

# Interactional Adequacy as a Factor in the Perception of Synthesized Speech

Timo Baumann and David Schlangen  
[baumann@informatik.uni-hamburg.de](mailto:baumann@informatik.uni-hamburg.de)

# Take-home message

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

# Take-home message

*Interactional Adequacy*

**xx** a Factor in the

Perception of Synthesized Speech

# Take-home message

*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

# Content

- 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

current point in time

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

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

# Speech Output in Typical Systems

current point in time



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

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



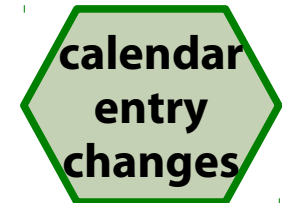
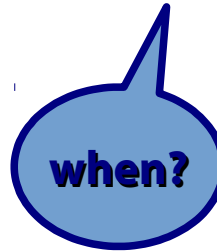
# Speech Output in Typical Systems

current point in time

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



**user feedback**



- inflexible: unable to change the ongoing utterance
  - no way to react to the listener or the environment

# Speech Output in Typical Systems

current point in time

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



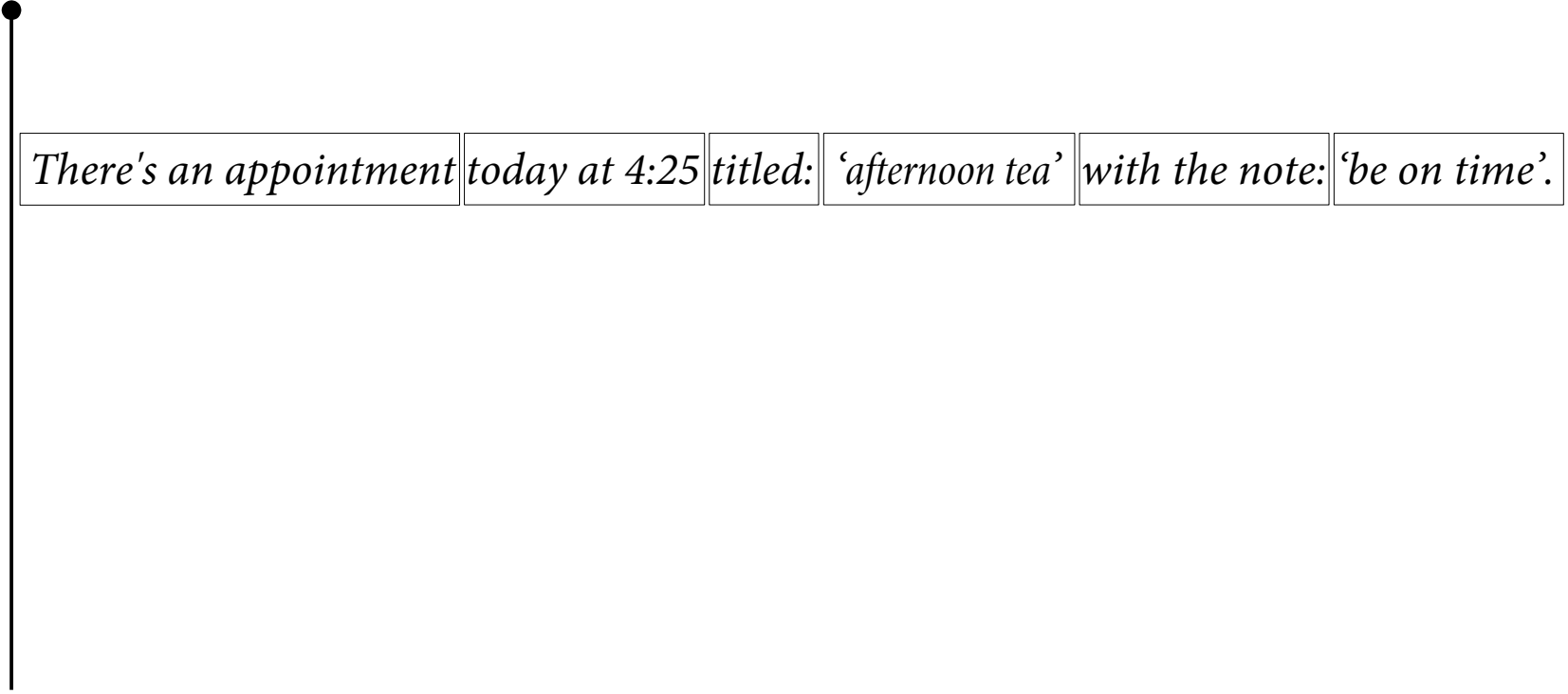
**user feedback**



- inflexible: unable to change the ongoing utterance
  - no way to react to the listener or the environment

# Potentially Better: Incremental Speech Output

current point in time

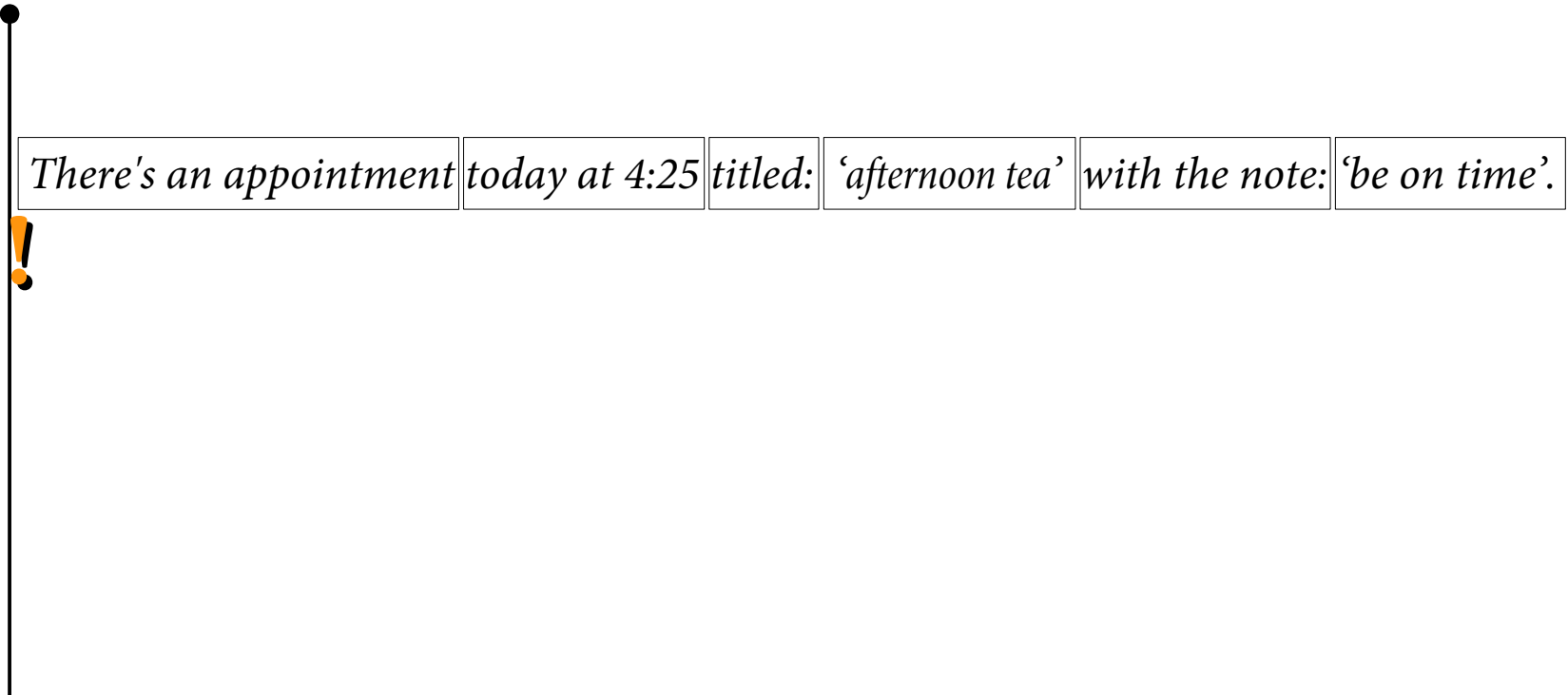


*There's an appointment today at 4:25 titled: 'afternoon tea' with the note: 'be on time'.*

- generate, synthesize and deliver the utterance in smaller *chunks*

# Potentially Better: Incremental Speech Output

current point in time

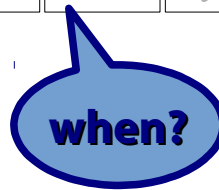


- less utterance-initial processing → faster onset

# Potentially Better: Incremental Speech Output

current point in time

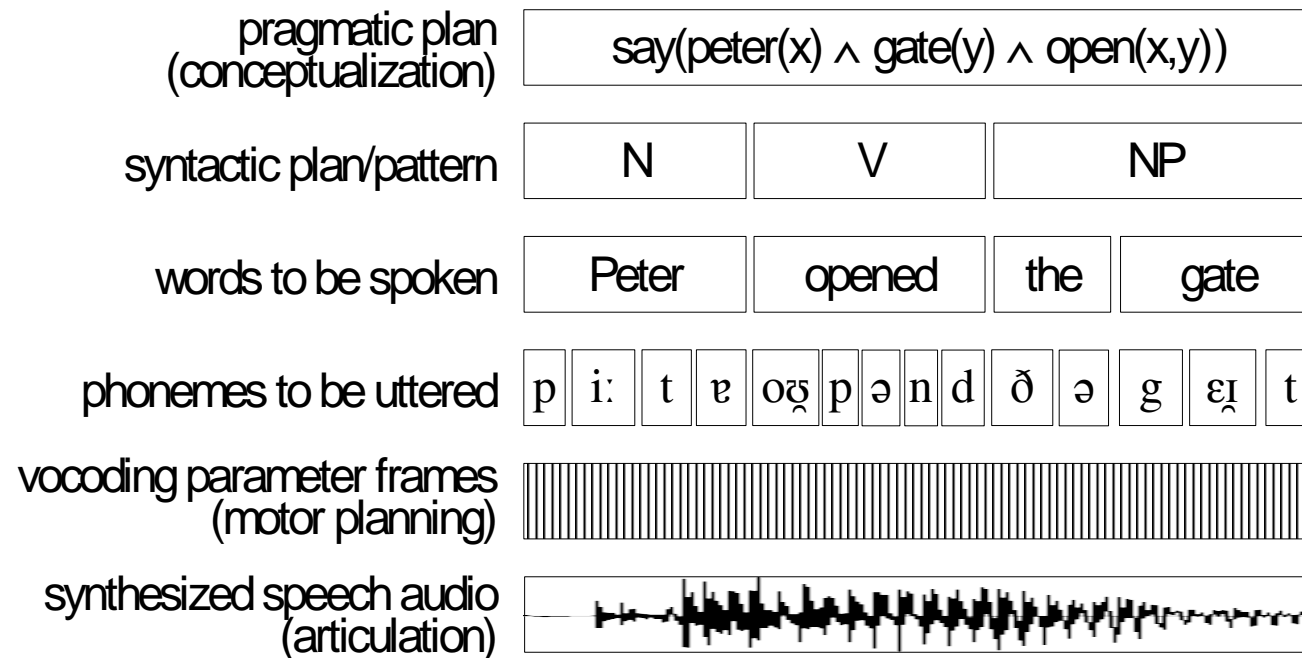
There's an appointment today at 4:25 titled: 'afternoon tea' with the note: 'be on time'.



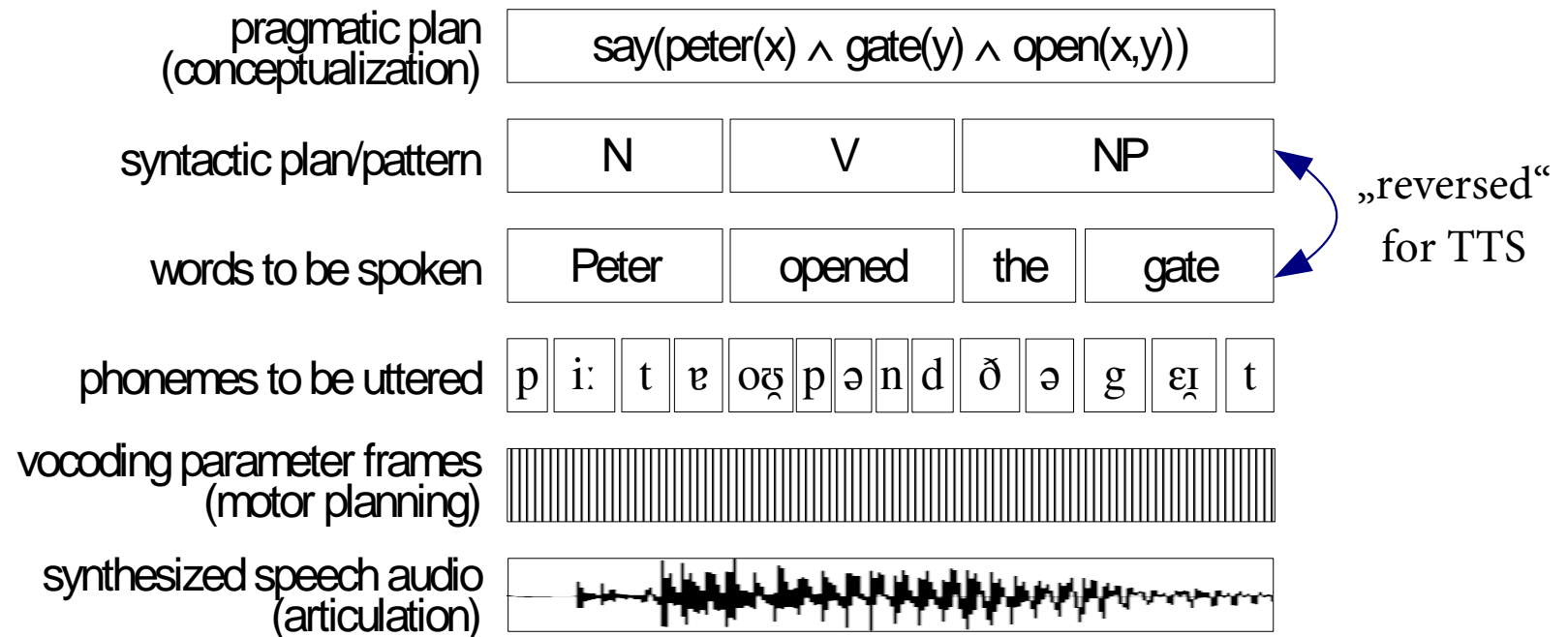
at 4:25, titled: 'afternoon tea' ...

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

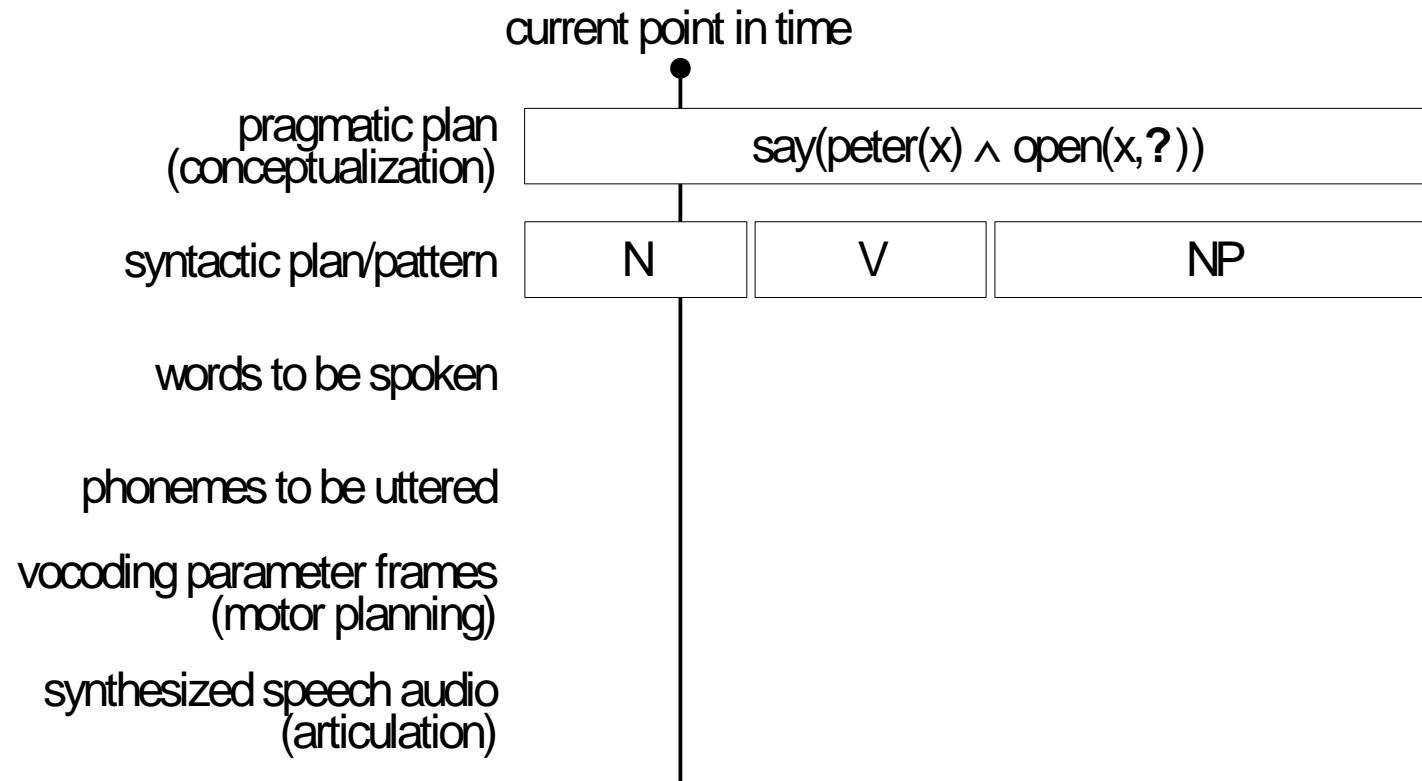
# Speech Output: Overall Architecture



# Speech Output: Overall Architecture

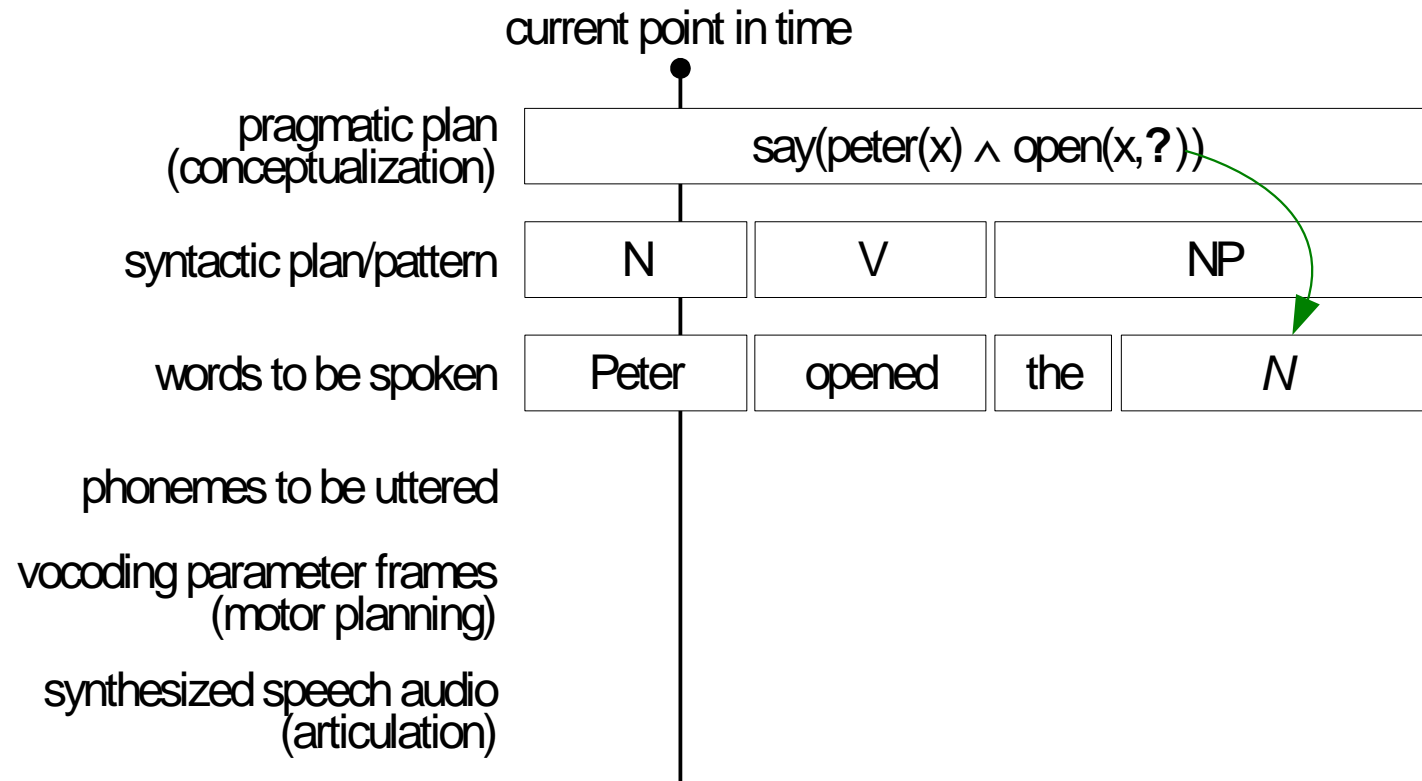


# A *Just-In-Time* Formulation for Incremental Speech Synthesis

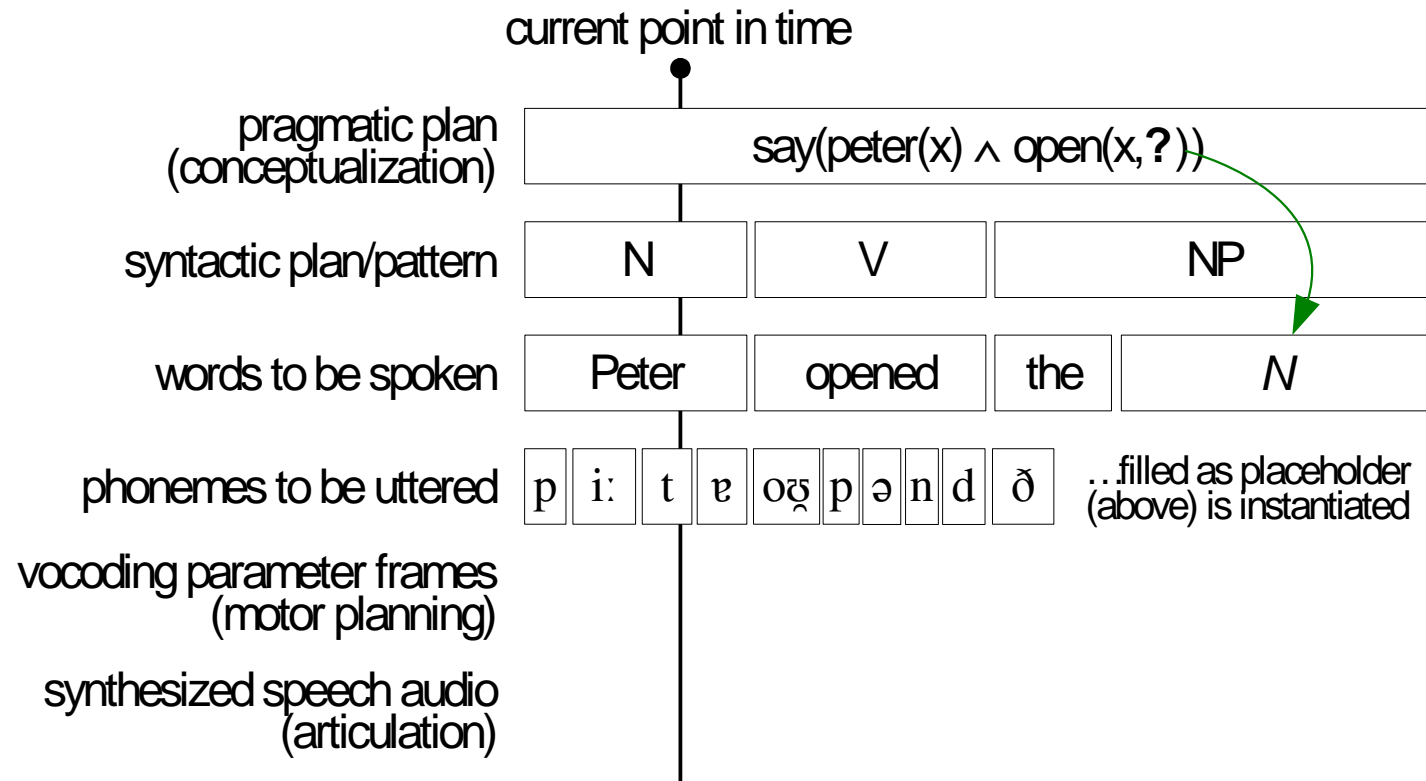




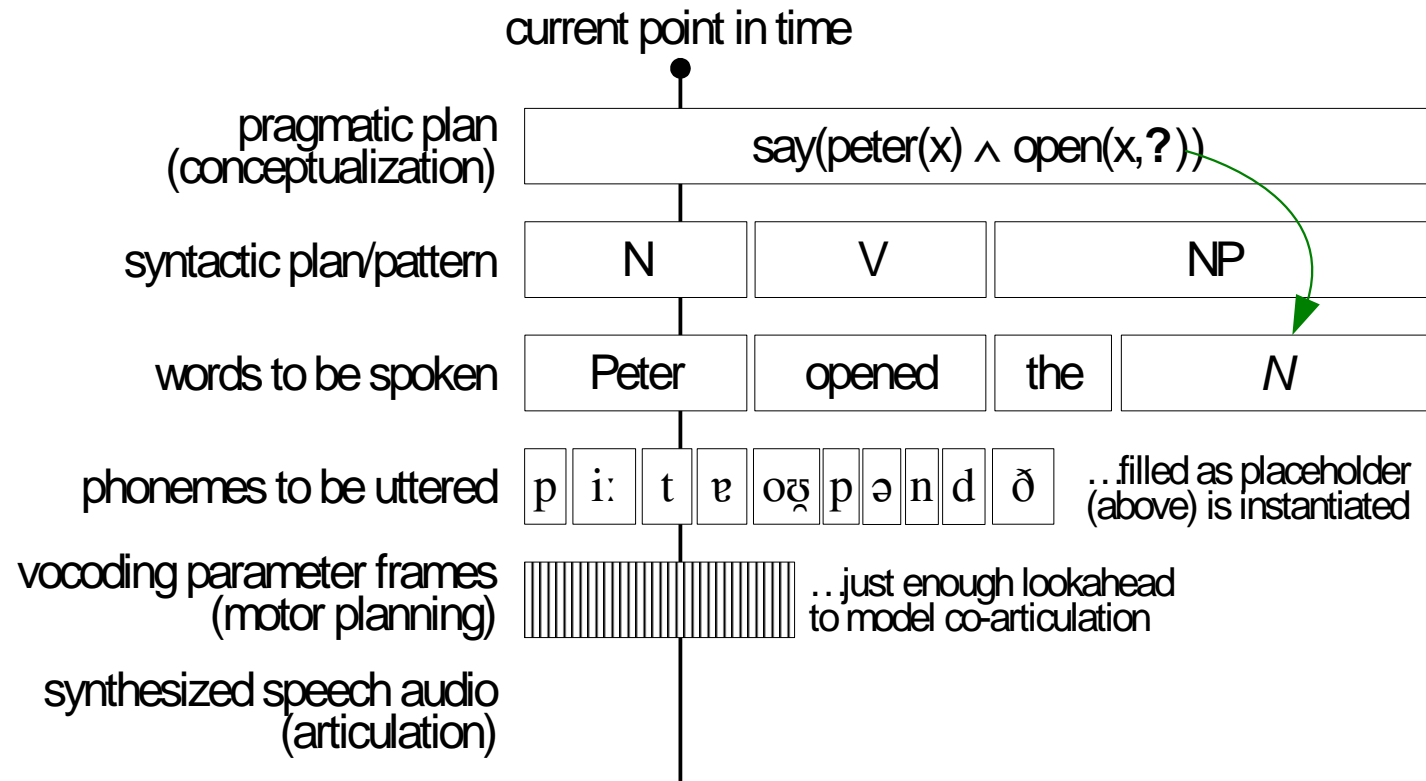
# A *Just-In-Time* Formulation for Incremental Speech Synthesis



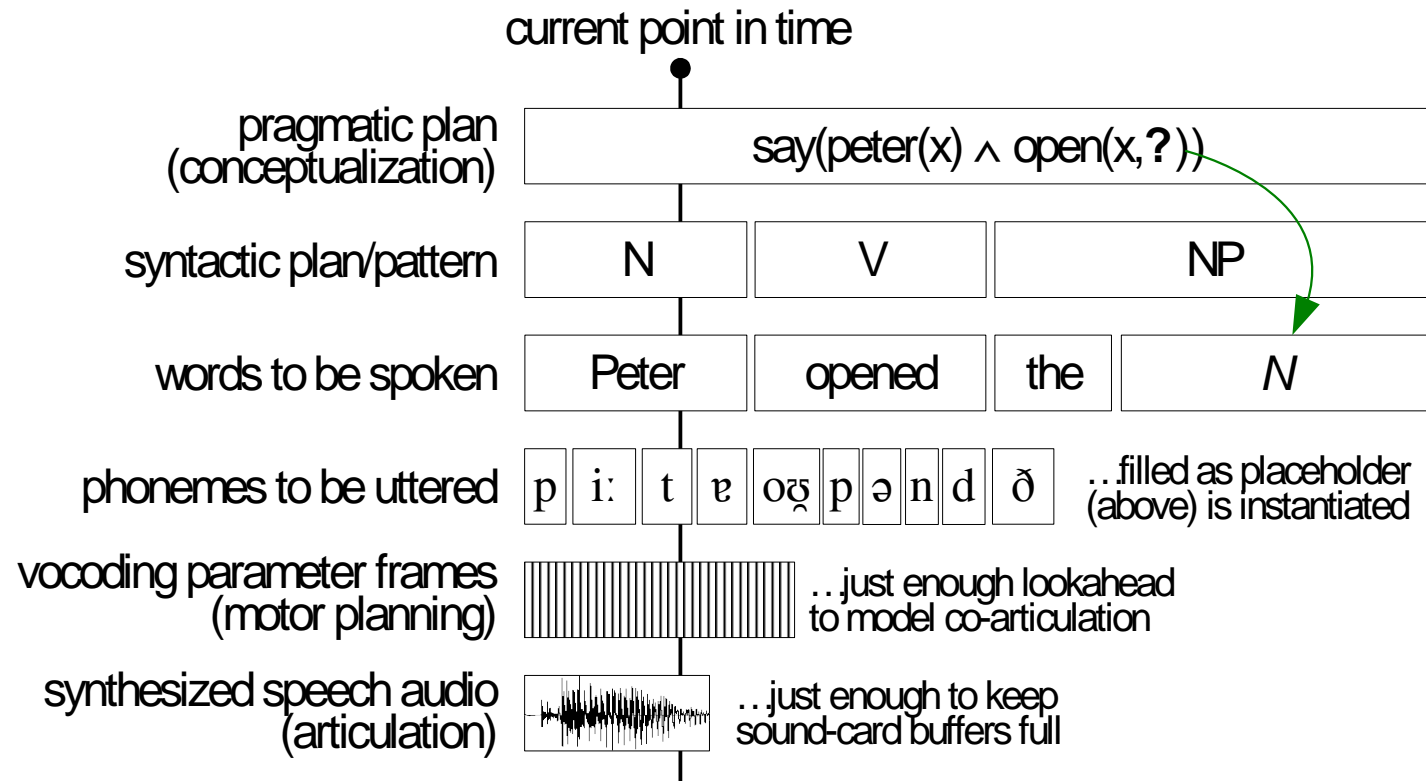
# A *Just-In-Time* Formulation for Incremental Speech Synthesis



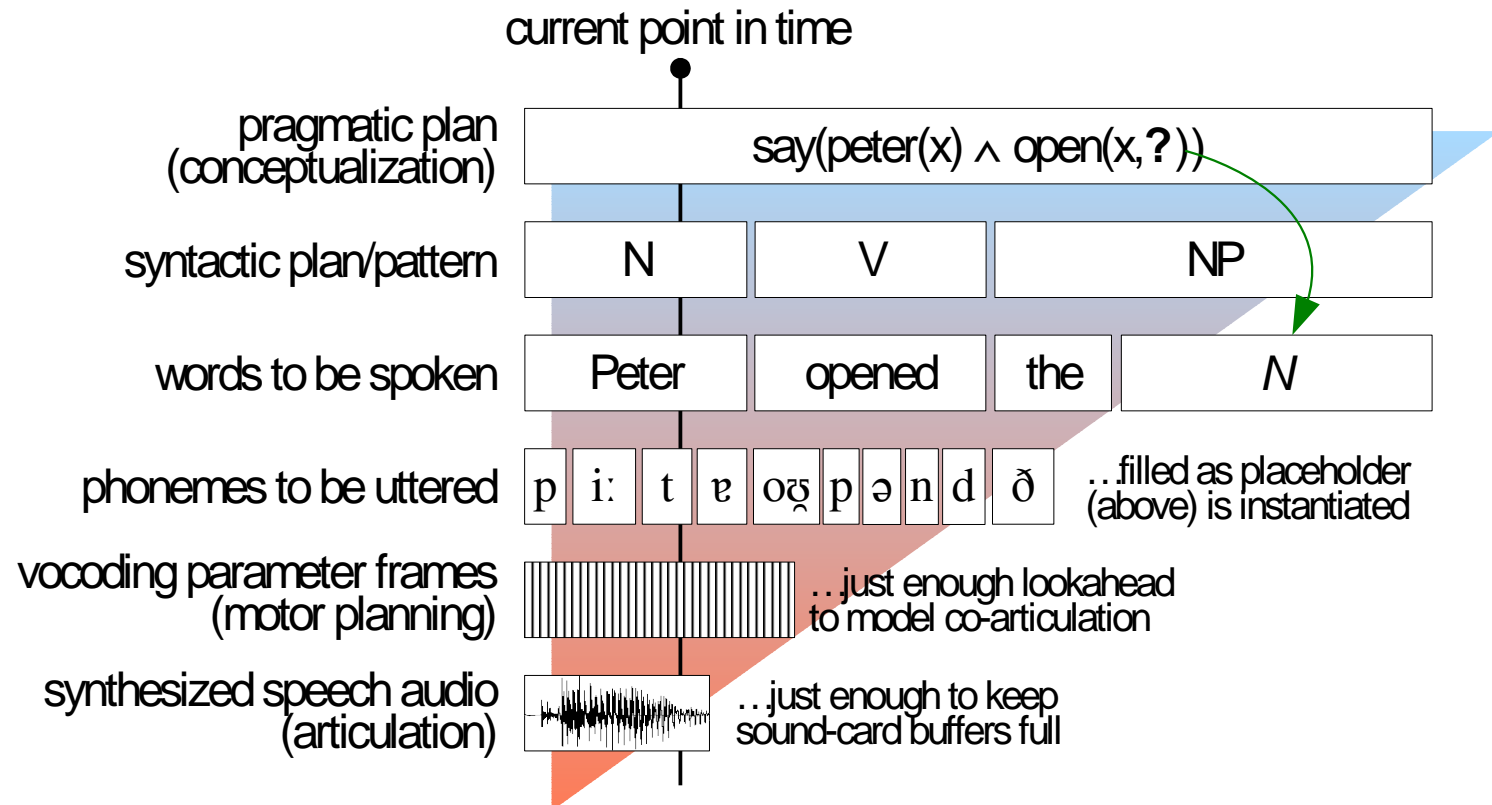
# A *Just-In-Time* Formulation for Incremental Speech Synthesis



# A *Just-In-Time* Formulation for Incremental Speech Synthesis



# 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

# Research question

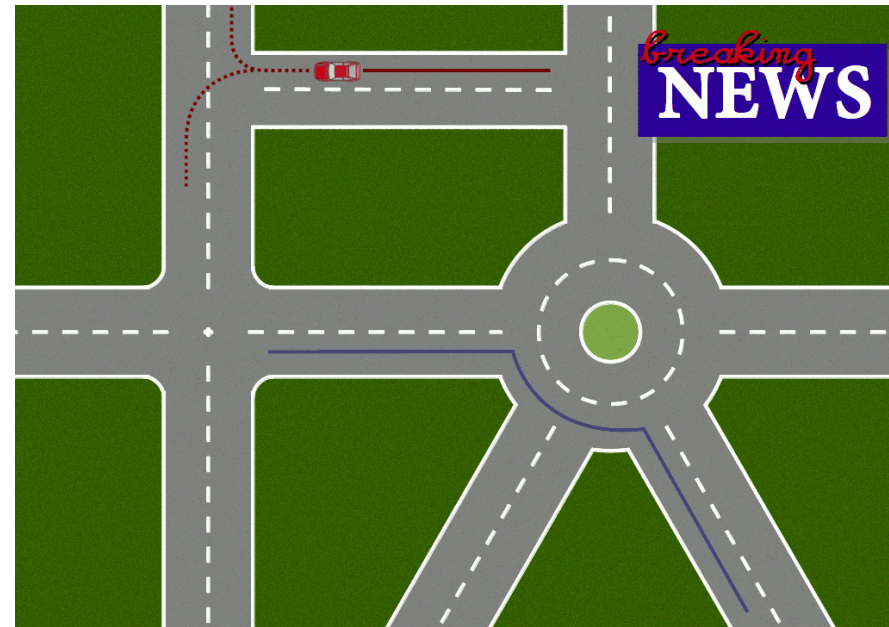
given that incremental speech synthesis  
measurable degrades prosodic parameters –  
→ **does this degradation matter to listeners?**

(based on our Interspeech'12 findings)

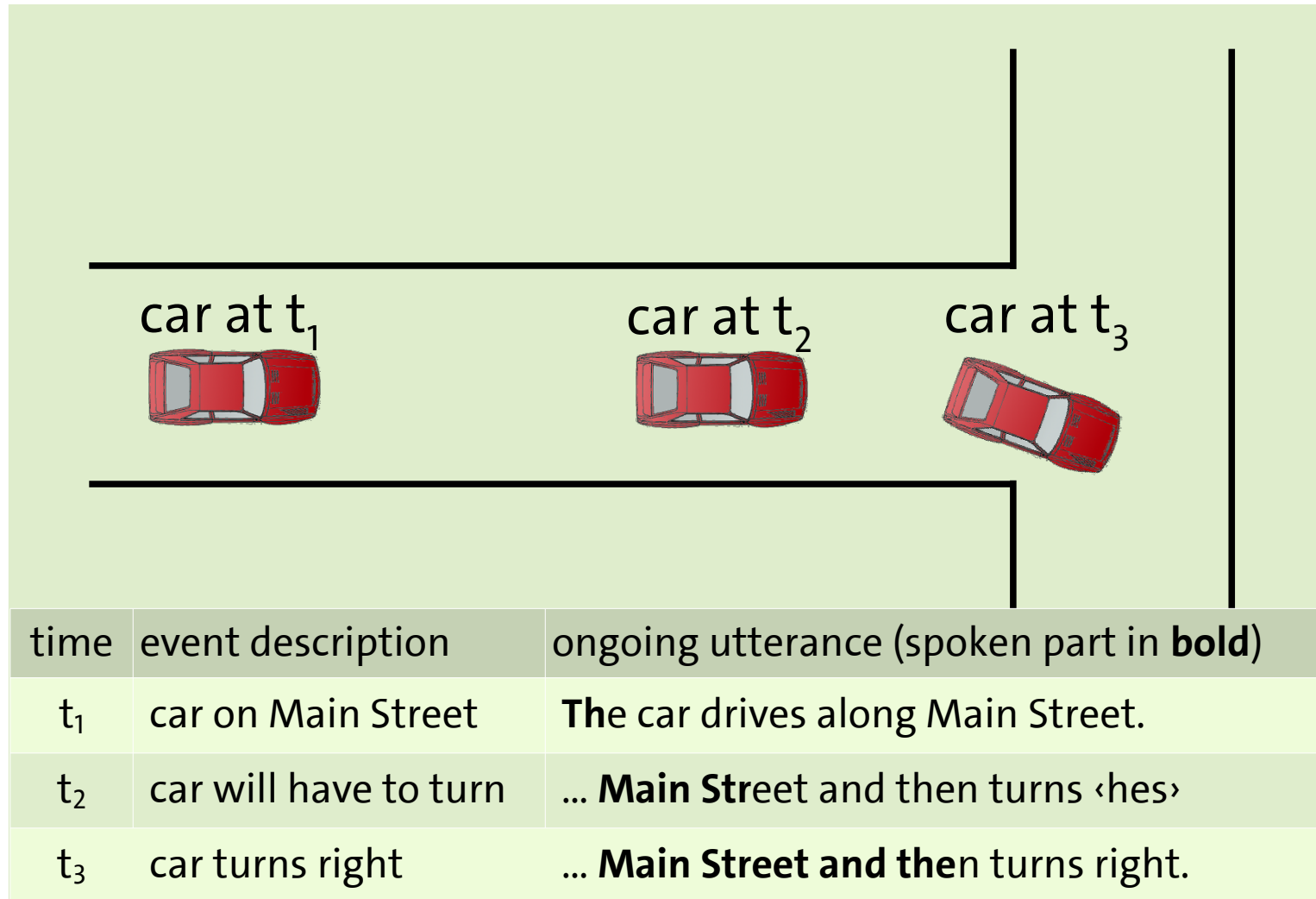


# Example: The CarChase domain

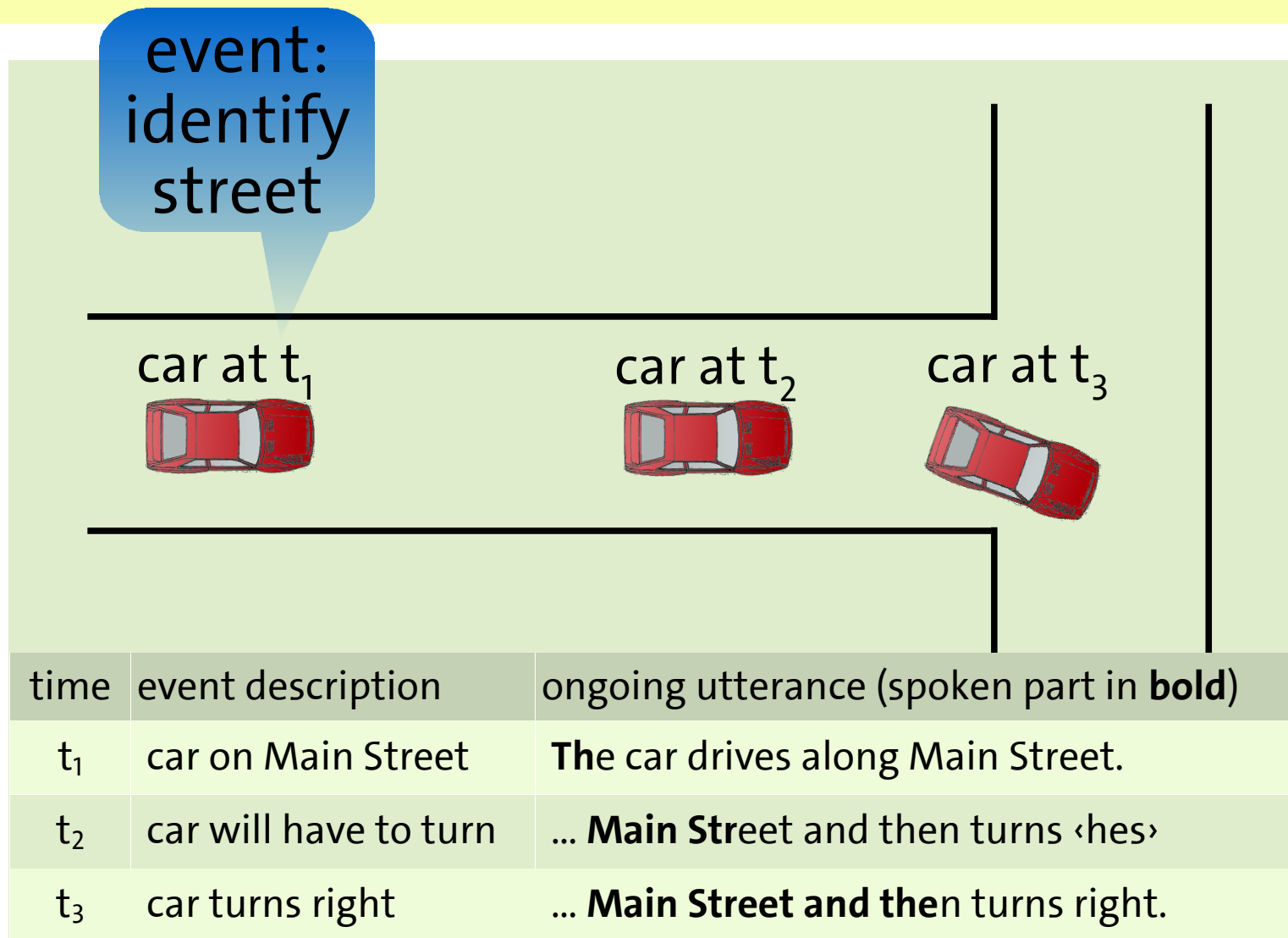
- system comments on events in the scene (car's motion)
- high event rate → impossible to speak isolated utterances
  - combine events into complex utterances (using incremental speech synthesis)
  - skip or abort event notifications in favour of more important information (baseline behaviour)
- simplification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



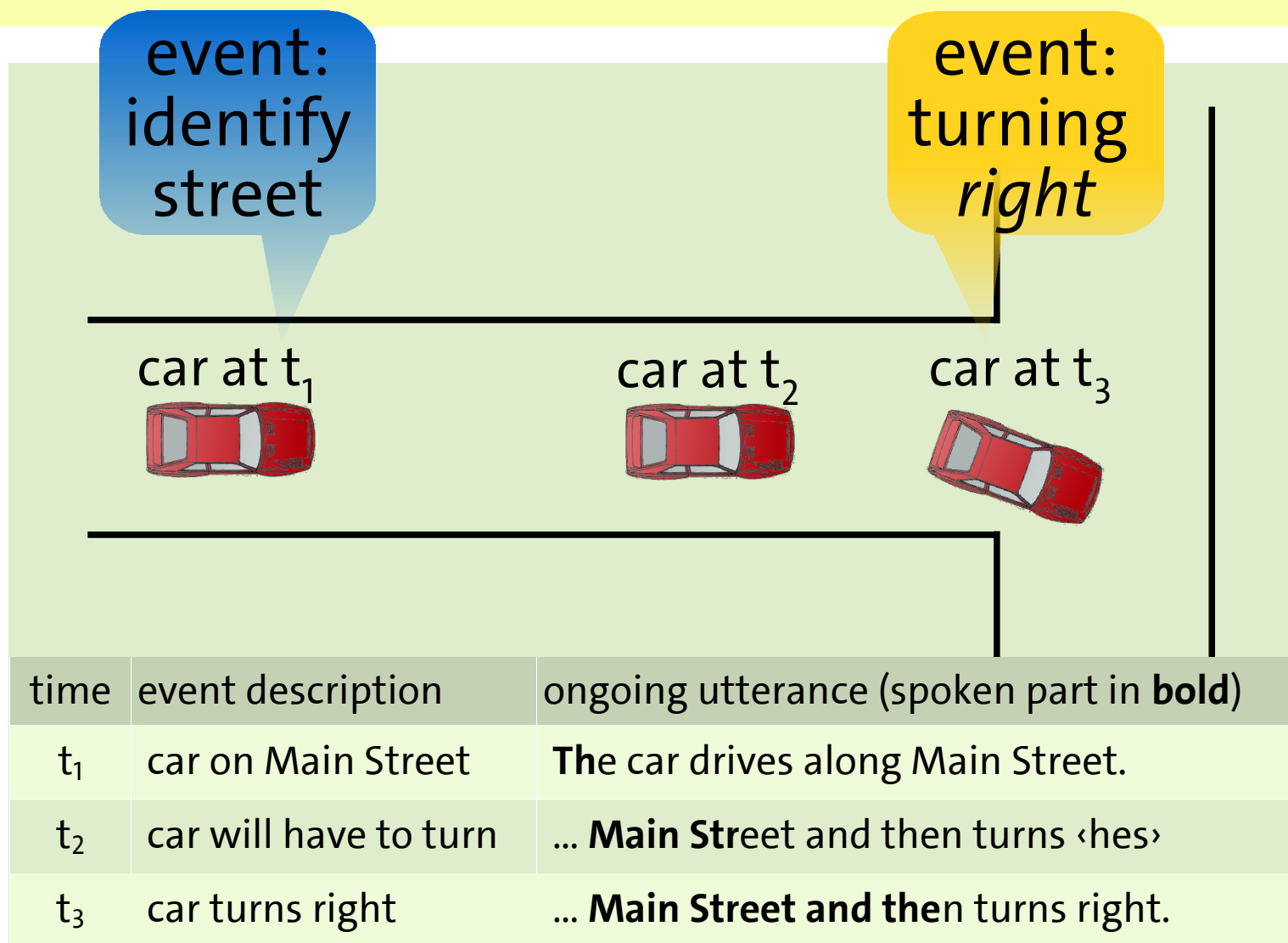
# Taking expectations into account



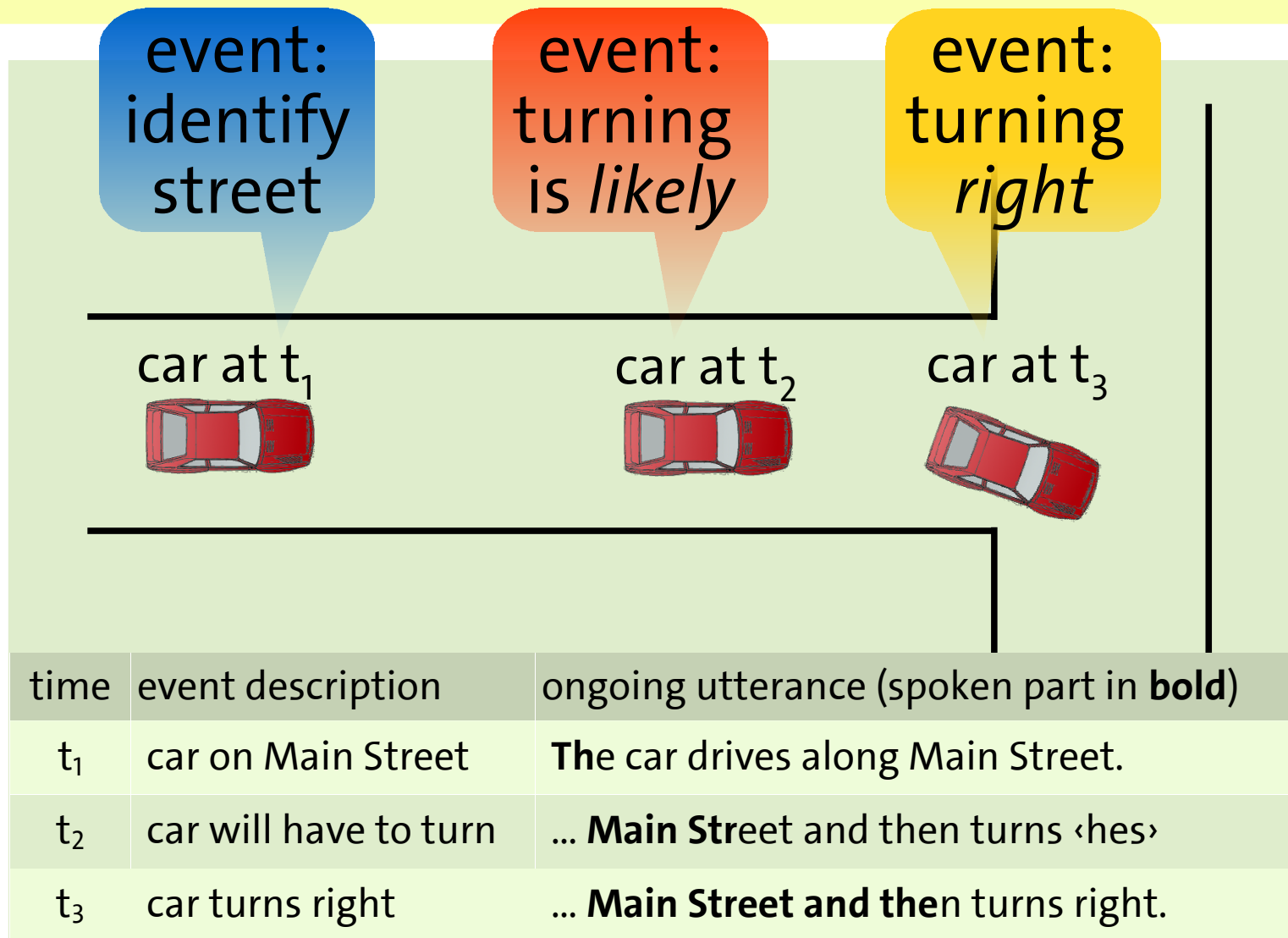
# Taking expectations into account



# Taking expectations into account



# Taking expectations into account



# 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

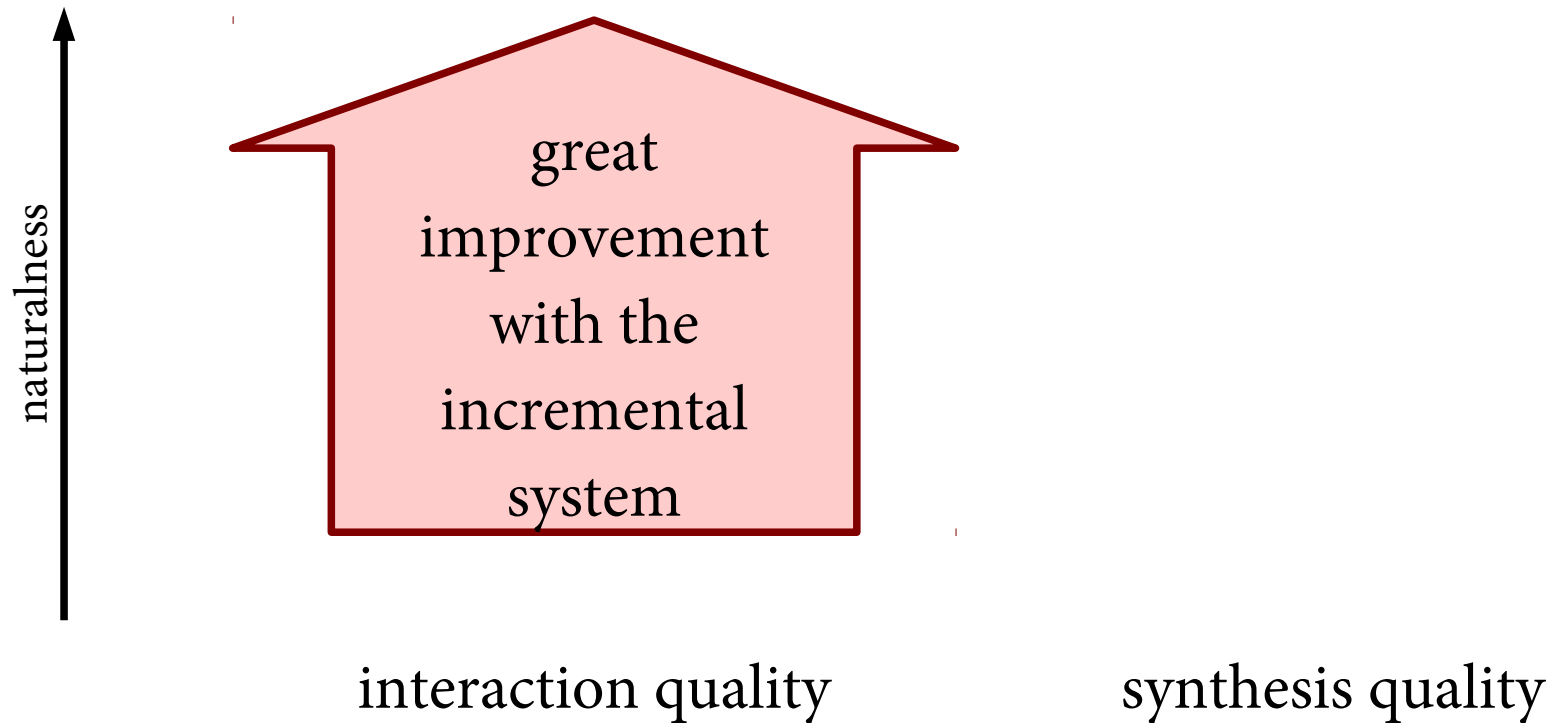
# Expected results

- we were hoping for a good trade-off:



# Expected results

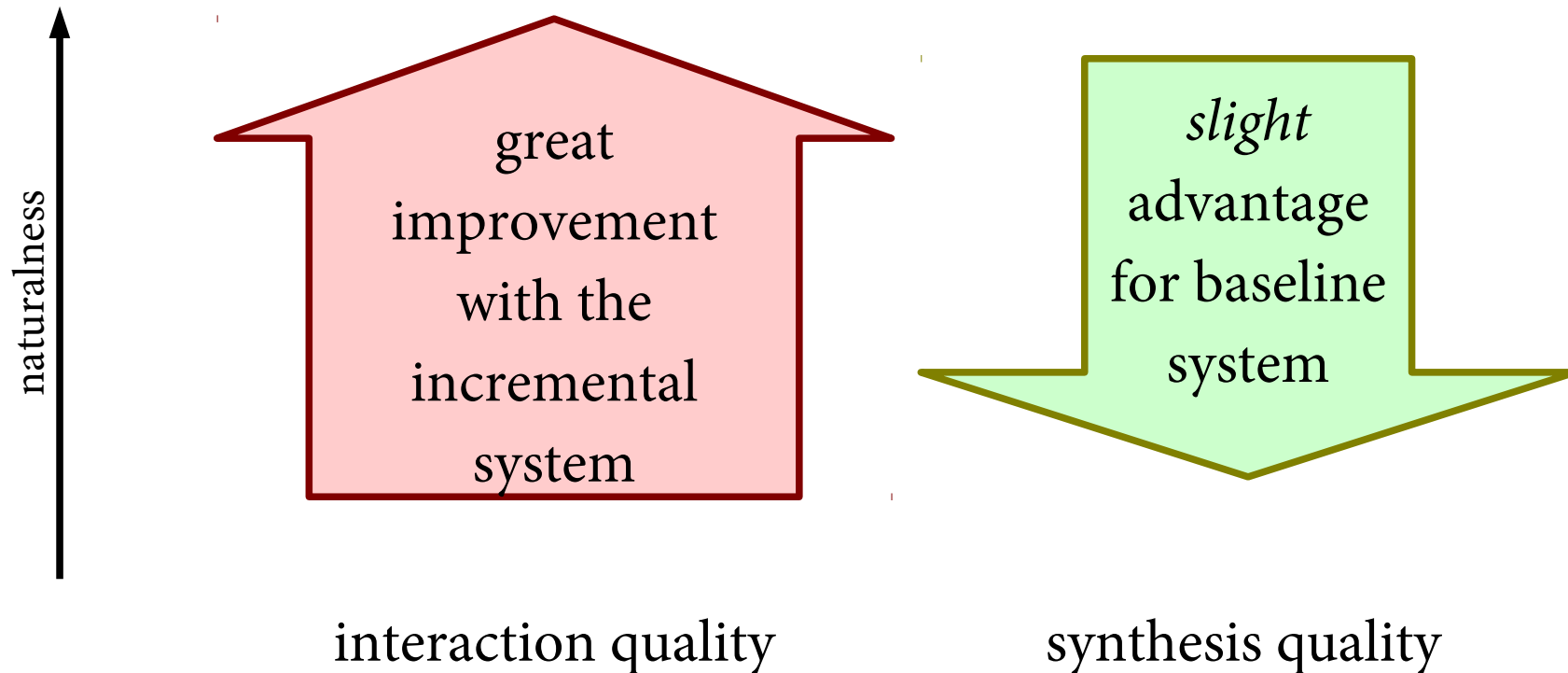
- we were hoping for a good trade-off:





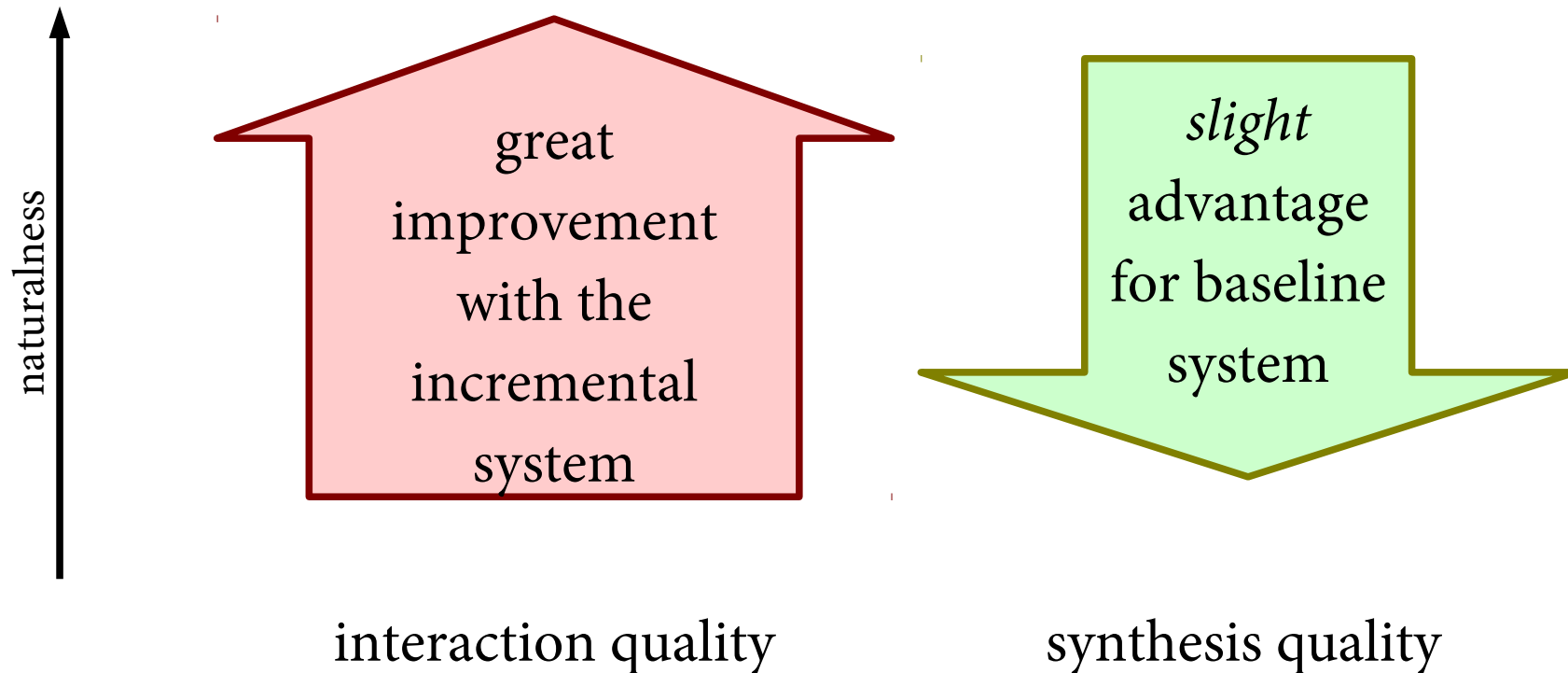
# Expected results

- we were hoping for a good trade-off:



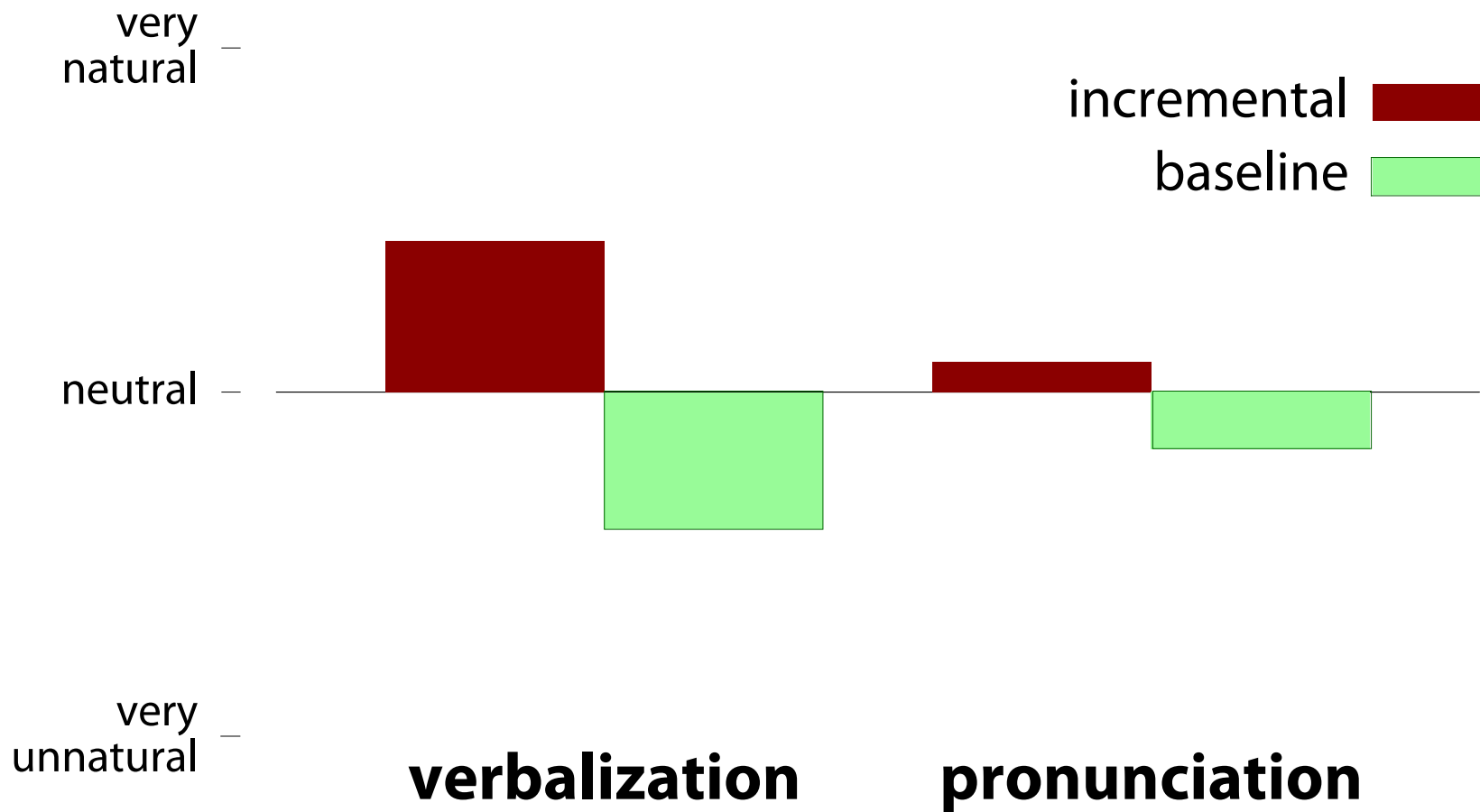
# Expected results

- we were hoping for a good trade-off:

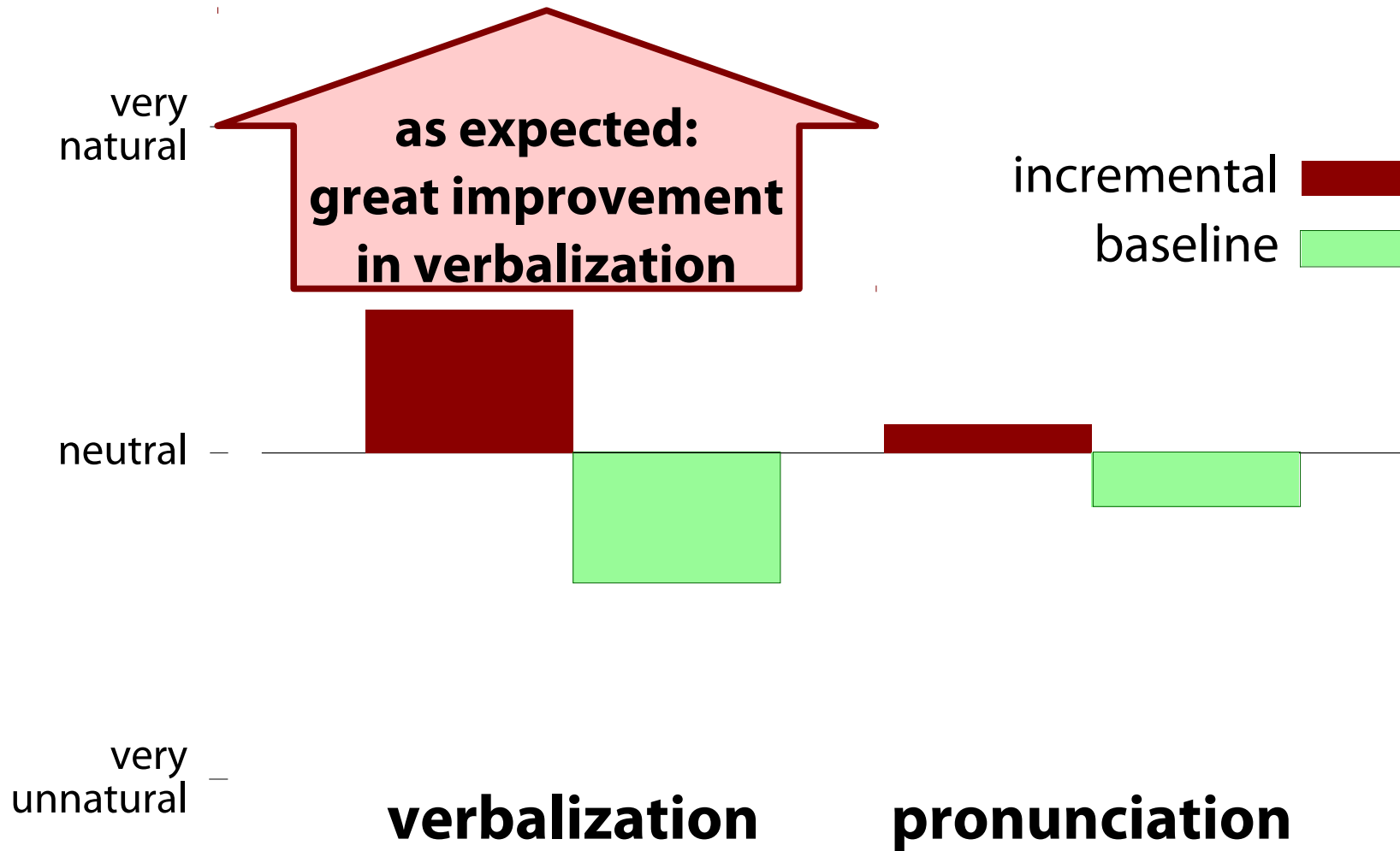


→ write paper: „Trade-off between incrementality of behaviour and speech synthesis quality“

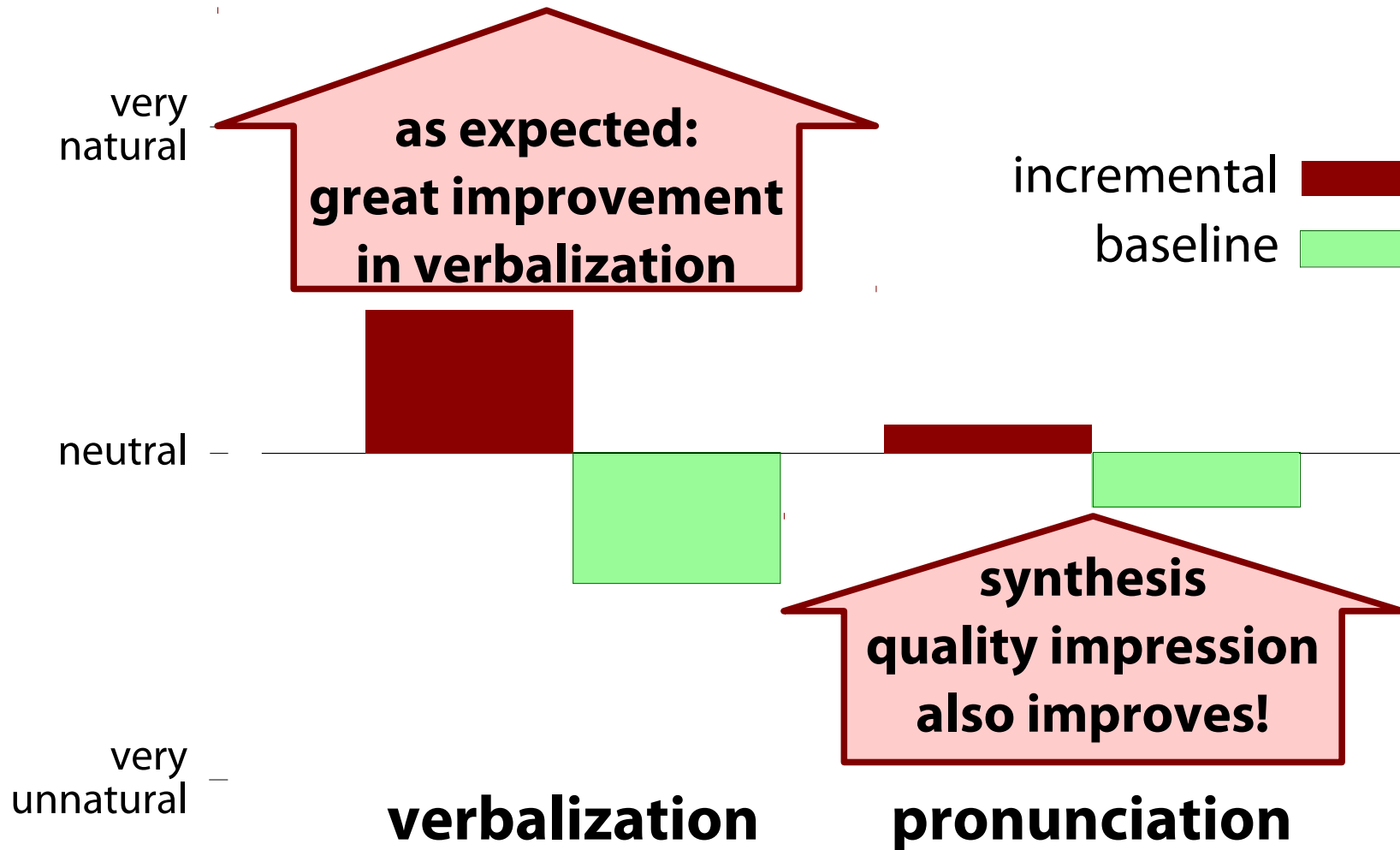
# Actual results



# Actual results



# Actual results



# 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](mailto:baumann@informatik.uni-hamburg.de)  
get the code at [inprotk.sf.net](http://inprotk.sf.net).

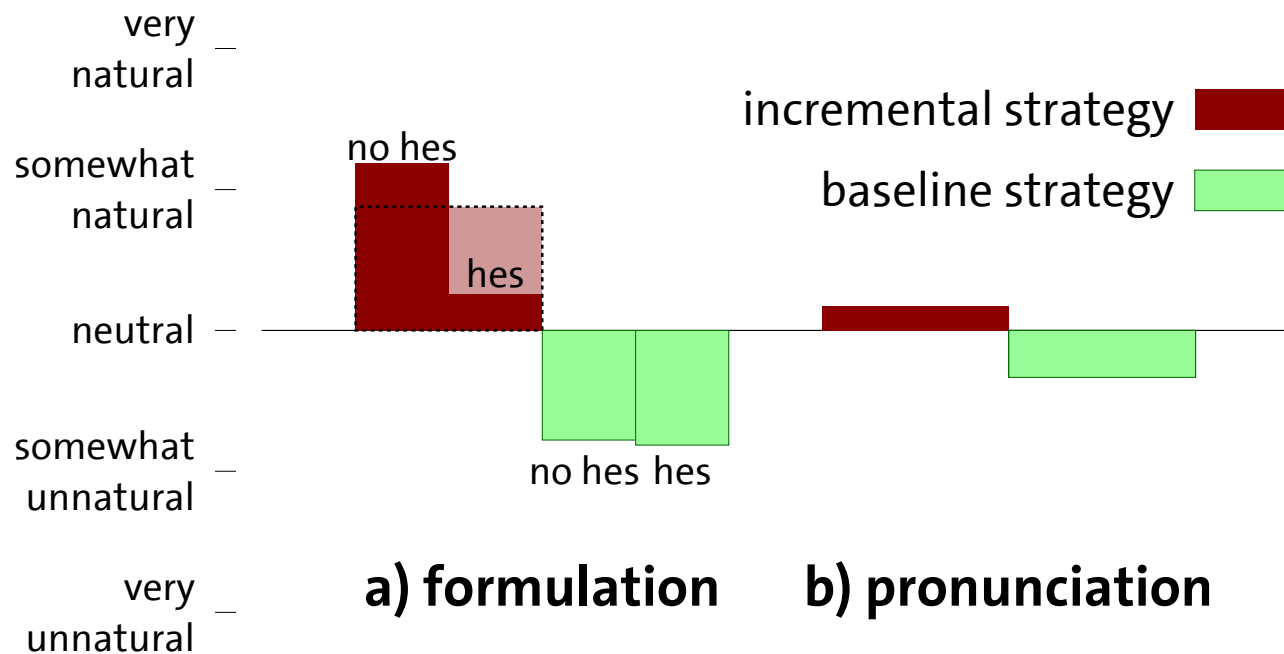
Thanks to Petra Wagner and Wolfgang Menzel.



page intentionally left blank

# „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