Prediction of Turn-wise User Judgments from Acoustic Features of User Utterances

Klaus-Peter Engelbrecht, Felix Burkhardt, Sebastian Möller

Deutsche Telekom Laboratories, TU-Berlin
klaus-peter.engelbrecht@telekom.de

Abstract. Knowledge about the users' satisfaction with a dialog is valuable when it comes to evaluation, or when the dialog strategy can be adapted. Predictions of user satisfaction have been made on the basis of interaction parameters for entire dialogs with moderate results. We try to enhance such models by including audio features describing the user utterance recording. As users were asked for a judgment after each turn, we make predictions on a turn-by-turn basis. We show that predictions of judgments on the basis of audio features are more accurate than the baseline.

Introduction

With increasing complexity of interactive systems available to customers, the assessment of quality becomes increasingly important. According to Jekosch ([1]), quality is a perceptual construct and thus subjective, which means that its measurement must involve human judges. It is common practice to invite users to interact with a system and have them fill out a questionnaire covering several quality aspects ([2], [3]). However, it is not always possible to ask the user for a judgment, e.g. when the user is in the middle of a dialog, or if it is inappropriate to molest the user with questions irrelevant to the actual task.

Therefore, methods were developed which allow automatic predictions of user judgments from the data acquired during the interaction. The most famous approach in this respect is the PARADISE framework ([4]), which assumes that user satisfaction can be modeled as a function of task success and dialog costs in terms of efficiency and dialog quality. The latter can be described with interaction parameters, such as the dialog length. A prediction model is obtained by training a linear regression function with the interaction parameters as predictors and user judgments as dependent variable. Such models are typically database-specific, and can predict about 50% of the variance in the observed judgments ([5]). Other machine learning techniques may be used for modeling the relation between parameters and judgments, e.g. HMMs ([6]). However, the relation between parameters and judgments still depends on the system and its specific problems ([7]). If the user judgment, however, could be derived from the acoustic features of the user utterance, this relation should be independent of the system or the database. Therefore, such relations are worth analyzing.

In a related line of research, unsuccessful dialogs (e.g. because of user hang-up) were predicted ([8], [9]). Like in PARADISE, typically parameters describing the interaction are used as predictor variables; however, attempts were made to incorporate emotion recognition from audio features in such predictions ([10]). Direct relations between audio features and user ratings were also analyzed ([11]), however, unlike in our study, features are averaged across entire dialogs, as judgments were collected only once after the interaction.

Data

In order to analyze the judgment behaviour of different users, we designed an experiment in which the users judged the quality of the dialog "so far" after each dialog turn. Judgments were provided on a number pad with an attached rating scale. As we were interested in the differences in the users' judgment behaviours, we forced the dialogs to follow predefined scripts, by letting a Wizard-of-Oz control the system actions. The dialog scripts were designed to contain a number of interaction problems observed in previous experiments or known from the literature, such as recognition errors, prompt wording problems, or task failure. In addition, we combined these problems with different confirmation strategies and different complexities of the possible user replies (i.e. 1 or 2 concepts at a time).

In order to keep up a plausible interaction scenario, we designed a consistent dialog strategy (i.e. system) for all tasks, which was flexible enough to cover a large part of the above problems. The resulting system mock-up resembles a typical slot filling strategy with system initiative in the domain of restaurant information.

25 users (13 f, 12 m) aged between 20 and 46 years (M=26.5; STD=6.6) participated in the experiment. Each user performed the same 5 interactions, preceded by a test run. The resulting dataset consists of 945 recorded turns with corresponding user judgments of the dialog up to this turn. The distribution of judgments is: 40 "bad", 120 "poor", 199 "fair", 371 "good", 215 "excellent". Further details on the experiment can be found in [12].

In order to analyze the classifiability of the data based on its acoustic features, we limited the complexity to a two-class problem. We merged the user satisfaction annotations into a binary decision: judgments from 1 to 3 were regarded as dissatisfied and 4 to 5 as satisfied. We also tried to group labels 3 to 5 as satisfied but this resulted in a much worse automatic classification.

Results

In a first step, we tried to predict the judgments from features describing the interaction. Speech recognizer errors, confirmation strategy and task success proved to be most useful for these models. We used a Hidden Markov Model to assess the development of the ratings over the course of the dialog. Features were modelled as emissions and ratings as states. This topology allows to predict the distribution of

judgments at each dialog turn. Comparison with the empirical distribution showed satisfactory results on independent test data (see [6] for detailed results).

In a second step, we investigated the classifiability based on the acoustic characteristics of the data. Whether the satisfaction can be predicted by acoustic measurements is a fascinating question, given that no evidence by manual inspection exists beforehand.

Based on the Praat toolkit [13], we extracted the following 64 features from the audio files:

- Pitch: Maximum, minimum, time of maximum and minimum, range, mean and standard deviation.
- Duration: total duration, relation of voice and unvoiced frames.
- Intensity: Maximum, minimum, time of maximum and minimum, range, mean and standard deviation.
- Formants: Maximum, minimum, time of maximum and minimum, range, mean and standard deviation of first five formants.
- Spectrum: Mean of 12 MFCCs.

In order to get a speaker-separated training and test-set, we used the leave-one-speaker-out method and did a 25-fold cross validation by using 25 times the data of one speaker for testing and the data of all other speakers for training.

Utilizing the WEKA classification toolkit [14], we classified the data with the logistic classifier [15], which minimizes a matrix distance based on regression functions.

Before classification, we selected the 25 best working features with WEKA's "SVMAttributeEval" method, which evaluates the worth of an attribute by using a Support Vector Machine (SVM) classifier. Attributes are ranked by the square of the weight assigned by the SVM.

Because more of the 945 turns were labelled as "satisfied" (586) than "dissatisfied" (359), the baseline for the trivial classifier that always votes the majority class is 0.619 overall accuracy.

Instead of using the overall accuracy to evaluate a classifier, we prefer the unweighted average F1 value (UAF), because it is invariant with respect to unequal distributions. F1 is the harmonic mean of precision and recall and computes for each class as twice the product of recall and precision divided by its sum. The UAF then computes as

$$UAF = \left(\frac{2r_{c1}p_{c1}}{r_{c1}+p_{c1}} + \frac{2r_{c2}p_{c2}}{r_{c2}+p_{c2}}\right)/2$$

with r_{cn} and p_{cn} being recall and precision of the nth class.

The baseline UAF is 0.382. With the optimized 25 feature set, the classifier reached a UAF of 0.449 while the overall accuracy was only slightly higher (0.622).

Conclusion

It is quite remarkable, given that the classification of this data is not based on listener impressions but on speaker introspection, that acoustical analysis can help to predict

the data, even if the accuracy is not very high. Note that, in contrast to other emotion detection tasks described in literature, in this case the emotional annotation was done by the user himself (assumed that user satisfaction is related to emotional states).

We admit though, that our test set is too small to allow for general conclusions. With only 25 persons for training and test, general assumptions can not be made and this investigation must be seen as preliminary.

As futures steps, beneath the acquisition of larger, and perhaps more natural, data sets, the modelling of the time dynamics of the dialog progression as well as speaker adoption can be investigated.

References

1. U. Jekosch, Voice and Speech Quality Perception. Assessment and Evaluation, Berlin: Springer, 2005.
2. ITU-T Rec. P.851, Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems, International Telecommunication Union, Geneva, 2003.
3. K. S. Hone, R. Graham, "Subjective Assessment of Speech-system Interface Usability," in Proc. of EUROSPEECH, 2001, pp. 2083-2086.
4. M. Walker, D. Litman, C. Kamm, A. Abella, "PARADISE: A Framework for Evaluating Spoken Dialogue Agents," in Proc. of ACL/EACL, 1997, pp. 271–280.
5. S. Möller, K.-P. Engelbrecht, R. Schleicher, "Predicting the Quality and Usability of Spoken Dialogue Services," Speech Communication, vol. 50, 2008, pp. 730-744.
6. K.-P. Engelbrecht, F. Gödde, F. Hartard, H. Ketabdar, S. Möller, „Modeling User Satisfaction with Hidden Markov Models," in Proc. of Sigdial, 2009.
7. K.-P. Engelbrecht, C. Kuehnel, S. Möller, "Weighting the Coefficients in PARADISE Models to Increase Their Generalizability," in Proc. of PIT, 2008, pp. 289-292.
8. A. Schmitt, C. Hank, J. Liscombe, "Detecting Problematic Calls With Automated Agents," in Proc. of PIT, 2008, pp. 72-80.
9. M. Walker, I. Langkilde, J. Wright, A. Gorin, D. Litman, "Learning to predict problematic situations in a spoken dialogue system: experiments with how may I help you?", in Proc. of the 1st North American chapter of the Association for Computational Linguistics conference, 2000, pp. 210-217.
10. O. Herm, A. Schmitt, and J. Liscombe, "When calls go wrong: How to detect problematic calls based on log-files and emotions?", in Proc. of the International Conference on Speech and Language Processing (ICSLP), 2008.
11. S. Möller, K.-P. Engelbrecht, M. Pucher, P. Fröhlich, L. Huo, U. Heute, F. Oberle, "TIDE: A Testbed for Interactive Spoken Dialogue System Evaluation," in Proc. 12th Int. Conf. Speech and Computer (SPECOM'2007), RU-Moscow, 2007.
12. K.-P. Engelbrecht, F. Hartard, F. Gödde, S. Möller, "A Closer Look at Quality Judgments of Spoken Dialog Systems," in Proc. of Interspeech, 2009.
13. P. Boersma and D. Weenink. Praat: doing phonetics by computer (version 5.1.04). 2009.
14. I. H. Witten and E. Frank. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco, 2005.
15. S. le Cessie and J. van Houwelingen. Ridge estimators in logistic regression. Applied Statistics, 41(1):191–201, 1992.