Improving Conversational Dynamics with Reactive Speech Synthesis

Mateusz Dubiel

The University of Strathclyde

mateusz.dubiel@strath.ac.uk

Martin Halvey

The University of Strathclyde

martin.halvey@strath.ac.uk

Leif Azzopardi

The University of Strathclyde

leif.azzopardi@strath.ac.uk

Matthew Aylett

CereProc Ltd.

matthewa@cereproc.com

Miriam Wester

CereProc Ltd.

mirjam@cereproc.com

David A. Braude

CereProc Ltd.

dave@cereproc.com

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

```
Player: Alexa, open the magic door.

Alexa: In front of us is an enchanted
land of adventures. We can explore this
land together.
Our journey will depend on the choices
that we make.
Your choice begins here with three roads:
one leads to the mountains, another leads
to the sea, the last is spooky path starting
in the dark forest.
Would you like to go to the mountains,
to the sea, or to the dark forest?

Player: To the dark forest.

Alexa: Ok, follow me up on the path
of the fallen leaves.
The path leads through skeleton-like trees
and we soon find ourselves
in the dark, cold forest.
It feels a bit creepy here.
We don't have to keep going.
I think it will only get creepier.
Would you like to keep going,
or get out of here?
```

Figure 1: Example of interaction between player and narrator from the adventure game *The Magic Door*.

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 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 chucks 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
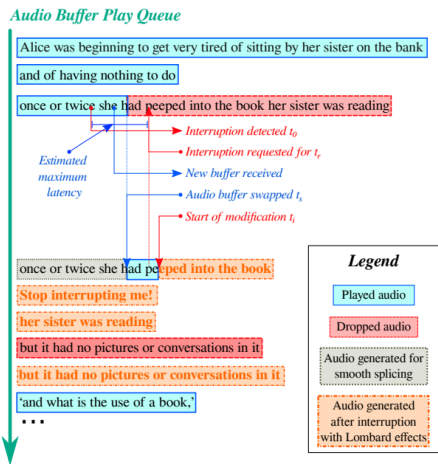
Figure 2: Example of the use of interruption API, showing the changes in audio buffers. First played audio is in blue and orange boxes, red and grey boxes are dropped. Note that $t_r - t_0$ must be larger than the maximum system latency.

conversational interfaces, and (3) suggest possible applications for reactive voice interfaces.

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.

```
Tutor: A flamboyant person is
someone who attracts attention
based on their exuberance,
con[fidence, and...

User: ↑ [What is exuberance?

Tutor: Exuberance means to be
full of energy.
```

Figure 3: Example interaction with a reactive tutor that can be interrupted by a user at any point.

Figure 4. Players making accusations in a game of _Werewolf_. Players are divided into villagers and werewolves. The game consists of two stages (1) _night_, when werewolves may covertly murder villagers, and (2) _day_ when the players who survived, debate the identities of the werewolves and try to eliminate them. The game continues until all the werewolves have been eliminated, or until the werewolves have outnumbered the villagers.

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. An example of what interaction with reactive tutor could look like is presented in Figure 3. The behaviour of an interactive tutor and the way it should handle interruptions will be modelled based on Novel Multilingual Database of Natural Dyadic Novice-Expert Interactions (NoXi) [11].

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[1]. Figure 4 explains the rules of the game.

---

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

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. Porcheron, Martin, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. Voice Interfaces in Everyday Life. In _Proceedings of the 2018 ACM Conference on Human Factors in Computing Systems (CHI'18)_. ACM, 2017.

2. Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In _Proceedings of the 2017 Conference on Conference Human Information Interaction and_

Retrieval (CHIIR '17). ACM, Oslo, Norway, 117-126 DOI: http://dx.doi.org/10.1145/3020165.3020183

3. Pierre Lison, and Meena Raveesh. Spoken Dialogue Systems: The Frontier in Human-Computer Interaction. In. *XRDS: Crossroads, The ACM Magazine for Students 21.1.2014. 46-51.*

4. Ewa Luger and Abigail Sellen. 2016. Like Having a Really Bad PA: The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16). ACM, New York, NY, USA, 5286-5297. DOI: http://dx.doi.org/10.1145/2858036.2858288

5. Robert J. Moore, Raphael Arar, Guang-Jie Ren, and Margaret H. Szymanski. 2017. Conversational UX Design. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '17). ACM, New York, NY, USA, 492-497. DOI: https://doi.org/10.1145/3027063.3027077

6. Roger Moore.2017. Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In *Dialogues with Social Robots*. Springer Singapore. 281-291.

7. David Schlangen and Gabriel Skantze. 2009. A general, abstract model of incremental dialogue processing. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09). Association for Computational Linguistics, Stroudsburg, PA, USA, 710-718.

8. Mirjam Wester, David A Braude, Blaise Potard, Matthew P Aylett, and Francesca Shaw. Real-time reactive speech synthesis: incorporating interruptions. In *Proceedings of Interspeech 2017*. ISCA, Stockholm, Sweden, 3996-4000 DOI: http://dx.doi.org/10.21437/Interspeech.2017-1250

9. Matthew Aylett and Chris Pidcock. 2007.The Cerevoice Characterful Speech Synthesiser SDK. In Intelligent Virtual Agents (IVA). 295-303.

10. Stephen C. Levinson, and Francisco Torreira. 2015. Timing in Turn-Taking and its implications for Processing Models of language. In Frontiers in Psychology 6.20.No.1 DOI: http://dx.doi.org/10.1016/tics2015.10.010

11. Angelo Cafaro, et al. 2017. The NoXi Database: Multimodal Recordings of Mediated Novice-Expert Interactions. In. *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI'17)*. ACM, Glasgow, The UK, 350-359. DOI: https://doi.org/10.1145/3136755.3136780

12. Hayley Hung and Gokul Chittaranjan. 2010. The idiap wolf corpus: exploring group behaviour in a competitive role-playing game. In Proceedings of the 18th ACM international conference on Multimedia (MM '10). ACM, New York, NY, USA, 879-882. DOI: https://doi.org/10.1145/1873951.1874102

13. The Magic Door LLC. 2016. *The Magic Door*.