STAIRS 2014
U. Endriss and J. Leite (Eds.)

doi:10.3233/978-1-61499-421-3-131

Human Speech Processing for Pedestrian Assistance: Towards Cognitive Error Handling in Spoken Dialogue Systems

Martin HACKER [a]
[a] *Interdisciplinary Center for Embedded Systems (ESI), Department of Computer Science, University of Erlangen-Nuremberg, Germany*

Abstract. Current spoken dialogue systems (SDS) often behave inappropriately as they do not feature the same capabilities to detect speech recognition errors and handle them adequately as is achieved in human conversation. Adopting human abilities to identify perception problems and strategies to recover from them would enable SDS to show more constructive and naturalistic behavior.

We investigated human error detection and error handling strategies within the context of a SDS for pedestrian assistance. The human behavior serves as a model for future algorithms that could yield reduced error rates in speech processing.

The results contribute to a better understanding which knowledge humans employ to build up interpretations from perceived words and establish their confidence in perception and interpretation. The findings provide useful input for SDS developers and enable researchers to estimate the potential benefit of future research avenues.

1. Introduction

Current spoken dialogue systems (SDS) often behave inappropriately when user utterances are incorrectly recognized by automatic speech recognition (ASR). This seems less to be a shortcoming of the acoustic processing but rather a consequence of the fact that SDS fail to detect and overcome these problems. In contrast, humans are able to employ a variety of different knowledge sources to estimate the reliability of hypotheses, interpret partially unreliable fragments and clarify missing or doubtful information during the subsequent course of the dialogue. Thus, to handle errors in the way that humans do would require SDS to implement the following functionalities [1]:

- FUNC1: establish the reliability of an ASR hypothesis and its constituents,
- FUNC2: build possible interpretations based on reliable parts of the hypothesis,
- FUNC3: choose an appropriate dialogue strategy to foster correct understanding, such as the *accept, ignore, clarify, reject* actions suggested in [2].

We believe that the challenge to achieve FUNC3 cannot be overcome without finding a solution for FUNC1/FUNC2. In particular, for the *clarify* strategy it is unclear what part of the transcript should be clarified and how the clarification request should be realized.

The knowledge sources employed by humans for reliability estimation (FUNC1) and interpretation (FUNC2) extend from phonetic and linguistic to situational [3] and common sense knowledge. To replicate human behavior in SDS, it is necessary to understand which information is utilized by humans in detail, and how this information is integrated during the interpretation process. These issues are subject to research in the fields of neuro- and psycholinguistics. On the other hand, we need to know how this information can be linked back and formalized for a particular SDS domain and whether it can in fact be beneficial for the performance of a SDS.

In this paper, we bridge this gap by considering a given dialogue system for pedestrian assistance and real user utterances collected with this system. By presenting ASR hypotheses in various forms to human subjects, we can activate cognitive processes and investigate human error handling behavior for these utterances. The study design is based on an experimental framework that overcomes methodological difficulties and provides experimental control over a variety of variables (cf. [1]).

The study aims at investigating the following research questions:

1. How well are humans doing in estimating the reliability of ASR hypotheses, depending on the type and amount of information provided by the experimenter? This investigation aims at evaluating which kind of information could contribute to improve error handling in SDS. By investigating the impact individual kinds of features have on the performance in human speech processing (HSP), we can estimate the potential benefit that incorporating such features in ASR may bring and establish if these are promising research avenues to explore.
2. Which criteria are applied to successfully estimate the reliability of hypotheses? How can such criteria be operationalized to improve confidence measures?
3. What strategies are applied by the subjects to interpret utterances that are believed to contain recognition errors? Which constituents and linguistic properties of erroneous transcripts are considered as reliable enough to use them as anchors for an interpretation? The insights are intended to inspire researchers to design novel formalisms for interpreting incomplete syntactical representations, and to be used to inform existing error-corrective mechanisms such as [4].
4. How confident are humans in their interpretations and how does this influence their choice of dialogue strategies? Such insights would help to establish a gold standard for naturalistic clarification strategies.
5. How can the human strategies be replicated algorithmically?

In this paper, we describe the study design and start to tackle some of these research questions. The paper is structured as follows: After summarizing related work and recapitulating the experimental framework, we introduce the dialogue system and the speech corpus the study is based on. Then we describe the experimental setting. We report and discuss the results of our analyses of reliability estimation performance as well as reliability criteria and interpretation strategies and conclude with an outlook on future work.

2. Related Work

Human error handling with respect to perception problems in SDS has been subject to several investigations before. Schlangen and Fernandez [5] simulated perception problems by introducing a noisy channel into human-human conversation. Single words in the audio signal were substituted by noise. The authors investigated which clarification requests were used by the subjects to retrieve the missing information. The results are valuable with regard to how naturalistic clarification requests can be generated. It though remains unclear, when explicit clarification should be preferred against other dialogue strategies. The choice of the dialogue strategy depends on the subject's interpretation of the corrupted utterance and his/her confidence therein – two variables that were not evaluated in the study. The experimental setting is targeted on non-perception instead of misperceptions that are prevalent in ASR. Applying the observed clarification strategies would require SDS to solve the tasks of FUNC1 and FUNC2 before in order to decide which words should be ignored and on which words the interpretation should be based on. A shortcoming of the use of the auditory channel is that experimenters loose control over the perception of the remaining words since it is unknown whether these words were correctly recognized by the subjects.

Skantze [6] applied a different method that resolves this shortcoming by substituting the noisy channel by an ASR module and visually presenting its output to the subjects that took on the role of the SDS operator. However, the study aimed at evaluating which dialogue stategies facilitated dialogue success after complete non-understanding.

Recent statistical approaches use POMDPs to optimize dialogue policies that allow to recover from misunderstanding [7]. The lack of understanding about why humans applied a particular strategy in the training corpus, however, still causes unnatural behavior.

Skantze and Edlund [8] ran a different experiment to establish a gold standard for ASR word error detection. The subjects were asked to correct ASR hypotheses that were shown together with varying extra information from the recognizer and the dialogue history. The human error detection performance was evaluated in terms of edit operations of correct and incorrect words. The results indicate that humans benefit from both contextual information and information provided by n-best recognition alternatives. As the study includes no qualitative analysis to establish why subjects decided to remove certain words from the transcripts, it remains unclear how the utilized information can be operationalized for their integration into SDS.

3. Study Design

3.1. Framework

For the study, we used the experimental framework described in [1]. The framework provides a method to control and evaluate the influence of different types of information on human speech processing and error handling. It overcomes a methodological difficulty that arises when trying to control the problem source of erroneous ASR transcripts, i.e. incorrect word recognition. Human word recognition is part of the more complex subconscious speech perception and understanding process that experimenters need to split up when trying to gain experimental control at the word recognition stage.

The framework proposes to preassign the results of word recognition by visually presenting ASR hypotheses to the subjects. In order to activate cognitive speech processing, the subjects are instructed to vocalize the given transcript in the head by subvocalization to envision the sound of the utterance without being biased by a concrete acoustic realization. Individual cognitive processes on different levels of perception can be activated and controlled separately by restraining the flow of information of other types:

- *Word recognition* can be controlled by presenting visual stimuli, consisting of single or competing word chains to vary the level of detail for phonetic information.
- *Pragmatic embedding* can be controlled by the amount of contextual information provided to the subjects.

- The *search process* underlying the interpretation of erroneous hypotheses can be controlled by introducing gaps to preassign the results of reliability estimation.

3.2. Dialogue Data

3.2.1. A Dialogue System for Pedestrian Assistance

We used a variant of the pedestrian assistance system ROSE [9] which offers a mixed-initiative spoken language interface. The systems provides the following functionalities:

- calculating routes with public transport connections,
- providing time-table information and live data about delays and cancelations,
- displaying outdoor as well as indoor maps for public transport station buildings,
- giving navigation instructions for pedestrian and public transport routes,
- recommending points of interest (POI) such as shops, cafes or ATMs.

3.2.2. Pedestrian Assistance Corpus (PAC)

For the experiments, we used a subset of a dialogue corpus collected with a Wizard-of-Oz variant of the above described system. The corpus contains utterances of 20 native German speakers (8f / 12m, age 16-68) with some of them speaking strong dialect. The participants were given up to 11 different tasks, resulting in 89 dialogues with an overall number of 544 on-talk user moves having an average length of 5.8 tokens (49% are of length 5 or longer). The tasks are ranging from information retrieval tasks such as public transport live timetable questions to more complex problem solving issues such as navigational assistance, re-planning or recommendation of nearby POI and activities.

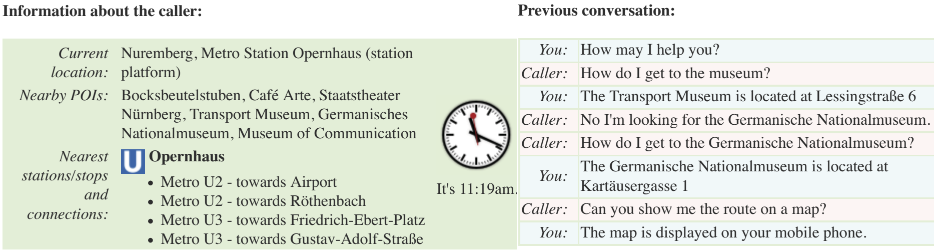
The language is characterized by a large amount of named entities denoting public transport stations, line numbers and POI. The domain vocabulary extends to 7351 words 542 of which were actually used in the recordings.

Linguistic Annotation The recordings from the PAC were manually transcribed and annotated. Among other things, colloquial words, clitics, abortions, self-corrections and slips of the tongue were marked and spelling variants were collected. This enables us to ensure a very high quality of our algorithms for alignment and ASR performance evaluation. For example, the phrase “wie viel Minuten sind es bis zu der U-Bahn-Haltestell’ ” and the ASR result “wieviel minuten sinds bis zur u bahn haltestelle” can be matched as identical, whereas standard evaluation algorithms would assign up to 8 word errors.

Speech Recognition and Evaluation For speech recognition we used Google Speech API [10] with open vocabulary and standard language model and, for comparison, Sympalog’s[1] recognition engine SymRec [11] with a bigram language model on the domain vocabulary of 7351 words. The language model was configured with classes for situation-dependent entities such as numbers, stations or POIs. By doing this, we ensured that the language model does not adapt to the concrete tasks used for building up the corpus.

Subset Selection The subset of the corpus that was selected for the study contains 25 utterances for which the ASR transcript contains recognition errors. For 15 of these utterances, we used the output of the Google recognizer, while using the SymRec output for the 10 remaining. The utterances were selected randomly but balanced with respect to length, speaker, underlying task, word error rate (WER) and position within the dialogue.

[1] http://www.sympalog.de



Information about the caller:

Current location:	Nuremberg, Metro Station Opernhaus (station platform)
Nearby POIs:	Bocksbeutelstuben, Café Arte, Staatstheater Nürnberg, Transport Museum, Germanisches Nationalmuseum, Museum of Communication
Nearest stations/stops and connections:	U **Opernhaus** • Metro U2 - towards Airport • Metro U2 - towards Röthenbach • Metro U3 - towards Friedrich-Ebert-Platz • Metro U3 - towards Gustav-Adolf-Straße

It's 11:19am.

Previous conversation:

You:	How may I help you?
Caller:	How do I get to the museum?
You:	The Transport Museum is located at Lessingstraße 6
Caller:	No I'm looking for the Germanische Nationalmuseum.
Caller:	How do I get to the Germanische Nationalmuseum?
You:	The Germanische Nationalmuseum is located at Kartäusergasse 1
Caller:	Can you show me the route on a map?
You:	The map is displayed on your mobile phone.

Figure 1. Situational knowledge as provided to the subjects

Context Representation Besides the dialogue history that is implicit in the corpus, the recordings are aligned with the following information representing situational context:

- logical location of the user (i. e. name instead of geographic coordinates),
- nearby points of interest,
- nearby stations and public transport connections therefrom,
- current date and time,
- recommended route if the system did provide one in a previous dialogue step.

Figure 1 shows a sample context representation as has been shown to the subjects.

3.3. Experimental Setting

We decribe the experimental setting of a web-based study based on the above described framework (sec. 3.1) and data (sec. 3.2). The setting has been tested before in a pilot study with 5 participants and accounts for some feedback given by these test participants.

3.3.1. Preparing the Subjects

The participants are instructed to imagine that they work as an operator of a phone hotline, assisting pedestrian customers that call from their mobile phone which makes it sometimes hard to understand what they say. This imaginary context is equivalent to the SDS application described above and is intended to help the subjects understand their task without being required to put themselves into a spoken dialogue system.

Before showing the questionnaires, the subjects are introduced into the information the hotline is intended for (cf. the functionalities of the SDS in section 3.2.1), followed by a thorough instruction about the course of the study and the information that will be provided in the questionnaires. To ensure that the subjects understand the instructions, an example task from another domain is shown (without a given solution to avoid bias).

3.3.2. Tasks

Each task to be done by the participants corresponds to one imaginary phone call and consists of two steps: In a first step, the subject imagines the given context by viewing a representation of the situation (cf. Figure 1) given as one of the variants in Table 1A.

In the second step, a stimulus is presented to the subject. The stimulus consists of a textual or acoustic variant of the ASR hypothesis. The variants used in the study are explained in Table 1B. The user has been instructed before to subvocalize the textual stimuli as proposed in the framework.

Table 1. Possible configurations of the tasks. **A**: context variants, **B**: stimulus variants.

A: *Context variant*	*Contextual information provided to the subject*
FULL_CONTEXT	complete context as depicted in Figure 1
FULL_DISCOURSE	only discourse history
LAST_DISCOURSE	only the last utterance from the discourse history
FULL_CALLER	only the situational information about the caller
REDUCED_CALLER	reduced situational information about the caller
NO_CONTEXT	no contextual information
MISLEADING	incorrect caller and discourse information

B: *Stimulus*	*Channel*	*Provided information*
SINGLE	visual	1-best ASR hypothesis
NBEST	visual	5-best ASR list
GAPS	visual	1-best ASR list with misrecognized words substituted by gaps (______)
AUDIO	auditive	original audio recording with misrecognized words substituted by noise
REFTRANS	visual	transcript without ASR errors
PSEUDO	visual	random word chain generated with the SymRec bigram language model

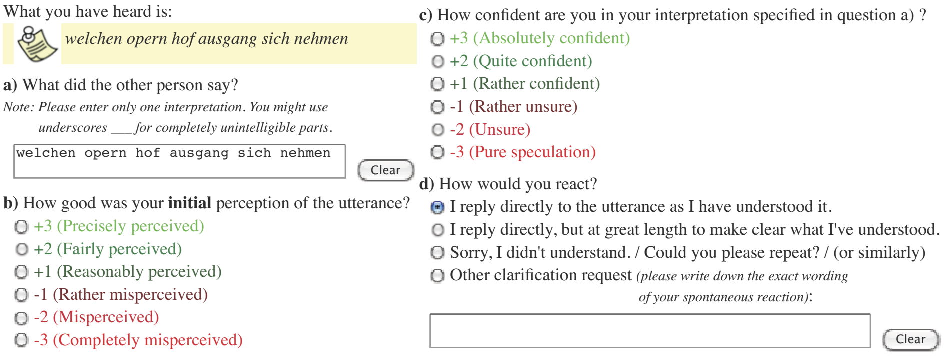
The presentation of the stimulus is followed by a questionnaire as depicted in Fig. 2. The user is asked to spontaneously interpret the stimulus, to estimate the reliability of the stimulus as perceived and to specify her confidence in the interpretation. The last question aims at evaluating the dialogue strategy the subject would choose as a response.

3.3.3. Participants and Configurations

The experiments were conducted with 36 human subjects (11w, 22m, 3 n/a) with age 19-69, mainly from the academic environment (students and university staff), each performing an individual subset of 9 tasks.

The subset of task configurations and underlying utterances per subject was balanced and rotated among the subjects in order to satisfy empirical standards. In particular, the following conditions are satisfied:

- The intentions of the speakers and the underlying situations (as well as the speakers themselves) are different for all tasks given to a subject.
- For every subject, the set of tasks is balanced with respect to sentence length and word error rate (WER) of the presented hypothesis.



What you have heard is:

welchen opern hof ausgang sich nehmen

a) What did the other person say?

Note: Please enter only one interpretation. You might use underscores ___ for completely unintelligible parts.

`welchen opern hof ausgang sich nehmen`

b) How good was your **initial** perception of the utterance?

○ +3 (Precisely perceived)
○ +2 (Fairly perceived)
○ +1 (Reasonably perceived)
○ -1 (Rather misperceived)
○ -2 (Misperceived)
○ -3 (Completely misperceived)

c) How confident are you in your interpretation specified in question a) ?

○ +3 (Absolutely confident)
○ +2 (Quite confident)
○ +1 (Rather confident)
○ -1 (Rather unsure)
○ -2 (Unsure)
○ -3 (Pure speculation)

d) How would you react?

● I reply directly to the utterance as I have understood it.
○ I reply directly, but at great length to make clear what I've understood.
○ Sorry, I didn't understand. / Could you please repeat? / (or similarly)
○ Other clarification request *(please write down the exact wording of your spontaneous reaction)*:

Figure 2. Example questionnaire for a misrecognized[2] utterance corresponding to the situation in Figure 1.

[2] Translation of the German hypothesis: *which opera yard exit oneself take*. The original utterance was: Welchen U-Bahnhof-Ausgang muss ich nehmen? (*Which metro station exit should I take?*)

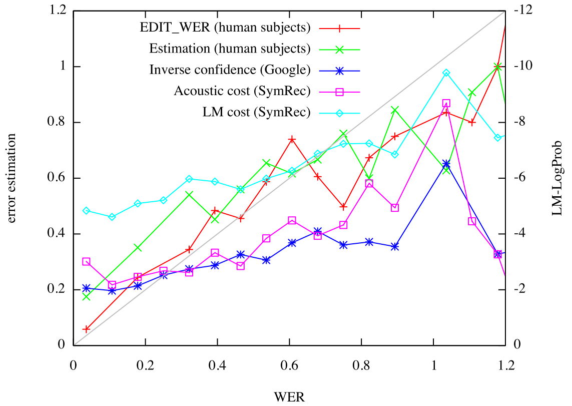


Figure 3. Correlation of the actual WER with human reliability estimation and ASR confidence values.

- For every subject, the task configurations (cf. Table 1) are balanced. Every subject is given 2 of each SINGLE/NBEST/GAPS/AUDIO tasks with 3 FULL_CONTEXT and 3 NO_CONTEXT variants and 2 of the partial context variants, and one additional task with one of the control configurations REFTRANS, PSEUDO or MISLEADING.
- The task ordering is rotated in order to avoid ordering bias.

3.3.4. Qualitative Analyses

At the end of the study, a separate questionnaire is shown where the participants are asked to reflect their answers. The tasks and answered questionnaires are shown again in read-only mode. The subjects are asked to describe what influenced their reliability estimation and what made them choose the specified interpretation.

The participants provided qualitative answers for 212 of the 287 completed tasks. These answers have proven to be extremely valuable during our analyses.

4. Results

4.1. Reliability Estimation

We investigated how well the subjects performed in estimating the correctness of the hypotheses. We excluded the auditory tasks, the tasks with misrecognitions marked (AUDIO/GAPS) and the tasks with misleading context from this analysis.

The estimated correctness can be calculated by normalizing the values the subjects used to indicate how well they perceived the utterance. Though the qualitative free text answers where subjects reflected and explained their decisions suggest that many subjects quantified the correctness of an hypothesis rather in terms of interpretability than in terms of word errors. With these observations in mind, we used the word-level edit distance of the subject's interpretation from the original utterance as alternative indication. From the mentioned edit distance, we can calculate a normalized score in a way analogous to word error rate (WER) calculation.

Figure 3 shows the mean correctness estimation for these two alternatives depending on the actual WER. The visualization suggests that the correlation for both alternatives is even higher as for the ASR confidence scores and the language model score that are also included in the figure. It should be mentioned that the human subjects – in contrast to the speech recognizers – had no access to acoustic information. Hence it can be assumed that the human subjects would perform even better if they could regard such information for their estimation – an assumption that is corroborated by the work of Skantze, who found out that human subjects benefit from ASR confidence information [8].

The human estimations in Figure 3 seem to oscillate compared to the computational measures. This is due to the fact that the latter were computed on the whole PAC corpus which contains about 20 times as many utterances as the subset used for the study[3].

The figure indicates that humans perform better in distinguishing completely correct hypotheses from those containing few errors. In the WER regions above 0.6, humans slightly underestimate the error rate, presumably because they start to be more creative in finding an interpretation.

To summarize, we can state that it is possible to reliably estimate the correctness of speech recognition output without auditory information only on the basis of linguistic and pragmatic knowledge.

4.2. Utilized Information

We used the qualitative answers (see section 3.3.4) to build up a grounded theory of the information utilized for reliability estimation. The resulting classes represent the most important knowledge sources and are listed in Table 2.

With the exception of L_1 and partially L_2, which can be covered by the language model to some degree, as well as A_2, which can be replicated by ASR confidence, the classes refer to knowledge sources that are either not covered by existing computational confidence measures or are considered by separate modules on higher levels of speech understanding without linking the information back to the recognition level. The hu-

[3] In contrast, the computational measures seem to be unreliable for word error rates above 1.0. This can be explained by the fact that there are only 6 utterances with such WER in the corpus while the study subset contained extra tasks of the type PSEUDO that show such high WER.

[4] The counts denote the number of cases where a corresponding influence has been mentioned as the most important reason, or has been mentioned at the first place.

Table 2. Classes of information used for reliability estimation.

ID	Description	Count[4]
P	*Pragmatic knowledge:*	
P_1	– Situational, domain and common sense knowledge	41
P_2	– Empathy, knowledge of or personal experience with the presumed speaker intention	9
I	*Interpretability:*	
I_1	– Interpretability based on important words	42
I_2	– Number of possible interpretations (nonambiguous;	11
	too many interpretations; missing contextual information to disambiguate)	11
C	*Completeness:*	
C_1	– Proportion of unreliable or unintelligible (gap tasks) words	15
C_2	– Missing important words	10
L	*Linguistic Knowledge:*	
L_1	– Grammar and syntax	23
L_2	– Combination of syntax and semantics	22
L_3	– Semantic coherence	15
A	*Auditory Information:*	
A_1	– Identification of unreliable words with high phonetic similarity to presumed target words	12
A_2	– Acoustic perceivability (listening tasks); phonetically confusable words (visual tasks)	12

man strategies reported below can inspire researchers to implement a tighter coupling of recognition and understanding modules, as can be found in human speech perception.

It would be of interest to quantify the benefit of the individual knowledge sources. The study design includes variables that enable us to control these sources (cf. Table 1). Our efforts to analyze the data in such a way have revealed that it would be beneficial to collect more data in order to allow for significant findings. We are currently planning to extend the study by recruiting additional subjects and using another subset of the corpus.

4.3. Interpretation Strategies

We analyzed the qualitative answers in which the subjects explained their interpretations. There seems to be a general behavior pattern for which we can find evidence in a large majority of the cases. This pattern can be outlined as the following 10-step strategy:

1. Identify reliable and important key words.
2. Try to capture the basic syntactic structure of the utterance.
3. Build possible interpretations based on these words, guided by situational and empathic expectations (P_1, P_2) or associations with the key words.
4. Try to assign missing information that is necessary to complete the interpretation.
5. If step 4 fails, specify the missing information in terms of syntactic and semantic categories.
6. Identify unreliable words (mainly by employing P, L_3, A_2).
7. Try to replace these words by phonetically similar words, augmented by
   - the syntactic and semantic categories of the missing information,
   - the identified syntactic structure,
   - domain- and situation-specific vocabulary,
   - semantic coherence (associations with or logical relations to reliable words).
8. Delete unreliable words that cannot be substituted.
9. Try to verify the resulting hypothesis by
   - trying to form a complete sentence with further substitutions or insertions of function words,
   - re-assessing the plausibility of the utterance considering all details,
10. Establish the confidence in the interpretation by regarding its plausibility and the number of alternative plausible interpretations.

Step 1 and step 4 are related to two basic spoken language understanding technologies used in many SDS: *keyword/keyphrase spotting* and *slot filling*. Although these mechanisms are often regarded as "unintelligent", they seem to be cognitively adequate in some respects, particularly when they are combined with rich confidence estimation. Unlike SDS, humans supplement these strategies with several verification and grounding steps in which information is passed back to the lower levels of processing. Instead of simply ignoring the segments not covered by key phrases, humans still try to recognize the correct words augmented by the information gained in the higher levels. Only if this grounding is successful, the interpretation is considered as highly reliable.

We should be aware that human speech processing is a highly interactive process where information is exchanged in any direction at any time. This complex process cannot be reduced to a plain script as outlined above. Our analyses though suggest that the script might reflect the most important flow of information with respect to speech processing in a limited domain with comparatively simple user statements.

5. Summary

We presented a study that aimed at investigating human speech processing of incorrectly recognized utterances as a model for error handling in spoken dialogue systems. In this paper, we reported our analyses how well the subjects performed in estimating the correctness of the given speech recognition hypotheses, what kind of information their estimations rely on and what strategies they apply in order to interpret the utterances.

We state that it is possible for humans to reliably estimate the correctness of speech recognition output only by utilizing non-acoustic information. The analyses of reliability criteria can be a first step towards developing confidence measures that integrate rich information from various knowledge sources and levels of processing.

The strategies humans applied to interpret the corrupted hypotheses disclose a way how future automatic speech understanding technologies might be designed. A cognitively adequate approach would combine the typical keyphrase spotting strategy with a kind of "grounded slot filling".

There is a number of open research questions that need to be investigated before the human behavior can be replicated in dialogue systems. As a next step, we will evaluate how well the subjects performed in correcting the hypotheses. Further effort will be put into the investigation of the dialogue and clarification strategies used by the subjects in response to the interpretation task.

Acknowledgements We would like to thank the Embedded Systems Initiative (ESI, `http://www.esi-anwendungszentrum.de`) for funding and David Elsweiler for his valuable advice in various empirical issues.

References

[1] Martin Hacker, David Elsweiler, and Bernd Ludwig. Investigating human speech processing as a model for spoken dialogue systems: An experimental framework. In *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI 2010)*, pages 1137–1138. IOS Press, 2010.

[2] Malte Gabsdil and Oliver Lemon. Combining acoustic and pragmatic features to predict recognition performance in spoken dialogue systems. In *Proceedings of 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 344–351. ACL, 2004.

[3] Martin Hacker. Context-aware speech recognition in a robot navigation scenario. In *Proc. of 2nd Workshop on Context Aware Intelligent Assistance (CAIA 2011)*, pages 4–17. CEUR-WS.org, 2011.

[4] Bernd Ludwig and Martin Hacker. Why is this wrong? – Diagnosing erroneous speech recognizer output with a two phase parser. In *Proceedings of ECAI 2008*, pages 323–327. IOS Press, 2008.

[5] D. Schlangen and R. Fernández. Speaking through a noisy channel: experiments on inducing clarification behaviour in human-human dialogue. In *Proc. Interspeech 2007*, pages 1266–1269. ISCA, 2007.

[6] Gabriel Skantze. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication*, 45(3):325–341, 2005.

[7] Jason D. Williams and Steve Young. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422, 2007.

[8] Gabriel Skantze and Jens Edlund. Early error detection on word level. In *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*, 2004.

[9] Bernd Ludwig, Bjørn Zenker, and Jan Schrader. Recommendation of personalized routes with public transport connections. *Intell. Interactive Assistance and Mobile Multimedia Comp.*, pages 97–107, 2009.

[10] Mike Schuster. Speech recognition for mobile devices at Google. In *Proceedings of the 11th Pacific Rim International Conference on Trends in Artificial Intelligence*, PRICAI'10, pages 8–10. Springer, 2010.

[11] Elmar Nöth, Axel Horndasch, Florian Gallwitz, and Jürgen Haas. Experiences with commercial telephone-based dialogue systems. *it–Information Technology*, 46(6/2004):315–321, 2004.