Robust Conversational System Design

Lou Boves

Department of Language & Speech
University of Nijmegen, the Netherlands
L.Boves@let.ru.nl

Abstract

Robustness of conversational systems is a multifaceted issue that involves factors such as the quality and robustness of the ASR module, but also the capabilities of the dialog manager and the interaction design. The techniques for output rendering also play a major role.
Furthermore, "design" per se has at least two different meanings, viz. the resulting system, but also the process that was used to produce that system. For the eventual system to be robust it is essential that the design process be user centered.

1. Introduction

It is not so easy to define what we mean by a 'robust conversational system', in such a way that the definition provides a basis for evaluating the robustness of a given system, or to design novel systems so that they will be robust. One reason why such a definition is not simple is that there are a large number of different types of conversations. In recent research we have investigated chats between friends and relatives [1, 2, 3] and interactions between uninformed subjects and an automatic system that supports (re-)design of bathrooms [4]. Unsurprisingly, the differences between human-human and human-system conversations is very large, but we have also found substantial differences between human-human conversations over the telephone and face-to-face.
The single most important reason for us to investigate human-human and human-system interactions in parallel is that we do not understand human-human interaction sufficiently to know why these interactions seem to be so smooth. Moreover, we think that 'robust conversational systems' should be able to entertain smooth interactions. We are convinced that only if we understand how people avoid misunderstandings and how they repair these if the do occur nevertheless, can we hope to be able to build systems that feel 'robust' when we use them.
One might argue that human-system conversation can be made robust despite that fact that the system falls short of human communicative competence, because the boundary conditions of the task help to avoid many, and perhaps most of the problems that occur in human-human conversations, and that therefore human-like competence is not needed for automatic systems. Moreover, research projects have demonstrated amazingly smooth interactions with systems that evidently lack capabilities without which adults would probably not be considered 'normal'. However, we strongly feel that these demonstrations are misleading, because they only show what is possible when the system designer, who knows exactly what her or his system is capable of doing, uses the system. In our own research we have always stressed that systems should also be able to entertain smooth interactions with uninformed users, who have no means of finding out the exact capabilities and limitations of a system. And perhaps even more importantly: who are not motivated to try and reverse engineer the systems that they need to use to accomplish daily tasks and chores.
From our experience with conversational systems it has appeared that unexpected user actions, triggered by system behavior that surprised the user, is probably the single most important cause of problems in human-system interaction. This does not only hold for conversational systems, but also for direct manipulation systems. Of course it is true that speech recognition errors are one of the major causes of unexpected system behavior. However, from our experience with systems that try to handle more complex applications than form filling it has appeared that ASR (and pen input) errors are certainly not the only source of communication problems. The more complex an application becomes, the larger is the role of (artificial) intelligence, expressed in the form of dialog management, supported by domain knowledge, user models, etc.
In this paper we first present a summary of recent results of investigations of human-human conversations. We believe that a better understanding of what goes on in those conversations will help building artificial conversational systems that can qualify as 'robust'. In the following section we briefly address the issue of robustness in ASR, and then we proceed to a summary of recent findings from experiments with multimodal human-system interaction.

2. Human-Human Conversations

Over the last couple of years we have spent substantial time and effort in building a large corpus of standard Dutch, spoken in a wide range of communicative setting, known as the Spoken Dutch Corpus, or CGN [5]. The CGN includes face-to-face and telephone conversations between friends and relatives. Already during the production of the CGN corpus, we have used the recorded and annotated conversations for our research.

2.1. Multiword expressions

We analyzed the transcriptions of the conversations (3.3 M words in total) for the presence of multiword expressions, primarily to investigate whether multiword expressions show pronunciation variation that differs from the variation that can be observed in sequences of words that do not follow each other with high frequency. The obvious long-term goal of this enterprise is to improve recognition performance of ASR systems, which are well known to suffer dramatically from the kind of pronunciation variation in conversational speech. In addition, previous research [6, 7] had shown that the addition of multiwords to the ASR lexicon does help to

increase performance. For practical reasons we limited the search for multiword expressions to sequences of between 3 and 6 words. In making the first inventory, we did not distinguish between lexicalized expressions (expressions of which the meaning can not simply be deduced from the meaning of the individual words) and sequences of words that happen to occur with sufficiently high frequency.

We found that slightly less than 3500 multiword expressions covered 21% of all words in the transcriptions. Apparently, conversational speech is to a very large extent predictable. In selecting multiword expressions we discarded all word sequences that comprised disfluencies, such as hesitations, filled pauses or repetitions. Therefore, we claim that it is quite likely that humans store these expressions as full units, and that they therefore have a status similar to 'words'. For an automatic system it would be extremely helpful to know these expressions, instead of trying to recognize these as a sequence of words.

Part of the CGN corpus comes with human-made broad phonetic transcriptions. We used the transcribed part of the spontaneous conversations to analyze pronunciation variation in multiword expressions. We found a difference between the type and amount of pronunciation variation between word sequences that qualify as 'lexicalized' and word sequences that happen to occur with high frequency for some other reasons. Specifically, the type and number of phone deletions is rather high in 'true' multiword expressions. The difference is large enough to motivate a different treatment of multiword expressions in developing procedures for dealing with pronunciation variation in conversational speech.

2.2. Unintelligible speech

The transcripts of the spontaneous conversations in CGN contain a substantial number of '*xxx*' codes, which stand for speech that the transcriber could not understand. Although a formal analysis of these situations remains to be performed, the results of the work on multiword expressions suggests that only a small proportion of these unintelligible intervals elicits 'say that again, please' replies from the interlocutor. This can mean to things: either the speakers, who are familiar with each other, have much less difficulty understanding each other than a third person, or the fact that one does not understand the interlocutor completely does not always affect the communication to such an extent that a repair meta-dialog is called for.

For automatic conversational systems both explanations imply problems. The familiarity issue can be solved to some extent if systems can be personalized. In that case, the system can try to adapt to one (or a small number) of speakers in a way similar to what is presently done in dictation systems. However, there are as yet no known techniques to automatically learn the kind of pronunciation variation that we have observed in human-human interaction. For a system to understand that it did miss part of the input, and continue the interaction without explicit repair actions requires dialog management technology and artificial intelligence well beyond what is presently available. In any case, ASR systems that are used as a module in a robust conversational system must be able to skip over unintelligible portions of the input speech, without affecting the recognition of the intelligible speech surrounding the unintelligible intervals.

2.3. Turn taking

We have made an attempt to analyze turn taking behavior in the face-to-face and telephone conversations. The CGN corpus supports this kind of research because it comes with transcriptions of all speakers on an individual tier, and with time markers that indicate the position of all word boundaries. Part of the boundary markers have been checked and corrected manually; for another part of the corpus these markers have been computed automatically on the basis of a forced alignment between the speech and the transcription.

It soon appeared that it was extremely difficult to divide the conversations in turns. One problem that is difficult to solve is to decide whether short utterances ('o.k', 'yes', 'no', 'mm', 'true', 'indeed', etc. but also some multiword expressions) should be considered as a true turn, or whether they actually function as back channels 2, 3]. In addition, we have observed a very substantial amount of overlap between the speech of the interlocutors. Here too, it appears to be difficult to decide whether overlapping speech is an attempt of one speaker to seize the floor while the other is not yet willing to yield, or whether the syntactic and prosodic features of the ongoing speech allow the interlocutor to predict the end of the present turn, so that it would be 'formally appropriate' to take the floor. To complicate things further, it is possible for short utterances such as 'yes' or 'no' and their likes to function as an answer to an explicit yes/no question, while the speaker who asked the question expected to keep the floor. In other words, it is possible to elicit responses without yielding the turn (or at least without intending to do so).

Figure 1, taken from [2], shows histograms of the time interval between the end of one turn and the start of the next one (called the Floor Transfer Offset) in a part of the CGN conversations that were analyzed in more detail.

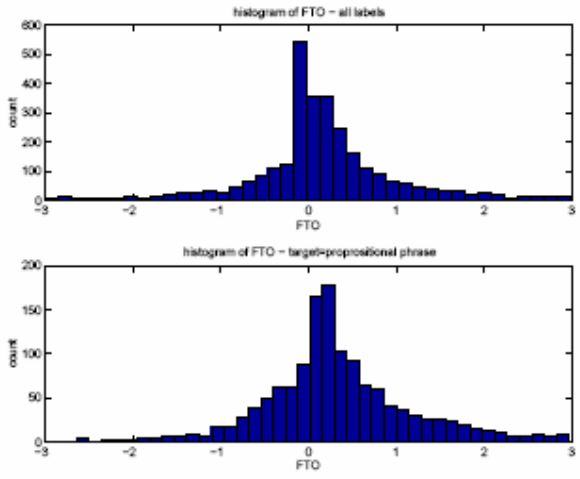

Figure 1: Floor Transfer Offset for all turn changes (upper panel, 3046 turns) and for 'true' turn changes only (lower panel, 1398 turns).

From Figure 1 it is clear that the negative modal time interval between successive turns is due to back channels. However, it is also clear from the lower panel that even for true turn changes a large proportion of overlap between the speakers is observed.

2.3.1. Backchannels, interrupts and echo cancellation

If we assume that a robust conversational system is a system that does not surprise and upset the user all the time by unexpected and unintuitive behavior, systems must be able to deal with interrupts and backchannel input from the user. Obviously, this implies that systems must come with the echo cancellation capabilities that are needed to allow interrupts and backchannel input. However, it may also be necessary that systems are able to generate backchannel output, since people will expect them to do so. Presently, we have virtually no knowledge about the dialog contexts where backchannels are appropriate, or even mandatory.

3. Robustness in ASR

The concept of 'robustness' for ASR systems can also be defined along the lines set out above: a robust ASR module is one that does make occasional mistakes, but not in ways that cannot be understood and 'reverse engineered' by non-expert users. It is goes without saying that we are far from being able to build such modules.

It is also common knowledge that ASR performance is affected by a large number of factors, and here again it is true that many of those factors relate to the behavior of the user. But there is also the unsolved issue of speech recognition in noisy environments.

3.1. Noise robustness

Noise robust ASR has received considerable attention over the last decade or so, as testified by projects such as AURORA and SPINE [8, 9]. From our own research [11], but also from independent work elsewhere [10] we have reached the conclusion that bottom-up only processing of noisy signals will not suffice to break the performance barrier between human and automatic ASR performance. Therefore, we are now trying to initiate a new line of research, inspired by ideas about active perception [12].

3.2. Pronunciation variation

Pronunciation variation is another problem that must be solved in order to arrive at ASR systems that do not upset users with unexpected recognition errors. Substantial effort has been spent in attempts to model pronunciation variation in terms of phonemic representations of the speech signal. However, the results of these efforts have generally failed to live up to the expectations [6, 7].

Although our recent work on multiword expressions has shed some new light on the problem of pronunciation variation, we believe that the eventual solution of the problem will have to be found in using sub-word units the size of syllables or perhaps demi-syllables. In this context we are exploring the options that seem to be offered by episodic (example-based) representations of the sub-word models, together with novel search techniques [13, 14]. Although episodic models can, in principle, be implemented with techniques reminiscent of conventional HMMs, they represent a very different point of view in the ongoing debate between symbolic and sub-symbolic approaches and representations in cognitive science and artificial intelligence.

3.3. User behavior

It has been observed that users who are confronted with ASR errors, and who do not have non-speech methods to correct the errors, spontaneously adapt their articulation behavior. Unfortunately, they do so in the direction of over-articulation which is clearly counter-productive [14]. In our own research using various versions of a train timetable information system, we have found that ASR performance in re-speaking drops from 71% accuracy to 29% [15]. Although this finding must be considered with some caution, because the ASR task was deliberately made very difficult by choosing confusable station names, and because for obvious reasons the names that are repeated belong to the most difficult ones, it was still evident that users adapted their pronunciation in ways that did not help the recognizer at all.

The human tendency to speak more clearly in case of misrecognition is so strong that we cannot hope that people will learn to avoid this behavior in their interactions with an automatic system, even if it also appeared from our research that non-expert users learn very quickly to choose more effective error correction techniques if these are available [15]. Coping with over-articulation is closely related to handling the wide range of pronunciation variation observed in conversational speech, be it that it only helps to increase the range of different variants substantially. For the time being, it seems more effective to design conversational systems in such a manner that re-speaking is avoided. Fortunately, it has been possible to design multimodal services to accomplish just that [15, 16].

4. Experiments with multimodal interaction

We have recently concluded a couple of experiments with multimodal interaction in which we wanted to investigate the way in which non-expert user interact with these systems, and how they evaluate multimodal interaction compared to speech-only and GUI-based interaction [4, 15, 17].

4.1. Timetable information – form filling

Not very surprisingly, we have found that users do not appreciate speech-only interaction with an automatic system, although it must be added immediately that we did not include IVR interfaces with touch-tone only input in the comparison. There are a number of reasons why users prefer interaction that is supported by a graphical display of the information and of the status and progress of the dialog. Most of these reasons are related to cognitive load that inevitably is larger in a situation where all information must be remembered. Moreover, it may well be that there is also an effect of the difficulty for the system to make the user aware of its functionality and limitations if the interaction is confined to speech, although this will only become apparent if users cannot be expected to know and understand the functionality from previous experience, perhaps with other implementations of the same service.

Both in [4] and in [17] we found that users are not very accurate in estimating the time it takes to complete a dialog with an automatic system. However, it seems to that time to

task completion is not always a good a predictor for user satisfaction. In [17] we found that users over-estimate the time it takes to complete a task with a multimodal system if they have to wait for the output of the ASR system after every input utterance. While the objective time needed to complete the task was longer with a GUI than with a multimodal system, users though that interaction with the GUI was faster. The only way to explain this is by assuming that the time spent waiting for the ASR system is felt as longer than the time spent in interacting with the GUI.
From [17] it also appeared that for a simple service like time table information the effectiveness of the three interfaces (GUI, VUI and Multimodal) did not differ from each other. Specifically, the GUI interface did not outperform the two interfaces that relied on ASR for entering all (VUI) or most (MM) of the information.

4.2. Architectural design

In [4] we reported preliminary results from a comparison of a pen-plus-speech based system for bathroom design with a GUI system. Recently, we completed a larger scale experiment with the same systems. The central research question was whether non-expert subjects prefer direct manipulation or a conversational system for a task that they perform seldom, in a domain in which they lack expert knowledge.
Figure 2 gives an overview of the COMIC system that was used in the experiments described in [4].

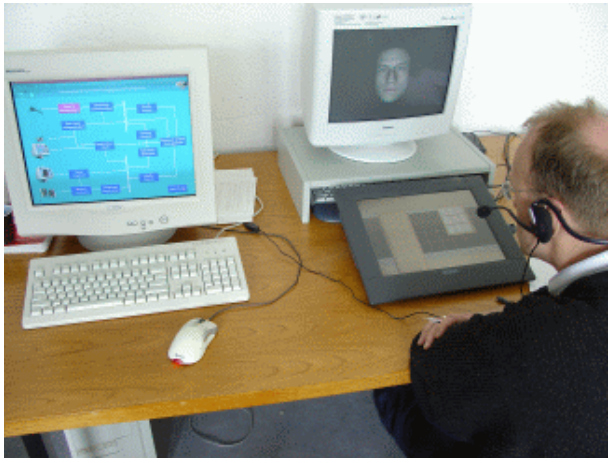

Figure 2: The COMIC system is shown on the right. The left screen shows the flow of active models for demonstration purposes.

We have found that it is not possible to predict whether non-expert users who have to perform a task in an unfamiliar domain prefer a conversational interface, that provides help and guidance, or a GUI, which leaves the user alone to find how the sub-tasks can be completed. We have seen substantial differences between subjects. Persons who had previous experience with using GUI systems for architectural design had much fewer problems in handling the direct manipulation system. At the same time it was clear that subjects who had little or no affinity with the domain appreciated the guidance from the conversational system. Another finding from this study is –unsurprisingly- that ASR performance has an enormous impact on the preference of the users for the input modality in the multimodal system.
The ASR module used in the COMIC system was a version of HTK, adapted to run as a module in a large architecture that supports multimodal interaction. One of the difficult issues that had to be tackled was turn taking. As is appears, turn taking in multimodal situations is much more complex than in speech-only dialogs, where it is already very complex (cf. section 2). If information can be exchanged in two or more independent and parallel channels, it is very difficult, if not impossible, to define 'turns' in the sense that has conventionally been used in models of spoken dialog (and machine communication). All the time both partners have the floor, at least in the sense that they can see each other and interpret the gestures made by their interlocutor. Moreover, it often appeared that the synchrony between the two input channels (speech and pen) was rather weak.
After the COMIC system was completely installed, and ready in the technical sense, it still took weeks of additional work to repair bugs and holes (in essentially all modules) that had never come to the surface during the first phase of system design and development. For ASR, most holes related to the vocabulary and the language model. Even if the task is relatively well described and 'small' (the users had to input the shape and size of a room, as well as the position of the window and the door – including the direction in which it opens) it remains difficult to predict how arbitrary users will formulate the information. Another major development effort was needed to adapt the dialog manager to user behavior that had never been used by the developers themselves. The experience with the development process has shown that the design process must involve the prospective users from the very beginning. This is especially important if the system under development is aimed at a user population that is (much) larger than the development team.
A central issue in the design of robust conversational multimodal systems is related to turn taking. Although we spent a substantial amount of effort to this issue, the turn taking protocol in the COMIC system is basically half duplex from the system's point of view: information provided by the user during the time the system is busy processing input and generating output is ignored. In the last version of the system we have tried to avoid turn taking problems by having the talking head display non-verbal information. The head is in a 'listening' mode as long as the system is willing to accept pen and speech input, and it switches to 'thinking' mode after it has detected an end-of-turn in the user's input. With few exceptions, this protocol was able to prevent major problems.
The smoothness of the interaction between the users and the COMIC system was affected by the performance of the HTK-based ASR system and the way in which it was integrated in the system architecture. Due to the limited amount of relevant training data, it was difficult to optimize the garbage model. At the same time, there was a considerable amount of out-of-vocabulary (and actually out-of-domain) speech, if only because the system did not always react in a way that was obvious to the users. But perhaps the most important problem that remains to be solved is the latency between the actual end of an input utterance and the visible or audible reaction of the system. We changed the HTK decoder in such a way that it can produce first-best output every 500 ms, but in

actual practice it appeared quite difficult to use the results, if only because it is not obvious how the Fusion and dialog management module can process preliminary and potentially incorrect ASR output, without consuming unacceptable amounts of CPU time.

5. Discussion and Conclusion

In this paper we have summarized the findings of our recent research in the field of conversational human-system interaction. Our experience shows that 'robustness' is a highly complex issue, that has as much to do with user behavior as with the technology used to build the system. Although it cannot be denied that 'raw' performance of the ASR (and the pen input recognition system) has a direct impact on the interaction, it is also evident that the impact of recognition errors on the interaction depends very much on the design of the interaction.

One conclusion that must be drawn from the experience gained in designing and building the COMIC system is that prospective users must be involved from a very early stage. Failing to do so will inevitable result in a system that may show impressive performance and behavior if it is demonstrated by its developers, but that breaks down dramatically as soon a person who does not know the functionality and limitation exactly. From experiments with form filling systems, such as the timetable information system used in part of our experiments, user centered design may seem less important, but that is a misinterpretation. If these systems can be designed successfully by a team of researchers, this is only because we now have a substantial knowledge about the ways in which users handle form filling applications. Moreover, users have experience in handling these applications as well. This lowers the thresholds at both sides of the interface. In designing applications for users who are not domain experts, re-use of existing design experience is much harder.

So far, we have not been able to prove the superiority of multimodal interaction over well designed unimodal systems. It appears that the open design and performance issues in multimodal interaction are still severe enough to annihilate the putative advantages. For multimodal systems to be really attractive the performance –both in terms of word error rate and response latency- of the ASR modules must be improved. Even then, we expect that the preference for multimodal systems will be limited to applications where conversational interaction has an inherent advantage in that it helps the user negotiating unknown territory.

6. Acknowledgements

This paper draws heavily on research carried out in two FP5 projects (SMADA, IST-1999-10667 and COMIC, IST-2001-32311). In addition, ample use has been made of research funded by the university and other Dutch funding agencies.

7. References

[1] Binnenpoorte, D., Cucchiarini, C., Strik, H. and Boves, L. ``Multiword expressions in Spoken Language: an exploratory study on pronunciation variation", Submitted to *Computer Speech and Language*, 2004.

[2] ten Bosch, L., Oostdijk, N., and de Ruiter, J.P. ``Turn-taking in social talk dialogues: temporal, formal and functional aspects'', *Proceedings SPECOM-2004*, 2004.

[3] ten Bosch, L., Oostdijk, N., and de Ruiter, J.P. ``Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues'', *Proceedings 7th International Conference on Text Speech and Dialogue*, Brno, Sept. 2004.

[4] den Os, E. and Boves, L. ``Natural multimodal interaction for design applications'', *Proceedings e-Challenges-2004*, Vienna, 2004.

[5] Oostdijk, N., et al. *Het Corpus Gesproken Nederlands*. Collection of papers about the Corpus Gesproken Nederlands. LOT Summer School, Netherlands Graduate School of Linguistics, 2002.

[6] Kessens J.M., Wester M., and Strik H. ``Improving the performance of a Dutch CSR by modelling within-word and cross-word pronunciation variation'', *Speech Communication*, 29 (2-4), p. 193-207, 2002.

[7] Kessens J.M., Cucchiarini C., and Strik H. ``A data-driven method for modeling pronunciation variation'', *Speech Communication* 40 (4), p. 517-534, 2003.

[8] Hirsch, H. G. and Pearce, D. ``The aurora experimental framework for the performance evaluations of speech recognition systems under noisy conditions'', *ICSA ITRW ASR2000*.

[9] Bilmes, J.A., Zweig, G., Richardson, T., Filali, K., Livescu, K., Xu, P., Jackson, K., Brandman, Y., Sandness, E., Holtz, E., Torres, J., Byrne, B. *Discriminatively Structured Graphical Models for Speech Recognition*, Final Report, JHU Summer Workshop, 2001.

[10] Barker, J., Cooke, M., en Ellis, D.P.W. Decoding speech in the presence of other sources http://www.dcs.shef.ac.uk/~martin/barker_crac_2002.pdf, 2002.

[11] de Wet, F. *Automatic speech recognition in adverse acoustic conditions*, PhD Thesis, University of Nijmegen, 2003.

[12] Bajcsy, R. ``Active perception'', *Proc. IEEE*, 76, pp. 996-1005, 1988. de Wachter, M., Demuynck, K., van Compernolle, D., Wambacq, P. ``Data Driven Example Based Continuous Speech Recognition'', *Proc. of Eurospeech-03*, Geneva, Switzerland: 1133-1136, 2003.

[13] Goldinger, S.D. ``Echoes of echoes? An episodic theory of lexical access.'', *Psychological Review*, 105: 251-279, 1998.

[14] Oviatt, S. and VanGent, R. "Error resolution during multimodal human-computer interaction". *Proceedings International Conference on Spoken Language Processing (ICSLP-96)*, pages 204-207, 1996.

[15] Sturm, J. and Boves, L. ''Effective error recovery strategies for multimodal form-filling applications'', to be published in *Speech Communication.*

[16] Oviatt, S. "Taming speech recognition errors within a multimodal interface". In: *Communications of the ACM*, vol. 43 (9), pages 45-51, 2000.

[17] Sturm, J. and Boves, L. `` Form-filling using a mobile device: pen, speech or both?'', submitted, 2004.