Coordinating Speech Delivery to Gesture Progress for Deictic Expressions with Incremental Speech Synthesis

Timo Baumann
Universität Hamburg
Informatics Department
Natural Language Systems Division
Hamburg, Germany
baumann@informatik.uni-hamburg.de

ABSTRACT

This paper presents ongoing work in incremental speech synthesis that enables a system to adapt speech delivery to unforeseen changes in the timing of motor events (e. g. a robot actuator working faster or slower than anticipated) in order to improve the coordination of speech and gestures for deictic expressions.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Speech synthesis*; I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Coherence and coordination*

General Terms

Design, Human factors

Keywords

Incremental processing, speech synthesis, online adaptation

1. INTRODUCTION

Deictic expressions can combine speech and gesture modalities to convey meaning (possibly encoded redundantly across modalities) and a lot of work has been done on interpreting and on producing speech and gesture alike (e. g. [14]).

While the role of coordination between speech and gesture is contested for human machine interaction [10], it is crucial for correct understanding that multiple deictic expressions be aligned with their corresponding gestures (e. g.: "*move **this** piece over **there** through **that** gate.*"). Above that, it is certain that humans are able to and often do finely coordinate their speech and gestures [6].

Speech and gesture co-planning and generation face the problem of variable actuator performance and hence unpredictable gesture timings, which result in deviations between planned and actual robot motion. At the same time, sensors may be available that inform about the actual state of the robot motion; the work presented here aims to make use of the resulting information about delivery discrepancies.

Speech synthesizers generate timings (and other aspects) of individual speech sounds, words, and phrases based on prosodic constraints or rules, but – to the best of the authors knowledge – largely lack interfaces to adapt timings to account for external requirements, especially if these external requirements change over the course of the utterance.

This paper presents ongoing work on enabling incremental speech synthesis to account for the timing difficulties of robotic gesture production by adapting to external timing requirements, as proposed by Lohse and Welbergen [9]. The next section describes the incremental speech synthesizer Inpro_iSS, Section 3 proposes a binding mechanism for speech and gesture and Section 4 explains how speech tempo adaptation could be performed. Finally, Section 5 discusses the approach and lays out future work.

2. INCREMENTAL SPEECH SYNTHESIS

INPRO_iSS [3] is a speech synthesis component for HSMM speech synthesis [15] that internally makes use of MaryTTS [12] and is implemented in the incremental processing toolkit INPROTK [4] and based on the IU model for incremental processing [11] in which all information is contained in increments that are linked to each other and that are changed in the IU network as the state of the system evolves.

As can be seen in Figure 1, the component performs all computationally expensive processing steps (such as waveform synthesis) as late as possible, while performing prosodic processing (which has non-local effects) as early as necessary [2], resulting in fast response times without sacrificing quality. Specifically, ongoing synthesis can still be changed and adapted prosodically with minimal latency and processing overhead, as re-computations can be kept to a minimum. For example, in Figure 1, at the time the system is finishing the word "*the*", vocoding is only a few frames of audio ahead and vocoding is only ahead one or two phonemes to account for co-articulation effects between speech sounds. Thus, it is trivial to change the timing of speech sounds (or other characteristics, even what is to be said) with very little delay. In this project, we make use of this characteristic, by changing the planned delivery duration of words to be synthesized to match the anticipated remaining duration of gestures to be

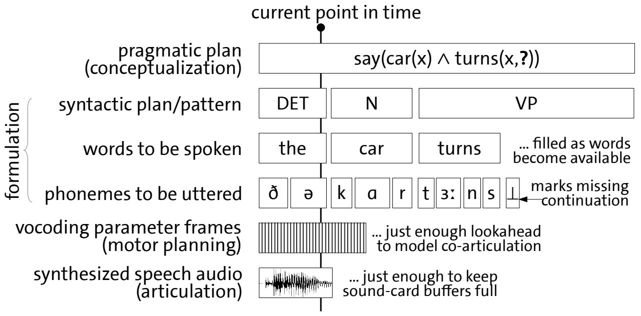
current point in time
pragmatic plan (conceptualization)
say(car(x) ∧ turns(x,?))
formulation
syntactic plan/pattern
DET
N
VP
words to be spoken
the
car
turns
... filled as words become available
phonemes to be uttered
ð
ə
k
ɑ
r
t
ɜː
n
s
⊥
marks missing continuation
vocoding parameter frames (motor planning)
... just enough lookahead to model co-articulation
synthesized speech audio (articulation)
... just enough to keep sound-card buffers full

Figure 1: Schematic view of 'lazy' just-in-time incremental speech synthesis as implemented in Inpro_iSS.

performed in synchrony to some stretch of speech.

3. GESTURE-SPEECH COORDINATION

As outlined above, we expect that robot sensors are available which enable the system to determine the diversion of actual from expected behaviour progress. This information should allow to compute a distribution of the expected remaining duration in terms of expected mean and standard deviation.

Furthermore, we propose that speech should be coordinated with gestures through some form of *anchoring* which may come in two forms: (a) *punctual* anchors that secure the coincidence of some speech segment (e. g. a syllable nucleus) to some gesture phase (e. g. the stroke); and (b) *progress* anchors that keep speech and gesture in synchrony over a certain timespan. It turns out that progress anchoring can be achieved through punctual anchors at the beginning and end of the synchronized timespans. Using incremental speech synthesis, the timings of the not-yet delivered speech can be adapted to meet the expected remaining duration that is to be spanned before the next anchor.

In addition, as Inpro_iSS allows access to the individual speech sounds involved, to the syllables, or to words (see Figure 1), the level at which anchoring takes place need not be pre-determined and could even vary in one system, based on the requirements of anchoring detail. Possible timing adaptation strategies are discussed in the following section.

4. ADAPTATION OF ONGOING SPEECH

Incremental synthesis allows to change the timing of speech as it is being uttered. Very simple approaches, such as redispatching individual words to meet start-time and duration demands [1], or simple linear scaling of speech sounds [3] show that it is possible to meet timing goals, however with a strong negative impact on speech quality. For this reason, this work proposes to use non-uniform scaling methods that are linguistically motivated. Technically, non-uniform scaling can be performed in one of two ways [13], either by integrating time-scaling into the TTS's linguistic pre-processing, or by post-processing already generated timings. The latter approach appears to be more practical in the incremental, minimal-delay use-case.

Previous work has analyzed the influence of speech rate on per-phoneme durations (i. e. estimated the *stretchability* from data), however only in the context of simple CV syllable structures [8]. Another approach, for time-scaling speech for computer-assisted language learning, uses different scaling factors for (in increasing order) plosives, vowels, voiced and unvoiced consonants and silence, with either linearly or exponentially increasing factors [7]. In listening experiments, their approach outperforms simpler scaling methods that do not take into account segmental information at all.

Finally, one could compute or derive from data non-linear scalings for all individual phonemes (possibly given their different contexts) e. g. using regression trees. However, the precise scaling of a stretch of speech (which would consist of a sequence of such non-linear scalings) would be computationally expensive and training such trees would require large amounts of data. This work hence proposes to use linear scaling factors that vary between phonemes (or classes of phonemes) with the respective factors estimated from a corpus of speech at different speech rates.

Time-scaling is a solution to small deviations between desired and planned timing. However, for longer gesture progress interruptions, slowing down speech is not a solution. Speech synthesis should be able to automatically give feedback about the expected synthesis quality degradation and/or combine tempo changes with automatically inserted hesitations [5] for larger timing revisions. Of course, one could imagine automatic rephrasing to take place (e. g. changing "***this** piece*" to "*the red piece **here***") in order to gain time, or to give up on speech-gesture coordination altogether ("*the red piece to my left*"). In principle, the IU architecture should support such behaviour.

5. DISCUSSION AND FUTURE WORK

This paper presents work in progress on realistic time-scaling of ongoing incremental speech synthesis in order to be able to coordinate speech delivery with gesture progress, e. g. for the precise alignment of deictic expressions.

Using incremental speech synthesis, upcoming speech can be scaled to meet timing demands, but the currently implemented scaling methods in Inpro_iSS result in a deterioration of naturalness. The current work aims to devise and implement better time-scaling methods for speech synthesis to allow speech-rate variations within the utterance with little or no delay.

One step not yet taken in the literature is the individual time-scaling of sub-phoneme units: there are multiple HMM states per phoneme and transition phases between phonemes should potentially be differently scaled than the phoneme's stable phase. It is unlikely that the decision trees for HMM state selection are suitable in the present case, as they are usually not trained on rate-varying speech.

Of course, once implemented, the actual coordination performance must be evaluated, both quantitatively (how accurate does the alignment perform?) as well as qualitatively (in terms of deterioration of speech naturalness, and in terms of improvement of rated speech-gesture coordination). Finally, task performance in utterances where coordination matters for task success (e. g. when multiple deictic gestures are employed) will be measured.

Acknowledgements

The author would like to thank Daimler-und-Benz-Stiftung for funding this research.

6. REFERENCES

[1] T. Baumann and D. Schlangen. Predicting the Micro-Timing of User Input for an Incremental Spoken Dialogue System that Completes a User's Ongoing Turn. In *Proceedings of SigDial 2011*, Portland, USA, 2011.

[2] T. Baumann and D. Schlangen. Evaluating prosodic processing for incremental speech synthesis. In *Proceedings of Interspeech*, Portland, USA, Sept. 2012. ISCA.

[3] T. Baumann and D. Schlangen. INPRO_iSS: A component for just-in-time incremental speech synthesis. In *Procs. of ACL System Demonstrations*, Jeju, Korea, July 2012.

[4] T. Baumann and D. Schlangen. The INPROTK 2012 release. In *Proceedings of SDCTD*, Montréal, Canada, 2012.

[5] T. Baumann and D. Schlangen. Interactional adequacy as a factor in the perception of synthesized speech. In *Proceedings of Speech Synthesis Workshop (SSW8)*, 2013. to appear.

[6] K. Bergmann, V. Aksu, and S. Kopp. The relation of speech and gestures: Temporal synchrony follows semantic synchrony. In *2nd Workshop on Gesture and Speech in Interaction (GeSpIn 2011)*, Bielefeld, Germany, 2011.

[7] O. Donnellan, E. Jung, and E. Coyle. Speech-adaptive time-scale modification for computer assisted language-learning. In *Advanced Learning Technologies, 2003. Proceedings. The 3rd IEEE International Conference on*, pages 165–169. IEEE, 2003.

[8] H. Kuwabara. Acoustic properties of phonemes in continuous speech for different speaking rate. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 4, pages 2435–2438. IEEE, 1996.

[9] M. Lohse and H. van Welbergen. Designing appropriate feedback for virtual agents and robots. Position paper at RO-MAN 2012 Workshop Robot Feedback in Human-Robot Interaction: How to Make a Robot Readable for a Human Interaction Partner, 2012.

[10] S. Oviatt. Ten myths of multimodal interaction. *Communications of the ACM*, 42(11):74–81, 1999.

[11] D. Schlangen and G. Skantze. A General, Abstract Model of Incremental Dialogue Processing. In *Proceedings of the EACL*, pages 710–718, Athens, Greece, 2009.

[12] M. Schröder and J. Trouvain. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(3):365–377, Oct. 2003.

[13] J. Trouvain. *Tempo Variation in Speech Production: Implications for Speech Synthesis.* PhD thesis, Universität des Saarlandes, Saarbrücken, Germany, 2003.

[14] W. Wahlster. Smartkom: Fusion and fission of speech, gestures, and facial expressions. In *Proceedings of the 1st International Workshop on Man-Machine Symbiotic Systems*, pages 213–225. MIT Press, 2002.

[15] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. A hidden semi-markov model-based speech synthesis system. *IEICE Transactions on Information and Systems*, 90(5):825–834, 2007.