

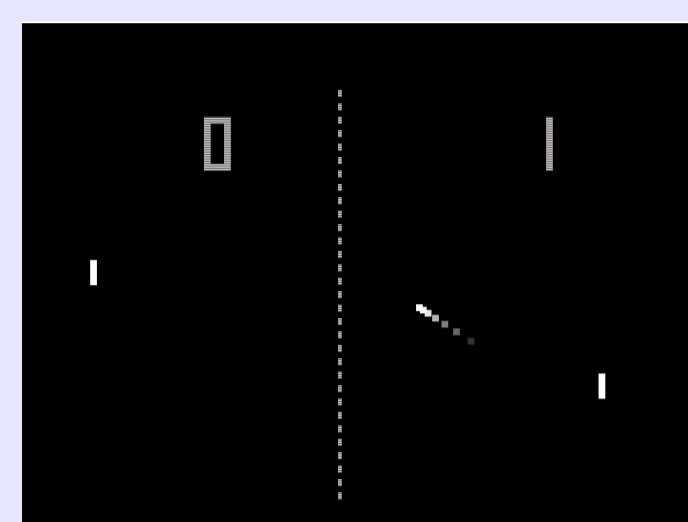
# Within-Turn Processing in Spoken Dialogue Systems

Timo Baumann

Linguistics Department, University of Potsdam, Germany  
mail@timobaumann.de

## 1. Within-Turn Processing vs. Ping-Pong Interaction

- \* Currently, SDSs use a *ping-pong style* of interaction  
→ interruptions/barge-ins are seen as exceptions
- \* In natural conversation, *floor sharing* (or turn-taking) is far less rigid:  
→ we constantly overlap, feedback, co-complete



Achieving this natural behaviour requires incrementality and predictive processing in order to allow for timely reactions.

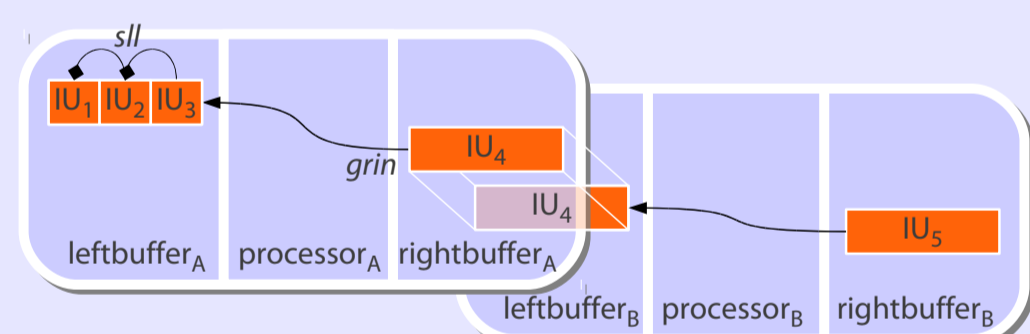
Related phenomena:

- \* back-channelling, short feedback, conversational grunts  
→ shallow processing gives limited results (e.g. how to choose the right back-channel)
- \* turn-taking  
→ shorter time-outs with incremental processing
- \* concurrent (multi-modal) output  
→ may be useful in information access (think: search-as-you-type)
- \* aggressive turn-grabbing  
→ to output urgent information, take-over at hesitations, ...
- \* co-completions  
→ e.g. to signal rapport (see also #3 on the right)

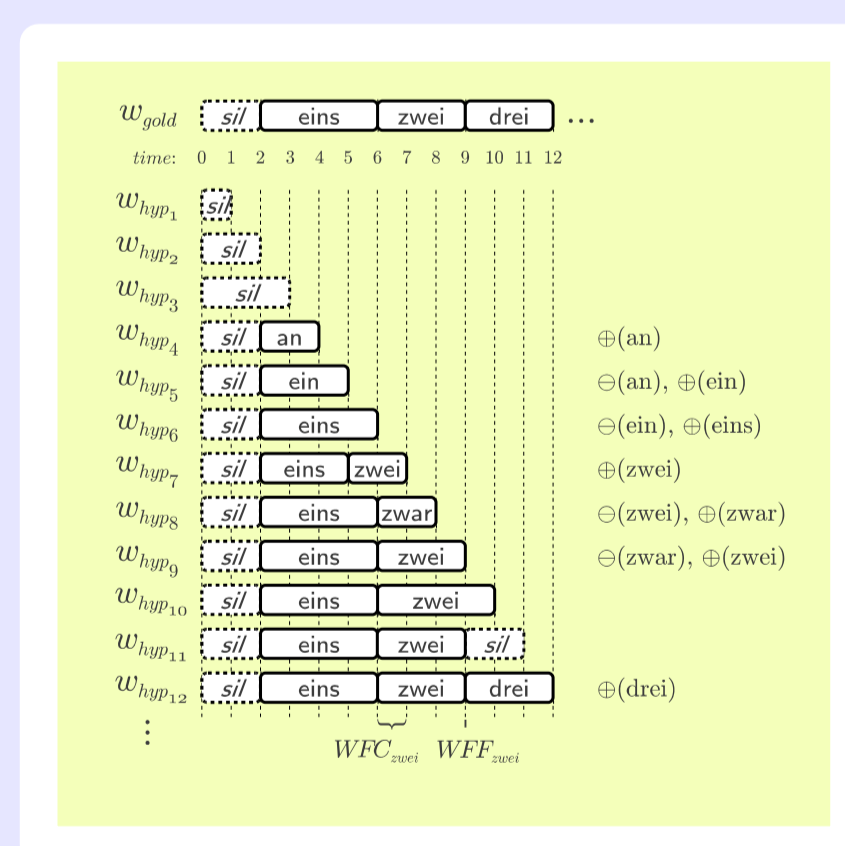
## 2. Incremental ASR and Incremental Evaluation

(Baumann et al., NAACL 2009, Schlangen et al., SIGDial 2010, Baumann et al., D&D 2011)

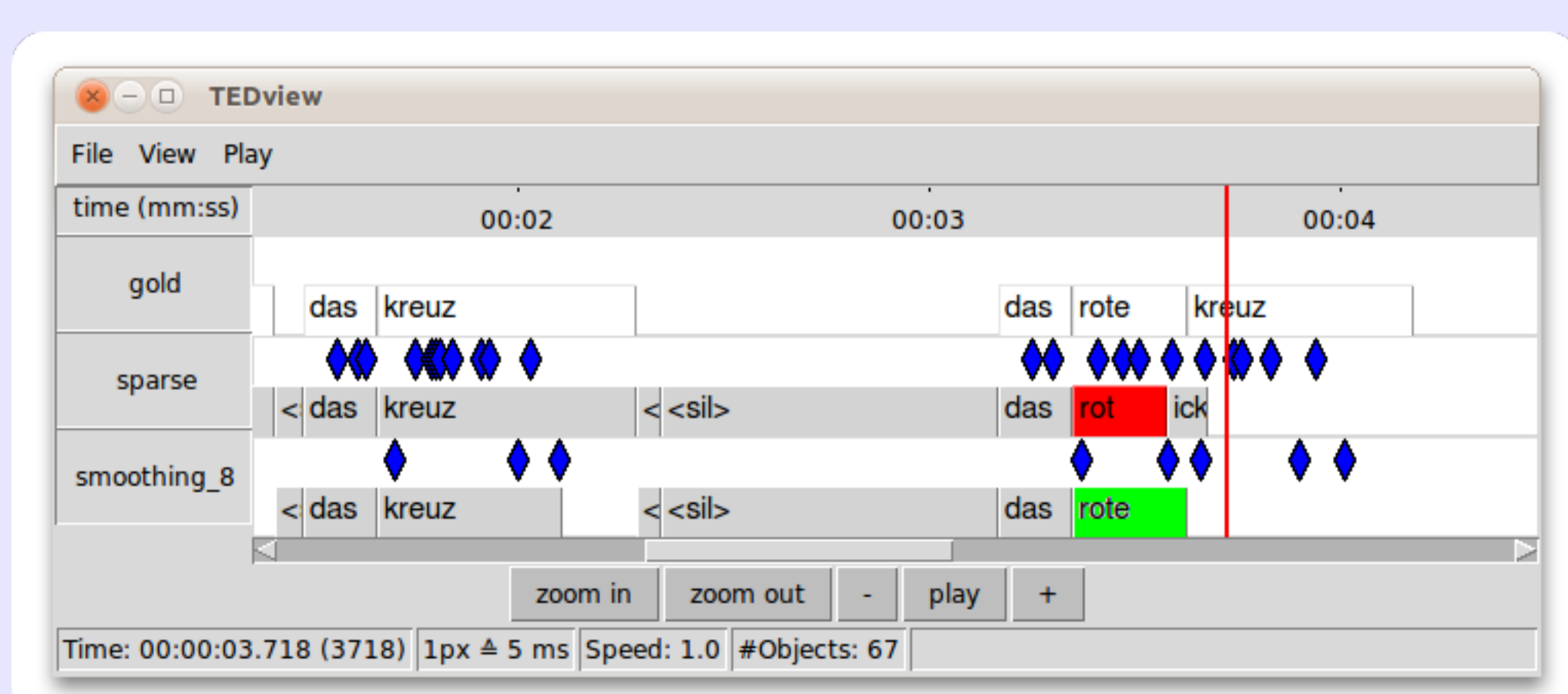
- \* Incremental Processing depends on an architecture which supports revisions (edits) to intermediate hypotheses (at least in a modular system)



- \* Incremental Evaluation – requirements:
  - hypothesize early* (i.e. as soon as a word begins)
  - finalize early* (i.e. as soon as a word is over)
  - edit as rarely as possible* (i.e. just once per word; to avoid overhead)



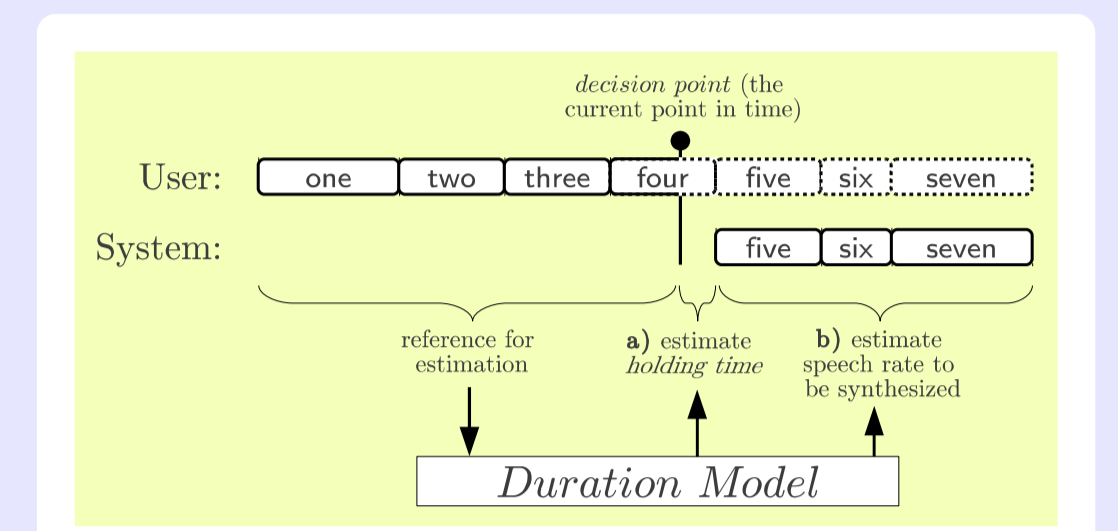
- \* Performance for basic iASR:
  - first hypothesis around 3/4 of the word
  - final recognition around end of the word
  - overhead from spurious edits: 90 % !!
- \* Optimizations to reduce spurious edits (*jitter*)
  - \* leave **right context**: only trust the older parts of recognition  
→ needs relatively high contexts (2-300 ms) and delays
  - \* do **hypothesis smoothing**: only trust changes once they are mature  
→ allows lower delays, because wrong edits usually die off quickly



## 3. Measuring the Micro-Timing of a User's Ongoing Words

(Baumann and Schlangen, SIGDial 2011)

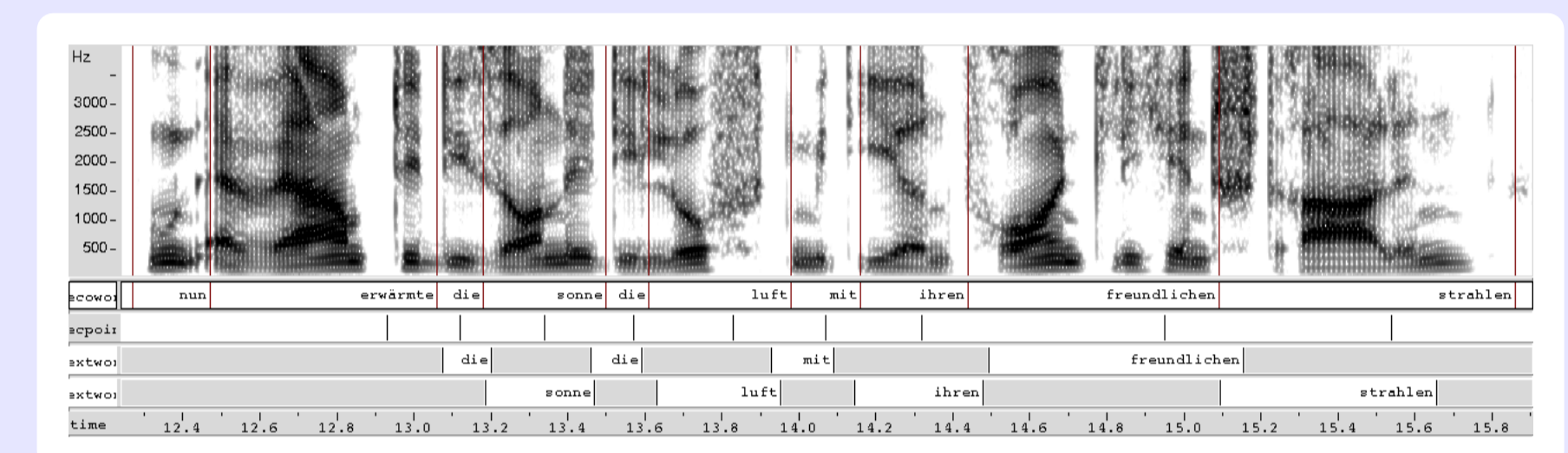
- \* we model the micro-timing of a user's ongoing (and next) words while she speaks
- \* the model could be used to monitor the speaker's fluency, to place back-channels, for turn-taking, or to align turn completions



- \* we demonstrate the micro-timing capabilities by *synchronously completing ongoing turns*  
→ rather a good technology demonstration than a useful application

Results:

- \* an ongoing word's end can be predicted with little error
- \* shadowing the speaker approaches human performance (MAE at 70 ms)



## 4. Next Step: Incremental TTS

- \* so far, we have mostly focussed on input/understanding  
→ solved: the system can be fast enough to co-complete with the user
- \* timing of TTS/output generation is becoming a limiting factor  
→ incremental systems need to constantly revise their output, taking into account the user's actions and re-actions
- \* next, I want to *incrementalize* speech synthesis
  - HMM synthesis algorithm to incrementally optimize state sequence
  - evaluate performance degradation in HMM synthesis quality vs. look-ahead at which unforeseen changes are incorporated
- \* possible features, use-cases, ...
  - on-the-fly adaptation of prosody when the user starts to speak (instead of cutting off at a barge-in)
  - more fine-grained copying/mimicking of the user during collaborative completions (and the like)
- \* is anyone interested in helping out or in using such an incremental TTS?

## Further Information

Please contact me at [mail@timobaumann.de](mailto:mail@timobaumann.de)

More information on this and related research is available on my website at <http://www.ling.uni-potsdam.de/~timo/> where you can find a PDF version of this poster in the publications section.



## Acknowledgements

This work was funded by a DFG grant in the Emmy Noether programme. The author would like to thank David Schlangen and Okko Buß for their cooperation in the research and encouragement with ideas presented here.

