Interactional Adequacy as a Factor in the Perception of Synthesized Speech

Timo Baumann and David Schlangen
baumann@informatik.uni-hamburg.de
Interactional Adequacy as a Factor in the Perception of Synthesized Speech
Take-home message

Interactional Adequacy

a Factor in the

Perception of Synthesized Speech
Interactional Adequacy is a Factor in the Perception of Synthesized Speech
Take-home message

*Interactional Adequacy* is a Factor in the Perception of Synthesized Speech

... and may be more important than synthesis quality in interactive systems
• Interactional Adequacy:
  shortcomings of speech output in spoken dialogue systems

• Possible Solution:
  incremental processing

• Experiment:
  is synthesis quality that important?

• Results & Conclusion
Speech Output in Typical Systems

- full utterances are generated, synthesized and delivered as a whole

There's an appointment today at 4:25 titled: ‘afternoon tea’ with the note: ‘be on time’.
Speech Output in Typical Systems

- potentially slow, as all processing is utterance-initial
  → reason for canned speech in deployed systems

There's an appointment today at 4:25 titled: ‘afternoon tea’ with the note: ‘be on time’.
Speech Output in Typical Systems

There's an appointment today at 4:25 titled: ‘afternoon tea’ with the note: ‘be on time’.

- inflexible: unable to change the ongoing utterance
  - no way to react to the listener or the environment
Speech Output in Typical Systems

There's an appointment today at 4:25 titled: ‘afternoon tea’ with the note: ‘be on time’.

- inflexible: unable to change the ongoing utterance
  - no way to react to the listener or the environment
Potentially Better: Incremental Speech Output

There's an appointment today at 4:25 titled: ‘afternoon tea’ with the note: ‘be on time’.

- generate, synthesize and deliver the utterance in smaller chunks
Potentially Better: Incremental Speech Output

current point in time

There's an appointment today at 4:25 titled: 'afternoon tea' with the note: 'be on time'.

- less utterance-initial processing → faster onset
Potentially Better:
Incremental Speech Output

- incremental output may take changes into account
- react and adapt to user feedback / requests / noise

current point in time

There's an appointment today at 4:25 titled: 'afternoon tea' with the note: 'be on time'.

at 4:25, titled: 'afternoon tea' …
**Speech Output: Overall Architecture**

<table>
<thead>
<tr>
<th>Pragmatic Plan (Conceptualization)</th>
<th>say(peter(x) ∧ gate(y) ∧ open(x,y))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic Plan/Pattern</td>
<td>N  V  NP</td>
</tr>
<tr>
<td>Words to be Spoken</td>
<td>Peter  opened  the  gate</td>
</tr>
<tr>
<td>Phonemes to be Uttered</td>
<td>p i t ə oʊ p ə n d ʊ t</td>
</tr>
<tr>
<td>Vocoding Parameter Frames (Motor Planning)</td>
<td></td>
</tr>
<tr>
<td>Synthesized Speech Audio (Articulation)</td>
<td></td>
</tr>
</tbody>
</table>
Speech Output: Overall Architecture

pragmatic plan (conceptualization) say(peter(x) ∧ gate(y) ∧ open(x,y))
syntactic plan/pattern N V NP
words to be spoken Peter opened the gate
phonemes to be uttered p iː t ə oʊ p ə n d ð ə g ɛɪ t
vocoding parameter frames (motor planning)
synthesized speech audio (articulation) "reversed" for TTS
A *Just-In-Time* Formulation for Incremental Speech Synthesis

Current point in time

- Pragmatic plan (conceptualization)
- Syntactic plan/pattern
- Words to be spoken
- Phonemes to be uttered
- Vocoder parameter frames (motor planning)
- Synthesized speech audio (articulation)

```latex
\text{say}(\text{peter}(x) \land \text{open}(x,?))
\text{N} \quad \text{V} \quad \text{NP}
```
A Just-In-Time Formulation for Incremental Speech Synthesis

- Pragmatic plan (conceptualization)
- Syntactic plan/pattern
- Words to be spoken
- Phonemes to be uttered
- Vocoder parameter frames (motor planning)
- Synthesized speech audio (articulation)

Current point in time:

- `say(peter(x) ∧ open(x, ?))`

Words to be spoken:

- Peter
- opened
- the
- N

Syntactic plan/pattern:

- NP

Phonemes to be uttered:

- N

Vocoder parameter frames (motor planning):

- (articulation)

Synthesized speech audio:

- (articulation)
A Just-In-Time Formulation for Incremental Speech Synthesis

Just-In-Time Formulation for Incremental Speech Synthesis

- Pragmatic plan (conceptualization)
- Syntactic plan/pattern
- Words to be spoken
- Phonemes to be uttered
- Vocoder parameter frames (motor planning)
- Synthesized speech audio (articulation)

Current point in time

Say(peter(x) \land open(x, ?))

NP

Peter opened the N

...filled as placeholder (above) is instantiated

p ə t ə oʊ pʊ̯ ə nd ð
A Just-In-Time Formulation for Incremental Speech Synthesis

```
say(peter(x) ∧ open(x, ?))
```

- **Pragmatic plan** (conceptualization)
- **Syntactic plan/pattern**
- **Words to be spoken**
  - Peter
  - opened
  - the
- **Phonemes to be uttered**
- **Vocoding parameter frames** (motor planning)
- **Synthesized speech audio** (articulation)

- **Current point in time**
- **Filled as placeholder** (above) is instantiated
- **Just enough lookahead** to model co-articulation
A Just-In-Time Formulation for Incremental Speech Synthesis

- **Pragmatic plan (conceptualization)**
  - say(peter(x) ∧ open(x,?))

- **Syntactic plan/pattern**
  - N V NP
  - Peter opened the N

- **Phonemes to be uttered**
  - p i t a oʊ p ə nd ð

- **Vocoding parameter frames (motor planning)**
  - Just enough lookahead to model co-articulation

- **Synthesized speech audio (articulation)**
  - Just enough to keep sound-card buffers full
A Just-In-Time Formulation for Incremental Speech Synthesis

more details on the implemented system in Baumann&Schlangen, ACL-Demo 2012.
Goals of Incremental Synthesis

- start speaking before processing has completed
  - *fold* processing time into delivery time
  - also: start before everything to be spoken about is known
- twiddle with vocoding parameters in real-time
  - all the amazing work done by MAGE/pHTS people
- accommodate change / extension of utterances
  - with minimal recomputation
  - but: need some lookahead / prediction for smooth prosody
Goals of Incremental Synthesis

- start speaking before processing has completed
  - *fold* processing time into delivery time
  - also: start before everything to be spoken about is known
- twiddle with vocoding parameters in real-time
  - all the amazing work done by MAGE/pHTS people
- accommodate change / extension of utterances
  - with minimal recomputation
  - but: need some lookahead / prediction for smooth prosody

more information on lookahead/prosody trade-off in Baumann&Schlangen, Interspeech 2012.
given that incremental speech synthesis measurable degrades prosodic parameters – → does this degradation matter to listeners?

(based on our Interspeech'12 findings)
Example: The CarChase domain

- system comments on events in the scene (car's motion)
- high event rate → impossible to speak isolated utterances
  - combine events into complex utterances
    (using incremental speech synthesis)
  - skip or abort event notifications
    in favour of more important
    information (baseline behaviour)
- simplification of similar
  real-world scenarios
  (like basketball commentary)
Taking expectations into account

<table>
<thead>
<tr>
<th>time</th>
<th>event description</th>
<th>ongoing utterance (spoken part in <strong>bold</strong>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>car on Main Street</td>
<td><strong>The</strong> car drives along Main Street.</td>
</tr>
<tr>
<td>$t_2$</td>
<td>car will have to turn</td>
<td>... <strong>Main Street</strong> and then turns ⟨hes⟩</td>
</tr>
<tr>
<td>$t_3$</td>
<td>car turns right</td>
<td>... <strong>Main Street and then</strong> turns right.</td>
</tr>
</tbody>
</table>

more details on interaction strategy in Baumann&Schlangen, SigDial 2013.
Taking expectations into account

<table>
<thead>
<tr>
<th>time</th>
<th>event description</th>
<th>ongoing utterance (spoken part in <strong>bold</strong>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>car on Main Street</td>
<td>The car drives along Main Street.</td>
</tr>
<tr>
<td>$t_2$</td>
<td>car will have to turn</td>
<td>... <strong>Main Street</strong> and then turns <em>(hes)</em></td>
</tr>
<tr>
<td>$t_3$</td>
<td>car turns right</td>
<td>... <strong>Main Street and then</strong> turns right.</td>
</tr>
</tbody>
</table>

more details on interaction strategy in Baumann&Schlangen, SigDial 2013.
### Taking expectations into account

#### event: identify street

<table>
<thead>
<tr>
<th>time</th>
<th>event description</th>
<th>ongoing utterance (spoken part in <strong>bold</strong>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>car on Main Street</td>
<td><strong>The car</strong> drives along Main Street.</td>
</tr>
<tr>
<td>$t_2$</td>
<td>car will have to turn</td>
<td>... <strong>Main Street</strong> and then turns <em>hes</em></td>
</tr>
<tr>
<td>$t_3$</td>
<td>car turns right</td>
<td>... <strong>Main Street</strong> and <strong>then</strong> turns right.</td>
</tr>
</tbody>
</table>

#### event: turning right

More details on interaction strategy in Baumann & Schlangen, SigDial 2013.
Taking expectations into account

<table>
<thead>
<tr>
<th>time</th>
<th>event description</th>
<th>ongoing utterance (spoken part in <strong>bold</strong>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t_1$</td>
<td>car on Main Street</td>
<td>The car drives along Main Street.</td>
</tr>
<tr>
<td>$t_2$</td>
<td>car will have to turn</td>
<td>... <strong>Main Street</strong> and then turns (hes)</td>
</tr>
<tr>
<td>$t_3$</td>
<td>car turns right</td>
<td>... <strong>Main Street and then</strong> turns right.</td>
</tr>
</tbody>
</table>

more details on interaction strategy in Baumann&Schlangen, SigDial 2013.
Experiment

- incremental system vs. baseline system
- 9 settings in the CarChase domain
- 9 subjects were asked to rate (5-point Likert)
  - naturalness of verbalization (to capture interactional adequacy)
  - naturalness of pronunciation (to capture synthesis quality)
- results in 81 paired samples

- incremental processing implemented in InproTK, using speech synthesis technology from MaryTTS

Expected results

- we were hoping for a good trade-off:

interation quality  synthesis quality
Expected results

- we were hoping for a good trade-off:

  ![Diagram with axes labeled naturalness and interaction quality vs. synthesis quality. The area under the curve represents a great improvement with the incremental system.]
Expected results

- we were hoping for a good trade-off:

  - great improvement with the incremental system
  - slight advantage for baseline system
Expected results

- we were hoping for a good trade-off:

  - great improvement with the incremental system
  - slight advantage for baseline system

→ write paper: „Trade-off between incrementality of behaviour and speech synthesis quality“
Actual results
Actual results

as expected: great improvement in verbalization
Actual results

as expected: great improvement in verbalization

very natural

incremental baseline

neutral

verbalization

very unnatural

synthesis quality impression also improves!

pronunciation
Pronunciation ratings

- Incremental processing cannot have systematically improved synthesis quality
  - incremental synthesis was previously shown to lead to a slight quality degradation (Dutoit et al., 2011)

- but:
  naïve listeners do not distinguish between interaction and synthesis quality (Pearson's r = .537)

- verbalization/wording adequacy seems to outweigh pronunciation/synthesis quality
Conclusions

- adequate verbalization / wording in a given context
  - may be more important than synthesis quality
  - may even lead to better synthesis quality ratings!

- applicability to interactive / multi-modal use is rarely an issue when valuating speech synthesis systems / approaches
  - good response timing and adequate behaviour can be crucial in interactive environments

- perceived synthesis quality can be improved by improving other (easier?) aspects of the system
Thank you.

baumann@informatik.uni-hamburg.de
get the code at inprotk.sf.net.

Thanks to Petra Wagner and Wolfgang Menzel.
“Covering up” with filled pauses

- synthesis may be faster than expected or development of events may be slower than anticipated
- we synthesize a filled pause („uhm“) in this case

- incremental formulations are still preferred in these cases

a) formulation  

b) pronunciation