Open-ended, Extensible System Utterances Are Preferred, Even If They Require Filled Pauses

Timo Baumann
Universität Hamburg
Department of Informatics
Germany
baumann@informatik.uni-hamburg.de

David Schlangen
University of Bielefeld
Faculty of Linguistics and Literary Studies
Germany
david.schlangen@uni-bielefeld.de

Abstract

In many environments (e. g. sports commentary), situations incrementally unfold over time and often the future appearance of a relevant event can be predicted, but not in all its details or precise timing. We have built a simulation framework that uses our incremental speech synthesis component to assemble in a timely manner complex commentary utterances. In our evaluation, the resulting output is preferred over that from a baseline system that uses a simpler commenting strategy. Even in cases where the incremental system *overcommits* temporally and requires a filled pause to wait for the upcoming event, the system is preferred over the baseline.

1 Introduction

In spontaneous speech, speakers often commit *temporally*, e. g. by starting utterances that they do not yet know how to complete (Clark, 1996), putting time pressure on them for the generation of a completion. While this may be for planning and efficiency reasons, it also enables them to start commenting on events for which the outcome is not yet known. For example when a ball is flying towards the goal, but it is uncertain yet whether it will hit, in sports commentary.

To accommodate this *incremental* behaviour, human speakers plan their utterances just somewhat ahead, typically in chunks of major phrases (Levelt, 1989), and remain flexible to change or abandon the original plan, or to hesitate, e. g. to adapt their timing. This flexibility is in contrast to speech output in spoken dialogue systems (SDSs) which typically generate, synthesize and deliver speech in units of full utterances that cannot be changed while ongoing, apart from being aborted or interrupted (Edlund, 2008).

Recently, incremental speech synthesis (iSS) has been presented (Dutoit et al., 2011; Baumann and Schlangen, 2012b) which allows to start partial utterances that are then smoothly extended during verbalization. Incremental spoken output for dialogue systems has been shown to improve naturalness (Buschmeier et al., 2012) and Skantze and Hjalmarsson (2010) have used filled pauses to hold a turn. Dethlefs et al. (2012) present an incremental NLG strategy to reduce the need for filled pauses in interactions.

We investigate the impact of incremental spoken output in a *highly dynamic* environment, that is, where the rate of external events is high enough to allow only few utterances to finish as planned. As an example, we choose an otherwise simple commentary domain, where incremental output enables the system to combine multiple events into one complex commenting utterance that takes into account predictions about upcoming events. If the system overcommits to the timing of future events, it autonomously uses a filled pause until more material becomes available.

2 Related Work

A paradigmatic example of a domain that uses open-ended utterances is sports commentary, which has received some attention in the NLG community. For example, Chen and Mooney (2008) present a system that learns from hand-annotated data what to comment on. However, attention seems to have been placed more on truthfulness of the content, as, judging from videos provided on their website,[1] the formulations that are produced are rather monotonic ("`pink7 dribbles towards the goal. pink7 shoots for the goal. pink7 passes to...`"). More importantly, the delivery of a produced utterance does not seem to be temporally tied to the occurrence of the event.

[1] `http://www.cs.utexas.edu/users/ml/clamp/sportscasting`

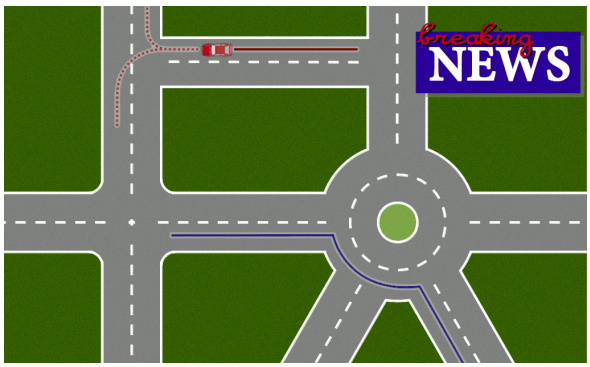


Figure 1: The map shown in the *CarChase* domain, including the car on one of its itineraries (red; another in blue). At the depicted moment we can assume that the car will take a turn, but do not know whether left or right.

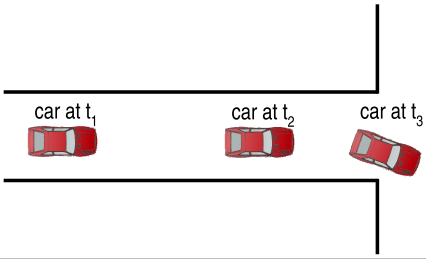
Repeatedly, utterances are synthesized long after the fact that they describe which sometimes has become obsolete at that point (for example, a goal is scored while the system still talks about a pass).

Lohmann et al. (2011) describe another domain that can be called highly dynamic: a system that adds spoken assistance to tactile maps for the visually impaired. In their settings, users can move around on a computer representation of a map with a hand-held haptic force-feedback device. Users are given spoken advice about the currently traversed streets' names, the relation of streets to each other, and to other map objects in the user's vicinity. Such exploratory moves by users can become rather quick, which in the system they describe can lead to output that comes late, referring to a position that has long been left.

3 A Highly Dynamic Commenting Domain

Our example domain combines properties of the sports commentary and map exploration domains mentioned above: the *CarChase* domain depicted in Figure 1. In the domain, a car drives around streets on the map and a commentator (supposed to be observing the scene from above) comments on where it is driving and what turns it is taking.

The car's itinerary in our domain simulator is scripted from a configuration file which assigns target positions for the car at different points in time and from which the motion and rotation of the car is animated. The speed of the car is set so that the event density is high enough that the setting cannot be described by simply producing one utterance per event – in other words: the domain is highly dynamic.



time	event description	ongoing utterance (already realized part in **bold**, newly appended continuation in *italic*)
t_1	car on Main Street	**The** car drives along Main Street.
t_2	car will likely turn	. . . **drives along Main Str**eet *and then turns* ‹hes›
t_3	car turns right	. . . **drives along Main Street and the**n turns *right.*

Figure 2: Example of incremental utterance production as a car drives along a street and turns. The ongoing utterance is extended as events unfold.

4 A Strategy for Incremental Commentary

We distinguish three types of events in the domain: *identification* (ID) events trigger the system to name the street the car is on, *turn* events fire when the car is taking a turn. Finally, *turn-prep* events fire when it is obvious that the car will turn but the direction of the turn remains open. These three event types are shown in Figure 2 at time t_1 (ID), t_2 (turn-prep), and t_3 (turn).

As can be seen in the example in Figure 2, the *turn-prep* event enables a system that is able to incrementally update its ongoing utterance to continue speaking about the anticipated future ("and then turns") without knowing the direction of the turn. This allows an incremental system to output efficient utterances that fluently combine multiple events and avoid repetition. Furthermore, *turn-prep* events enable the system to output the direction of the turn (the most important information) very shortly after the fact.

A non-incremental system, in contrast, must output individual utterances for every event and utterances can only start after the fact. Furthermore, a non-incremental system cannot extend ongoing utterances, rendering *turn-prep* events useless.

5 Implemented System

The system used for the experiment reported below uses an early version of incremental speech synthesis as implemented in INPROTK (Baumann and Schlangen, 2012c), a toolkit for incremental spoken dialogue processing based on the IU model (Schlangen and Skantze, 2009). The system allows to extend ongoing utterances, enabling the

incremental commenting strategy outlined above.

In addition, we implemented a capability to synthesize a hesitation if no more content is specified, and to continue as soon as content becomes available. (Thus, in contrast to (Skantze and Hjalmarsson, 2010), hesitations do not consume additional time.) By using hesitations, the system gracefully accommodates temporal *over-commitment* (i. e. the obligation to produce a continuation that is not fulfilled in time) which may occur, e. g. when the car drives slower than anticipated and a turn's direction is not yet known when the system needs it.

In the preliminary version of iSS used for the experiments, no prosodic integration of continuations takes place, resulting in prosodic discontinuities; see (Baumann and Schlangen, 2012a) for a detailed assessment of prosodic integration in iSS.

As we focus on the merit of iSS in this work, we did not implement a scene analysis/event detection nor a NLG component for the task.[2] Instead, the commentary is scripted from the same configuration file that controls the car's motion on the board. iSS events lag behind slightly, ensuring that visual analysis would be possible, and event/text correspondence is close, matching NLG capabilities.

6 Experiment

To evaluate the incremental system, we compared it to a non-incremental baseline system which is unable to alter speech incrementally and hence cannot smoothly extend ongoing partial utterances. Instead, the baseline system always produces full utterances, one per event. To ensure the temporal proximity of delivery with the causing event in the baseline system, utterances can be marked as optional (in which case they are skipped if the system is still outputting a previous utterance), or non-optional (in which case an ongoing utterance is aborted in favour of the new utterance). All 'turn' events in the domain were marked as optional, all street ID events as non-optional.

We devised 4 different configurations (including the itineraries shown in Figure 1), and the timing of events was varied (by having the car go at different speeds, or by delaying some events), resulting in 9 scenarios; in 3 of these, the incremental system generated one or more hesitations. Both systems' output for the 9 scenarios was recorded with a screen-recorder, resulting in 18 videos that were played in random order to 9 participants (university students not involved in the research). Participants were told that various versions of commentary-generating systems generated the commentary based on the running picture in the videos and were then asked to rate each video on a five-point Likert scale with regards to how natural (similar to a human) the spoken commentary was (a) formulated, and (b) pronounced. In total, this resulted in 81 paired samples for each question.[3]

The assumption (and rationale for the second question) was that the incremental system's formulations would result in higher formulation ratings, while we hoped the acoustic and prosodic artefacts resulting from the coarsely implemented incremental synthesis would not significantly hurt pronunciation ratings. In order to not draw the subjects' attention towards incremental aspects, no question regarding the timeliness of the commentary was asked for explicitly.

7 Results

The mean ratings for both formulation quality and pronunciation quality for the incremental and baseline systems is shown in Figure 3. The median differences in the ratings of the two conditions is 2 points on the Likert scale for question (a) and 0 points for question (b) (means of 1.66 and 0.51, respectively), favouring the incremental system. The sign test shows that the advantage of the incremental system is clearly significant for questions (a) (68+/9=/4-; $p < .0001$) and (b) (38+/30=/13-; $p < .0007$)[4].

Thus, it is safe to say that the production strategies enabled by incremental speech synthesis (i. e. starting to speak before all evidence is known and extending the utterance as information becomes available) allows for formulations in the spoken commentary that are favoured by human listeners.

Incremental behaviour in the 3 scenarios that required hesitations was rated significantly worse than in those scenarios without hesitations for both questions (t-tests, $p < .001$ (a) and $p < .01$ (b)). This

[2] However, Lohmann et al. (2012) present an incremental NLG strategy for a similar task.

[3] The experiment was conducted in one language (German) only, but we believe our results to carry over to other languages. Specifically, we assume that most or all languages cater for commenting, and believe that human commenters universally use their ability to integrate events late in the utterance. However, practices of commenting may work differently (and differently well) among languages.

[4] We also conducted a non-paired t-test for question (b), as the different formulations of the systems might have effects on pronunciation quality; this test was also significant ($p < .0012$).

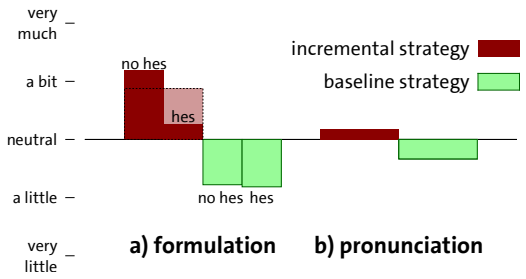


Figure 3: Mean ratings of formulation and pronunciation for the incremental and baseline systems; the formulation rating differs for utterances with and without hesitations in the incremental system.

is a clear indication that a system should try to avoid over-commitment, as users do not accept hesitations as inevitable (given that there was simply no evidence yet where the car would turn, for example). However, even in those scenarios that require filled pauses, the incremental commentary's formulation is rated as significantly better than the baseline system's (sign test, 18+/5=/4-; $p < .005$) while there is no effect on pronunciation in these cases.

8 Discussion & Outlook

The results indicate a clear user preference for open-ended, extensible utterances that grow as events unfold. Furthermore, this preference is stronger than the negative impact of filled pauses that are needed to cover temporal over-commitment, and despite the poor quality of filled pauses in the current system, which we plan to improve in the future.

Similarly to spoken commentary in dynamic domains, conversational speech requires revisions and reactions to events such as listener feedback, or the absence thereof (Clark, 1996). Thus, we believe that our results, as well as iSS in general, also apply to a broad range of conversational SDS tasks.

Finally, synthesis quality appears to be less important than interaction *adequacy*: we found no difference in rating of perceptual quality ('pronunciation') between the variants, even though in isolation iSS sounded noticeably worse in the prototype. This result calls for interactive adequacy as an optimization target over (isolated) perception ratings for speech synthesis, and also challenges the use of canned speech in conversational SDSs, which does not adapt to the interaction.

Acknowledgements The first author would like to thank Wolfgang Menzel for fruitful discussions on the topic, and permanent encouragement.

References

Timo Baumann and David Schlangen. 2012a. Evaluating prosodic processing for incremental speech synthesis. In *Procs. of Interspeech*, Portland, USA.

Timo Baumann and David Schlangen. 2012b. INPRO_iSS: A component for just-in-time incremental speech synthesis. In *Proceedings of ACL System Demonstrations*, Jeju, Korea.

Timo Baumann and David Schlangen. 2012c. The INPROTK 2012 release. In *Proceedings of SDCTD*, Montréal, Canada.

Hendrik Buschmeier, Timo Baumann, Benjamin Dorsch, Stefan Kopp, and David Schlangen. 2012. Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In *Procs. of SigDial*, pages 295–303, Seoul, Korea.

David L. Chen and Raymond J. Mooney. 2008. Learning to sportscast: A test of grounded language acquisition. In *Proceedings of 25th Int. Conference on Machine Learning (ICML)*, Helsinki, Finland.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press.

Nina Dethlefs, Helen Hastie, Verena Rieser, and Oliver Lemon. 2012. Optimising incremental generation for spoken dialogue systems: Reducing the need for fillers. In *Procs. of the Seventh Int. Natural Language Generation Conf.*, pages 49–58, Utica, USA.

Thierry Dutoit, Maria Astrinaki, Onur Babacan, Nicolas d'Alessandro, and Benjamin Picart. 2011. pHTS for Max/MSP: A Streaming Architecture for Statistical Parametric Speech Synthesis. Technical Report 1, numediart Research Program on Digital Art Technologies.

Jens Edlund. 2008. Incremental speech synthesis. In *Second Swedish Language Technology Conference*, pages 53–54, Stockholm, Sweden. System Demo.

William J.M. Levelt. 1989. *Speaking: From Intention to Articulation*. MIT Press.

Kris Lohmann, Carola Eschenbach, and Christopher Habel. 2011. Linking spatial haptic perception to linguistic representations: assisting utterances for tactile-map explorations. In *Spatial information theory*, pages 328–349, Berlin, Heidelberg. Springer.

Kris Lohmann, Ole Eichhorn, and Timo Baumann. 2012. Generating situated assisting utterances to facilitate tactile-map understanding: A prototype system. In *Procs. of SLPAT 2012*, Montréal, Canada.

David Schlangen and Gabriel Skantze. 2009. A General, Abstract Model of Incremental Dialogue Processing. In *Procs. of the EACL*, Athens, Greece.

Gabriel Skantze and Anna Hjalmarsson. 2010. Towards incremental speech generation in dialogue systems. In *Procs. of SigDial*, pages 1–8, Tokyo, Japan.