# Human pause and resume behaviours for unobtrusive humanlike in-car spoken dialogue systems

**Jens Edlund**
KTH Speech, Music and Hearing
Stockholm
Sweden
edlund@speech.kth.se

**Fredrik Edelstam**
KTH Speech, Music and Hearing
Stockholm
Sweden
freede41@kth.se

**Joakim Gustafson**
KTH Speech, Music and Hearing
Stockholm
Sweden
jocke@speech.kth.se

## Abstract

This paper presents a first, largely qualitative analysis of a set of human-human dialogues recorded specifically to provide insights in how humans handle pauses and resumptions in situations where the speakers cannot see each other, but have to rely on the acoustic signal alone. The work presented is part of a larger effort to find unobtrusive human dialogue behaviours that can be mimicked and implemented in-car spoken dialogue systems within in the EU project Get Home Safe, a collaboration between KTH, DFKI, Nuance, IBM and Daimler aiming to find ways of driver interaction that minimizes safety issues,. The analysis reveals several human temporal, semantic/pragmatic, and structural behaviours that are good candidates for inclusion in spoken dialogue systems.

## 1 Introduction

In-car spoken dialogue systems face specific challenges that are of little or no relevance for systems designed for other environments. The two most striking of these are (1) the very strong focus on safety in the driving situation and (2) the fact that the person who speaks to the system – its user, in other words the driver in the majority of cases – does so in an environment that may change quite drastically from the beginning of an interaction to its completion. The most straightforward source for this change is the fact that the car (and the user) moves through the environment while the dialogue progresses. The dynamic and mobile nature of the surrounding traffic adds to the complexity. Generally speaking, safety is the key concern when designing spoken dialogue systems for in-car use. While poor performance in spoken dialogue systems can clearly be a nuisance to a driver, the promise of using properly designed spoken dialogue instead of other interfaces is increased safety. This promise is based in the nature of speech: it does not require the driver to divert the use hands and eyes from the driving, and it is a mode of communication that most are quite used to and comfortable with, so should not induce great amounts of cognitive load.

We present a corpus consisting of a set of human-human dialogues recorded specifically to provide insights in how humans handle interruptions - how they pause and resume speaking - in situations where the speakers cannot see each other, but have to rely on the acoustic signal alone, and a preliminary analysis of these which reveals several candidates for inclusion in in-car spoken dialogue systems. Finally, we discuss how these can be implemented and how a selection of them are included in the Get Home Safe experiment implementation.

## 2 Background and related work

In a government-commissioned survey from 2011, the Swedish National Road and Transport Research Institute reviews several hundred research publications on traffic safety and the use of mobile phones and other communication devices [Kircher et al., 2011]. Amongst the most striking findings: although there is a broad consensus that visual-manual interactions (e.g. using social media or texting) with communication devices impair driving performance, bans have not had any measurable effects in terms of lowered accident rates or insurance claims. Ban compliance statistics show



*Proceedings of the of the EACL 2014 Workshop on Dialogue in Motion (DM)*, pages 73–77,
Gothenburg, Sweden, April 26-30 2014. 

that bans have an effect on driver behaviour the first year, after which drivers return to their former habits. With bans being virtually ineffective, solutions must be sought elsewhere. Allowing drivers to manage more tasks using speech, which does not occupy hands and eyes, would decrease the time spent in visual-manual interaction while driving, provided that the drivers can be persuaded to use the systems.

Clearly, the systems must work well - a large proportion of errors may well put the driver at risk (e.g. Kun et al., 2007). It is also unlikely that drivers can be persuaded to use systems that do not work well. But using hand-free and eyes-free controls may not suffice. Kircher et al. (2011) notes that there is virtually no evidence that hands-free telephony is less risky than hand-held use, suggesting that the conversations in themselves may be a risk factor. Speaking to a person who is present in the car and who shares the driver's situation, however, is much safer (Peissner et al., 2011), suggesting that a system that is perceived as and behaves like a co-present human is a sensible aim. In the EU project Get Home Safe, of which this research is a part, we call such systems *humanlike proactive systems*. Where a traditional spoken dialogue system bases its decisions largely on (1) whether it has something to say, (2) what the user has just said, and (3) whether the user is speaking or is silent, a humanlike proactive system will also consider (4) the (traffic) situation, (5) the user's (driver's) estimated attention, and (6) the urgency of the task at hand, much like a passenger might.

This paper focusses on two broad types of proactive humanlike behaviours: *user controlled pacing*, referring to the ability to pause at the whim of the user in the middle of a conversation, or even an utterance, and then resume the conversation; and *situation sensitive speech*, the ability to allow the situation to affect the manner in which the system speaks. We are searching for behaviours that people use when interrupted, either by their interlocutor or by some event in their environment, and when they resume the original dialogue again. We are specifically interested in behaviours that can be implemented in the Get Home Safe architecture without major changes to existing applications. The architecture allows a central manager to instruct applications to stop where they are and maintain their inner state until instructed to either exit or continue where they were.

The task has been approached by others, albeit in different manners. Villing (2010) presents an analysis of interruptions and resumptions in human-human in-vehicle dialogues, as well as implications for future in-car dialogue systems, and Yang et al. (2011) used human-human multi-tasking dialogues that involved a poker game as the main task, and a picture game as an interrupting real-time task.

## 3 Method

Our goal is to collect and analyse data that will provide an insight to how a human speaker deals with interruptions in in-car dialogue (our target setting) and to find relevant behaviours that can be successfully mimicked in an in-car human-computer environment. The question can be subdivided: 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?

### 3.1 Data Collection

**Setting.** Collecting data from a real driving situation is time consuming, not to say dangerous when adding a secondary task. We have instead opted to simulate the key elements of interest in our dialogue recording studio – a safe recording environment consisting of several physically distinct locations that are interconnected with low and constant latency audio and video. The interlocutors were placed in different rooms, and communicated through pairs of wireless close-range microphones and loudspeakers.

**Subjects**. The purpose of this data collection is not for example training a recognizer, 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 in-car 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 an in-car dialogue (our target setting); 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.

### 3.2 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 and resumptions. In this initial analysis, we looked at temporal statistics (e.g. the durations between interruption from the listener and silence from the speaker), semantics/pragmatics (e.g. lexical choices, insertions, repetitions) and syntax (e.g. where in an utterance resumption begins).

## 4 Results

A categorical difference was found in the distribution of speaker response times (from the onset of a listener interruption to the offset of speaker speech) depending on whether the interruption occurred in the middle of a phrase or close to the end of the phrase. In the first case, the vast majority of the response times are

distributed between 300 and 700 ms, with a clear mode around 400 ms. Only a fraction of response times are slower than 700 ms, and none except one is faster than 300 ms. Phrase final interruptions show an almost flat response time distribution, with only a very weak mode around 500 ms, and a large proportion with response times longer than 700 ms.

For lexical/pragmatic choices, we find a categorical variation for the insertion of vocalizations we somewhat lazily term filled pauses (e.g. "eh", "em") and what we equally lazily term lexical cue phrases (e.g. "right", "ok") before resumption. The existence of such insertions, as well as the choice of vocalization, is straightforwardly dependant on the contentfulness of the interruption. For short interruptions of light content, filled pauses are nearly never inserted before resumption. Lexical cue phrases are inserted, but rarely. In the typical case, the speaker goes straight back to the informational text. For long, contentful interruptions, resumption is initiated by an insertion in an overwhelming majority of cases. If the insertion consists of one vocalization only, this is nearly always a filled pause. If more than one vocalization is present, then lexical cue phrases occur frequently, but overall, lexical cue phrases are no more common here than in the case of the short interruptions.

In the case of structural comparisons, the one clear distinction we found has to do with what, if any, material is repeated at resumption, a characteristic that varies strongly with the type of interruption. For long interruptions, in every instance but a handful, the speaker either repeats the entire utterance in which the interruption occurs, or - in the few cases where an interruption occurred just as an utterance came to an end - with the next utterance. For short interruptions, resumptions also start most regularly from either the start of the current utterance or from the start of the next one. However, starts from the beginning or end of the current phrase, word, or even part of word are also frequent.

## 5 Discussion

We think that the three main findings presented in the results are all good candidates for implementation. The different distributions of response times suggest that if an interruption occurs centrally, in the midst of a production, the speaker stops as fast as possible - the distribution is largely consistent with reaction time distributions. Towards the end of phrases, the distribution is flat and quite different to what one would expect if reaction time was the main governing factor. The larger proportion of long response times suggests that when the speaker is close to the end of a phrase, finishing the phrase first might be preferable to stopping as soon as reaction permits. From an implementation perspective, this is quite encouraging. In order to create a behaviour consistent with this, we need to halt system speech with a reaction time of around 3-500ms. If possible (i.e. if the system knows how much time remains of its production), we may instead complete the utterance if less than, say, 700ms remains.

Seemingly, short light content interruptions need no specific signalling of resumption. If such signalling is made, it is in the form of a lexical cue phrase, such as "ok" or "right". Resumptions following longer, contentful interruptions are routinely initiated by a filled pause. This may be solely due to the speaker's need to find the correct place in the script to start over, but it is noteworthy that instead of doing this in silence, the speaker opts to vocalize. For implementation, resumptions following contentful subdialogues should start with a filled pause and perhaps a lexical cue phrase.

The straightforward interpretation of the third finding is that in the case of short interruptions, both speaker and listener have the point of interruption in fresh memory, and need no reminder, while long interruptions require the speaker to help the listener out by recapitulating what was last said. In the latter case, the system can simply start over with its last utterance (provided that it produces its synthesis on a granularity of at least utterance level).

**Acknowledgments**

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

## References

Kircher, K., Patten, C., & Ahlström, C. (2011). *Mobile telephones and other communication devices and their impact on traffic safety: a review of the literature*. Technical Report VTI 729A, Stockholm.

Kun, A., Paek, T., & Medenica, Z. (2007). The effect of speech interface accuracy on driving performance. In *Proc. of Interspeech 2007*. Antwerp, Belgium.

Peissner, M., Doebler, V., & Metze, F. (2011). *Can voice interaction help reducing the level of distraction and prevent accidents? Meta-Study on Driver Distraction and Voice Interaction*. Technical Report, Fraunhofer, Germany and CMU, USA, Aachen, Germany.

Villing, J. (2010). Now, where was I? Resumption strategies for an in-vehicle dialogue system. In *The 48th Annual Meeting of the Association for Computational Linguistics* (pp. 798-805). Sweden.

Yang, F., Heeman, P. A., & Kun, A. L. (2011). An investigation of interruptions and resumptions in multi-tasking dialogues. *Computational inguistics, 27*(1), 75-104.