

Investigating Human Speech Processing as a Model for Spoken Dialogue Systems: An Experimental Framework

Martin Hacker and David Elsweller and Bernd Ludwig¹

1 MOTIVATION AND BACKGROUND

Current spoken dialogue systems (SDS) suffer from a lack of user acceptance, which has been repeatedly reported to be a result of high error rates in speaker-independent automatic speech recognition (ASR). It seems unlikely that researchers will overcome this problem using current approaches as ASR performances have plateaued in the last decade. However, mishearing and failure to understand are not unique to ASR; they are phenomena ubiquitous in human conversation. Humans are, however, able to perceive speech to a very high level, but this has probably less to do with their acoustic skills than with their abilities to interpret fragmented or partially unreliable information and remove doubt through a clarification dialogue.

With this in mind, we believe that the major performance bottleneck in SDS is not ASR errors themselves, but rather a lack of capabilities to detect and deal with them. If SDS could incorporate similar error handling behaviour to those in human conversation this would be a major step towards more naturalistic interaction in SDS. This would require SDS to implement the following functionalities:

- F1: establish the reliability of an ASR hypothesis, of its parts and structure,
- F2: build possible interpretations based on the reliable information,
- F3: choose an appropriate dialogue strategy to foster correct understanding, e. g. from the *accept-ignore-clarify-reject* actions suggested in [1].

To replicate human behaviour in SDS, we need to understand some cognitive processes involved in human speech perception (HSP). This raises a number of research questions, examples of which are:

- Q1: Which non-acoustic features influence human reliability estimation?
- Q2: Which hypothesis parts are believed to be correctly recognised?
- Q3: If a listener regards a hypothesis as insufficiently reliable, what is his strategy to create further hypotheses from the available information?
- Q4: Which types of error handling strategies are used in which situation?
- Q5: Do the individual strategies facilitate or hinder dialogue success?

Q1 aims to establish the kinds of features – e. g. syntactic, semantic or contextual – which are helpful to achieve F1. By investigating their effect on HSP performance, we can estimate the potential benefit of future research avenues which can be inspired by insights gained from the investigation of Q2. Q3 relates to the problem of interpreting both lexically and semantically (F2). Finally, examining Q4-5 will provide a much more detailed and extensive set of actions than we currently have and help to establish criteria that determine which dialogue strategy should be chosen in which situation (F3).

Answering these kinds of questions would allow researchers to build SDS that detect errors in their ASR module and react in a constructive, naturalistic way. However, designing experiments to investigate individual HSP processes is challenging. We outline the reasons and describe an experimental framework to overcome the difficulties and to investigate a wide range of research questions.

2 RELATED WORK

Collaborative dialogue aims at aligning the mental states of dialogue participants. Conflicting alignments – often caused in SDS by ASR errors – can be resolved by clarifications. However, it remains unclear *when*, *what* and *how* to clarify – questions that correspond to F1-3. Recent approaches apply dialogue strategies learned from human dialogue corpora, attempting to implement F3 – how to clarify – without considering the need to achieve F2 – what to clarify. Poor understanding of why humans applied a particular strategy in a recorded situation leads to unpredictable system behaviour [6].

Researchers tried to tackle this by estimating the reliability of ASR transcripts (F1), e. g. [5], but recent proposals address the integration of new utterances as a multidimensional problem related to all levels of communication [3], corresponding to the interpreting role of F2 which serves as a link between reliability and dialogue strategy.

The studies in [5, 4] employ similar methods to those suggested here. In [5], subjects estimated which words in a given ASR hypothesis were erroneous. [4] investigated human reactions to non-understanding with subjects playing the role of operators in a SDS. The types of reactions that facilitated dialogue success were evaluated, but as only complete non-understanding was regarded, the set of three reaction types was much smaller than that we aim to establish. These studies investigated how humans achieve F1 and F3, but disregard the role of F2 as a link. The authors provide no qualitative analysis to establish why subjects displayed a particular behaviour.

Human clarification speech acts refer to subcategorization information from phrases with the situation taken into account [2]. Understanding such behaviour, however, requires the investigation of the underlying cognitive processes that align the well understood acoustic processing with information from other levels of communication.

3 A NEW EXPERIMENTAL FRAMEWORK

When investigating human mishearings, in the sense of erroneous word recognition hypotheses, two major methodological difficulties are encountered. It is not possible for the experimenter to observe and control HSP subprocesses, as word recognition or interpretation, as he sees only the end result of the complex process, nor is it possible for the subject to achieve this by means of introspection as speech perception is subconscious. The second problem is that it is difficult to understand or control contextual variables that may play a role. As these problems make investigations in natural situations impossible, we must create artificial environments to gain experimental control and ensure that the results transfer to natural environments.

As low-level acoustic and lexical processing has been extensively investigated with a view to improving ASR, we focus on the less considered higher levels of syntacto-semantic and pragmatic processing.

¹ Univ. of Erlangen-Nuremberg, Germany, email: martin.hacker@cs.fau.de

What you have heard is:
I wanna take a flight to January week
What did the other person say?
I wanna take a flight to ? this week
How good was your perception of the utterance?
+3 (precise) +2 (fair) +1 (ok) -1 (rather poor) -2 (poor) -3 (completely failed)
How would you react?
trust my interpretation
"Sorry, I didn't understand." (or similar)
other clarification request.
Where would you like to go this week?

What you have said is:
I wanna take a flight to Genoa this week
Response of the dialogue partner:
Where would you like to go this week?
How do you find this response?
+3 (highly facilitates conversation)
+2 (facilitates conversation)
+1 (rather facilitates conversation)
-1 (rather complicates conversation)
-2 (complicates conversation)
-3 (highly complicates conversation)

Figure 1. (l) Example task from a questionnaire, (r) Investigating the outcome of dialogue strategies with a separate questionnaire

These considerations led to the following **Transcript Paradigm**:

- By visually presenting speech recognition hypotheses, acoustic information can be suppressed and the results of word recognition preassigned².
- The higher levels of HSP are activated by the cognitive reconstruction of the acoustic signal: The subjects are instructed to read the transcripts silently while vocalizing them in the head.
- By restraining the access to information of certain types we can activate and constrain individual subprocesses separately.

Based on this paradigm, our framework provides methodical guidelines for studies to investigate research questions of the kind outlined. HSP can be compared with a specific SDS, using user turns that are transcribed manually and automatically by the ASR of the SDS.

Preparing the participants Subjects should be asked to imagine a situation where they are in control of a device while another person expresses their wishes. They should be informed of the device type, i. e. SDS domain, and told that communication takes place in a noisy environment. This ensures that the tasks, as interpreted by the subjects, are comparable with the automatic speech processing task³.

The participants will then be assigned a series of tasks, each corresponding to one user turn. To investigate how HSP depends on situational knowledge, the subject could be primed with such information at the start of a task, e. g. from the discourse history or the device state. We suggest to systematically vary the amount of information provided or even display misleading information in some tasks.

Activating cognitive processing The subjects will then be exposed to a particular stimulus to activate cognitive processing. The stimuli can be of various types, allowing experimental control over further cognitive subprocesses aside from situational embedding, as postulated in Section 3. To give an example, for the user turn

I wanna take a flight to Genoa this week

the corresponding ASR output could be presented as a word chain

I wanna take a flight to January week,

as a word chain with gaps (substituting words by underscores)

I wanna take a flight to _____ week,

or as a list of n best word chains:

1: *I wanna take a flight to January week*

2: *I wanna make a flight to gender are this week*

3: *I want a take alright to chain they are this week*

These variants allow to control the search process for alternative hypotheses by hiding or adding word hypotheses, to investigate which parts of an utterance are crucial for understanding and to simulate recognition results with very low confidence at some points.

² As the suppressed prosody is likely to influence not only word recognition, but also the high-level processes to be investigated, this influence needs to be evaluated to ensure transferability to natural environments. Our framework allows to do this by inserting listening tasks and controlling prosody.

³ Domain knowledge induces vague presuppositions on possible words, constructions and requested actions – analogous to SDS that base their interpretations on specialised language models and limited sets of system operations.

A baseline can be established by inserting control tasks, using a sequence of random words as a pseudo hypothesis:

month I airport please ticket

Instead of visual stimuli, experimenters might wish to use acoustic stimuli in some tasks by presenting the original speech or a text-to-speech (TTS) synthesis of the human transcript. Gaps can be introduced by substituting the corresponding time interval by noise. By comparing results from the visual, original and TTS versions, the influence of acoustic and prosodic information can be evaluated.

The tasks After perceiving the stimulus, subjects are given a task such as in Fig. 1 (l). They should be asked to

- intuitively guess what was said originally,
- estimate on a 6-point scale how good their perception was,
- indicate how they would react in natural human conversation: trusting their interpretation of the perceived signal or giving some feedback to the dialogue partner (the exact wording needs to be transcribed).

The information retrieved in a) can be used to answer Q3 which is about interpretation and Q2 as the words which the subjects rely on will be reused in the interpretation. The responses from b) can be evaluated to answer Q1. Finally, the use of dialogue strategies as Q4 is concerned with can be examined in the data obtained from c).

Qualitative analyses In order to facilitate qualitative analyses, we suggest a follow-up survey where subjects would be encouraged to reinspect their answers from the questionnaire and reflect how they decided to give particular answers, i. e. why they regarded a particular hypothesis as unreliable or how they established their interpretation.

Evaluating the dialogue strategies To evaluate the outcome of the clarifications from c) (pursuing Q5), the subsequent dialogue can be simulated by a separate questionnaire presented to additional subjects: Thereby the correct transcript should be displayed, followed by the clarification to be examined. Without knowing the ASR hypothesis the listener had to deal with, subjects quantify how much the clarification facilitates or complicates the conversation (Fig. 1 (r)).

4 SUMMARY & FUTURE WORK

We have described an experimental framework to investigate human speech processing and error handling from a computer science perspective. It provides experimental control over many variables in order to overcome methodological difficulties and to obtain results that can be related to applications in ASR and SDS. We are in the process of preparing a web-based study within the framework using German user turns from four domains (TV, household, robotics, pedestrian navigation). Most of the evaluation work can be automatised, which allows a large number of participants. We can demonstrate that a wide range of research questions can be investigated that are of high relevance for the development of future spoken dialogue technology.

REFERENCES

- [1] M. Gabsdil and O. Lemon, 'Combining acoustic and pragmatic features to predict recognition performance in spoken dialogue systems', in *Proceedings of ACL 2004*, pp. 344–351, Barcelona, (2004).
- [2] M. Purver, *The Theory and Use of Clarification Requests in Dialogue*, Ph.D. dissertation, King's College, University of London, 2004.
- [3] A. Roque and D. Traum, 'Improving a virtual human using a model of degrees of grounding', in *Proc. of IJCAI 2009*, pp. 1537–1542, (2009).
- [4] G. Skantze, 'Exploring human error recovery strategies: Implications for SDS', *Speech Communication*, **45**(3), 325–341, (2005).
- [5] G. Skantze and J. Edlund, 'Early error detection on word level', in *Proc. of ITRW on Robustness Issues in Convers. Interaction*, Norwich, (2004).
- [6] J. D. Williams, 'The best of both worlds: Unifying conventional dialog systems and POMDPs', in *Proceedings of Interspeech*, Brisbane, (2008).