

Applications of Distributed Dialogue Systems: the KTH Connector

Jens Edlund & Anna Hjalmarsson

Centre for Speech Technology, KTH Sweden
{edlund, annah}@speech.kth.se

Abstract

We describe a spoken dialogue system domain: that of the personal secretary. This domain allows us to capitalise on the characteristics that make speech a unique interface; characteristics that humans use regularly, implicitly, and with remarkable ease. We present a prototype system – the KTH Connector – and highlight several dialogue research issues arising in the domain.

1. Introduction

Many of the commercial speech technology applications that have been successfully deployed share a common characteristic: they use the speech modality as an *alternative* to some other modality that is unavailable or impractical. Speech is a substitute for GUI menu systems on handheld computers in hands-free or eyes-free situations (e.g. voice dial, navigation). In telephony applications, speech is an alternative to touch tone menus in call routing (e.g. bank services). Speech also provides cheap telephone access to services that are also web based (e.g. many booking services). It is often unclear why speech would be a preferred modality in these systems – sometimes blatantly so, as in Example 1.

S: Welcome to XXX.
For current account information, say "one".
For mortgage information, say "two".
[...]

Example 1: Approximate transcript of an authentic UK bank service

Whereas the characteristics of spoken interfaces that these systems benefit from (e.g. hands-free and eyes-free operation) are oft-noted advantages of speech (e.g. [1]), the systems are not particularly successful in exploiting many of the more exciting characteristics commonly attributed to spoken interfaces. For instance, spoken interfaces are said to be easy to learn, since we already know how to speak; to provide great instruments for error handling, such as redundancy and grounding; to be flexible in that users do not need to study a new interface in order to move to a new domain; and to be responsive and efficient (e.g. [2], [3]).

The *alternative modality* application of speech technology may also be burdened by a drawback of another type. To be commercially viable, a product must either create revenues or cut costs. When substituting some existing means of interaction (e.g. a menu) for speech, we by definition provide an alternative to an *existing* service. Occasionally, this may cut costs by reducing the need for staff (e.g. some call routing applications). More frequently, it may allow a given service to be accessed more widely. Rarely, however, do these applications create altogether new revenues. Nevertheless, alternative modality applications are quite prevalent in speech technology.

The alternative modality applications are justified, not least within the field of accessibility. If, however, speech

technology is to create new revenues in new, uncharted areas, it is likely that we are going to have to present applications where speech is the *primary modality* – applications that *cannot* be created *without* a human language interface.

2. The domain

The domain described here is that of the telephone based personal assistant, or *secretary*. From the users' perspective, the domain is motivated by our constantly increasing availability – the extended use of mobile phones, e-mail, and wireless networks makes us potentially accessible wherever we go. However useful, the possibility of being contacted when on a train, in a meeting, or on holiday in another country is often a nuisance. We feel invaded by constantly ringing mobile phones and, as a result, turn them off and rely on our answer phones. Personal assistants, or secretaries, would alleviate the problem, but precious few can afford that luxury. Spoken dialogue systems in the secretary domain address this issue.

The actors in the domain are a telephony based secretary – the *system* (S); its owner – normally the intended *recipient* of a call (R); and an arbitrary number of people wishing to get in contact with S – the *callers* (C_n). The implementation described in this paper is part of CHIL, an Integrated Project (IP 506909) funded by the European Union under its 6th Framework Program, and has a somewhat more complex set of actors. In addition to reaching recipients on their phones or other traditional communication devices, a group of recipients can be reached through fixed communication equipment in CHIL enabled rooms (see Figure 1).

The secretary domain as described here bears similarity to the appointment negotiating domain of VERBMOBIL [4], but it differs in that the function of S is to be a filter between the user and the stream of incoming calls – although *incoming calls* may be extended to include text messages, e-mails, et cetera. It is the task of the system to establish intelligent communication links between people, connecting one or more persons in an appropriate manner at an appropriate time. The conversations involve multiple simultaneous dialogues between the system on the one hand and the recipient or the caller on the other. The system is activated either by the recipient opting to redirect calls, or by some caller calling the system directly on a separate number.

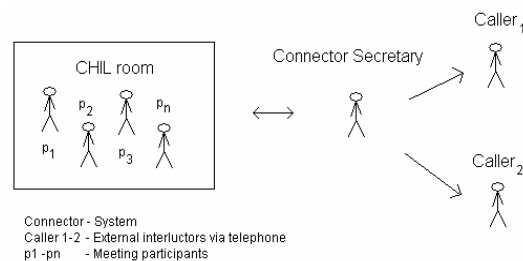


Figure 1: Schematic view of the dialogues in domain

In a simple scenario, a caller (C) wants to talk to the re-

ipient (R). The system (S) takes C’s name (unless it is already known by other means), calls R, and finds out whether this is a good time. If it is, S simply patches the two calls together after informing C. Otherwise, S may get information about how to proceed from R: that R will return the call; that C should call back after a certain time has passed; or that C should leave a message. Example 2 shows a dialogue snippet taken from the initial scenario document for this domain in the CHIL project [5].

In a more complex scenario S may find out what would be a suitable time for R and C to talk and inform both R and C that the call will be re-established by S at that time.

R:	[Occupied in a meeting, discussing the topic of the day vigorously]
C:	[calls R on the phone, gets redirected to CHIL]
S:	R is busy in a meeting at the moment. May I take a message?
C:	Tell R that I will pick her up after the meeting.
S:	I'll do that. Thank you for calling.
C:	Thanks

Example 2: Part of a simple dialogue

Another more complex scenario involving a meeting in a CHIL room is presented in Example 3, also from [5]. A video demonstrating another, similar scenario is available at the CHIL project web server (<http://chil.server.de/servlet/is/2765/>).

R:	[Occupied in a meeting, discussing some topic vigorously]
C:	[calls S]
S:	TMH Connector, how may I help you?
C:	Hello, this is Anna calling from Paris. I have an appointment with the CHIL meeting this morning.
S:	The meeting has just started and according to the agenda your participation is scheduled in ten minutes. Do you want me to call back or try to reschedule you?
C:	Which topic is being discussed right now?
S:	The current topic is emotion recognition.
C:	What will be discussed next?
S:	The spoken dialogue in the TMH Connector.
C:	Can you try to fit me in right before the next topic?
S:	Just a minute. [...]

Example 3: A more complex dialogue

As the examples above suggest, a very compelling characteristic of this domain is that although it can be implemented quite simply for prototyping and testing, it can also be expanded almost indefinitely. The complexity can be varied in a number of different ways, including:

- Dialogues may involve the system and the caller only, or may require the system to talk alternatively to the recipient and the caller.
- Complex systems may be able to suggest alternative actions if the caller’s request cannot be fulfilled, for example by offering the caller alternative channels to reach the recipient.
- The complexity of the communication between the system and the recipient may vary from simple text message notification of missed calls to full conversations.
- The system may be given access to meeting rooms (through for example loudspeakers) for multiparty dialogues, for example to give notification about latecomers.

3. The KTH Connector

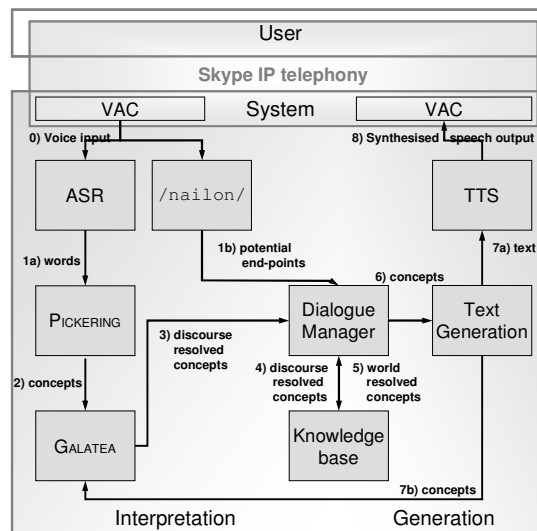


Figure 2: The KTH Connector: System overview

Figure 2 shows a system overview of the KTH Connector in its present incarnation. The prototype uses a third-party stochastic ASR, with interaction control (e.g. end-pointing) augmented by /nailon/ prosodic analysis [6]. For interpretation, discourse management, and text generation, we use components from the generic spoken dialogue system HIGGINS [7], for which the KTH Connector is one of the test implementations. Text-to-speech is currently done with MBROLA [8]. Dialogue management is performed by a custom component, as is the knowledge back-end. The latter, however, is currently being updated to utilise CHIL situation modelling. Finally, for the telephony we currently use a custom built solution on top of the Skype API [9] (which also provides a bridge to the standard telephone net) combined with VAC [10] to manage audio streams.

The current prototype of the KTH Connector speaks English only. The following is a simplified description of the processing of a dialogue turn in the system (cf. Figure 2). It is simplified in that in reality, the KTH Connector processes are incremental and work on smaller chunks than “utterances”.

- 0) Voice input: The user (be it a caller C or a potential recipient R) speaks. The signal is divided and sent to ASR and interaction control analysis (e.g. end-pointing) separately.
- 1a) The ASR results, consisting of words and word level confidence measures, are sent to Pickering for interpretation [11]. Currently, only the first hypothesis is used.
- 1b) /nailon/ sends an event every time the analysis shows that the present time may be an occasion for the system to speak.
- 2) The interpretation performed by Pickering results in deep semantic tree structures. Pickering uses only the context provided by the surrounding words themselves and the static world knowledge provided by grammars and vocabulary. The results are sent to Galatea for further processing [12].
- 3) Galatea is a discourse modeller – it performs interpretation of the Pickering results in dialogue, or discourse, context. Put simply, the resulting structures are similar to those produced by Pickering, but with references,

anaphora, and ellipsis resolved. The results are sent to the dialogue manager.

- 4-6) The dialogue manager maps the input semantics to one of a number of possible plans. It can consult the knowledge base as needed in order to interpret results in the context of the real world, or at least a model of relevant parts of it, such as where the recipient is at the moment. It then decides on an action, which in the case of speech acts is encoded as the same kind of deep semantic tree structure that describes the interpreted input. The dialogue manager uses the information from /nailon/ (1b) to time these utterances.
- 7a) The text generation is a prototype, which currently is able to build sentences from semantic structures according to a grammar, and to use referential expressions such as pronouns and ellipses in a generic manner. It produces text with some mark-up, which is sent to the TTS.
- 7b) As the text is sent to the TTS, the original structure is sent to Galatea so that the discourse model includes the system's utterances as well as the user's.
- 8) The synthesised system response is sent to the user and the turn is complete.

The current prototype handles several system-user dialogues concurrently by engaging in one dialogue at a time whilst the others are put on hold. Each dialogue is modelled separately as far as the discourse is concerned, but the dialogue manager, which is responsible for the planning and execution of any actions, has access to all dialogues. Example 4 illustrates what the current system can do.

The system has been tested in Wizard-of-Oz experiments where certain components had human supervision, and the sharp system was demonstrated live at the CHIL general assembly in Paris in June 2005, where the interaction in Example 4 took place. A first round of user experiments is under preparation presently.

C:	[Dials the Connector on his mobile phone]
S:	Welcome to the TMH Connector. Who am I talking to?
C:	This is Rolf – I want to talk to Jens.
S:	Jens is busy at the moment – can I take a message?
C:	No, this is important.
S:	Jens is in a CHIL meeting in Paris. Are you sure you want me to disturb him?
C:	Yes, I know about the meeting, but this really is important.
S:	OK, hold on for a moment.
S:	[Dials Jens' mobile phone]
R:	[Phone rings. Picks up, annoyed]
R:	Yes, this is Jens?
C:	I've got Rolf on the line. I have told him about the meeting, but he says it is urgent.
R:	OK, patch him through then.
C:	[Connects Rolf's and Jens' phone calls and retreats.]

Example 4: Live demonstration at the CHIL general assembly in Paris 2005 (approximate transcript)

4. Innovation and research

A number of interesting research issues emerge from the secretary domain. Some of these are of particular relevance for spoken dialogue systems where speech is the primary modality and where the most exciting features of spoken human-

human dialogue are exploited. The following is a listing of such issues. It is not exhaustive, nor does it attempt to be. Rather, it consists of issues currently addressed in the TMH Connector development.

4.1. Responsiveness

Human-human dialogue is typically very fast and responsive compared to the average spoken dialogue system. If a system such as the KTH Connector is to be successful in performing some of the duties of a real secretary, the issue of responsiveness must be addressed. Increased responsiveness can be achieved in a number of ways. For example, if the system is able to model feedback and backchannels, grounding and clarification will be more expeditious. Likewise, better timing of the system's dialogue contributions will make it more responsive. Such interaction control can be improved by the use of prosodic cues [6], and semantic as well as syntactic completeness has been shown to help [11], [13]. Combining these, and other, sources of information for interaction control purposes is likely to yield better results than using any one of them.

4.2. Incrementality

Incrementality is a closely related concept. A system with an architecture like the one described above will always be unresponsive until *end of utterance* is reached. In order to react faster, incoming speech must be processed incrementally [14]. Incremental processing is also instrumental to allow the system to cut its own utterances short if the user barges in, whilst still keeping track of what it *actually* said, not just of what it *intended* to say.

The different events in a responsive spoken dialogue system occur with different average frequency. Prosodic end-of-utterance cues occur roughly once per utterance; words are produced several times per utterance; and semantic completeness may not be reached even once per utterance. A responsive dialogue manager needs to be able to merge input that comes not only from different sources, but that comes at entirely different paces. To be responsive, a dialogue system may have to give reflex responses to prosody, for example. Such back-channels could be performed before the system has a clear understanding of the verbatim meaning of the input. Preliminary tests by the authors indicate that this may be a viable solution. However, the system then needs to know how to choose between conflicting (simultaneous) output.

4.3. Unobtrusiveness

In CHIL-enabled rooms, the KTH Connector may take the role of a meeting secretary. Situations will occur where the system needs to notify one or more of the meeting participants of something, such as one of the participants being late. A real secretary would perhaps 1) carefully open the door, 2) peek in, 3) wait for the right moment, and 4) quickly pass the message on. How an automatic system should perform actions 1) to 3) is an interesting issue. Animated agents or physical objects may help the system attract attention in an unobtrusive manner [15].

4.4. Adaptivity

Many of the users will use the system frequently and gradually become more accustomed to the system and its

features. A user with no system experience will likely need extra guiding. An experienced user, on the other hand, will probably experience such extra guiding as irritating and a waste of time. A system which only gives guidance when needed – when a lack of guidance leads to errors – will improve the system’s performance and usability. The system has access to the caller’s identity – either from the telephone number or by simply asking – which helps the system adapt to individual users.

4.5. Multi-party conversation

Although conversations in the secretary domain chiefly take place between the system and one user at a time, it is often necessary for the system to engage in several such dialogues in order to solve a task (e.g. Example 4). Several opportunities for innovative dialogue design arise from this. For example, a fully incremental system may be able to run some of these dialogues more or less simultaneously, asking partial questions in one dialogue whilst delivering partial answers in the other. Each dialogue may also take place not only in different languages, much like in VERBMOBIL [4], but also using different media – the system may for example speak to the user over a telephone line whilst contacting the recipient over a chat channel. Such a system should obviously also be able to talk about what media and languages are available, both to the system and for person-to-person communication.

4.6. World modelling

Special consideration must be taken when modelling the systems knowledge in the secretary domain. Whereas it is commonly enough to keep track of the state of the world – be it static train table data or more dynamic data – the secretary domain must also keep track of what data is available. It is important to model knowledge flexibly when access to context varies: the recipient may be in a CHIL-enabled room, in which case a lot of real-time information is available; the recipient may have provided the system with a lot of information previously; or the system may not have any information about the recipient’s whereabouts at all, and have to resort to calling the recipient to find out.

4.7. Evaluation

Methods for evaluation of spoken dialogue systems, often include *task success* and *user satisfaction* (c.f. PARADISE [16]) – metrics that are notoriously difficult (e.g. [17], [18]). In domains with more complex tasks, such as the one presented here, new issues arise. For example, it is not obvious from whose perspective task success and user satisfaction should be seen – the caller’s or the recipient’s? For instance, is a blocked call successful from the caller’s point of view if the errand could be delivered in another way or if information when to call back was provided, and moreover is it satisfactory? In addition to these questions, we need methods to merge success measures based on the caller’s wishes with those based on the recipient’s.

5. Acknowledgements

The research is supported by the EU project CHIL (IP506909). This research was carried out at the Centre for Speech Technology, a competence centre at KTH, supported

by VINNOVA (The Swedish Agency for Innovation Systems), KTH and participating Swedish companies and organisations.

6. References

- [1] Martin, T. B. (1976): “Practical applications of voice input to machines”, in *Proceedings of the IEEE*, 64 (4),
- [2] Cohen, P. (1992): “The role of natural language in a multimodal interface”, in *Proceedings of UIST ’92*, Monterey, CA
- [3] Cohen, P. & Oviatt, S. (1995): “The role of voice input from human-machine communication”, in *Proceedings of the National Academy of Sciences*, v. 92(22)
- [4] Wahlster, W. (ed.) (2000): “Verbmobil: Foundations of speech-to-speech translation”, Springer, Heidelberg, Germany
- [5] Edlund, J., & Hjalmarsson, A. (2005): Connector dialogue scenarios v. 2.0, CHIL Report, <http://www.speech.kth.se/~edlund/publications/>, last available 2005-09-26
- [6] Edlund, J., & Heldner, M. (2005): “Exploring prosody in interaction control”, in the thematic issue of *Phonetica* entitled *Progress in Experimental Phonology*
- [7] Edlund, J., Skantze, G., & Carlson, R. (2004): “Higgins – a spoken dialogue system for investigating error handling techniques”, in *Proceedings of ICSLP 2004*
- [8] MBROLA Project, <http://tcts.fpms.ac.be/synthesis/mbrola.html>, last available 2005-09-26
- [9] Hinrikus, T. (2005): Skype API v. 1.2, downloaded from http://share.skype.com/media/1.2_api_doc_en.pdf, last available 2005-09-26
- [10] Virtual Audio Cable 3.10, software
- [11] Skantze, G. & Edlund, J. (2004): “Robust interpretation in the Higgins spoken dialogue system”, in *Proceedings of ROBUST 2004*, Norwich, UK
- [12] Skantze, G. (2005): “GALATEA: A Discourse Modeller Supporting Concept-level Error Handling in Spoken Dialogue Systems”
- [13] Bell, L., Boye, J., & Gustafson, J. (2001): “Real-time Handling of Fragmented Utterances”, in *Proceedings of NAACL 2001*
- [14] Allen, J., Ferguson, G., & Stent, A. (2001): “An architecture for more realistic conversational systems”, in *Proceedings of IUI-01*, Santa Fe, NM
- [15] Marti, S. & Schmandt, C. (2005): “Physical embodiments for mobile communication agents”, in *Proceedings of UIST 2005*, Seattle, Washington
- [16] Walker, M., Kamm, C., and Litman, D. (2000): “Towards developing general models of usability with PARADISE. Natural Language Engineering”, *Special Issue on Best Practice in Spoken Dialogue systems*
- [17] Hjalmarsson, A. (2002): “Evaluating AdApt, a multimodal conversational dialogue system, using PARADISE”, M.Sc. Thesis, KTH Royal Institute of Technology, Stockholm
- [18] Larsen, LB (2003): “On the Usability of Spoken Dialogue Systems”, Ph.d. Thesis, Aalborg University