Integrating prosodic modelling with incremental speech recognition

Timo Baumann

Department for Linguistics, University of Potsdam, Germany

Introduction

Rationale: Incremental spoken dialogue systems process while a user is still speaking.

Incremental ASR (Baumann et al., 2009) and prosody analysis (Edlund and Heldner, 2006) **modules** already **exist separately.**

We **integrate** both for **mutual benefits.**

This is **work in progress,** no final results yet.

Related Work

Some SDSs that use prosody in a similar way:

Soeda and Ward (2001) show a system for a very similar setting, featuring "sub-second responsiveness" using **prosodic analysis only.**

Skantze and Schlangen (2009) integrate ASR and prosodic analysis but **don't use** a prosody model motivated by **phonologic theory.**

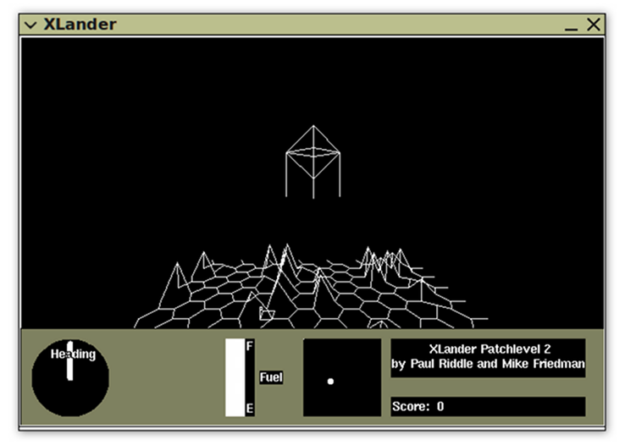

Figure 1: Domain of (Soeda and Ward, 2001)

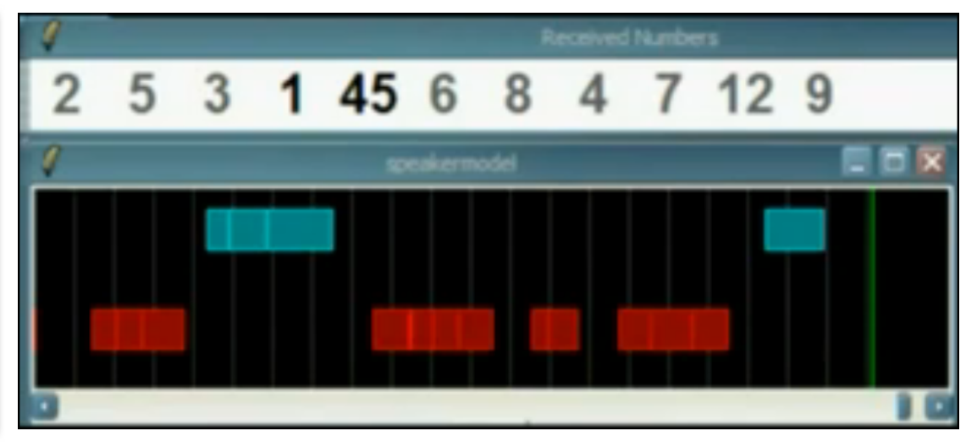

Figure 2: Screenshot of the System from (Skantze and Schlangen, 2009)

Prosody Model

Prosody is determined by:

- pitch and loudness contours and
- duration proportions over time

Prosody can be modelled as (Pierrehumbert, 1980):

- **accentuation** tones on syllables
- **juncture** of adjacent words

Acoustic prosodic features per frame:

- fundamental frequency
- frame-energy
- we look into FFV (Laskowski et al., 2008)
- advanced loudness metering (ITU-R, 2006)
- possibly spectral tilt

→ features must be normalized

(please read on at the top of the center column)

Integration with incremental ASR

- ASR supplies **partial hypotheses** about words & phones
  - hypothesis-filtering as described in (Baumann et al., 2009))
- **syllabification** via dictionary or on the fly
  → **duration proportions** of syllable and nucleus, speech rate
- **incremental** pitch tracking (right-reduced dynamic programming)
  - other features can be calculated independently for each frame
- **curve-fitting**, similar to PaIntE, (Möhler, 1998), or TILT (Taylor, 1998)
  → descriptive contour parameters
- use **regression** or **classification** for syllables and words
  → phonologically sound **accentuation** and **juncture measures**

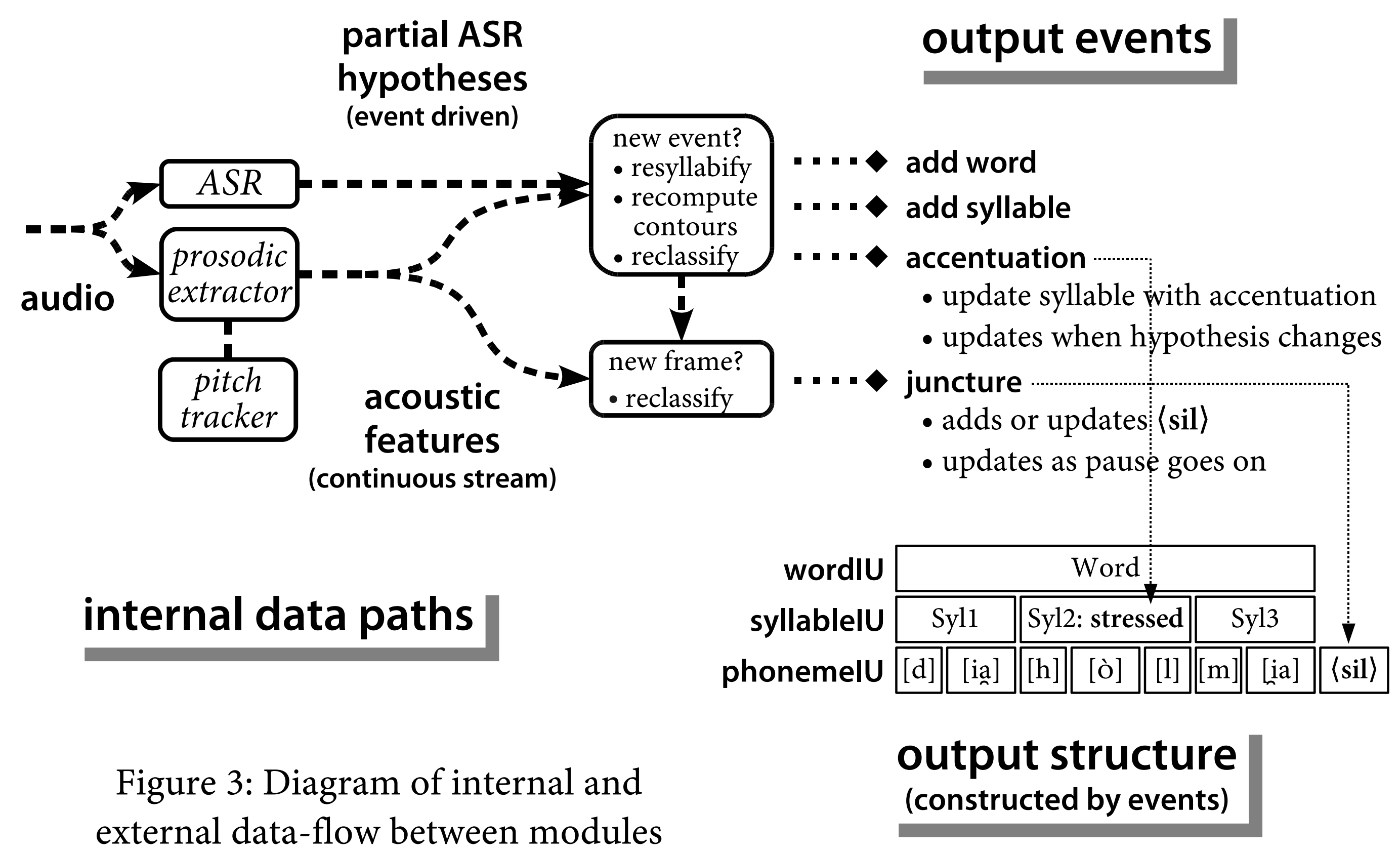

Figure 3: Diagram of internal and external data-flow between modules

Our Prototype

We design a **micro-domain** (Edlund et al., 2008) to fit our research agenda:

- **elicit** interesting expressive **prosody**
- require **quick reaction** (to show-off incrementality)
- **prosody** should be **helpful but not necessary** to understanding
- **restricted** domain to make things (dialogue management, …) easier

Interactive control of a robot arm (see Figure 3):

- **1-dimensional motion** control
- **final drop** signal
- **actions** (moving, stopping, dropping) require **different levels of certainty** (as dropping cannot be undone)

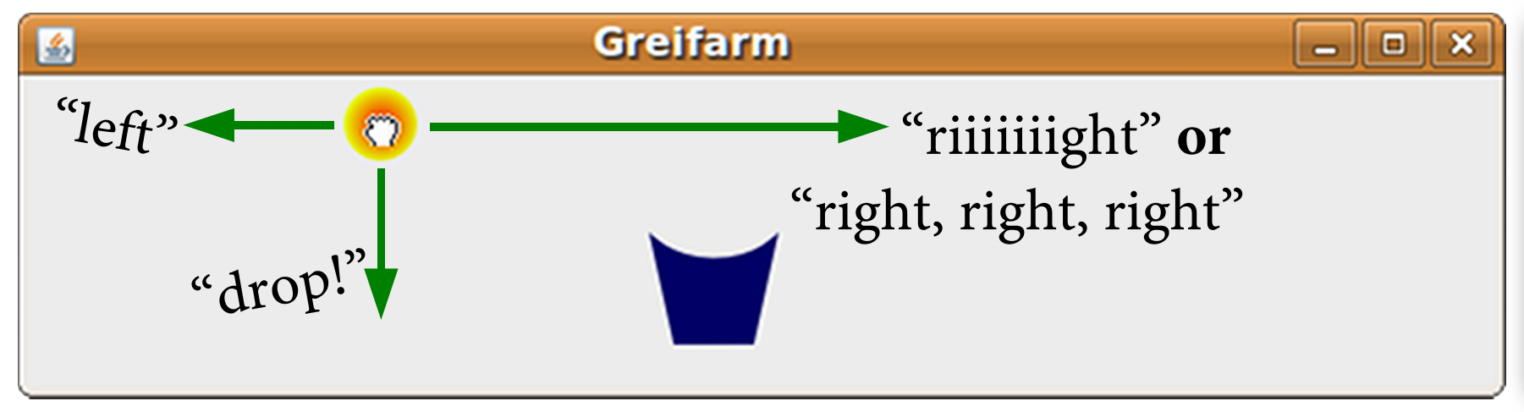

Figure 5: User-Interface of the prototype; some possible actions are indicated by arrows.

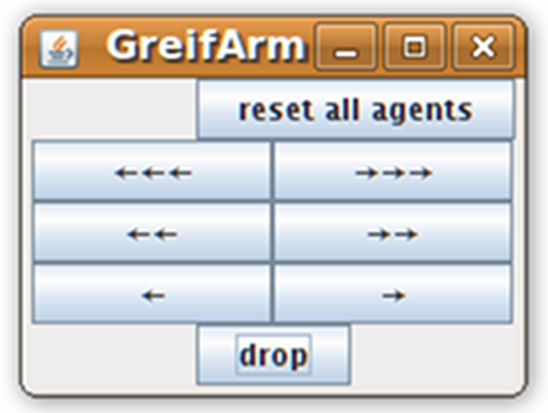

Figure 6: WoZ-Interface

Advantages of the Integration

ASR supplies phonemes and word boundaries:

- **no need for external** (p)syllabification, silence detection
- can be used in loudness and pitch **normalization**

Prosodic **information can be fed back** to the ASR:

- allow lengthening of syllables when noticing emphasis (leeeft)
- prosodically detect and handle within-word self-interruptions

Coordinated output of word- and prosody-information

- **no later input fusion** for consumers which could cause problems

Extensible to n-best or lattice recognition (easily?)

- each recognition trellis has its matching prosodic analysis

As flexible as non-integrated approach:

- integrate "non-linguistic" feature abstractions, like **linear regressions**
- integrate **classifiers for specific** complex **decisions:**
  - end-of-turn/hesitation, barge-in/back-channel, …

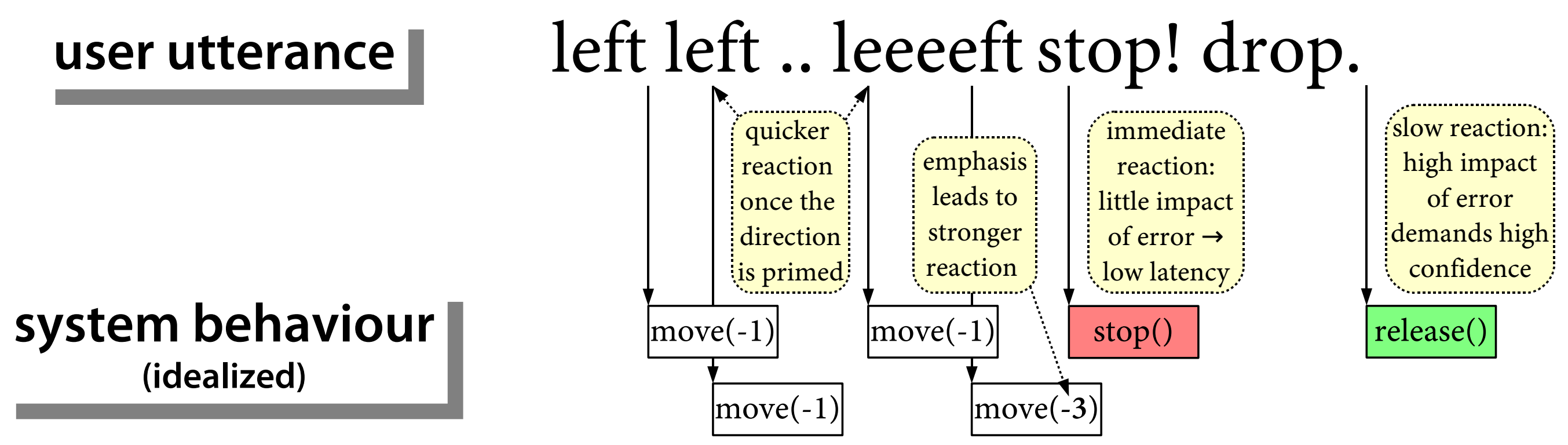
user utterance
left left .. leeeeft stop! drop.
system behaviour
(idealized)
quicker reaction once the direction is primed
emphasis leads to stronger reaction
immediate reaction: little impact of error → low latency
slow reaction: high impact of error demands high confidence
move(-1)
move(-1)
move(-1)
move(-3)
stop()
release()

Figure 4: Idealized system behaviour (mock-up)

WoZ Corpus

We use a Wizard-of-Oz setup to **analyse users' system interaction:**

- corpus contains 12 subjects, 40' audio, 1500 words
- only 1 wizard for higher system consistency

Wizard controls **three degrees of** directional **motion** & **drop** action.

- exact distance is a normal distribution
  → according to users, the **motion seems very natural**
- we forgot the "stop" action :-(

Data shows the **expected behaviour:**

- **repetition, lengthening** or **waiting** to express distance
- **marking** of corrections **through prosody**
- very **quick commitment** (for drop-action) by the wizard

Further Steps

Our **model** implements **more** than strictly **necessary** for the task.

There are, however, **more use-cases for** incrementally available **prosody information:**

- use juncture in **language modelling**
- use prosodic patterns in **ASR rescoring**
- juncture and accentuation in **parsing** and
- **semantic** and **pragmatic interpretation**
- **extend to** more **complex domains**

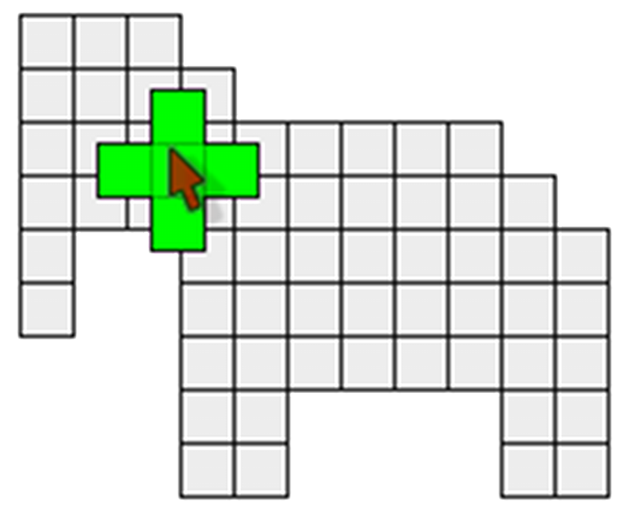

Figure 7: Fine-positioning
"left .. a little fur-, that's good."

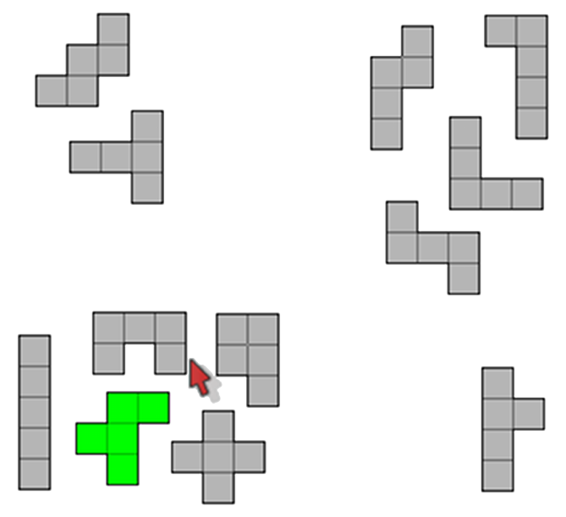

Figure 8: Interactive selection
"in the bottom left .. yes, the center one"

Bibliography

Timo Baumann, Michaela Atterer, and David Schlangen. 2009. Assessing and Improving the Performance of Speech Recognition for Incremental Systems. In *Proceedings of NAACL-HLT 2009*, Boulder, USA.

Timo Baumann. 2008. Simulating Spoken Dialogue With a Focus on Realistic Turn-Taking. In *Proceedings of the 13th ESSLLI Student Session*, Hamburg, Germany.

Jens Edlund and Mattias Heldner. 2006. /nailon/ - Software for Online Analysis of Prosody. In *Proceedings of ICSLP 2006*. Pittsburgh, USA.

Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication*, **50**:630-645.

ITU-R. 2006. *ITU-R BS. 1770-1. Algorithm to measure audio programme loudness and true-peak audio level.* international Telecommunication Union.

Kornel Laskowski, Mattias Heldner, and Jens Edlund. 2008. The fundamental frequency variation spectrum. In *Proceedings of FONETIK 2008*, Gothenburg, Sweden.

Gregor Möhler. 1998. *Theoriebasierte Modellierung der deutschen Intonation für die Sprachsynthese*. Ph.D. thesis, Universität Stuttgart, Germany.

Janet B. Pierrehumbert. 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. thesis, MIT, Cambridge, USA.

Shunsuke Soeda and Nigel Ward. 2001. Design for a System able to use Time-Critical Spoken Advice. In *Proceedings of the 15th Annual Conference of JSAI*, Matsue, Japan.

Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proceedings of EACL 2009*, Athens, Greece.

Paul Taylor. 1998. The TILT Intonation Model. In *Proceedings of ICSLP 1998*, Sydney, Australia.

Further Infos

Please contact **timo@ling.uni-potsdam.de**

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

This work is funded by a DFG grant in the Emmy Noether programme, and a short-term grant by the German Academic Exchange Service (DAAD).

The author would like to thank David Schlangen and Jens Edlund for valuable input on the ideas and plans presented.