On NoMatchs, NoInputs and BargeIns: Do Non-Acoustic Features Support Anger Detection?

Alexander Schmitt, Tobias Heinroth
Dialogue Systems Research Group
Institute for Information Technology
Ulm University, Germany
alexander.schmitt@uni-ulm.de
tobias.heinroth@uni-ulm.de

Jackson Liscombe
SpeechCycle, Inc.
Broadway 26
New York City, USA
jackson@speechcycle.com

Abstract

Most studies on speech-based emotion recognition are based on prosodic and acoustic features, only employing artificial acted corpora where the results cannot be generalized to telephone-based speech applications. In contrast, we present an approach based on utterances from 1,911 calls from a deployed telephone-based speech application, taking advantage of additional dialogue features, NLU features and ASR features that are incorporated into the emotion recognition process. Depending on the task, non-acoustic features add 2.3% in classification accuracy compared to using only acoustic features.

1 Introduction

Certainly, the most relevant employment of speech-based emotion recognition is that of a telephone-based Interactive Voice Response System (IVR).

Emotion recognition for IVR differs insofar to “traditional” emotion recognition, that it can be reduced to a binary classification problem, namely the distinction between angry and non-angry whereas studies on speech-based emotion recognition analyze complete and relatively long sentences covering the full bandwidth of human emotions. In a way, emotion recognition in the telephone domain is less challenging since a distinction between two different emotion classes, angry and non-angry, is sufficient. We don’t have to expect callers talking to IVRs in a sad, anxious, happy, disgusted or bored manner. I.e., even if a caller is happy, the effect on the dialogue will be the same as if he is neutral. However, there still remain challenges for the system developer such as varying speech quality caused by, e.g., varying distance to the receiver during the call leading to loudness variations (which emotion recognizers might mistakenly interpret as anger). But also bandwidth limitation introduced by the telephone channel and a strongly unbalanced distribution of non-angry and angry utterances with more than 80% non-angry utterances make a reliable distinction of the caller emotion difficult. While hot anger with studio quality conditions can be determined with over 90% (Pittermann et al., 2009) studies on IVR anger recognition report lower accuracies due to these limitations. However, there is one advantage of anger recognition in IVR systems that can be exploited: additional information is available from the dialogue context, the speech recognizer and the natural language parser.

This contribution is organized as follows: first, we introduce related work and describe our corpus. In Section 4 we outline our employed features with emphasis on the non-acoustic ones. Experiments are shown in Section 5 where we analyze the impact of the newly developed features before we summarize our work in Section 6.

2 Related Work

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 (Yacoub et al., 2003), training and testing data in later studies has been more and more based on real-life data, see (Burkhardt et al., 2008),(Burkhardt et al., 2009). Most studies are limited to acoustic/prosodic features that have been extracted out of the audio data. Linguistic information was additionaly exploited in (Lee et al., 2002) resulting in

Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pages 128–131,
Queen Mary University of London, September 2009. ©2009 Association for Computational Linguistics

a 45.7% accuracy improvement compared to using only acoustic features. In (Liscombe et al., 2005) the lexical and prosodic features were additionaly enriched with dialogue act features leading to an increase in accuracy of 2.3%.

3 Corpus Description

For our studies we employed a corpus of 1,911 calls from an automated agent helping to resolve internet-related problems comprising 22,724 utterances. Three labelers divided the corpus into *angry*, *annoyed* and *non-angry* utterances (Cohen's $\kappa = 0.70$ on whole corpus; L1 vs. L2 $\kappa = 0.8$, L1 vs. L3 $\kappa = 0.71$, L2 vs. L3 $\kappa = 0.59$). The reason for choosing three emotion classes instead of a binary classification lies in the hope to find clearer patterns for strong anger. A distinction between non-angry and somewhat annoyed callers is rather difficult even for humans. 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. 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.

4 Features

We created two different feature sets: one based on typical acoustic/prosodic features and another one to which we will refer as 'non-acoustic' features consisting of features from the Automatich Speech Recognition (ASR), Natural Language Understanding (NLU), Dialogue Manager (DM) and Context features.

4.1 Acoustic Features

The acoustic/prosodic features were extracted with the aid of Praat (Boersma, 2001) and consist of power, mean, rms, mean harmonicity, pitch (mean, deviation, voiced frames, time step, mean slope, minimum, maximum, range), voiced pitch (mean, minimum mean, maximum mean, range), intensity (mean, maximum, minimum, deviation, range), jitter points, formants 1-5, MFCC 1-12. The extraction was performed on the complete short utterance.

4.2 Non-Acoustic Features

The second, i.e. non-acoustic, feature set is based on features logged with the aid of the speech platform hosting the IVR application and is presented here in more detail. They include:

ASR features: raw ASR transcription of caller's utterance (*Utterance*) (unigram bag-of-words); ASR confidence of returned utterance transcription, as floating point number between 0 (least confident) and 1 (most confident) (*Confidence*); names of all grammars active (*GrammarName*); name of the grammar that returned the parse (*TriggeredGrammarName*); did the caller begin speaking before the prompt completed? ('yes', 'no') (*BargedIn*); did the caller communicate with speech ('voice') or keypad ('dtmf') (*InputModeName*); was the speech recognizer successful ('Complete') or not and if it was not successful, an error message is recorded such as 'NoInput' or 'NoMatch' (*RecognitionStatus*)

NLU-Features: the semantic parse of the caller utterance as returned by the activated grammar in the current dialog module (*Interpretation*); given caller speech input, we need to try and recognize the semantic meaning. The first time we try to do this, this is indicated with a value of 'Initial'. If we were not returned a parse then we have to reprompt ('Retry1' or 'Timeout1'). Similar for if the caller asks for help or a repetition of the prompt. Etc. (*LoopName*)

DM-Features: the text of what the automated agent said prior to recording the user input (*PromptName*); the number of tries to elicit a desired response. Integer values range from 0 (first try) to 7 (6th try) (*RoleIndex*); an activity may request substantive user input ('Collection') or confirm previous substantive input ('Confirmation') (*RoleName*); within a call each event is sequentially organized by these numbers (*SequenceID*); the name of the activity (aka dialog module) that is active (*ActivityName*); type of activity. Possible values are: Question, PlatformValue, Announcement, Wait, Escalate (*ActivityType*)

Context-Features: We further developed additional cumulative features based on the previous ones in order to keep track of the NoMatch, NoInputs and similar parameters serving as an indicator for the call quality: number of non-empty NLU parses (*CumUserTurns*); number of statements and questions by the system (*CumSysTurns*); number of questions (*CumSysQuestions*); number of help requests by the user (*CumHelpReq*); number of operator requests (*CumOperatorReq*); number of NoInput events (*CumNoInputs*); number of NoMatch events (*CumNoMatchs*) number of BargeIns (*CumBargeIns*).

5 Experiments

In order to prevent an adaption of the anger model to specific callers we seperated the corpus randomly into 75% training and 25% testing material and ensured that no speaker contained in training was used for testing. To exclude that we receive a good classification result by chance, we performed 50 iterations in each test and calculated the performance's mean and standard deviation over all iterations.

Note, that our aim in this study is less finding an optimum classifer, than finding additional features that support the distinction between angry and non-angry callers. Support Vector Machines and Artificial Neural Networks are thus not considered, although the best performances are reported with those learning algorithms. A similar performance, i.e. only slightly poorer, can be reached with Rule Learners. They enable a thorough study of the features, leading to the decision for one or the other class, since they produce a human readable set of if-then-else rules. Our hypotheses on a perfect feature set can thus easily be confirmed or rejected.

We performed experiments with two different classes: 'angry' vs. 'non-angry' and 'angry+annoyed' vs. 'non-angry'. Merging angry and annoyed utterances aims on finding all callers, where the customer satisfaction is endangered. In both tasks, we employ a) only acoustic features b) only ASR/NLU/DM/Context features and c) a combination of both feature sets. The number of utterances used for training and testing is shown in Table 1.

	Test A		**Test B**	
	angry+ annoyed	non-a.	angry	non-a.
Training	∼ 320	∼ 320	∼ 80	∼ 80
Testing	∼ 140	∼ 140	∼ 40	∼ 40

Table 1: Number of utterances employed for both tests per iteration. Since the samples are selected randomly and the corpus was separated by speakers before training and testing, the numbers may vary in each iteration.

As result we expect acoustic features to perform better than non-acoustic features. Among the relevant non-acoustic features we assume as an indicator for angry utterances low ASR confidences and high barge-in rates, which we consider as signal for the caller's impatience. All tests have been performed with the machine learning framework RapidMiner (Mierswa et al., 2006) featuring all common supervised and unsupervised learning schemes.

Results are listed in Table 2, including precision and recall values. As expected, Test B (angry vs. non-angry) has the highest accuracy with 87.23% since the patterns are more clearly separable compared to Test A (annoyed vs. non-angry, 72.57%). Obviously, adding non-acoustic features increases classification accuracy significantly, but only where the acoustic features are not expressive enough. While the additional information increases the accuracy of the combined angry+annoyed task by 2.3 % (Test A), it does not advance the distinction between only angry vs. non-angry (Test B).

5.1 Emotional History

One could expect, that the probability of an angry/annoyed turn following another angry/annoyed turn is rather high and that this information could be exploited. Thus, we further included two features *PrevEmotion* and *PrevPrevEmotion*, taking into account the two previous hand-labeled emotions in the dialogue discourse. If they would contribute to the recognition process, we would replace them by automatically labelled ones. All test results, however, did not improve.

5.2 Ruleset Analysis

For a determination of the relevant features in the non-acoustic feature set, we analyzed the ruleset generated by the RuleLearner in Test A. Interestingly, a dominant feature in the resulting ruleset is 'AudioDuration'. While shorter utterances were assigned to non-angry (about <2s), longer utterances tended to be assigned to angry/annoyed. A following analysis of the utterance length confirms this rule: utterances labeled as angry averaged 2.07 (+/-0.73) seconds, annoyed utterances lasted 1.82 (+/-0.57) s and non-angry samples were 1.57 (+/- 0.66) s in average. The number of NoMatch

Test A: Angry/Annoyed vs. Non-angry	only Acoustic	only Non-Acoustic	both
Accuracy	70.29 (+-2.94) %	61.43 (+-2.75) %	72.57 (+-2.37) %
Precision/Recall Class ’Ang./Ann.’	71.51% / 61.57%	68.35% / 42.57%	73.67% / 70.14%
Precision/Recall Class ’Non-angry’	69.19% / 73.00%	58.30% / 80.29%	71.57% / 75.00%
Test B: Angry vs. Non-angry	only Acoustic	only Non-Acoustic	both
Accuracy	87.06 (+-3.76) %	64.29 (+-1.32) %	87.23 (+-3.72) %
Precision/Recall Class ’Angry’	87.13% / 86.55%	66.0% / 58.9%	86.88% / 87.11%
Precision/Recall Class ’Non-angry’	86.97% / 87.53%	62.9% 69.9%	87.55% / 87.33%

Table 2: Classification results for angry+annoyed vs. non-angry and angry vs. non-angry utterances.

events (CumNoMatch) up to the angry turn played a less dominant role than expected: only 8 samples were assigned to angry/annoyed due to reoccurring NoMatch events (>5 NoMatchs). Utterances that contained ’Operator’, ’Agent’ or ’Help’ were, as expected, assigned to angry/annoyed, however, in combination with high AudioDuration values (>2s). Non-angry utterances were typically better recognized: average ASR confidence values are 0.82 (+/-0.288) (non-angry), 0.71 (+/- 0.36) (annoyed) and 0.56 (+/- 0.41) (angry).

6 Conclusion and Discussion

In IVR systems, we can take advantage of non-acoustic information, that comes from the dialogue context. As demonstrated in this work, ASR, NLU, DM and contextual features support the distinction between angry and non-angry callers. However, where the samples can be separated into clear patterns, such as in Test B, no benefit from the additional feature set can be expected. In what sense a late fusion of linguistic, dialogue and context features would improve the classifier, i.e. by building various subsystems whose opinions are subject to a voting mechanism, will be evaluated in future work. We will also analyze why the linguistic features did not have any visible impact on the classifier. Presumably a combination of n-grams, bag-of-words and bag of emotional salience will improve classification.

7 Acknowledgements

We would like to take the opportunity to thank the following colleagues for contributing to the development of our emotion recognizer: Ulrich Tschaffon, Shu Ding and Alexey Indiryakov.

References

Paul Boersma. 2001. Praat, a System for Doing Phonetics by Computer. *Glot International*, 5(9/10):341–345.

Felix Burkhardt, Richard Huber, and Joachim Stegmann. 2008. Advances in anger detection with real life data.

Felix Burkhardt, Tim Polzehl, Joachim Stegmann, Florian Metze, and Richard Huber. 2009. Detecting real life anger. In *Proc. of ICASSP*, April.

Chul Min Lee, Shrikanth Narayanan, and Roberto Pieraccini. 2002. Combining Acoustic and Language Information for Emotion Recognition. In *International Conference on Speech and Language Processing (ICSLP)*, Denver, USA, October.

Jackson Liscombe, Guiseppe Riccardi, and Dilek Hakkani-Tür. 2005. Using Context to Improve Emotion Detection in Spoken Dialog Systems. In *International Conference on Speech and Language Processing (ICSLP)*, Lisbon, Portugal, September.

Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. 2006. Yale: Rapid prototyping for complex data mining tasks. In *KDD ’06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, August.

Johannes Pittermann, A. Pittermann, and Wolfgang Minker. 2009. *Handling Emotions in Human-Computer Dialogues*. Text, Speech and Language Technology. Springer, Dordrecht (The Netherlands).

Sherif Yacoub, Steven Simske, Xiaofan Lin, and John Burns. 2003. Recognition of emotions in interactive voice response systems. In *Proc. Eurospeech, Geneva*, pages 1–4.