Improving Conversational Dynamics with Reactive Speech Synthesis

Abstract
The active exchange of ideas and/or information is a crucial feature of human-human conversation. Yet it is a skill that present-day ‘conversational’ interfaces are lacking, which effectively hampers the dynamics of interaction and makes it feel artificial. In this paper, we present a reactive speech synthesis system that can handle user’s interruptions. Initial results of evaluation of our interactive experiment indicate that participants prefer a reactive system to a non-reactive one. Based on participants’ feedback, we suggest potential applications for reactive speech synthesis systems (i.e. interactive tutor and adventure game) and propose further interactive user experiments to evaluate them. We anticipate that the reactive system can offer more engaging and dynamic interaction and improve user experience by making it feel more like a natural human-human conversation.

Author Keywords
Conversational agents; UX design; Usability testing; Reactive HCI

ACM Classification Keywords
H.5.2. Information interfaces and presentation (e.g., HCI): User Interfaces
Introduction
The Cambridge Dictionary of English defines conversation as: 'a talk between two or more people in which thoughts, feelings and ideas are expressed, questions are asked and answered, or news or information is exchanged'. The above definition underlines the dynamic character of a conversation, which, contrary to a monologue, requires active exchange of ideas between interlocutors. We believe that the same principle should apply to human-computer interaction (HCI) to make it a more dynamic and a more realistic experience than the one offered by present day 'conversational' agents.

Porcheron et al. [1] claim that 'Conversational Interface' is a misnomer, since, unlike in human-human dialogue, responses provided by 'conversational' agents lack interactivity and are based on a pre-defined script. Indeed, present-day interactions with Voice User Interfaces (VUI) remain a rather limited and artificial activity. In order to illustrate this point, let us consider the Amazon Echo adventure game The Magic Door [13]. In the game, the narrator (Alexa) provides a player with choices that determine how the story develops. As presented in Figure 1, the interaction lacks dynamics, as the player is forced to listen to long chunks of text being read out without the possibility to interrupt the narrator. In contrast, as presented in [12], role-playing games between humans are very interactive with participants expressing a variety of emotions.

Making VUI More Reactive
Previous research in the field of voice interfaces identified adaptability [2,3,4,5] and incrementality [3,6] as prerequisites for creating a truly interactive conversational system.

Adaptability.
In conversation, adaptability is an ability of the system to adjust its behaviour based on actions of the interlocutor and changes in the environment (e.g. noise, interruptions etc.), and react accordingly. Lison et al. [3] pointed out that, a conversational system needs to behave like an active listener that is receptive and responsive to user's verbal and non-verbal behaviour while, at the same time, one that is not too interruptive or intimidating. In a similar vein, Radlinski et al. [2], noted that, in order to feel natural, a conversation with a conversational system should be shaped by the context in which it takes place.

Incrementality.
In the context of conversation, incrementality is an ability to actively process interlocutor's input as words are being spoken and gradually update the interpretation as more information becomes available. Schlangen et al. [7] proposed an incremental dialogue-processing model that allows the system for processing input at sub-utterance level. The merit of incremental system is that it reduces latency consequently making the interaction feel more reactive.

In current this paper, we expand on the previous work on VUIs [2,3,4,5,6] by proposing a reactive system that is both adaptable and incremental and explore its impact of perceived realism of interaction. The system uses reactive speech synthesis as its output. Our goals are threefold: (1) investigate the effect of more reactive voice interfaces on user experience, (2) offer design guidelines to improve the design of future
Real-Time Reactive Speech Synthesis
A reactive VUI is one that adapts to interlocutors’ behaviour and responds accordingly. This can be achieved via use of reactive speech synthesis which allows for re-planning the output of a system and splicing it in at the right time.

In their recent work, Wester et al. [8] explored the impact of a system that mimicked human speech output responses to interruptions via ‘Reactive Synthesis’ demo. In their demo, 6 participants (3 males and 3 females) were asked to interrupt the system which was reading out text. The system used three different reaction styles to respond to participants’ interruptions, i.e. (1) No Reaction Whatsoever: the system carried on reading, (2) Courteous Reaction: the system stopped reading, having detected participant’s interruption and waited for them to finish speaking before continuing with the reading task, and (3) Lombard Effect Reaction: the system became increasingly annoyed after each interruption of the participant (i.e. with pitch and tempo being adjust to reflect anger). The reactive speech synthetic used in the demo was generated with CereProc SKD [9] (illustrated diagrammatically in Figure 2).

The feedback received from the participants during the focus group that followed the demo, indicated that participants preferred a reactive system to a non-reactive system. The majority of participants commented that the interaction style 2 (Courteous Reaction) was the most suitable for a reactive system, bearing in mind that such a system should be cooperative. However, it was also pointed out that the reaction style 3 (Lombard Effect Reaction) made the reactive system seemed more personable.

Following the evaluation three design recommendations for a reactive system were proposed. In order to be considered reactive the voice interface should:

Be Fast Enough.
System needs to synthesise replacement chunk with the required latency – at least 200ms. This is based on human-human conversation where a typical gap between switching the speakers is 200ms [cf.10].

Splice Audio in.
Audio output should be closely controlled – it must be possible to alter the queued audio as it is waiting to be played and it must be clear what audio was played.

Know how to Respond.
The appropriate response of the system would vary on the application. Versatile systems should be able to react in a variety of ways: while helpful systems should be polite and rephrase, game characters should be able to display a range of human responses to interruptions such as rudely continuing and speaking over the interlocutor.

Potential Applications of Reactive Voice Interfaces
Based on the results of evaluation of ‘Reactive Synthesis’ demo we would like to suggest the following applications for reactive speech synthesis.
Tutoring System
The content of Massive Open Online Courses (MOOCs) is currently generally transcribed in English, which opens a possibility of implementing a reactive tutor. The tutor would allow learners to ask questions at any time, or request repetition, which, arguably, would make a learning experience more interactive and personalised. For instance, the tutor could facilitate the learning process by providing alternative definitions or paraphrasing the content of the course to make it easier to understand.

Interactive Role-Playing Game
In order to make user’s experience more engaging, reactive speech synthesis could be applied to change the way in which virtual character interacts with the player by allowing them opportunity ask more questions and, consequently, offering a more engaging interaction. For instance, based on player’s reactions, the characters could change their speaking style by responding emotionally or talk over the user. The character’s interaction can be modelled on behaviour of players of The Werewolf (presented in [12]) who changed their speech patterns to deceive their opponents. Figure 4 explains the rules of the game.

Further Evaluation of Interactive VUI
In order to evaluate the effectiveness of our proposed reactive interfaces we will create a prototype of interactive Tutoring System and an Adventure Game test them in interactive user experiments where users will interact with each system and then asked to rate their experience in terms of engagement and perceived naturalness of interaction.

Conclusion
In this paper, we have identified some potential benefits of reactive speech synthesis for making interaction with VUIs more dynamic. We outlined some initial guidelines for designing reactive speech interfaces, and suggested some potential applications. We anticipate that our future evaluation experiments will enable us to further explore the impact of real-time speech synthesis on user’s experience and perceived reactivity of the system. We believe that behavioural insights obtained from our proposed experiments will benefit the HCI community and, eventually, help to foster development of more engaging and dynamic VUIs.

References

---

[1] The Idiap WOLF Dataset with recordings of games is available at: https://www.idiap.ch/dataset/wolf/download-proc


