Pauses and resumptions in human and in computer speech

Jens Edlund, Fredrik Edelstam, Joakim Gustafson
KTH Speech, Music and Hearing, Sweden
edlund@speech.kth.se, freede41@kth.se, jocke@speech.kth.se

Abstract

We present a study in which 16 subjects were recorded while interacting with a human narrator acting the part of a spoken dialogue system (SDS). Interruptions to the narrator's speech were added systematically. The recordings were analysed to find pause and resume behaviours that may be suitable for implementation in SDSs. The first results show that resumptions are initiated, on average, with a higher pitch than other utterances.

Introduction

We present a study in which 16 subjects were recorded while interacting with a human narrator acting the part of a spoken dialogue system (SDS). Interruptions to the narrator's speech were added systematically. The recordings were analysed to find pause and resume behaviours that may be suitable for implementation in SDSs.

Background

Pauses and resumptions in spoken dialogue systems

There are a number of good reasons to equip SDSs – be they robots, machines or computers that communicate using speech – with the ability to cut themselves short – to stop speaking before they have finished saying what the planned to say.

The reason most frequently discussed in SDS design is to handle so-called user barge-ins – user appearing in the middle of the system's speech. In current SDSs, barge-ins are often disallowed (e.g. ignored) by simply turning the microphone off when the system is speaking. SDSs that do handle barge-in listen (keep the microphone on) and, if needed, are trained to disregard their own voice while they are speaking. The reaction when user speech is detected is generally the most basic imaginable: simply cancel the current speech output. There are exceptions. A particularly interesting approach by Ström & Seneff (2000) takes inspiration from human-human dialogue to design a system which increases its voice intensity when barge-ins occur at dialogue states where interruptions are undesirable, signalling that barge-ins are disallowed at this stage. When a barge-in occurs at a less critical point in the dialogue, they propose that the system reduce its intensity, but continue to speak, which allows the system to verify that the detected barge-in was indeed speech from the user before cutting itself short.

A less frequently implemented, but equally important and practical, reason is to handle temporary fluctuations in the ambient noise environment. If a lorry drives by, or someone suddenly starts drilling in a near-by wall, people respond by either raising their voices or by simply pausing until the noise recedes before finishing. An SDS that copies this behaviour will be easier to understand in adversary conditions. To our knowledge, this simple functionality has not been implemented in any published system, but there is a surge of research into a related area: the use of Lombard speech from SDSs to overcome adverse noise conditions.

A third example concerns in particular *humanlike* SDSs that strive to achieve spoken communication in a manner that is similar to how humans use speech to communicate (Cassell, 2007; Edlund et al., 2008). Human con-

versations are emergent, and humans often reconsider their plan while speaking, and may pause briefly to consider before finishing. A humanlike SDS that treats the conversation as an emergent phenomenon and senses the environment incrementally and continuously should also be able to halt and to change its mind mid-utterance. Although this type of pause behaviour can be used in a pre-planned, simulated manner to achieve focus, a more interesting challenge is for the system to pause thoughtfully when it actually needs processing time. Skantze & Hjalmarsson (2013) successfully use utterance-initial filled pauses to this end.

Finally, situated SDSs – systems designed to monitor and model the context and the environment as well as the emergent dialogue – may pause if they notice an unreceptive listener. This may happen if the listener becomes disturbed by another person, or if some external task of greater importance gets in the way. Such systems may also implement pausing as a proactive behaviour and pause when they anticipate that the user will become otherwise occupied presently. Kousidis et al. (2014) show that an in-car SDS that pauses when complicated driving situations occur leads to improvements both in the driving and in the driver's recall of what the SDS said.

Where it fails

Although it can be shown that well-positioned pauses can improve the usability of an SDS as it affords user barge-ins or the efficiency and safety as for example an in-car SDS can adapt to changes in for example the driving situation, there are drawbacks. A system implementing user barge-in is likely to halt in the wrong places, as it will misinterpret non-speech sound such as coughs or external noises for barge-ins. If these mistakes occur frequently, user trust in the system will diminish, and the pausing behaviour is likely to do more harm than good. It is worth noting that current machines are at a great disadvantage here. In the most primary and common form of spoken communication amongst humans –face-to-face conversation – a speaker has access to a whole slew of indications when an interlocutor intends to take the floor. Commonly discussed speech preparatory events include in-breath, smacks and mouth opening, posture and head pose shifts; and gaze patterns. A barge-in supporting SDS generally senses nothing but sound.

There is evidence that users can be dissatisfied with pausing systems even if these are objectively better. The adaptive (and objectively safer and more efficient) system presented by Kousidis et al. (2014) received poorer subjective judgements from its users than a non-adaptive (non-pausing) counterpart, with comments suggesting that users thought the system might have paused because of programme errors.

In order to allow our SDSs to pause when need be without being second-guessed by their users, it seems important to clearly signal that the pause is intentional and planned. This way, users will be confident that they can tell intended pauses – features – from bugs. We think it likely that if the subjects in the Kousidis et al. (2014) study had felt confident that the adaptive system knew what it was doing, and that is was all for their benefit, they would have graded the system's performance higher.

The choice of signal, however, is important and not trivial. A signal that is clear and easy to perceive may not be sufficient. Edlund & Nordstrand (2002) compare an SDS which signals that it is thinking with a spinning hourglass (as in a popular computer operating system) with one where the SDS's avatar (an animated talking head) simply looks away. The system with the more obvious hour-glass results in slightly more efficient dialogues, but is strongly disliked by the users, who suspected the computer was having problems running the SDS. The system that looked away

while thinking was almost as efficient (and outperformed a system with no indicators at all) as well as liked by the users. The baseline system without indicators resulted in quite poor dialogue efficiency, but was slightly better liked by users than the hour-glass version.

The way forward

We believe that we would benefit from finding out how people behave when they pause and when they resume speaking, and attempt to implement these behaviours in humanlike SDS. In other types of SDSs, mimicking human behaviours may not be a good option (Edlund et al., 2008). In this paper, we present a first step towards this goal.

Method

The target of our experiment can be formulated in three questions: How does a human speaker stop speaking when faced with an (possible) interruption? How does a human speaker resume speaking after such an event? Which of these behaviours are plausible candidates for inclusion in a spoken dialogue system?

Data Collection

Setting

We recorded the dialogues in our dialogue recording studio – a recording environment consisting of several physically distinct locations that are interconnected with low and constant latency audio and video. Pairs of interlocutors were placed in different rooms, and communicated through pairs of wireless close-range microphones and loudspeakers. Video was not used here, since we are interested in behaviours that are triggered by the acoustic cues that are available to most SDSs.

Subjects

The end-goal of this data collection is not to train a recognizer or a recognition or categorisation device, but the generation of a consistent set of candidate behaviours for implementation in a spoken dialogue system – one that contains behaviours that could all plausibly be used by the same speaker. To achieve this, we consistently use the same single male speaker in the role as the system ("speaker", hereafter) for all recordings. For the user role ("listener", hereafter), a balanced variety of speakers were used: two sets of 8 listeners, both balanced for gender, were used. None of the listeners had any previous knowledge of this research. All listeners were rewarded with one cinema ticket. They were told that those who performed the task best would earn a second ticket, and the top performers from each setup received a second ticket after the recordings were completed.

Task

The data collection was designed as a dual task experiment. The main task for the speaker was to read three short informative texts about each of three cities (Paris, Stockholm, and Tokyo), arranged so that the first is quite general, the second more specific, and the third deals with a quite narrow detail with some connection to the city. This task is equivalent to what one might expect from a tourist information system. For the listener, the main task is to listen to the city information. The listener is motivated by the knowledge that the reading of each segment – that is each of the nine informative texts –is followed by three questions on the content of the text. Their performance in answering these questions and in completing the secondary task counted towards the extra movie ticket. The secondary task was designed as follows. At irregular, random intervals, a clearly visible coloured circle would appear, either in front of the speaker or the listener. When this happened, the speaker was under obligation to stop the narration and instead read a sequence of eight digits from a list. The listener must then to repeat the digit sequence back to the speaker, after which the speaker could resume the narration.

Conditions

We considered two characteristics of interruptions that we assumed would have an effect on how humans react to the interruption and to how they resume speaking after it: the source of an interruption can be either internal or external in a dialogue; and the duration and content of an interruption varies: they can be brief or even the result of a mistake, or they can be long and contentful. The condition mapping to the first of these characteristics was designed such that the coloured circle signalling an interruption was presented randomly to either the speaker, mapping to en external event visible to the system but not the driver, or to the listener, mapping to an interruption from the driver to the system (the listener had to speak up to inform the speaker that the circle was present). The second condition was designed such that in one set of eight dialogues, the coloured circle would start out yellow, and as soon as the speaker became silent, it would randomly either disappear (causing only a short interruption with light or no content, corresponding to e.g. a false alarm) or turn red, in which case the sequence of digits would be read and repeated (a contentful interruption). In the other set of eight recordings, the circle always went straight to red, and always caused digits to be read and repeated.

Analysis

Each channel of each recording was segmented into silence delimited speech segments automatically, and these were transcribed using Nuance Dragon Dictate. The transcriptions were then corrected by a human annotator, and labelled for interruptions – either from the listener (who was prompted by a light indicator) or from the reader being interrupted by a similar light indicator. Resumptions from the pauses caused by these interruptions were coded as well.

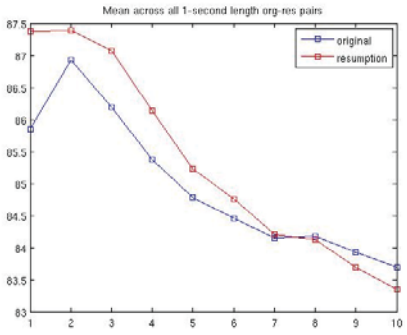

Figure 1. The average pitch (in semitones, Y-axis) for the first through tenth (X-axis) voiced 100 ms segment in the original readings, and in the resumptions following interruptions into these readings, over 59 pairs.

We take the resumption to be the point where the reader returns to reading the script. Any material after the interruption, and in applicable cases the completion of the embedded, secondary task, and the resumption was coded as pre-resumption material.

In this initial analysis, we looked the pitch of resumptions. For each resumption in the material, we also took the first sentence from the script preceding the interruption to get a pair. These pairs are matched in time, at least within a minute, so the voice characteristics of the reader should be similar. We extracted the first 10 100 ms frames containing voiced speech from each of these pairs and analysed them for pitch. 59 pairs were found were there was at least 10 frames of voiced speech. Only these were used in the analysis.

Results

We had anticipated results showing that the end of speech following an interruption is different, but so far our analysis have come up empty. Furthermore, the pauses follow near-instantly in most cases, with a delay that is just slightly above the minimum reaction time.

Prosodically, we see a clear difference between the beginning of the resumptions and the non-resumption utterances. The pairwise difference be-

tween the average pitch over the first 10 voiced 100ms segments of each part of the original-resumption pairs is plotted in Figure 1.

Figure 1 shows that the resumptions start on average about 1.5 semitones higher. They then drop, and after about 0.5 seconds, approximately 2 syllables, they are on a level comparable to original readings.

Discussion

We think that the pitch difference we found is a good candidate for implementation in current systems. The finding is consistent with the impressionistic finding that the resumptions are often characterized by a stronger initial stress, and suggests that increasing initial stress in resumptions is a candidate behavior for humanlike resumption.

Future work

The pitch finding is not straightforward to implement in current systems, as they normally do not grant control over pitch or initial stress. The finding, however, can be implemented and tested in research systems.

The data recorded will be annotated and analysed further. In particular, the short interruptions that originated on the reader side are interesting, as it seems that the listener in many cases never even noticed the interruption, as the reader masked it using several strategies such as coughing or drawing for breath slowly.

Acknowledgments

This work was funded by the GetHomeSafe (EU 7th Framework STREP project 288667).

References

Cassell, J. (2007). Body language: lessons from the near-human. In Riskin, J. (Ed.), *Genesis Redux: Essays on the history and philosophy of artificial life* (pp. 346-374). University of Chicago Press.

Edlund, J., & Nordstrand, M. (2002). Turn-taking gestures and hourglasses in a multi-modal dialogue system. In *Proc of ISCA Workshop Multi-Modal Dialogue in Mobile Environments*. Kloster Irsee, Germany.

Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. *Speech Communication, 50*(8-9), 630-645.

Kousidis, S., Kennington, C., Baumann, T., Buschmeier, H., Kopp, S., & Schlangen, D. (2014). Situationally aware in-car information presentation using incremental speech generation: safer, and more effective. In *Procs. of the EACL 2014 Workshop Dialogue In Motion* (pp. 68-72). Gothenburg, Sweden.

Skantze, G., & Hjalmarsson, A. (2013). Towards Incremental Speech Generation in Conversational Systems. *Computer Speech & Language, 27*(1), 243-262.

Ström, N., & Seneff, S. (2000). Intelligent barge-in in conversational systems. In *Procedings of ICSLP-00*.