Eliciting interactional phenomena in human-human dialogues

Joakim Gustafson

KTH Speech Music & Hearing
jocke@speech.kth.se

Miray Merkes

KTH Speech Music & Hearing
miray@kth.se

Abstract

In order to build a dialogue system that can interact with humans in the same way as humans interact with each other, it is important to be able to collect conversational data. This paper introduces a dialogue recording method where an eavesdropping human operator sends instructions to the participants in an ongoing human-human task-oriented dialogue. The purpose of the instructions is to control the dialogue progression or to elicit interactional phenomena. The recordings were used to build a Swedish synthesis voice with disfluent diphones.

1 Background

Our research group have a long-standing interest in human conversational behaviour and a special interest in its mimicry and evaluation in spoken dialogue systems (Edlund et al., 2008). In human-human conversations both parties continuously and simultaneously contribute actively and interactively to the conversation. Listeners actively contribute by providing feedback during the other's speech, and speakers continuously monitor the reactions to their utterances (Clark, 1996). If spoken dialogue systems are to achieve the responsiveness and flexibility found in human-human interaction, it is essential that they process information incrementally and continuously rather than in turn sized chunks (Dohsaka & Shimazu, 1997, Skantze & Schlangen, 2009). These systems need to be able to stop speaking in different manners depending on whether it has finished what it planned to say or if it was interrupted mid-speech by the user. In order to be responsive, the system might also need to start talking before it has decided exactly what to say. In this case it has to be able to generate interactional cues that restrain the user from start speaking while the system plans the last part.

To date very few spoken dialogues systems can generate crucial and commonly used interactional cues. Adell et al. (2007) have developed a set of rules for synthesizing filled pauses and repetitions with PSOLA. Unit selection synthesizers are often used in dialogue systems, but a problem with these is that even though most databases have been carefully designed and read, they are not representative of "speech in use" (Campbell & Mokhiari, 2003). There are examples of synthesizers that have been trained on speech in use, like Sundaram & Narayanan (2003) that used a limited-domain dialogue corpus of transcribed human utterances as input for offline training of a machine learning system that could insert fillers and breathing at the appropriate places in new domain-related texts. However, these were synthesized with a unit selection voice that had been trained on lecture speech.

When modelling talk-in-use it is important to study representative data. The problem with studying real dialogues is that the interesting interactional phenomena often are sparsely occurring and very context dependent. When conducting research on spontaneous speech you have the option to use controlled or uncontrolled conditions. Anderson et al., 1991) recorded unscripted conversations in a map task exercise that had been carefully designed to elicit interactional phenomena. When using controlled conditions in a study you risk to manipulate the data, while in uncontrolled conditions there's a risk that the conversation goes out of hand which leads to a lot of unnecessary material (Bock, 1996). Bock suggests a set of eliciting methods to be used when studying disfluent speech. If the goal is to study speech errors and interruptions, a situation with two competing humans is useful. If the goal is to study hesitations and self-interruptions, distracting events can be used to disrupt the flow of speech.

This paper presents a new method for elicitation of interactional phenomena, with the goal of reducing the amount of necessary dialogue recordings. In this method an eavesdropping human operator sends instructions two subjects as they engage in a task-oriented dialogue. The purpose of these instructions is either to control the dialogue progression or to elicit certain interactional phenomena. The recordings from two sessions were used to build a synthesis voice with disfluent diphones. In a small synthesis study on generation of disfluent conversational utterances this voice was compared with a commercial Swedish diphone voice based on read speech. The subjects rated the created voice as more natural than the commercial voice.

2 Method

A dialogue collection environment has been developed that allows a human operator (Wizard) to eavesdrop an ongoing computer-mediated human-human conversation. It also allows the Wizard to send instructions to the interlocutors during their conversation, see Figure 1. The purpose of the instructions is to control the progression of the task-oriented dialogue and to elicit interactional phenomena, e.g. interruptions and hesitations. The Wizard has access to graphical and textual instructions. Graphical instructions are pictures that are manipulated or text labels that are changed. Textual instructions are scrolled in from the right at the bottom of the screen. They can be of three categories: *Emotional* instructions that tell the receiver to act emotional (e.g. act grumpy); *Task-related* instructions that require the receiver to initiate a certain subtasks (e.g. buy a red car); and *Dialogue flow related* instructions that tell the receiver to change his way of speaking, (e.g. speak fast, do not pause).

3 The pilot study

The DEAL system is a speech-enabled computer game currently under development, that will be used for conversational training for second language learners of Swedish (Hjalmarsson, 2008). In this system an embodied conversational character (ECA) acts as a shopkeeper in a flea trade-market and the user is a customer. The developed environment was adapted to the DEAL domain, and in a pilot study two human subjects were instructed to act as shopkeeper and customer. They were given written persona descriptions and were then placed in separate rooms. They interacted via a three-party

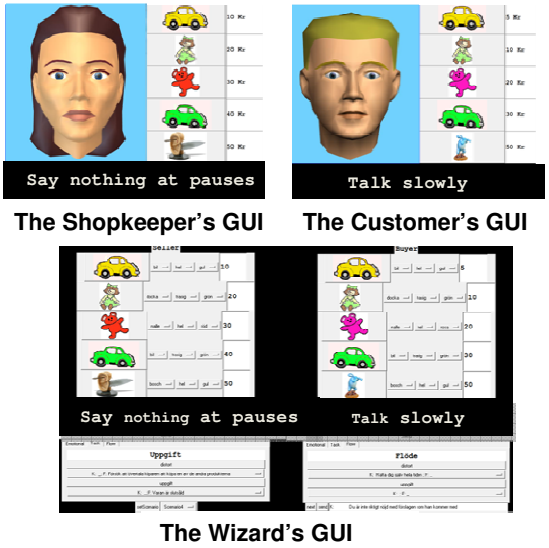

Figure 1. The GUIs used by the wizard and subjects.

Skype call, which allowed the Wizard to eavesdrop their conversation. In order to get a situation that was similar to the DEAL system, the subjects saw an avatar with lip movements driven by, and in synchrony with, the other subjects' speech. In order to achieve this, the SynFace system was used, which introduced a 200 ms delay in each direction (Beskow et al., 2004). Apart from the avatar the interfaces also contained pictures of objects currently for sale with accompanying prices, see Figure 1. At the bottom of the screen there was a black area where the subjects got the textual instructions from the Wizard.

The eavesdropping Wizard was placed in a third room, with an interface that allowed her to control the current set of objects and prices on the subjects' screens. The Wizard interface also contained an area for the textual instructions. In order to distort the dialogue flow some of the instructions involved sending instructions to both subjects at the same time. A main idea is to instruct one of the interlocutors to display a verbal behavior that will elicit interactional phenomena in the other dialogue partner's contributions. Table 1 shows some examples of the different types of textual instructions to the subjects and their intended effect on the shopkeeper party in an ongoing conversation. The Wizard interface also gave access to automated instructions that follows a pre-scripted manuscript in order to facilitate consistent instructions across different sessions. This also made it possible to transmit multiple successive instructions with high speed and a minimum risk of mistakes.

Shopkeeper reaction	Graphical	Emotional	Task related	Dialog flow related
Hesitation	Show an ambiguous picture (S)	Be wining and talk about how unfair life is (S)	Sell blue car (S) Buy red car (C)	Talk slowly (S) Say nothing at pauses (C)
Interruption	Change picture in mid speech (S)	Be a annoying customer (C)	Tell your price (S) Tell your price (C)	Speak without pauses (S) Try to speak all the time (C)
Change of sub-task	Show a picture (S)	Discuss the advantages of a certain item (S)	Sell the red car (S)	Ask a lot of questions (C) Answer with questions (S)

Table 1. Examples of instruction types and their intended reaction in the shopkeeper's subsequent turn(s). The receiver of the instruction is indicated by S (Shopkeeper) and C (Customer).

4 The effect of the Wizard's instruction

Two half-hour conversations were recorded where the same male subject (acting as shopkeeper) interacted with two different female subjects (acting as customers). The audio recordings were synchronized with the instructions that had been submitted by the Wizard during the conversation. The effects of the instructions were analyzed by inspecting both subjects' turns following an instruction from the Wizard. The analysis was focused on the disruptive effect of the instructions, and it showed that they often lead to turns that contained hesitations, interruptions and pauses. The task-related instructions lead to disfluent speech in half of the succeeding turns, while the dialogue flow related instructions, the emotional instructions and the graphical instructions led to disfluent turns in two thirds of the cases. The analysis of the instructions' effect on the disfluency rates revealed that the ones that changed the task while the subjects talked were very efficient, e.g. changing the price while it was discussed. The effect on the disfluency rates was most substantial when contradictive instructions were given to both subjects at the same time.

In order to get a baseline of disfluency rates in human-human dialogues in the current domain, the dialogue data was compared with data recorded in a previous DEAL recording. In this study 8 dialogues were recorded where two subjects role-played as a shopkeeper and a customer, but without the controlling Wizard used in the present study (Hjalmarsson, 2008). In these recordings approximately one third of the turns contained disfluent speech. This indicates that the disfluency rates found after the instructions in the current study are a higher than in the previous DEAL recording. Finally we analyzed the effect of the instructions on the dialogue progression. The instructions were very helpful in keeping the discussion going and the task oriented instructions provided useful guidance to the subjects in their role-playing.

5 A speech synthesis experiment

In a second experiment the goal was to evaluate two methods for collecting conversational data for building a corpus-based conversational speech synthesizer: collecting a controlled human-human role-playing dialogue or a recording a human that reads a dialogue transcription with tags for interruptions and hesitations. In this experiment the recordings of the male subject that acted as shopkeeper were used. 20 of his utterances that contained hesitations, interruptions and planned pauses were selected. New versions of these utterances were created, where the disruptions were removed. In order to verify that the disruptive sections could be synthesized in new places a set of test sentences were constructed that included their immediate contexts. Finally, new versions of the new test sentences were created, that had added tags for disruptions. All types of utterances were read by the original male speaker. Both the original dialogue recordings and the read utterances were phonetically transcribed and aligned in order to build a small diphone voice with the EXPROS tool (Gustafson & Edlund, 2008). This diphone voice contained fillers, truncated phonemes and audible breathing.

All types of utterances were re-synthesized with the newly created voice and with a Swedish commercial diphone voice that was trained on clear read speech. While re-synthesizing the original recordings all prosodic features (pitch, duration and loudness) were kept. The main difference between the two voices was the voice quality: the commercial voice is trained on clear read speech, while the new voice was created from the dialogue recordings contains both reduced and truncated diphones.

Secondly, a number of utterances were synthesized, where disfluent sections were inserted into fluently read sentences. For both voices the disfluent sections' original pitch, duration and loudness were kept. As in the previous case the main differ-

ence between the two cases is that the newly created also made use of its disfluent diphones. The disfluent sections were either taken from the original dialogue recordings or from the set of read sentences with tags for disfluencies.

6 Preliminary synthesis evaluation

16 subjects participated in a listening test, where they were told to focus on the disrupted parts of the utterances. They were instructed to indicate when they could detect the following disruptions: hesitation, pause, interruption and correction. They were also asked to assess on a six-graded likert scale how natural these sounded and how easy it was to detect the disrupted parts. Results show that disrupted utterances that were synthesized with the new voice were rated as natural in two thirds of the cases, while the ones that were generated with commercial synthesis voice, that lacked disfluent diphones, was rated as natural in half of the cases. Kruskal-Wallis rank sums were performed, and the interrupted utterances generated by new voice was significantly more natural than those generated with the commercial voice ($p=0.001$). When comparing how easy it was to detect the disrupted parts both versions are comparable (90% of them were easy to detect, with no significant difference).

In order to analyze the difference between real and pretended disruptions, the subjects were asked to compare re-synthesis of the of disrupted dialogue turns with corresponding read versions. They were asked to judge which of the two they thought contained a pretended disruption. When comparing re-synthesis of complete utterances from either of these types they were able to detect the version with pretended disruptions in 60% of the cases. In cases where the disfluent parts were moved to new fluently read sentences the users could not tell which version contained a pretended disruption. This is probably because they rated how the whole sentence sounded, rather than only the disrupted part. These differences were significant according to a chi-square test. Finally, the subjects' ability to identify the different types of disfluencies when synthesized by the two voices was compared. For both voices, about 80% of the hesitations and interruptions were correctly identified, while only 70% of the planned pauses were correctly identified. For both voices about 85% of the missed pauses were instead identified as hesitations or interruptions. For the new voice most of them were identified as hesitations, while they were mostly misinterpreted as interruptions for the commercial voice. The share of inserted interruptions is the only significant identification difference between the two voices. This is not surprising since they both used the pitch, power and durations from the original human recordings, while only the new voice also had access to truncated diphones.

This pilot study showed that the instructions from the Wizards were useful both to control the dialogue flow and to elicit interactional phenomena. Finally, the male participant reported that it was hard to pretend to be disfluent while reading dialogue transcripts where this was tagged.

Acknowledgements

This research is supported by MonAMI, an Integrated Project under the European Commission (IP-035147).

References

Adell, J., Bonafonte, A., & Escudero, D. (2007). Filled pauses in speech synthesis: towards conversational speech. In *Proc. of Conference on Text, Speech and Dialogue(LNAI 07)*

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., & Weinert, R. (1991). The HCRC Map Task corpus. *Language and Speech, 34*(4).

Beskow, J., Karlsson, I., Kewley, J., & Salvi, G. (2004). SYNFACE - A talking head telephone for the hearing-impaired. In Miesenberger, K., Klaus, J., Zagler, W., & Burger, D. (Eds.), *Computers Helping People with Special Needs*. Springer-Verlag.

Bock, K. (1996). Language production: Methods and methodologies. In *Psychonomic Bulletin and Review*.

Campbell, N., & Mokhiari, P. (2003). Using a Non-Spontaneous Speech Synthesiser as a Driver for a Spontaneous Speech Synthesiser. In *Proceedings of ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*. Tokyo, Japan.

Clark, H. H. (1996). *Using language*. Cambridge, UK: Cambridge University Press.

Dohsaka, K., & Shimazu, A. (1997). System architecture for spoken utterance production in collaborative dialogue. In *Working Notes of IJCAI 1997 Workshop on Collaboration, Cooperation and Conflict in Dialogue Systems*.

Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech Communication, 50*(8-9).

Gustafson, J., & Edlund, J. (2008). expros: a toolkit for exploratory experimentation with prosody in customized diphone voices. In *Proceedings of PIT 2008*.

Hjalmarsson, A. (2008). Speaking without knowing what to say... or when to end. In *Proceedings of SIGDial 2008*.

Skantze, G., & Schlangen, D. (2009). Incremental dialogue processing in a micro-domain. In *Proceedings of EACL-09*.

Sundaram, S., & Narayanan, S. (2003). An empirical text transformation method for spontaneous speech synthesizers. In *Proceedings of Interspeech 2003*, Switzerland.