INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Real-time reactive speech synthesis: incorporating interruptions

Mirjam Wester, David A. Braude, Blaise Potard, Matthew P. Aylett, Francesca Shaw

CereProc Ltd., Edinburgh, UK

{mirjam,dave,blaise,matthewa,francesca}@cereproc.com

Abstract

The ability to be interrupted and react in a realistic manner is a key requirement for interactive speech interfaces. While previous systems have long implemented techniques such as 'barge in' where speech output can be halted at word or phrase boundaries, less work has explored how to mimic human speech output responses to real-time events like interruptions which require a reaction from the system. Unlike previous work which has focused on incremental production, here we explore a novel re-planning approach. The proposed system is versatile and offers a large range of possible ways to react. A focus group was used to evaluate the approach, where participants interacted with a system reading out a text. The system would react to audio interruptions, either with no reactions, passive reactions, or active negative reactions (i.e. getting increasingly irritated). Participants preferred a reactive system.

Index Terms: speech synthesis, reactive, interruption

1. Introduction

"Situated Interaction", the theme of Interspeech 2017 is addressed in this paper from a synthesis perspective. Being able to react in a realistic manner is a key requirement for an interactive speech synthesis system. There are situations when the output of a speech synthesis system (virtual agent) needs to be interrupted, for example, when a noise event occurs (banging door, passing train, or other environmental noises) that make it unlikely that the user will be able to understand the synthesis. Additionally, there will be times when the user wants to interrupt the system. While previous systems have long implemented techniques such as 'barge in' where speech output can be halted at word or phrase boundaries, less work has explored how to mimic human speech output responses to real-time events like interruptions at which point the system is expected to react, ideally in a convincingly authentic manner.

There is an implicit assumption that reactive synthesis has to be incremental. This is not the case, it just needs to be stoppable. To be reactive, the synthesis has to be fast enough to re-plan content (re-planning) and insert it (splicing). It is true that incremental systems offer locations for insertion but, given that any system has full timings described, such insertion points can be chosen without the need for incremental generation.

In contrast, there are clear examples where incremental systems are required for example synthesising typed text as the user is typing it [1] or performative synthesis where the interruption rate can be seen as being continuous and focussed on spectral and duration properties rather than linguistic context [2]. In this work, we are focussed on producing a system that can allow a conversational agent the ability to react rapidly and naturally to external interruptions and stimuli. Therefore, we present a system that incorporates re-planning as a strategy for dealing with user-initiated interruptions.

We created a demonstration of 'Reactive Synthesis' for the British Science Museum Lates in London (March 2017) as an educational tool to provide a general introduction to speech technology. Prior to the presentation at the museum, we were interested in obtaining feedback from the general public as to their opinions and attitudes to speech synthesis in general and to the demo more specifically. The approach we took in this work was to evaluate by means of a focus group. This paper describes the 'Reactive Synthesis' demo and how the focus group responded to it. The contributions of this paper are:

- Clarifying the relationship between incremental processing, re-planning and splicing.
- A novel approach using low-level audio splicing to implement re-planning.
- Design guidelines – what needs to be taken into account when designing a reactive speech synthesis system.
- Exploring a new way of evaluating a system in a reactive and interactive setting.

1.1. Previous work

Most previous reactive speech synthesis work falls under the umbrella of incremental speech synthesis (iSS) in the context of spoken dialogue systems. We argue, however, that the studies have typically used a mixture of re-planning and incremental planning, and that for many of the systems the term incremental is slightly misconstrued.

Edlund [3] describes specific requirements that incremental speech synthesis should meet for use in spoken dialogue systems. A system needs to be responsive and flexible like human interlocutors and it needs to process information incrementally and continuously rather than in large chunks. To achieve this, the synthesis must know what has been said, must be able to halt, then continue/break as well as be able to stop. Furthermore, it must be real-time and online. Baumann et al. [4] expand on this by adding that the synthesis should be able to begin speaking before utterance processing has finished; it should make edits to (as-yet unspoken) parts of the utterance and adaptations of delivery parameters such as speaking rate or pitch should be possible.

The processing paradigm deployed in inproTK [4] is one where processing takes place just-in-time. The processing steps are taken as late as possible to avoid re-processing when assumptions change. [5] extend the work of [4] and integrate incremental natural language generation (NLG) and speech synthesis and demonstrate the flexibility that an incremental approach to output generation for speech systems offers by implementing a system that can repair understanding problems. [5] found, in their system, that incremental output generation was able to greatly speed up system response time. Furthermore, the behaviour of their fully incremental and adaptive system was perceived as significantly more human-like than the non-incremental but responsive baseline systems. However, we would argue this is due to the responsiveness of the system

http://dx.doi.org/10.21437/Interspeech.2017-1250

and not the incremental nature of the processing; the adaptation strategy does not require incremental speech synthesis. It is not that incremental processing is irrelevant, just that it is an engineering solution to allow re-planning for slower systems. Replanning is still a requirement for reactiveness.

Astrinaki et al. [2] do not approach the issue of incremental speech synthesis from a spoken dialogue perspective but rather from a performative and context-reactive perspective. Their performative HTS system is reactive and allows for changes to the pitch and tempo of upcoming phonemes. A lookahead of only a few phones is sufficient to achieve speech quality indistinguishable from a regular system. Although [5] argues as the pHTS system is fed from a (non-incrementally produced) label file, it cannot easily be used in an incremental system. We argue that what [5] is referring to as incremental is in actual fact re-planning. Changes to the label file are not what defines the incremental nature of a system. We argue that the system described in [2] is precisely incremental because the interruption rate can be seen as being continuous and focussed on spectral and duration properties rather than linguistic context not despite it. There is no reason re-planning cannot be introduced to cover the responsiveness of the system at a linguistic level.

In contrast, Pouget et al. [1] describes a scenario where speech synthesis needs to be produced while the user is **still** in the process of typing. In this scenario re-planning is difficult to carry out as the system would have to guess the intentions of the user. Therefore incremental synthesis, as a strategy for online estimation of the target prosody from an incomplete sentence, is required.

To summarize, incremental approaches are required in some scenarios, they also can greatly reduce latency which improves responsiveness. However, re-planning is sufficient if synthesis can be carried out fast enough, and if re-planned audio can seamlessly replace output audio before it has to be sent to the output device.

2. System description

The re-planning and splicing approach is as follows: given a required latency, e.g. 200 ms, the system must operate fast enough to resynthesise the current chunk of speech with an alternative ending within that time. The initial part of the synthesis must match exactly the initial part of the current chunk. The new audio can then be seamlessly re-splicing into the audio stream replacing the original planned output. This requires tight control of audio playback, but has the advantage of being agnostic to the type of synthesis system you are using.

CereProc's SDK [6] synthesises on a phrase-by-phrase basis, firing a callback between phrases. During the callback a special audio buffer is available which contains the audio of the phrase as well as some metadata, this buffer is queued for playback. We created new functionality in the SDK that takes as input one of these buffers, a minimum interruption time, t_r, and an interruption type, and returns a new buffer. In this buffer the audio up to t_r is guaranteed to be identical to the original buffer. After that it will be interrupted at $t_i \geq t_r$. t_i will be a natural point for interruption, i.e. a syllable nucleus or boundary. Once this buffer is available the agent can seamlessly swap the audio buffer that is being played at some point $t_s < t_r$. By setting this time slightly in the future of when the interruption is needed some latency for processing can be added. This is illustrated diagrammatically in Figure 1.

Depending on the call the system has multiple strategies for finishing the phrase:

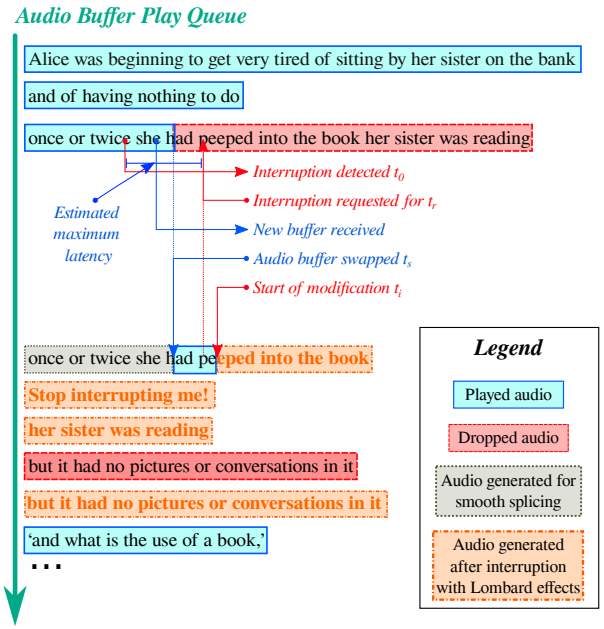

Figure 1: *Example of the use of the interruption API, showing the changes in audio buffers. Final played audio is in blue and orange boxes, red and grey boxes are dropped. Note that $t_r - t_0$ must be larger than the maximum system latency.*

- stopping immediately,
- tailing off over a few words (a polite turn pass),
- adding Lombard effects for a few words (an angry turn pass),
- completing the original phrase with Lombard effects added.

The system can then add additional speech before returning to the original queue if appropriate. Otherwise it may need to drop some phrases that have been resynthesised differently, or empty the queue entirely, depending on the application.

2.1. Demo set-up

As the demonstrator was intended for a science exhibit (the British Science Museum Lates), and more specifically to provide a general introduction to speech technology to the general public, the voice of the system was chosen for its clarity and apparent patience. The text to be read is a general introduction about speech synthesis, which is available online: `http://derstandard.at/1325485448039/Talking-Technology`.

It is well known that combining speech synthesis with a strong visual modality has the side effect of reducing the overall impact of the audio modality. To maximise the impact of the speech synthesis, the visual aspects of the system were thus kept to a minimum, and only displays the text and the current position within it, and a basic interface reflecting its internal state (see Figure 2).

Rather than an exhaustive study of the whole range of possible reactions to interruptions, we selected a small subset in order to simulate three distinct system reaction styles.

- First, a baseline system that simply did not react whatsoever when the user attempted to interrupt, it simply continued reading.

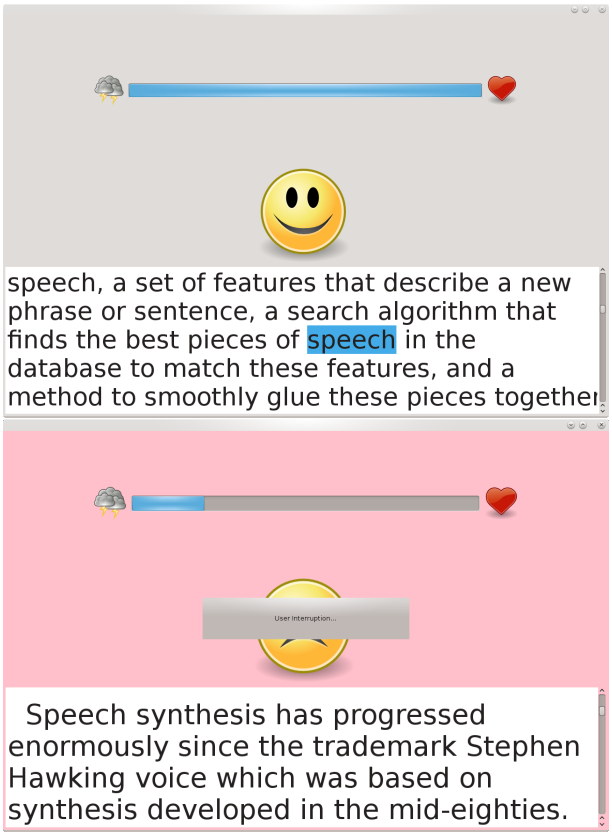

Figure 2: *Visualisation of the system in action. The bar at the top (and the figure in the centre) indicates the system's mood. The text being read is displayed at the bottom, with the current word being read highlighted. When an interruption occurs, the background colour is modified and an explicit dialogue window is displayed.*

- The second system demonstrated a basic reactive system: it detected user interruptions and stopped its speech gracefully, almost instantly. It then waited for the user to stop speaking before resuming the reading task.
- The third system was intended to demonstrate a more complex behaviour of the system: it initially reacted gracefully to users' interruptions, but as the interactions continued, the type of reactions evolved to demonstrate an increased level of irritation of the system when interrupted. After about 10 interruptions, the system simply decided it had had enough and it would leave the interaction.

For the demo, Voice Activation only was used to detect interruptions. No linguistic or recognition processing was carried out. More specifically, the system used a Voice Activity Detection (VAD) module [7] to detect speech-based interruptions from the user. The system did not distinguish between speech feedback from the user or actual voluntary interruptions, considering any event that triggered an active state of the VAD module as an interruption. When an interruption was detected, a reaction was generated, comprising 5 stages:

1. Modify the currently playing phrase to surrender the floor to the user as described above; normally stopping at a natural point or tailing off politely (maximum latency is 50 ms).
2. Pause to let the user speak; the system would stay silent until a silence of at least 200 ms long was detected.
3. Generate a reaction from the system about the interruption that had just happened. At the beginning of the interaction, reactions were rather polite and invited the user to carry on speaking, but as the interaction progressed the reactions turned into vocal gestures, and in the final phase clearly irritated remarks.
4. Optionally pause to let the user express him/herself; the duration of the pause becomes shorter as the system becomes more irritated.
5. Resume the reading task.

In practice, all three systems were built the same way, simply disabling some of the functionalities of the "full" system:

- The first system had the VAD module disabled
- The second system had its mood change disabled, and always performed the most polite of reactions

3. Evaluation

There is no established approach to evaluating the effectiveness of reactive synthesis. Many studies look at the quality or naturalness of the output synthetic speech after applying incremental processing compared to a standard non-incremental version of their system [1, 2]. However, in such cases the responsiveness of the synthetic speech is not considered in the evaluation. In [5], an evaluation was carried out in which subjects were asked to rate the behaviour of a system. The samples presented were interrupted by noise, to which the system did or did not react. The reactive system was rated considerably more human-like. However, [8]'s multi-modal evaluation experiment illustrated that users not only rate formulations that are enabled by incremental speech synthesis higher, but also judge such a system's synthesis quality as higher, despite the fact that it was actually worse than non-incremental synthesis. This result highlights that pure listening evaluations may be inadequate and that evaluation in context may lead to more relevant judgements.

Our system of re-planning and splicing has not yet been incorporated in a truly relevant, meaningful context as it is still at a development stage. Rather than try and obtain judgements on naturalness and quality of the synthetic speech we were interested in obtaining feedback from the general public on a higher level regarding their opinions and attitudes to speech synthesis in general and to the 'reactive synthesis' demo created for educational purposes more specifically. The approach we took to evaluation was by means of a focus group.

3.1. Focus Group

In setting up the focus group we followed the Guidelines for Conducting a Focus Group [9]. In addition to the 'Reactive Synthesis' demo there was another demo 'Bot or Not' which had also been designed for the Science Museum Lates. All the various stages in the focus group meeting are described below. However in terms of feedback we only present responses relevant to the 'Reactive Synthesis' demo. Six people (3M/3F) were selected to take part in the focus group. They were recruited through the Edinburgh University Careers Service MyCareerHub. The only requirements were that the participants were able to speak English to a native level and were able to attend the meeting.

Two facilitators ran the focus group meeting. The meeting included the following stages:

- Welcome to participants – filling in of consent forms.
- Playing with interactive (speech technology related) toys.
- Filling in of a general speech synthesis questionnaire.
- The first demo 'Bot or Not' was carried out in pairs .
- Group conversation discussing the interactive toys and 'Bot or Not'.
- 'Reactive Synthesis' demo was run in three different ways.
    1. System talked – no reaction to interruptions at all.
    2. System stopped talking when interrupted – then continued.
    3. System stopped talking and reacted every time it was interrupted – reactions started polite and ramped up a step each interruption until very irritated.
- Group conversation discussing the reactive synthesis demo.
- End of focus group meeting - subjects thanked and remunerated for their time and effort.

3.2. Feedback on reactive synthesis demo

The goal of the focus group meeting was to get a more general, higher level of input from people who may have previously not considered speech synthesis. One of the first striking elements that became clear during the discussions and when considering the responses of the participants (P) to the speech synthesis questionnaire was that speech recognition and speech understanding were seen as part and parcel of speech synthesis, for example answers to the questions "Do you use synthetic speech? If no, why not?" included:
P3:*"I'd be worried I'd find because I talk fast and with a broad Scottish accent, I'd not get understood."*
P4. *"I find it's too laborious and time-consuming. You have to repeat yourself over and over irritating."*

The gist of the feedback that was given after the reactive synthesis had been demonstrated was that overall participants preferred the reactive mode over the non-reactive. And of the two reactive modes they preferred the more simple approach, in which the system stopped when interrupted and when it resumed it went back and repeated the spurt[1] it was interrupted in. Furthermore, there was agreement that emotion is not the point of synthetic speech and that strong emotions get in the way of efficiency. A few of the participants' quotes that illustrate the above summary:
P3: *"The second seemed the most appropriate for most situations, given that the third one is it's great that it's maybe more emotion, but it's often not going to react in exactly the right way, and getting that is so difficult"*
P1: *"I agree that the second version is best, cause it's reactive but it doesn't get in the way."*
P4: *"I quite like the other one actually (the third one), the fact that it kind of spoke back with you. Maybe, like, it being a bit sassy was a bit not appropriate at the time, but I think I'd quite like that to have on my phone. "*
P6: *"If it's personable it's fine, but if it's angry then probably not."*

Some of the participants also mentioned that they had been concentrating more on how the synthesis sounded in the demo rather than paying attention to the interface and the interruptions.
P3: *"cause I was listening to the voice and him getting more annoyed, I didn't notice that there was a bar of increasingly, like, one end was thunderstorm and one was happy, right?"*
P1: *"I didn't notice the face or anything because I was trying to work out, well what is good and what's bad about this voice."*

Furthermore, they mentioned they had never given reactive speech synthesis any thought nor ever dealt with it.
P1: *"I've probably never dealt with reactionary stuff before so perhaps the fact that it's stopping, I don't know quite how to judge what is good and what is bad"*

Finally, the focus group liked the idea of synthesis as an aid in human-human conversation rather than an interference.
P5: *"It could be used in a lecture type setting."*
P2: *"like it could work with a lecturer or somebody as opposed to in place of them. Yeah I think that's a good idea."*

4. Discussion

The focus group gave some useful feedback regarding reactive synthesis, and a renewed perspective of a what a group of people from the general public think about speech synthesis. In our focus group, participants had not previously come across reactive speech synthesis. Their first reaction –objecting to the system being annoyed– makes sense as they assumed the system should be cooperative. However, having annoyed or obstructive agents has a clear use case in a virtual environment for games or training purposes. These are types of use cases that people have yet to really come across.

We have discussed the relationship between incremental processing, re-planning and splicing. It should be clear that we consider reactiveness to depend on being able to re-plan the output of a synthesis system and splice it in at the right time. Incremental speech synthesis can be seen as a way of speeding up processing, but is not sufficient to achieve a reactive system. This leads to the following design guidelines for a reactive speech synthesis system:

1. **Fast enough** Whatever the chunk is (utterance, phrase, etc) the system must be able to synthesise replacement chunk within the required latency (we would suggest 200 ms as a minimum). For example, in our system where the chunk is a spurt (or intonational phrase) and 95% are less than 1850 ms for a 200 ms latency the system has to be at least 10x real-time.
2. **Splice audio in** You need to be able to tightly control audio output, i.e. be able to alter queued audio while it is waiting to be played and to know almost exactly what audio has been played. For multi-modal systems this would extend to the video output as well.
3. **Know how to respond** The appropriate response to an interruption varies considerably by application, as we found in our focus group, helpful systems should pause politely and rephrase, however for virtual characters a whole set of human responses to interruptions including rudely continuing may need to be implemented.

5. Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 645378 (Aria VALUSPA).

[1] A spurt is a section of speech with silence before and after it. The silence can be very short (as little as 20 ms). The CereProc spurt is a modified version of spurt as defined by Shriberg [10].

6. References

[1] M. Pouget, T. Hueber, G. Bailly, and T. Baumann, "HMM training strategy for incremental speech synthesis," in *Interspeech*, 2015, pp. 1201–1205.

[2] M. Astrinaki, N. d'Alessandro, B. Picart, T. Drugman, and T. Dutoit, "Reactive and continuous control of HMM-based speech synthesis," in *IEEE Spoken Language Technology Workshop (SLT)*, 2012, pp. 252–257.

[3] J. Edlund, "Incremental speech synthesis," in *Second Swedish Language Technology Conference*, 2008, pp. 53–54.

[4] T. Baumann and D. Schlangen, "INPRO_iSS: A component for just-in-time incremental speech synthesis," in *ACL 2012 System Demonstrations*, 2012, pp. 103–108.

[5] H. Buschmeier, T. Baumann, B. Dosch, S. Kopp, and D. Schlangen, "Combining incremental language generation and incremental speech synthesis for adaptive information presentation," in *13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2012, pp. 295–303.

[6] M. P. Aylett and C. J. Pidcock, "The Cerevoice characterful speech synthesiser SDK," in *Intelligent Virtual Agents (IVA)*, 2007, pp. 413–414.

[7] A. Bergkvist, C. Jennings, D. C. Burnett, A. Narayanan, and B. Aboba, *WebRTC 1.0: Real-time Communication Between Browsers*. W3C, 2012–2017. [Online]. Available: https://www.w3.org/TR/webrtc/

[8] T. Baumann and D. Schlangen, "Interactional adequacy as a factor in the perception of synthesized speech," in *8th ISCA Speech Synthesis Workshop*, 2013.

[9] "Guidelines for conducting a focus group," https://assessment.trinity.duke.edu/documents/How_to_Conduct_a_Focus_Group.pdf, 2005, accessed: 2017-03-01.

[10] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: findings and implications for automatic processing of multi-party conversation." in *Interspeech*, 2001, pp. 1359–1362.