with PCA analysis and testet different dimensionalities.

For initial experiments we fitted the model using 16 Gaussians for the estimation of each category operating on diagonal covariance matrices.

4. RESULTS

The data contained 647 dialogs with 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 containing no anger. The training set contains 1761 angry and 2502 non-angry turns, the test set 190 angry turns and 302 non-angry.

For the experiments we excluded the "garbage" turns as they would have resulted in a third class. An experimental automatic classification of the 2711 "garbage" turns resulted in 431 angry (16 %), 2099 non-angry (77 %) and 181 turns that were rejected by the classifier (7 %). This indicates that a system that does not detect garbage would be biased towards non-angry class decisions.

The results of the evaluation are presented in table 2. In order to be able to compare our results with those reported in [8] and [2], we compute class average accuracy and class average f1, which computes as $(\frac{2r_a p_a}{r_a+p_a} + \frac{2r_{na}p_{na}}{r_{na}+p_{na}})/2$ with r_a and p_a being recall and precision for anger and r_{na} and p_{na} for non-anger respectively. The use of absolute accuracy doesn't make much sense for anger detection in voice portals, as, because the vast majority of the data is non-angry, a naive classifier always pleading for non-angry would already achieve 80 % accuracy.

Table 2. *Class-averaged recall and f1 for different classifier configurations.*

Classifier	average f1	average recall
Exp.1: GMM w. other training	0.46	0.55
Exp.1: GMM w. duration feat.	0.58	0.57
Exp.1: SVM w. duration feat.	0.66	0.67
Exp.1: GMM reduced feat.	0.61	0.61
Exp.1: SVM reduced feat.	0.70	0.69
Exp.2: GMM w/o priors	0.68	0.67
Exp.2: GMM with priors	0.70	0.69

First we trained our original GMM system with old data from a different voice portal [9], it performs worse than all other new trained classifiers, underlining once again the data dependence of statistically based classification approaches.

As can be seen in the other results from the first experiment, the SVM based classifier generally performed better than the GMM. The scores increase when we use the reduced feature set for the GMM as well as for the SVM. The difference for f1 between GMM and SVM for the feature vector with phonemic features amounts to 0.8 and for the reduced feature set between GMM and SVM amounts to 0.9. A smaller but also clear improvement appears when using the reduced features set, that excludes duration features based on voiced speech parts. The difference for f1 between the two feature sets for the GMM approach amounts to 0.03 and in the case of the SVM to 0.04. Thus, despite our findings reported in [3], where the duration related features resulted in a higher accuracy, for this data better results were obtained without them.

The table also shows the results of the second experiment, where, beneath prosody related features, cepstral information was included. We determined an optimum when using ten dimensions after PCA dimension reduction. In the proposed system we included the prior and obtained optimal result when setting its weight factor to 2.6. The prior information inclusion led to increased non-anger recall while at the same time recall of anger decreased. Operating the system with no a-priori information leads to an absolute drop in f1 of 0.02, retaining 7 dimensions after PCA. To summarize the table, using an SVM classifier as well as adding cepstral information improved the results in our case and our logical next step will be to combine these two approaches.

Figure 3 shows the trade-off between the recall of anger and the recall of non-anger in the second experiment depending on the prior weights for different numbers of dimensions after PCA dimension reduction. For the prior weight we tested values for zero to three. The points at the lower ends of

the lines illustrate a weight of zero what practically excluded the prior knowledge from the likelihood estimation. The blue dashed-dotted line represents the impact of prior tuning when keeping only one dimension after PCA, i.e. the one with highest variance explanation. The green dotted line was obtained using two dimensions. The red dashed line resulted from three dimensions, aqua solid line resulted from ten dimensions. Finally the black solid line shows the decrease of accuracy score when including too many dimensions, which is congruent with adding non relevant information to the system.

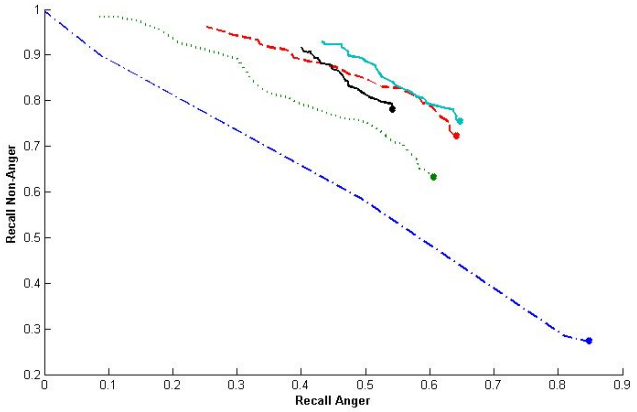

Fig. 3. Dimensionality and prior weighting for recall of anger and non-anger (blue dashed: 1 dimension, green dotted: 2 dimensions, red dashed: 3 dimensions, aqua solid: 10 dimensions, black solid: 20 dimensions).

5. DISCUSSION AND OUTLOOK

We experimented on the improvement of our anger detection component for voice portals and achieved enhancements by using an adapted features set, adding cepstral analysis and SVM instead of GMMs and achieved higher f1 values than reported in [8] or [2].

The following section lists some ideas for further development. 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 and it probably makes more sense to model the exploitable speech data with a kind of universal background model. In order to provide for a linguistic detection of anger, the training of "emotion salient" words as described in [4] seems promising. As a precondition, the application of a large vocabulary grammar in the voice portal is required. 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.

6. REFERENCES

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

[2] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using gmms," *Proc. ICSLP, Pittsburgh*, 2006.

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

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

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

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

[7] N. Sato and Y. Obuchi, "Emotion recognition using mel-frequency cepstral coefficients," *Information and Media Technologies*, vol. 2, no. 2, pp. 835–848, 2007.

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

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

[10] 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. ICASSP, Philadelphia*, 2005.

[11] T. Polzehl and F. Metze, "Using prosodic features to prioritize voice messages," *Proc. Searching Spontaneous Conversational Speech Workshop, SIGIR, Singapore*, 2008.