TIDE: A Testbed for Interactive Spoken Dialogue System Evaluation

Sebastian Möller [1], Klaus-Peter Engelbrecht [1], Michael Pucher [2], Peter Fröhlich [2]
Lu Huo [3], Ulrich Heute [3], Frank Oberle [4]

[1] Deutsche Telekom Laboratories, Berlin University of Technology, Germany
[2] Forschungszentrum Telekommunikation Wien (ftw.), Austria
[3] LNS, Christian-Albrechts-Universität Kiel, Germany
[4] T-Systems Enterprise Services GmbH, Berlin, Germany

Abstract

Telephone-based spoken dialogue platforms usually log a multitude of information for each interaction between user and system. Due to its large amount and specificity, this information is very difficult to interpret. In this paper, we present a new testbed for the analysis and evaluation of log-file information collected with spoken dialogue platforms. On the signal level, parameters are extracted which allow the sources of recognition errors to be allocated to channel and user characteristics. On the symbolic level, phonemic similarities are computed which allow the confusability of lexicon and grammar to be evaluated. The algorithms are integrated into a graphical user interface for an effective analysis of system performance. It is evaluated on data collected within a usability test of a prototype implementation for an automated pre-qualifying application of Deutsche Telekom, and correlations between extracted parameters, recognition rates and user judgments are pointed out.

1. Introduction

Commercial spoken dialogue telephone services for tasks such as timetable or tariff information, reservation, or telephone-banking are usually implemented on dialogue platforms which possess speech recognition and understanding, dialogue management, speech generation, and speech output components. These platforms are able to log a multitude of information upon request, such as the audio signals recorded from the user, the signals played to the user, the recognition and understanding results, or information on the state of the dialogue machine. Because of its large amount and diversity, this information is very difficult to interpret. However, it may be very useful for service analysis, optimization, and monitoring.

Different types of information may be distinguished: The primary information exchanged between user and system is reflected in the *audio signals* from the user and the system. Because the system is not always "listening", i.e. the speech recognition channel is not always open, the audio signal recorded from the user is influenced by the voice activity detection (VAD) algorithm implemented on the platform. It may also be enhanced by noise and echo suppression algorithms. The system's speech signal is commonly available as a clean pre-recorded audio file, or as a clean file generated by a text-to-speech (TTS) module.

The interaction between user and system can additionally be quantified on a symbolic level, in terms of *interaction parameters*. These parameters describe the behavior of the user and the one of the system, and reflect the performance of the system in the interaction. Examples include the number and length of user and system utterances, timing information, the frequencies of help requests, cancel attempts, or barge-ins, recognition and understanding accuracies, or task success parameters. In ITU-T Suppl. 24 to P-Series Recommendations [1], a large number of such interaction parameters are described. The determination of most of these parameters requires a manual transcription and annotation by a human expert.

Additional information is hard-coded in the system, e.g. in terms of the interaction logic, the vocabulary, or the grammar which is available at each state of the dialogue. These *system characteristics* have a significant impact on the dialogue flow and on the quality of the interaction, as it is perceived by the human partner. The latter, however, can only be determined by collecting *subjective judgments* in controlled experiments. ITU-T Rec. P.851 [2] describes methods for carrying out such subjective interaction experiments. They are usually very expensive, and as a consequence, system designers try to avoid them as far as possible.

In the present paper, we make use of information which can easily be collected in log-files for analyzing, optimizing and monitoring the quality of interactions. Our focus is on parameters which can be extracted automatically, without requiring a tedious and time-consuming transcription or labeling process. We extract information both on the signal and on the symbolic level. The algorithms used for this purpose are described in Sections 2 and 3, respectively. With the help of these algorithms, we analyze data collected within a usability test of a prototype implementation for a pre-qualifying application of Deutsche Telekom, see Section 4. The test was performed in cooperation with Siemens AG, Corporate Technology, Competence Center "User Interface Design". The extracted parameters are analyzed in Section 5, first by addressing their relationship to the recognition performance of the system, and then to the subjective judgments obtained from the test users. We finally describe how the algorithms are integrated into a graphical tool for an easy and efficient log-file analysis in Section 6. Some conclusions and an outlook on possible use cases and extensions of the tool are given in Section 7.

2. Information on the Signal Level

The audio signals which can be recorded at the dialogue platform include

- the user's speech signal, transmitted through the telephone channel and cut out by the VAD,
- the system's speech signal available as a clean audio file.

No audio signals are available on the user's side. Thus, we do not have any information on how the system's speech signal is degraded by the transmission channel before it is perceived by the user, nor on the clean speech signal uttered by the user.

The user's speech signal is analyzed with respect to the characteristics of the user, as well as of the transmission channel. For this purpose, the following parameters are extracted:

1. *Active Speech Level (ASL)*: This level is calculated from the segments extracted by the VAD, following the algorithm described in ITU-T Rec. P.56 [3]. It consists of a histogram analysis with multiple variable thresholds and results in a level relative to the overload point of the digital system.
2. *Noise Level*: Noise mainly stems from background noise present at the user's side and picked up by the telephone handset, as well as from circuit noise induced by the subscriber line. Two algorithms have been compared to determine the noise level: (a) Speech pauses have been extracted with the help of the GSM VAD [4], and then a smoothed noise power spectrum is determined from the windowed non-speech segments. (b) Alternatively, minimum statistics [5] has been used to extract the noise power during speech segments. Method (b) tended to overestimate the noise level on a set of controlled test data, and therefore we decided to use method (a) for noise level determination.
3. *Signal-to-Noise Ratio (SNR)*: With a similar processing as for the noise power, a smoothed power spectral density of the speech signal is determined during speech activity. The SNR is calculated as the ratio between both power densities, calculated per utterance.
4. *Mean Cepstral Deviation (MCD)*: It is known that multiplicative noise is introduced in telephone channels by logarithmic PCM or ADPCM coding. In order to determine the level of degradation introduced this way, we assume that the recorded speech signal $y(n)$ is determined by the clean speech signal $s(n)$ and a white Gaussian noise component $n(n)$ with a certain ratio Q:

$$y(n)=s(n)+s(n)\cdot 10^{-Q/20}\cdot n(n) \qquad (1)$$

Falk et al. [6] proposed to measure the noise in the degraded speech signal via the flatness of the output speech signal $y(n)$. The underlying idea is that – because the multiplicative noise of Eq. (1) introduces a fairly flat noise in the spectral domain – the lower the Q value, the less the spectrum of $s(n)$ can be preserved in $y(n)$, and the flatter the spectrum of $y(n)$ is. We use the MCD as a measure of the amount of multiplicative noise present in the degraded speech signal, because analyses have shown that the correlation between Q and MCD is about -0.93 [6]. To calculate MCD, we calculate cepstral coefficients for the speech frames (decided by the GSM VAD), and average their standard deviations.

5. *Single-ended Speech Quality Estimate*: Multiplicative noise is not the only degradation introduced by modern telephone channels. In particular, non-waveform speech codecs generate distortions which have different perceptual and signal correlates, and which have shown to degrade recognition performance [7]. In order to cover these channel degradations, we used the single-ended model described in ITU-T Rec. P.563 [8] to obtain an indication of the overall speech quality degradation introduced by the channel. This model generates a clean speech reference from the degraded speech signal by means of an LPC analysis and re-synthesis. Both the generated clean and the recorded degraded speech signals are transformed to a perceptually-motivated representation. An estimate of the overall quality, MOS, is then determined from a comparison of both representations. The approach can also be applied to TTS signals generated by the system, as it has been shown in [9].
6. *Active Speech Duration*: We use the GSM VAD to cut off pauses at the beginning and at the end of the speech signals which remain after the VAD of the dialogue platform.
7. *Fundamental Frequency (F0)*: Hirschberg et al. [10] have shown that mean and maximum F0 can be useful in predicting the recognition error. We adopted the autocorrelation analysis from Rabiner [11] and some simple smoothing algorithm for F0 estimation. For each user utterance, the mean, the standard deviation, and the 95% percentile of F0 are calculated.

Parameters 1-5 mainly reflect the characteristics of the telephone transmission channel, whereas 1, 6 and 7 address the characteristics of the user. Thus, by determining these parameters, it may be decided whether recognition failures are due to the transmission channel or to user particularities.

3. Information on the Symbolic Level

Apart from the audio signals, the dialogue platform logs information related to the dialogue state the system is in, as well as to the speech recognizer. For the dialogue state, the system logs the ID of the respective state, a time stamp when the system enters the state, the prompt which is played in this system state, as well as the ID of the vocabulary and the grammar used by the speech recognizer in the state.

The vocabulary of the system has been analyzed with respect to its confusability, as this is expected to be predictive of the recognition performance of the system in that specific state. We calculate four types of confusion measures: (1) The Minimum Edit Distance (MED) with an equal weight for substitutions, insertions and deletions; (2) an articulatory phonetic distance, where the substitution costs of the MED have been weighted according to an articulation-based phonetic distance; (3) a perceptual phonetic distance, where the MED has been weighted according to perceptual similarities; (4) an HMM distance based on the Kullback-Leibler divergence between two Gaussian mixture models. Details on the measures are given in [12]. They show that methods 2-4 all provide reasonable correlations with word confusion. Because the acoustic models of the speech recognizer are not accessible on the dialogue platform, only the articulatory (2) and the perceptual phonetic distances (3) are discussed here.

The speech recognizer of the dialogue platform provides a status flag indicating “recognition” (correct or incorrect), “rejection” (due to the rejection threshold of the system), “hang up” (the user hangs up), “speech too early” (the user speaks too early), “no speech timeout” (the user did not speak in the time interval where the recognizer was open), or “left-overs” (the user leaves the dialogue without speaking). The labels “recognition” and “rejection” refer to the behavior of the recognizer and can be evaluated with respect to the recognition performance.

4. Data Acquisition

The described algorithms have been applied to data collected with a telephone-based system. This system is implemented on a Nuance dialogue platform. It provides information on fixed and mobile telephone and internet tariffs, and allows internet problems to be reported. It generates log-files and records the user speech signal in each dialogue state.

25 native German test subjects interacted with a prototype implementation of this system. The subjects were recruited according to 5 groups: AF (adult female), AM (adult male), SF (senior female), SM (senior male), and C (child). All subjects had to carry out a minimum of 4 interactions

with the system, targeting on fixed telephone tariff enquiry, mobile telephone tariff inquiry, internet tariff inquiry, and internet problem report. This resulted in a set of 280 dialogues and 1672 user audio files in the database.
After each interaction, the users rated a questionnaire with several items related to their current experience. Here we consider user judgments on the four general items

- overall impression (1…very good – 6…unsatisfactory)
- system wording comprehensible (1…yes – 5…no)
- system understood what user wanted (1…yes – 5…no)
- dialogue should be changed (1…yes – 2…no)
- which have been collected after each dialogue.

5. Data Analysis

All user audio files have been transcribed by a human expert. The transcriptions and the recognizer's hypotheses have been compared with the help of NIST's "sclite" software [13], determining the number of correct words, substitutions, deletions and insertions for each user utterance classified with the "recognition" or "rejection" label.

Table 1. Summary of user speech characteristics.

Parameter	Mean	STD
ASL (dB)	-24.5	7.8
Noise level (dB)	-57.4	7.9
SNR (dB)	32.1	13.3
MCD	0.103	0.008
Single-ended estimate	2.46*	0.85*
Active speech duration (s)	1.53	1.87
F0 mean (Hz)	165.0	45.2

*) Due to the short utterance length, P.563 estimations had to be derived from artificially concatenated segments per dialogue, which might have caused low MOS values.

A general analysis of all user utterances is given in **Table 1**. It shows that the utterances have a high signal-to-noise ratio and are of relatively high speech quality. We think that this is due to the test set-up where subjects interacted with the system from two test cabinets equipped with a good wireline telephone. In addition, the utterances are relatively short, indicating that the subjects preferred to use a simple command language towards the system.

5.1 Analysis with respect to Subject Groups

An analysis by user group was carried out for parameters related to the signal quality (ASL, noise level, MCD, recognizer status labels). Data was aggregated on a dialogue level (mean values for utterance-wise variables) and ANOVAs were calculated taking into account that each test person contributed 3-13 cases to the data set. As was expected, the signal quality parameters do not differ between person groups, while the label "recognition" shows a significant effect (ANOVA parameters F=3.9; p=0.02), which however does not hold for the label "rejection".
Groups differed significantly with respect to their commanding style as parameterized in active speech duration, timeout and barge-in frequency (F>4.3; p<0.01). SF (senior female) show highest means for duration and barge-ins and directly follow C for timeouts, while M, F and SM have comparatively homogeneous means for duration and timeout.
Interestingly, the judgment on overall impression differs significantly depending on the groups (F=8.4; p=0.00), SM rating worst, followed by SF, M, F and C. The judgment on system understanding shows the same tendency, however, SF and F each gain one rank, which agrees with the observation that recognition performance is lower for F than for M and for SF than for SM.

5.2 Signal Level Parameters and Recognition Performance

In order to represent recognition performance, the recognition results are classified into two classes "correctly recognized" and "error", where the first term refers to the complete match between the transcript and the recognition result and the second refers to any mismatch or rejected situation. After deleting SF data from the database, 29.14% of the 1105 remaining recognition results are

classified as “error”. So without any parameters in the prediction model, we can predict the recognition error with an error rate of 29.14%. Any model that reduces this error rate can help to predict the recognition performance.
Inspired by the work of Hirschberg [10], we also adopt the rule-learning program “RIPPER” from Cohen [14] to generate plausible rules based on the parameters extracted from the audio signal that can be used to predict recognition performance.
It turns out that active speech duration and MCD are the most significant parameters. With the help of these parameters we can decrease the prediction error from 29.14% to 20.69%.

5.3 Relation of User Judgments and Parameters

A method to automatically predict usability judgments is provided by the PARADISE framework [15], in which a linear regression (LR) function is trained on interaction parameters with subjective judgments as targets. The potential of our parameters for such a prediction has been evaluated by correlating them with the questionnaire items acquired in the experiment. The strongest, however, still moderate relations with overall impression were found for channel-related signal parameters plus ASL ($|\rho|$=[0.25;0.31], p<0.01) and barge-ins (ρ=0.28, p<0.01), followed by recognizer labels ($|\rho|$=[0.14;0.22], p<0.05). Parameters describing confusability on the word level correlate weakly with $|\rho|$=[0.14;0.16], p<0.05.
For perceived system understanding, highest correlations were found with dialogue-related parameters and recognizer status labels ($|\rho|$=[0.22;0.46], p<0.01). Furthermore, phonetic similarities across concepts available in the system grammar are weakly correlated with the same item ($|\rho|$=[0.17;0.25], p<0.01). Ironically, signal-related parameters show no significant correlations with system understanding except the noise level ($|\rho|$=0.22, p<0.01).
An LR model has been trained on the z-transformed parameters (stepwise inclusion), predicting perceived system understanding with an accuracy of R^2_{adj}=0.354. Parameters included in the model are “correctly recognized” (-0.399), turns (0.277), noise level (0.202) and barge-ins (0.147). Training the function on overall impression leads to inclusion of “correctly recognized” (-0.277), ASL (-0.44) and rejection (0.217), describing the data with R^2_{adj}=0.225 accuracy.

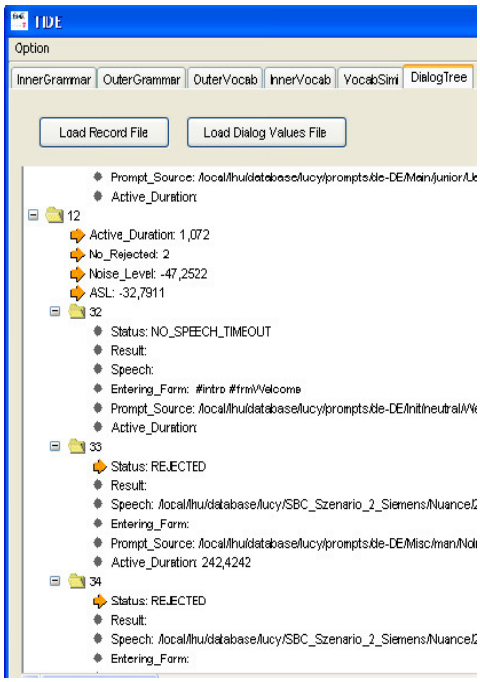

Figure 1 *TIDE dialogue browser*

6. Testbed Integration

The algorithms can be used in a graphical user interface. Figure 1 shows one tab of the user interface that allows a user to browse dialogs and find problematic dialogue turns, based on “RIPPER” rules derived from signal level parameters.
The other tabs of the interface allow for grammar analysis in terms of word/interpretation confusability derived from phonetic distance measures.

7. Conclusions and Future Work

For the experience of quality of voice dialogue systems the performance of the automatic speech recognition is of vital importance. Problems with speech recognition performance may have many causes. But at present complex, mixed initiative, natural-language-understanding, interruptible voice dialogs with very large and complex grammars (where often more than one grammar is active at the same time) make the analysis of ambiguity and confusion the most important issue. We described methods that are effective for analyzing and predicting such ambiguities. Furthermore we integrated some of these methods into a graphical user interface.

Speech recognition errors introduce wrong or conflicting information. The fact that the same word can have several meanings within one grammar, or within different grammars running at the same time, can have serious consequences. Those homonyms have to be tagged with different semantic meanings, depending on the spoken context.
Even more problematic are phonetic similarities between different keywords. They are hard to identify but can result in a collapse of the whole application. In short, developers and designers of dialogue systems should be aware of ambiguities. It is therefore essential to have a tool which alerts them and keeps track of all the potential problems before launching a system. The TIDE testbed which enables the designer and developer to automatically detect the potential problems can significantly shorten the nerve-racking and time-consuming test and error search periods. It thus contributes to a decrease of development costs, optimizes time-to-market, supports quality monitoring, and helps to increase system acceptance.

8. Acknowledgements

The described work was supported by the TIDE project funded by Deutsche Telekom AG. The authors would like to thank all colleagues who supported the run of the experiment and data annotation. The authors gratefully acknowledge Siemens AG, Corporate Technology, Competence Center "User Interface Design", for providing the project team with the usability test results.

References

[1] ITU-T Suppl. 24 to P-Series Rec., Parameters Describing the Interaction with Spoken Dialogue Systems, ITU, Geneva, 2005.

[2] ITU-T Rec. P.851, Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems, ITU, Geneva, 2003.

[3] ITU-T Rec. P.56, Objective Measurement of Active Speech Level, ITU, Geneva, 1993.

[4] ETSI ETS 300 040, European Digital Cellular Telecommunications System (Phase 1); Voice Activity Detection (GSM 06.32), ETSI, Sophia Antipolis, 1992.

[5] Martin, R., "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics", IEEE Trans. Speech and Audio Process. 9(5), 504-512, 2001.

[6] Falk, T.H., and Chan, W.-Y., "Single-Ended Speech Quality Measurement Using Machine Learning Methods", IEEE Trans. Audio Speech Language Process. 14(6):1935-1947, 2000.

[7] Möller, S., Quality of Telephone-Based Spoken Dialogue Systems, Springer, NY, 2005.

[8] ITU-T Rec. P.563, Single-ended Method for Objective Speech Quality Assessment in Narrow-band Telephony Applications, ITU-T, Geneva, 2004.

[9] Möller, S., Heimansberg, J., "Estimation of TTS Quality in Telephone Environments Using a Reference-free Quality Prediction Model", Proc. 2nd ISCA/DEGA Tutorial and Research Workshop on Perceptual Quality of Systems, Berlin, 56-60, 2006.

[10] Hirschberg, J., Litman, D., and Swerts, M., "Prosodic and Other Cues to Speech Recognition Failures", Speech Communication 43:155-175, 2004.

[11] Rabiner, L., "On the Use of Autocorrelation Analysis for Pitch Detection", IEEE Trans. Acoustics, Speech and Signal Process. 25(1):24-33, 1977.

[12] Pucher, M., Türk, A., Ajmera, J., Fecher, N., "Phonetic Distance Measures for Speech Recognition Vocabulary and Grammar Optimization", submitted to: Proc. 10th Int. Conf. on Spoken Language Processing (Interspeech 2007 – Eurospeech), Antwerp, 2007.

[13] NIST Speech Recognition Scoring Toolkit (SCTK) Version 2.2.1, National Institute of Standards and Technology, Gaithersburg MD, http://www.nist.gov/speech/ tools/index.htm.

[14] Cohen, W., "Learning trees and rules with set-valued features", 13th Conference of the American Association of Artificial Intelligence (AAAI), Portland, 709-716, 1996.

[15] Walker, M.A., Litman, D.J., Kamm, C.A. and Abella, A., "PARADISE: A Framework for Evaluating Spoken Dialogue Agents", Proc. ACL/EACL 35th Meeting, Madrid, 271-280, 1997.