Incremental Spoken Dialogue Processing: Architecture and Lower-level Components

Vortrag zur Disputation am 16. Mai 2013

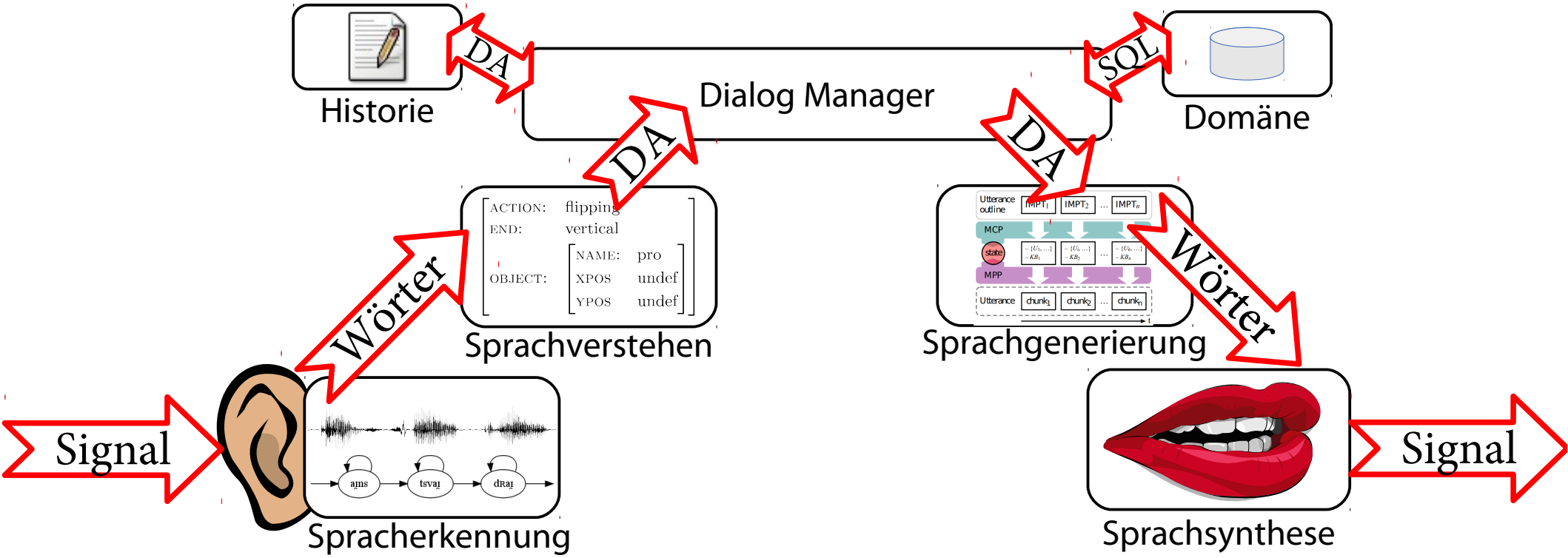
Timo Baumann

baumann@informatik.uni-hamburg.de
www.timobaumann.de/work

Incremental Spoken Dialogue Processing: Architecture and Lower-level Components

Inkrementelle Sprachdialogverarbeitung: Architektur und signalnahe Komponenten

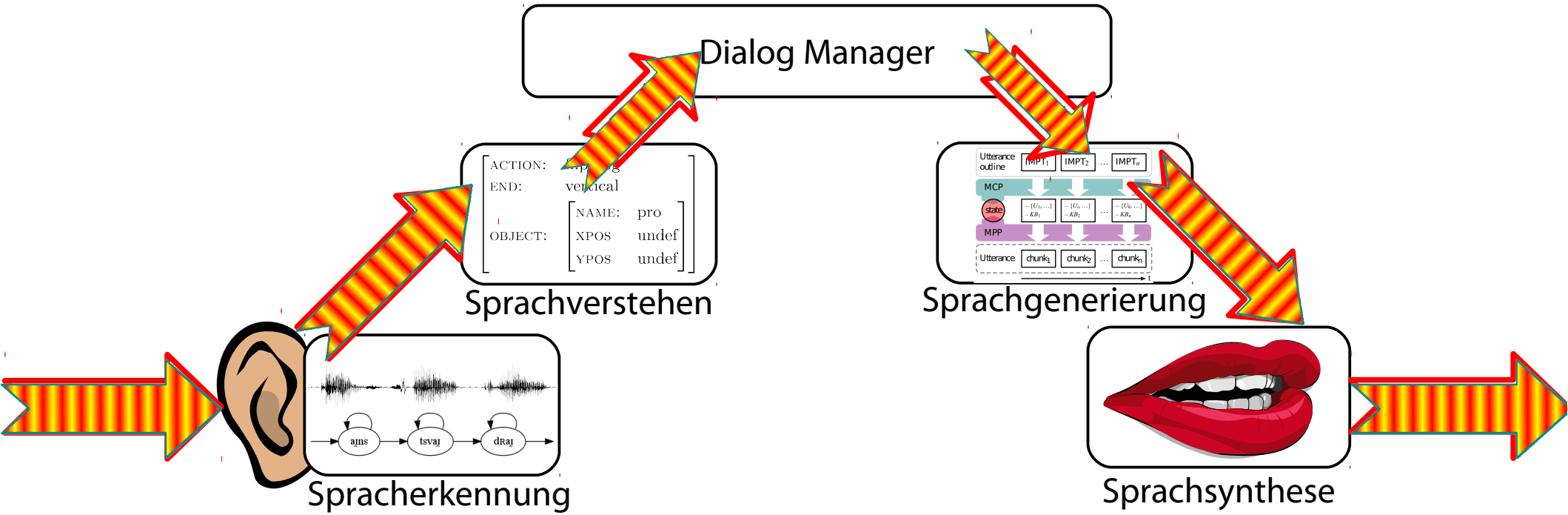
Verarbeitung in einem herkömmlichen Sprachdialogsystem



- Verarbeitungsschritte laufen nacheinander ab

Verarbeitung in einem **inkrementellen** Sprachdialogsystem

Verarbeitungsschritte synchron:



- dadurch wird kontinuierliches Verhalten möglich

Gegenüberstellung

herkömmliches System

- Verarbeitung *zwischen* den Redebeiträgen (turns)
- Verzögerungen durch Verarbeitungsablauf
- rein reaktive Ping-Pong-Interaktion

inkrementelles System

- Verarbeitung *just-in-time* *während* der Redebeiträge
- Verarbeitungszeit während der Sprechzeit
- weniger starre Muster: Feedback, Kollaboration

Gründe für inkrementelle Verarbeitung

- Menschen …
  - produzieren Sprache inkrementell (Levelt 1989)
  - nehmen Sprache inkrementell wahr (Tanenhaus et al. 1995)
  - kollaborieren inkrementell im Dialog (Clark 1996)
- bisherige Dialogsysteme „funktionieren“ auch ohne
- inkrementelle Verarbeitung …
  - erleichtert Mensch-Maschine-Interaktion (Aist et al. 2007)
  - Interaktivität in SDS ist ein *Usabilityproblem* (Ward et al. 2005)

Ward et al. (2005): „Root causes of lost time and user stress“

- 7 Ursachen für Ineffizienz und Unbehagen in Mensch-Computer-Dialogen
- 3 davon lassen sich direkt auf *Ping-Pong-Interaktion* und mangelnde Interaktivität zurückführen:
  - Time-outs, Responsivität, Feedback
  - diese machen Mensch-Computer-Dialoge unnatürlich
- **Inkrementalität in SDS könnte also helfen, ein tatsächlich drängendes Problem zu lösen**

Leitfrage der Arbeit

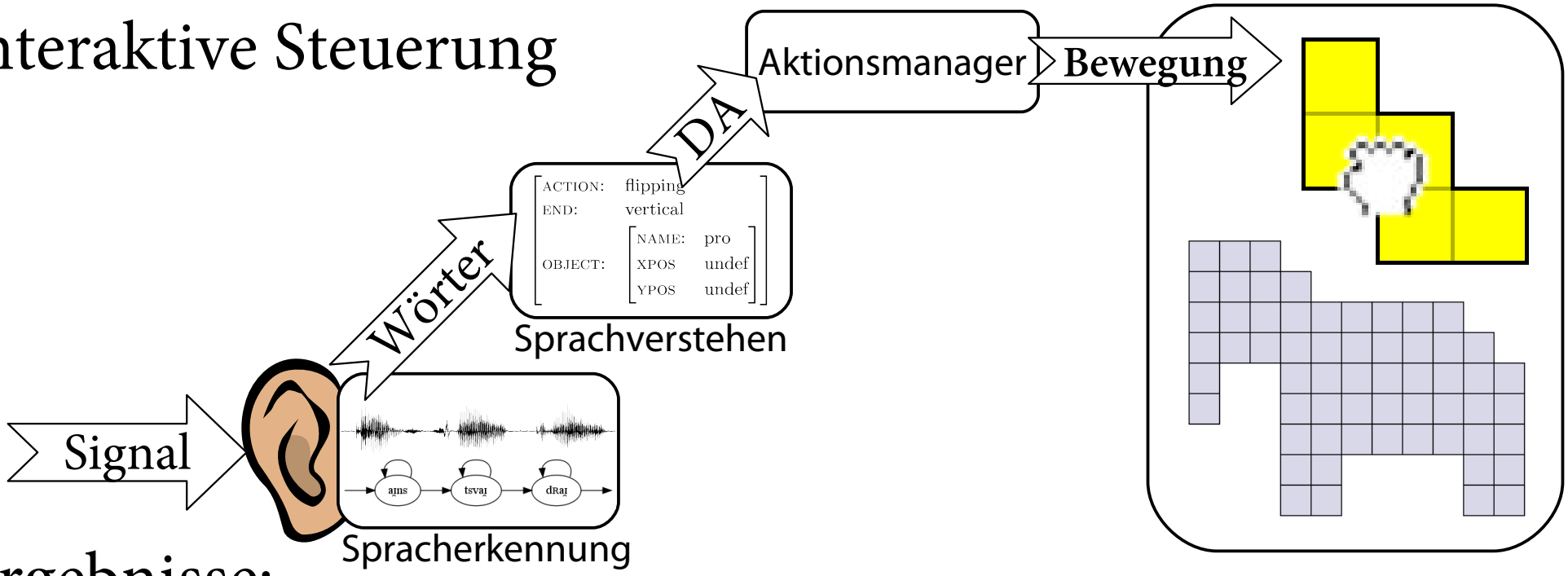
Ist feingranulare, inkrementelle Verarbeitung für natürlichere Dialogsysteme praktikabel *und hilfreich*?

Methode

- Qualitätsmaße für inkrementelle Verarbeitung
  - Evaluation von einzelnen Modulen
- reduzierte Prototyp-Systeme
  - Teilsysteme für ausgewählte Teilprobleme
- Fokussierung auf signalnahe Komponenten
  - Spracherkennung und Sprachsynthese
  - übergeordnete Komponenten nur simuliert/vereinfacht

Behandelte Teilprobleme (1)

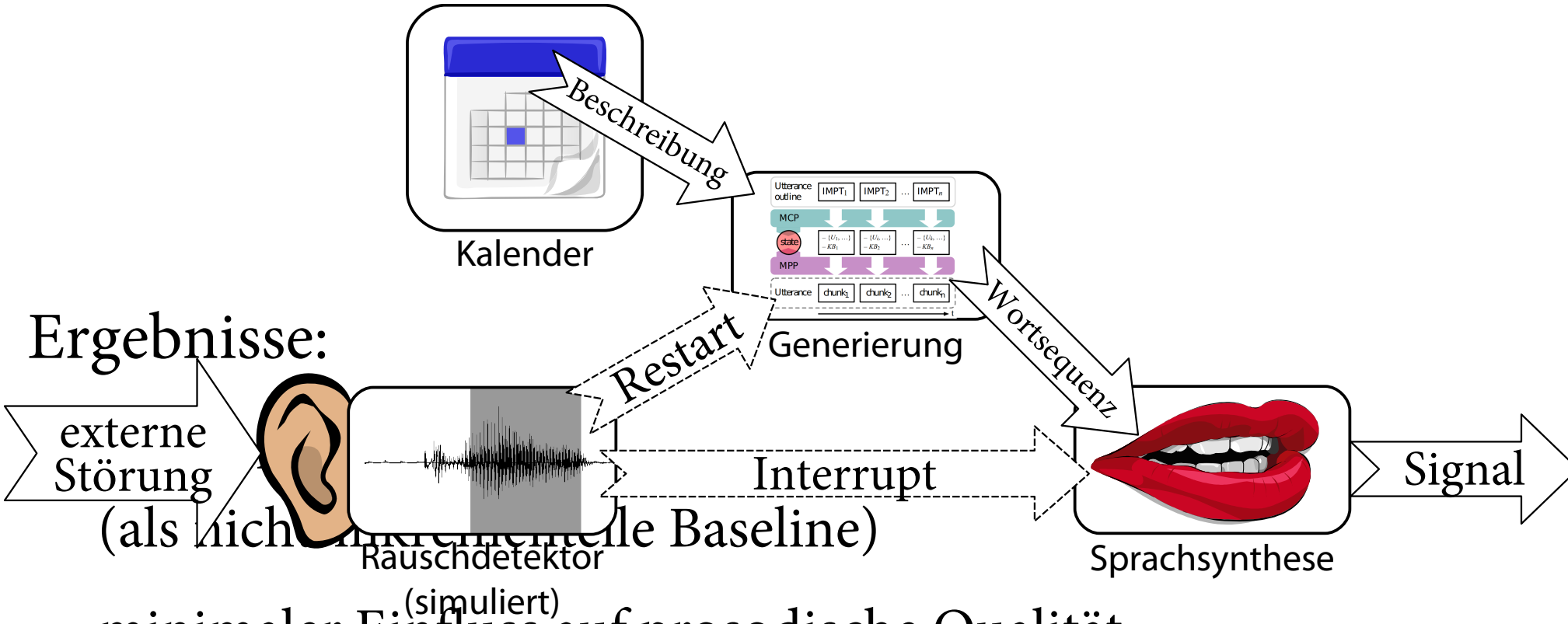
- inkrementelle *Spracherkennung*, zum Beispiel für interaktive Steuerung



- Ergebnisse:
  - zeitgerechte, stabile inkrementelle Zwischenergebnisse
  - einfache, robuste und intuitive Steuerung (Baumann etal 2013)

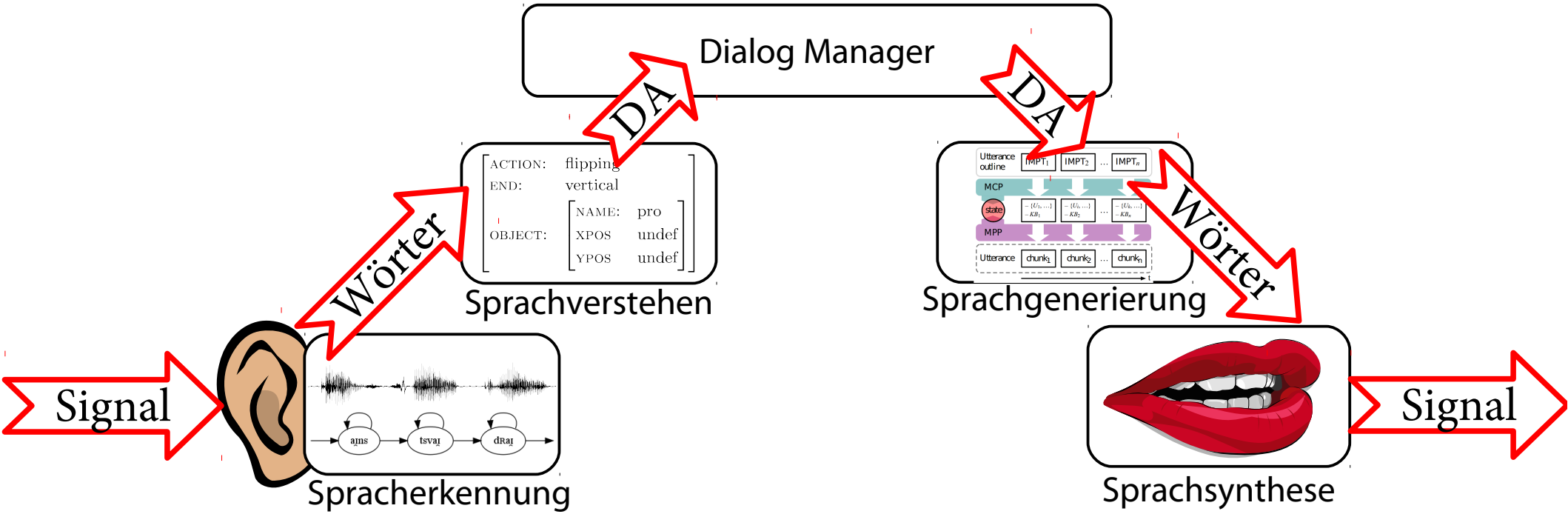
Behandelte Teilprobleme (2)

- inkrementelle *Sprachsynthese* für reaktive Ausgaben



- Ergebnisse:
  - (als nicht-inkrementelle Baseline)
  - minimaler Einfluss auf prosodische Qualität
- Interaktiv: flexible Anpassung an externe Störungen

Sprachdialogsysteme jenseits der Ping-Pong-Interaktion



- kontinuierliche Verarbeitung

  ➔ *kontinuierliche Interaktionssteuerung*

Interaktionssteuerung und Dialogfluss-Analyse

ohne Ping-Pong-Schema gewinnt
Dialogfluss-Analyse an Bedeutung:

- ist der Nutzerbeitrag jetzt (gleich) zuende?
- ist dies ein guter Moment für Feedback?
- sollte das System den Nutzer unterbrechen?
  (und: wird das System gerade vom Nutzer unterbrochen?)
- jeweils möglichst zügige und genaue Entscheidung

Echtzeit-Verhalten

- Dialogfluss-Analyse muss sinnvollerweise *(r)echtzeitig* erfolgen
  - Prognosefähigkeiten sind notwendig, denn
  - Verzögerungen im System sind unvermeidbar
- Proof-of-Concept:
  - *synchrones Mitsprechen* (bei bekanntem Text)
    - zum Beispiel Co-Completion/Shadowing
      (notwendige Textprädiktion siehe DeVault et al. 2009)

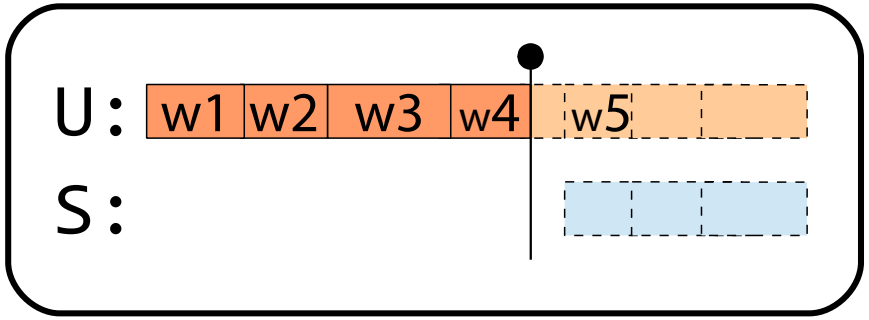
Synchrones Mitsprechen

- zeigt die Echtzeitfähigkeit des Verarbeitungssystems
  - alle *Verzögerungen* (Erkennung, Ausgabe, …) werden durch *entsprechende Vorhersagefähigkeiten* ausgeglichen
- die Echtzeitfähigkeit ist auch für weitere synchrone Verhaltensweisen notwendig:
  - präzise aligniertes Backchannel-Feedback (oder visuell)
  - Disfluenzerkennung (→ Abweichung von Vorhersage)
  - verbesserte End-of-Turn-Vorhersage
  - koordiniertes Verhalten, z. B. bei *Turn-Fights*

Synchrones Mitsprechen (2)

- Aufgabe:
  - Erkennung der gesprochenen Wörter in Echtzeit
    (jedes Wort *bevor* es zuende gesprochen wurde)
  - Vorbereitung des nächsten Wortes (insb. Tempo)
  - Beginn und Aussprache des nächsten Wortes synchron
- Micro-Timing-Abschätzung:
  - Schätzung wie lange das aktuelle Wort andauern wird
  - Schätzung der Sprechdauer (Tempo) des nächsten Wortes

Micro-Timing-Abschätzung

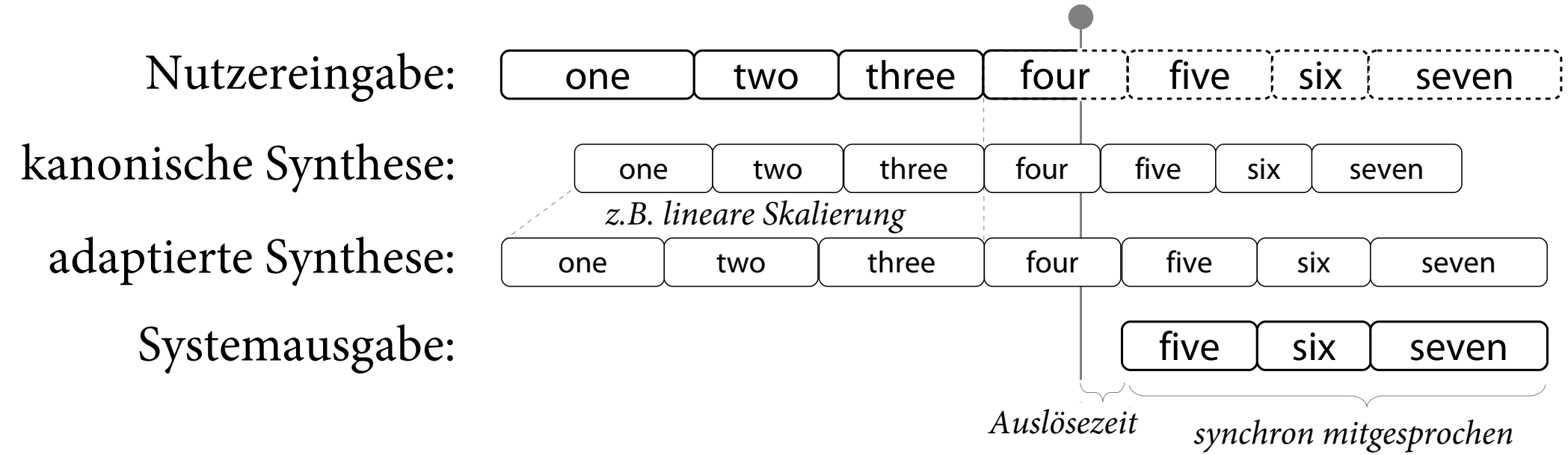


- Zeitverlauf in der *Vergangenheit* ist (prinzipiell) bekannt
  - Ergebnis der inkrementellen Spracherkennung
- Inhalt in der Zukunft ist bekannt (Annahme)
- Zeitverlauf in der Zukunft soll vorhergesagt werden

Analysis-by-Synthesis-Strategie zur Dialogfluss-Vorhersage

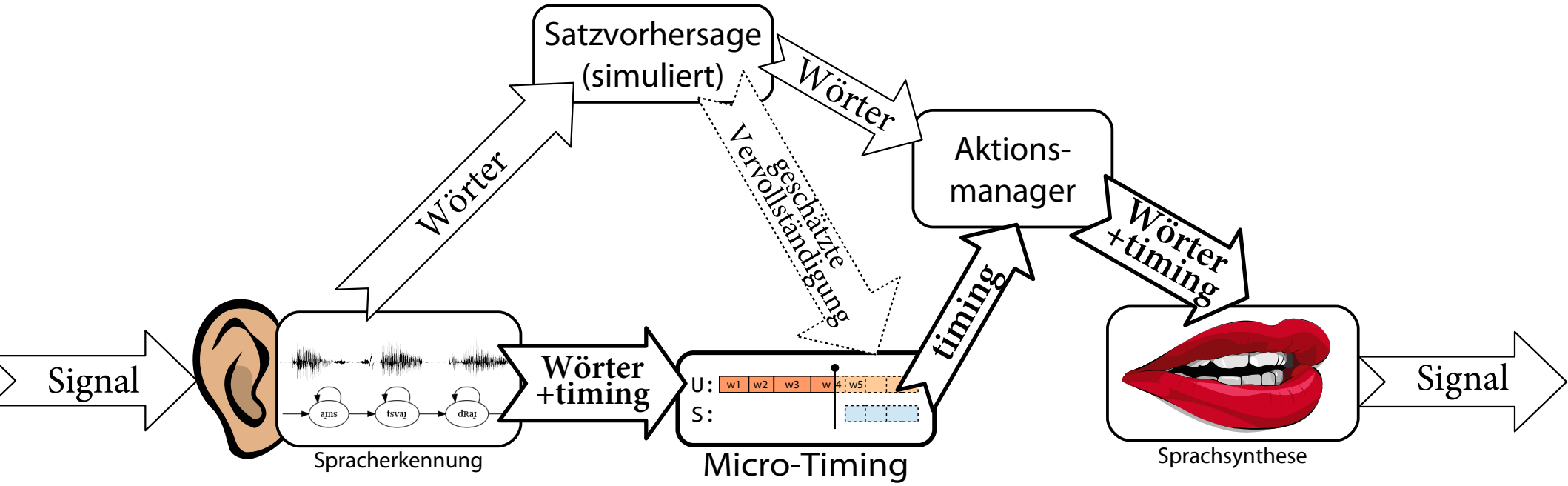
- für jedes neu erkannte Wort des Nutzers:
  - Sprachsynthese liefert kanonische Realisierung der ganzen Äußerung (bekannt oder vorhergesagt)
    - ➔ beachtet syntaktische, morphologische, phonologische Daueffekte
  - Anpassung der kanonischen Realisierung an Nutzer *auf Basis des bekannten Teils der Äußerung* (hier: lineare Skalierung)
  - nächstes Wort ist korrekt skaliert (Annahme: Sprechertempo ändert sich nur langsam)
  - kann also synchron mitgesprochen werden

Analysis-by-Synthesis Strategie zur Dialogfluss-Vorhersage



- Vorhersagegenauigkeit des Modells ähnlich gut wie Menschen bei synchronem Lesen (Cummins 2008)

Behandelte Teilprobleme (3): Mitsprechen in Echtzeit



- Vorhersagegenauigkeit des Modells
  in etwa so gut wie Menschen beim Mitsprechen

Hörbeispiel: Mitsprechen in Echtzeit

Endlich gab der Nordwind den Kampf auf.
Nun erwärmte die Sonne die Luft mit ihren freundlichen Strahlen
und schon nach wenigen Augenblicken zog der Wanderer seinen Mantel aus.

1. nur die mitzusprechende Passage
2. nur die zeitgleich erzeugte Passage
3. beide gemeinsam

Anmerkungen:

- nicht jedes Wort wird mitgesprochen
  (nur wenn Modellsicherheit hoch ist)
- benutzt Wort-für-Wort-Synthese
  (keine vollständig inkrementelle Synthese)

Zusammenfassung

- feingranulare inkrementelle Verarbeitung funktioniert
- und erweitert das Interaktionsinventar von SDS

- proaktive Prädiktion kann Systemverzögerungen an anderer Stelle ausgleichen und erlaubt dadurch *synchrone Interaktionsfähigkeiten*

Vielen Dank für Ihre Fragen!

- feingranulare inkrementelle Verarbeitung funktioniert
- und erweitert das Interaktionsinventar von SDS

- proaktive Prädiktion kann Systemverzögerungen an anderer Stelle ausgleichen und erlaubt dadurch *synchrone Interaktionsfähigkeiten*

These 1

- Inkrementelle Verarbeitung stellt bisherige Muster für Sprachdialogsysteme infrage
  - *Verarbeitung*smuster gängiger SDS muss massiv umgestellt werden (von Turn-basierter auf inkrementelle Verarbeitung)
    - wie lässt sich dies in standardbasierten, verteilten Systemen umsetzen?
  - *Interaktion*smuster könnten sich stark verändern
    - durch das System: wie reagieren Nutzer? wie weit gehen? Adaption?
    - durch die Nutzer: System ermöglicht neue Interaktionsformen, muss mit diesen auch umgehen können (zum Beispiel „Auto-Completion“)
    - bleibt Natürlichkeit das Ziel, oder wird sie durch *Supernatürlichkeit* abgelöst?

Inkrementalität und VoiceXML-Stack

- Erweiterung der Synthesefähigkeit vergleichsweise am einfachsten umzusetzen (vgl. BML-IS-Erweiterung)
  - potentiell kürzere Turnwechsel-Pausen
- inkrementell partielle SISR-Ergebnisse?
  - wie ließe sich das
- VoiceXML als Interface-Sprache ist in sich eher nicht inkrementalisierbar
  - am ehesten durch inkrementelle *fat states*, die Dialogteile externalisiseren (vgl. Baumann et al. 2013)

These 2

- Inkrementelle Sprachdialogverarbeitung ermöglicht neuartige Interaktionsforschung
  - DiET-Toolkit ermöglicht interaktiv manipulierten Chat zur Dialogforschung (z.B. Split-Utterances)
  - inkrementelle Verarbeitung könnte interaktiv & synchron manipulierte Mensch-Mensch-Dialoge ermöglichen
  - was wären Anwendungsmöglichkeiten?
  - was wären ethische Schwierigkeiten?
  - besteht bei solcher Forschung die Gefahr der Dual-Use?

Synchrone Manipulation von Mensch-Mensch-Dialogen

- in steigender Schwierigkeit
  - Pitch-Exkursion (verstärken/schwächen/ändern)
    - zum Beispiel in Kombination mit Zeigegesten, oder bei Korrekturen
  - Spektralverteilung, Tempo, Vokal/Konsonant-Verhältnis
  - Einfügen/Unterdrücken von Feedback, Häsitationen, …
  - in Multi-Party-Dialog: Vertauschung von Sprecher/Hörer-Relationen

These 3

- zukünftig kombinierte Sprach-Ein-/Ausgabekomponente für Dialogsysteme
  - bisher: weitgehend getrennte Komponenten
    - aber ähnliche Daten&Modelle (Lexikon, HMMs, Prosodie, …)
  - *maximale Entfernung* zwischen den Modulen
    - dabei haben sie beide (als einzige) mit dem Sprachsignal zu tun
- sinnvoll wäre eine kombinierte Sprachkomponente

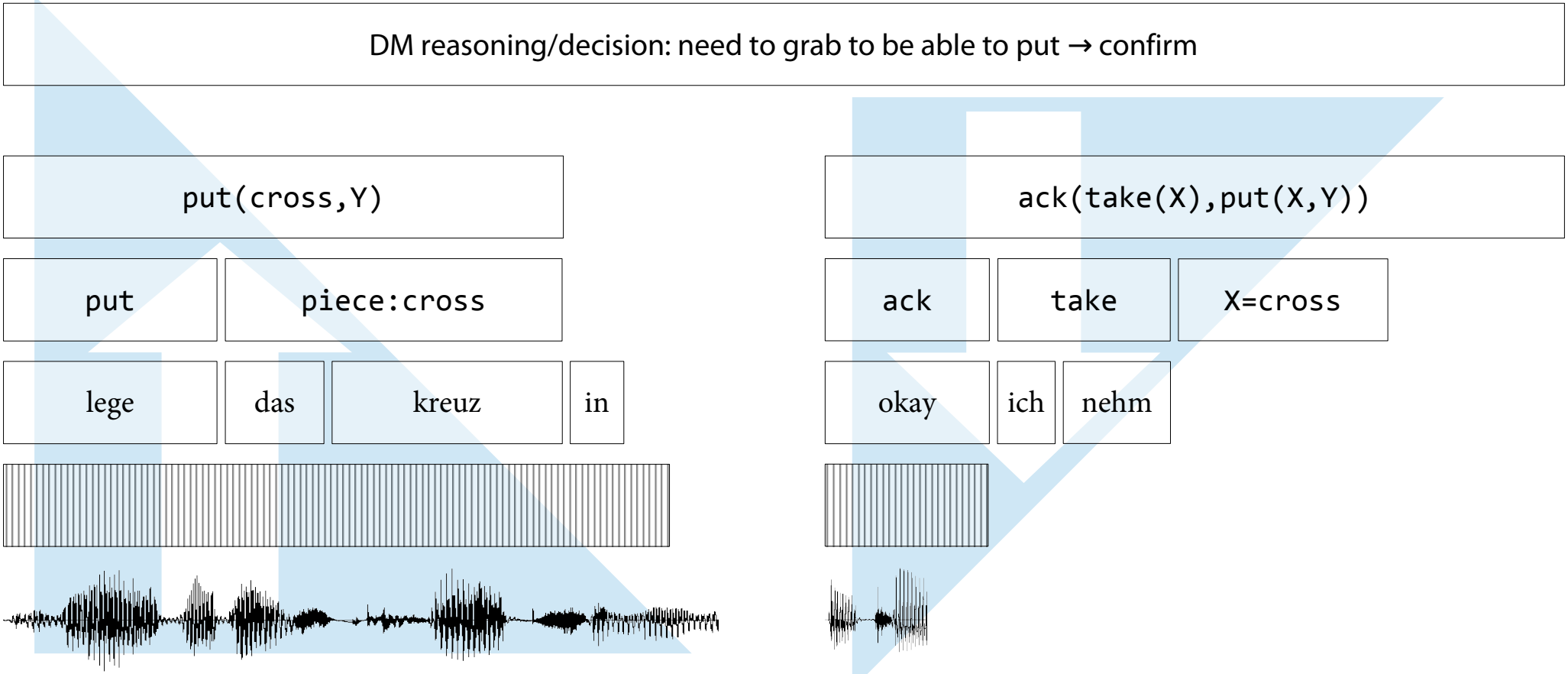
Kombinierte Sprach-Ein/Ausgabe

- vereinfacht Analysis-by-Synthesis-Ansätze
- vereinfacht engere Kopplung zwischen Systemausgabe und Nutzeräußerungen
  - das Modul könnte Feedback, genaues Timing, … autonom übernehmen (sollte jedoch weiterhin Steuerung erlauben)
- Gegenargumente:
  - für Sprachverstehen/-generierung gilt dasselbe Argument
  - Spannungsfeld Modularisierung/Konnektionismus

Übersicht über die Arbeit

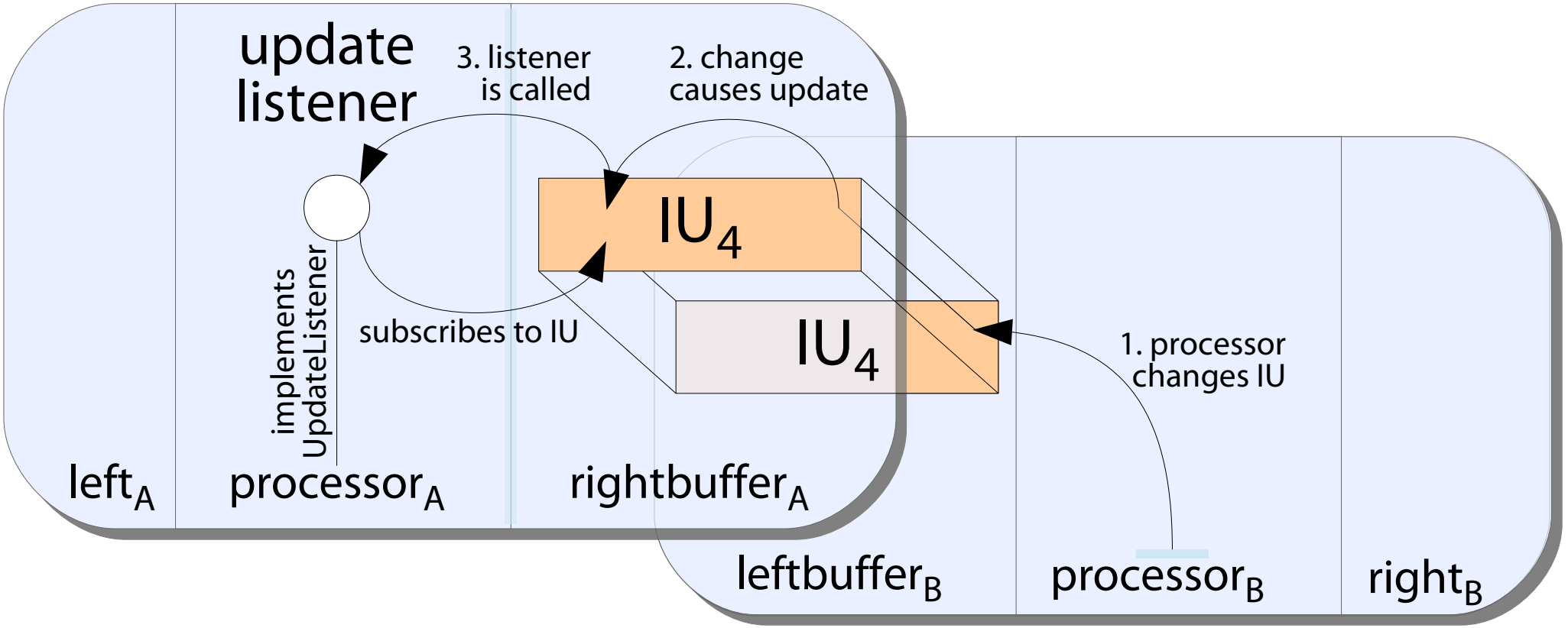
- Formalismus und Qualitätsmaße schritthaltender Verarbeitung für inkrementell revidierte Hypothesen auf Basis des IU-Modells (Schlangen&Skantze 2009) → Kapitel 3
- Software-Toolkit für modulare, inkrementelle Verarbeitung (InproTK), Daten- und Verarbeitungsmodelle → Kapitel 4
- Inkrementelle Spracherkennung → Kapitel 5
- Inkrementelle Sprachsynthese → Kapitel 7
- proaktives, kontinuierliches Agieren im Dialogfluss → Kapitel 6
  - Sub-Turn-Kollaboration
  - synchrones Mitsprechen: *durchgängige Inkrementalität in Echtzeit!*

Datenmodell für inkrementelle Just-in-Time-Verarbeitung



- unterspezifiziert in die Zukunft
- kompatibel mit Lattice-Verarbeitung

Top-Down Feedback trotz unidirektionaler Modulverbindung



- Incremental Units unterstützen das Notification-Muster

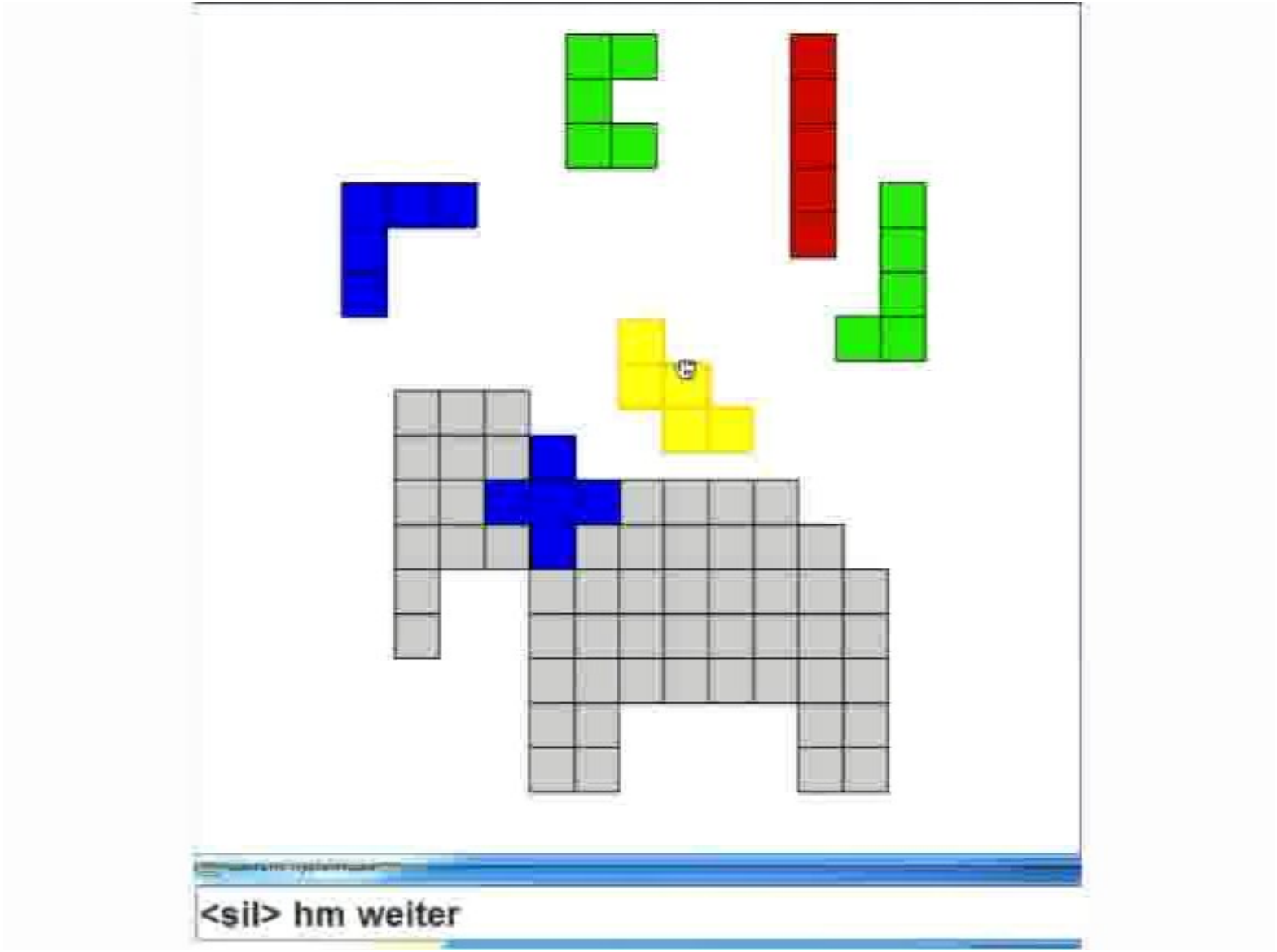
kontinuierliche Steuerung mit inkrementeller Spracherkennung

Video:

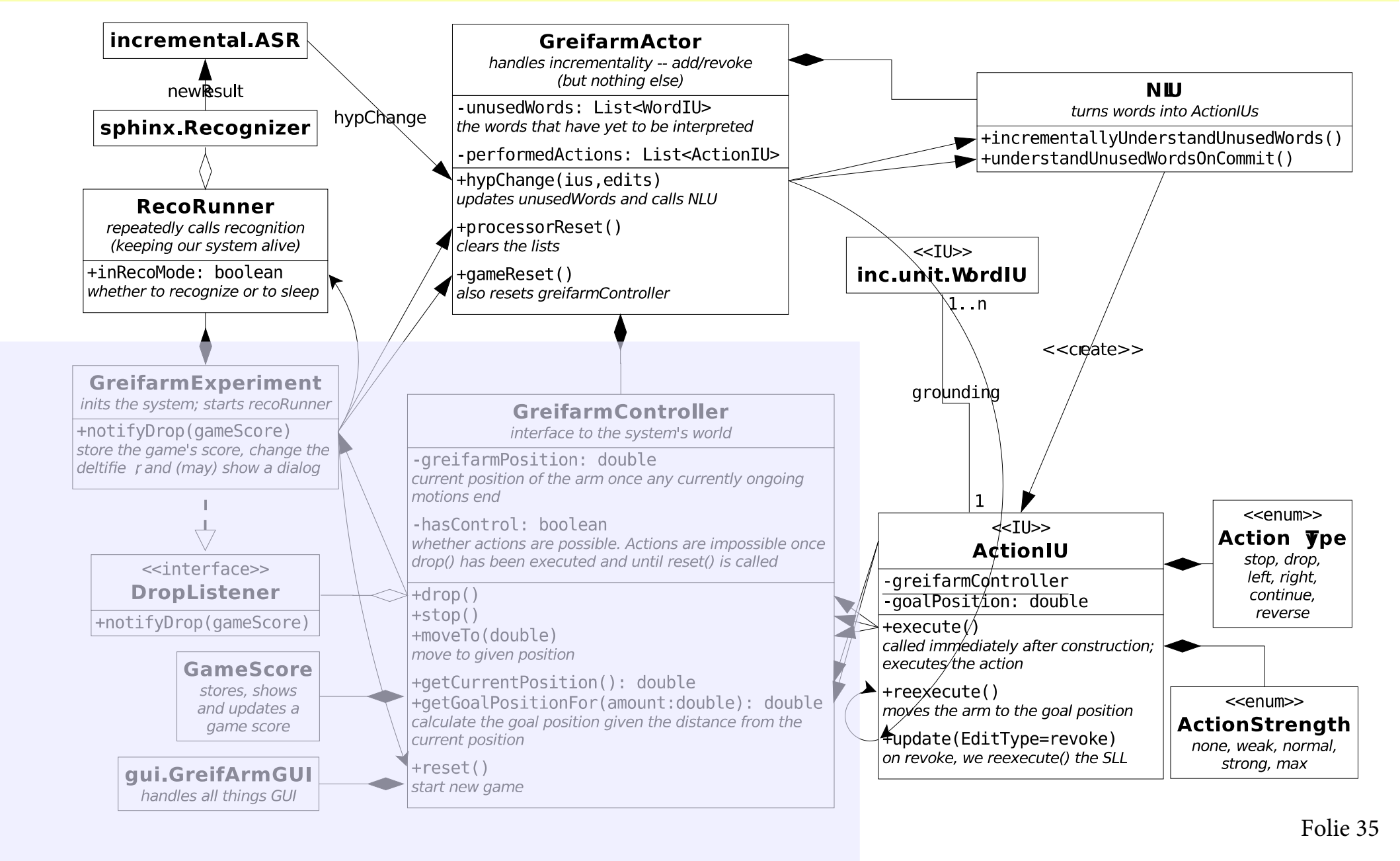
kontinuierliche, flüssige Steuerung einer 2D-Bewegung durch inkrementelle Spracherkennung (Baumann et al. 2013)

- neuartige Interaktionsmöglichkeit
- besonders fehlerrobust da Korrekturen schnell und einfach möglich sind

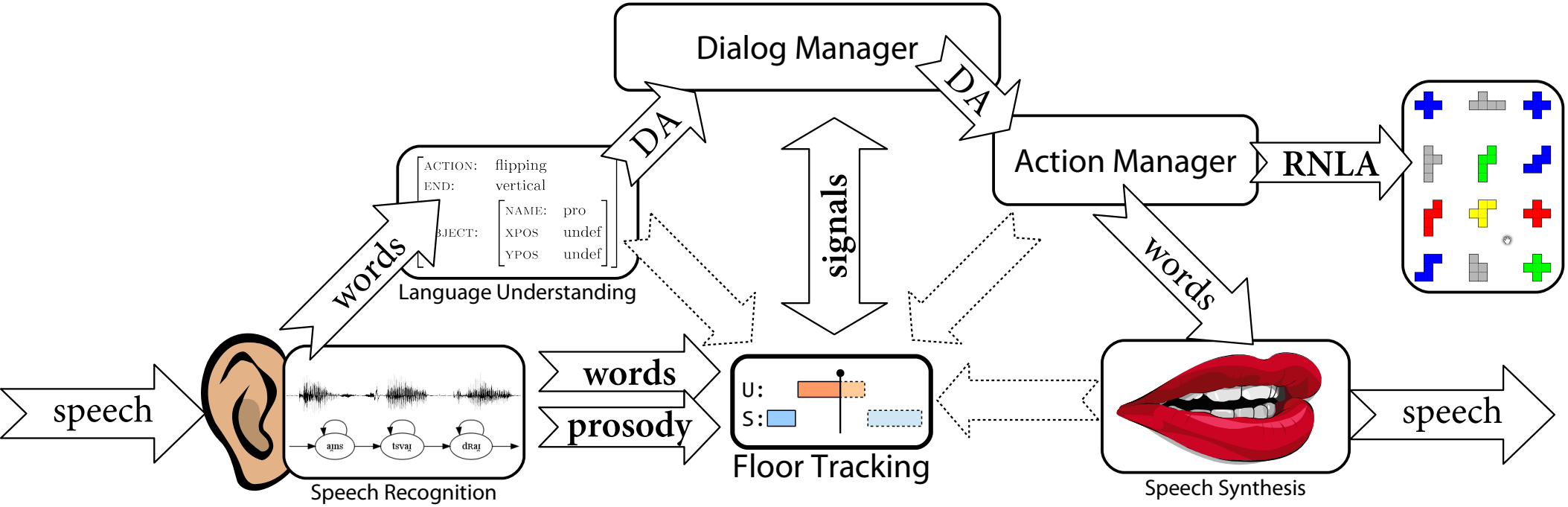
kontinuierliche Steuerung mit inkrementeller Spracherkennung



UML-Diagramm der Steuerung

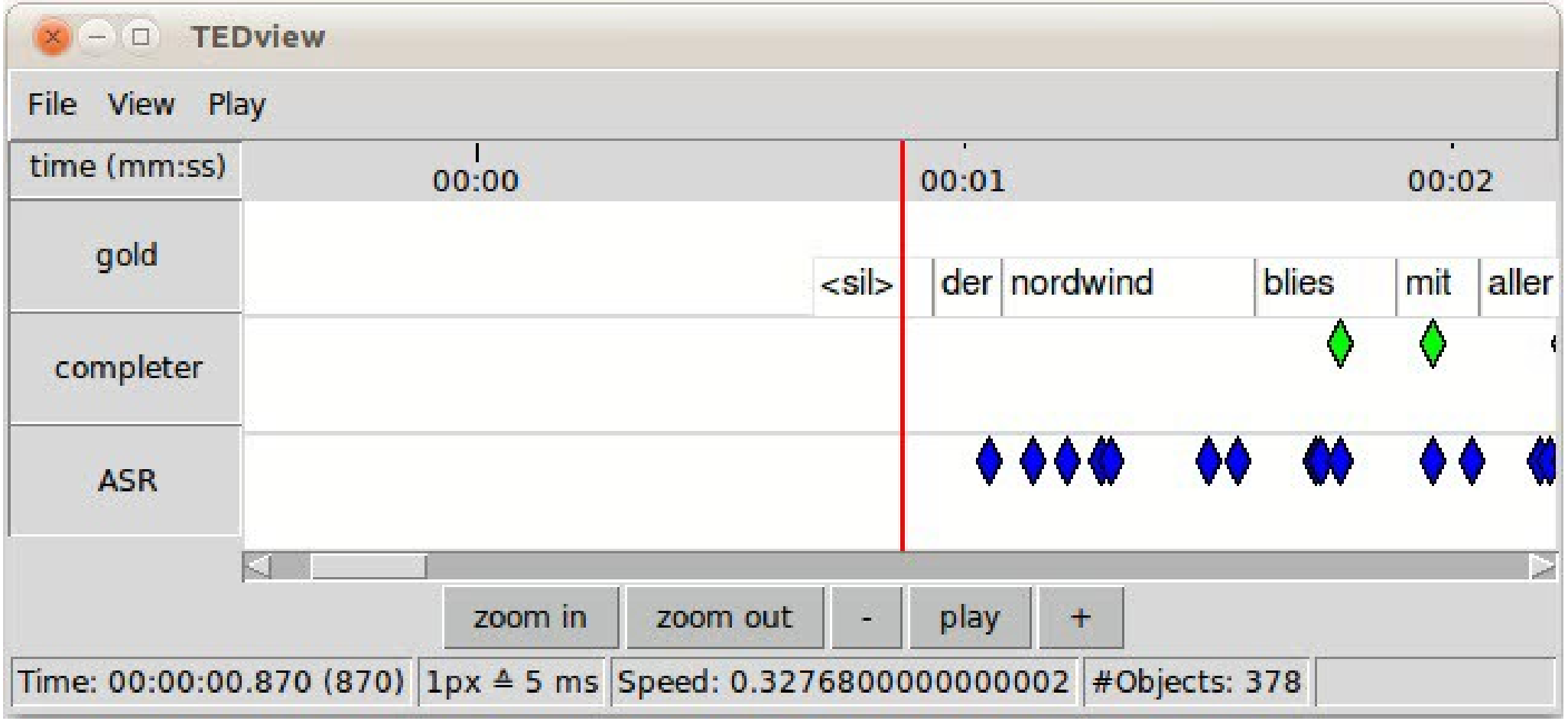


Der Floor-Tracker in InproTK



- Parallelpfad im Verarbeitungsnetz

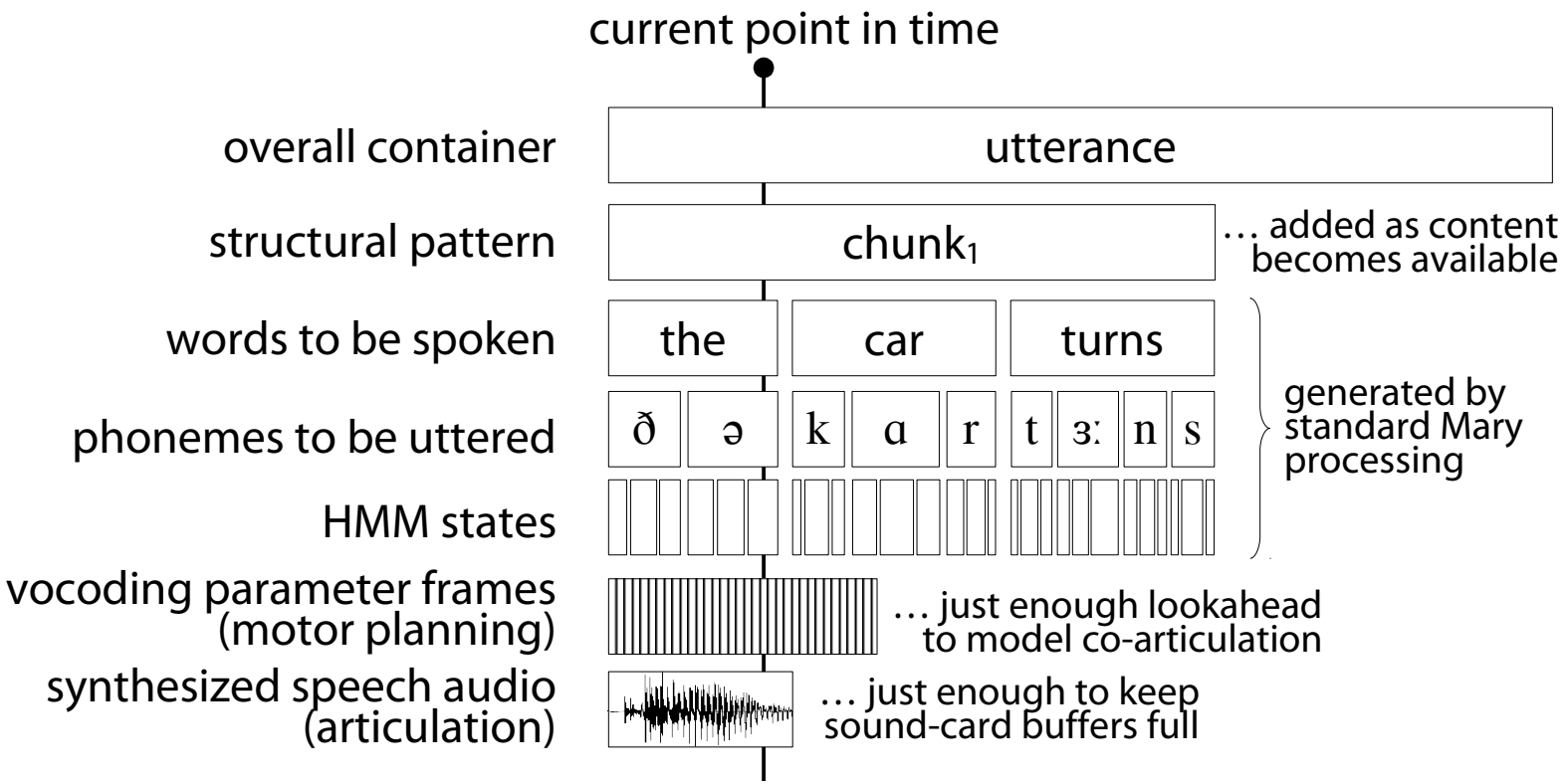
Synchrones Mitsprechen: Beispiel



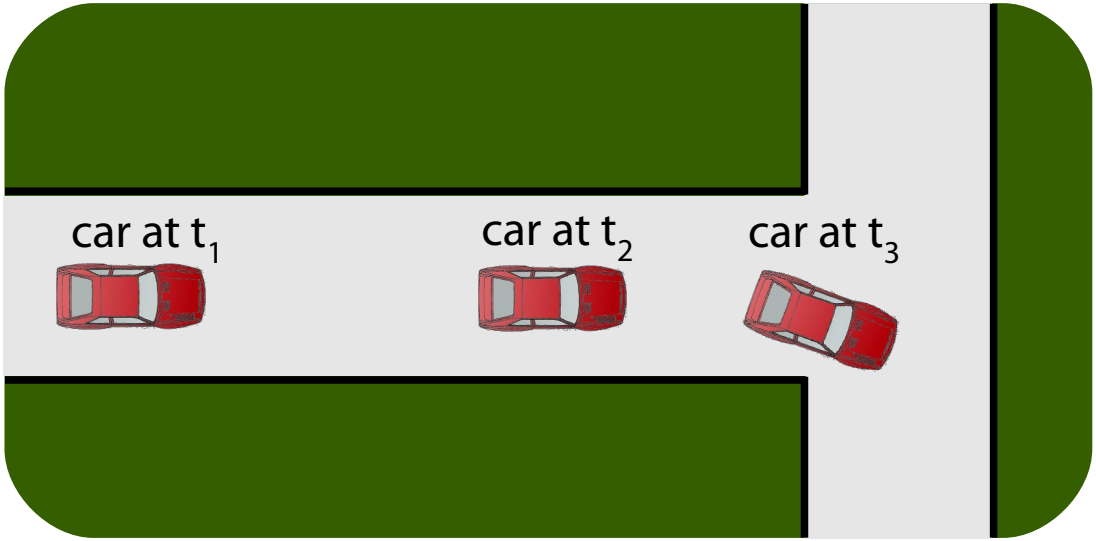
Inkrementelle Sprachsynthese

- bisher wurde inkrementelle Verarbeitung in der Sprachausgabe weitgehend ignoriert
- in interaktiven Anwendungen ist sie aber elementar
  - relevante Information kommt tröpfchenweise hinzu
  - Reaktion auf den Zuhörer oder auf die Umgebung

Inkrementelle Sprachsynthese

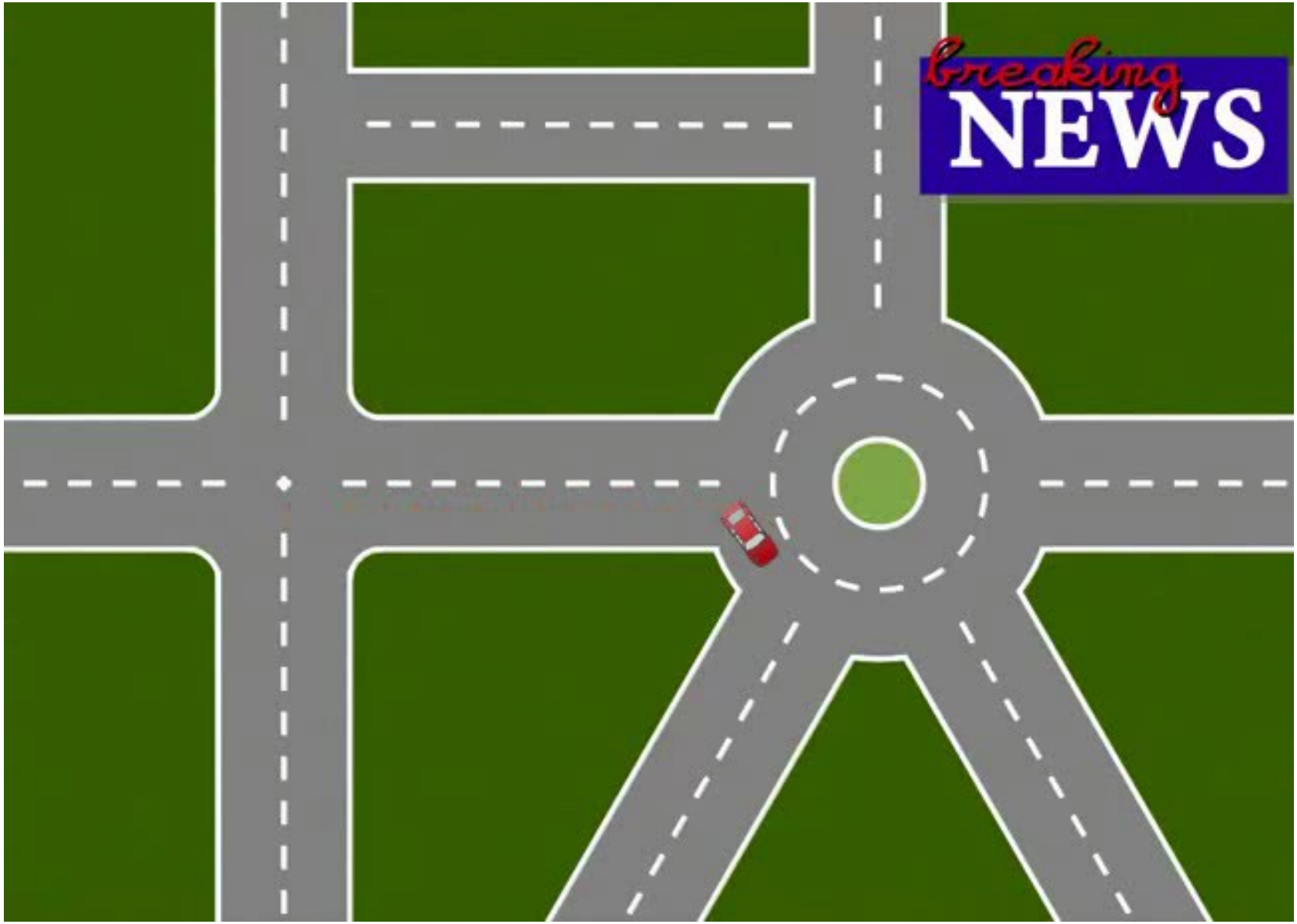


Inkrementelle Sprachsynthese in interaktiver Umgebung

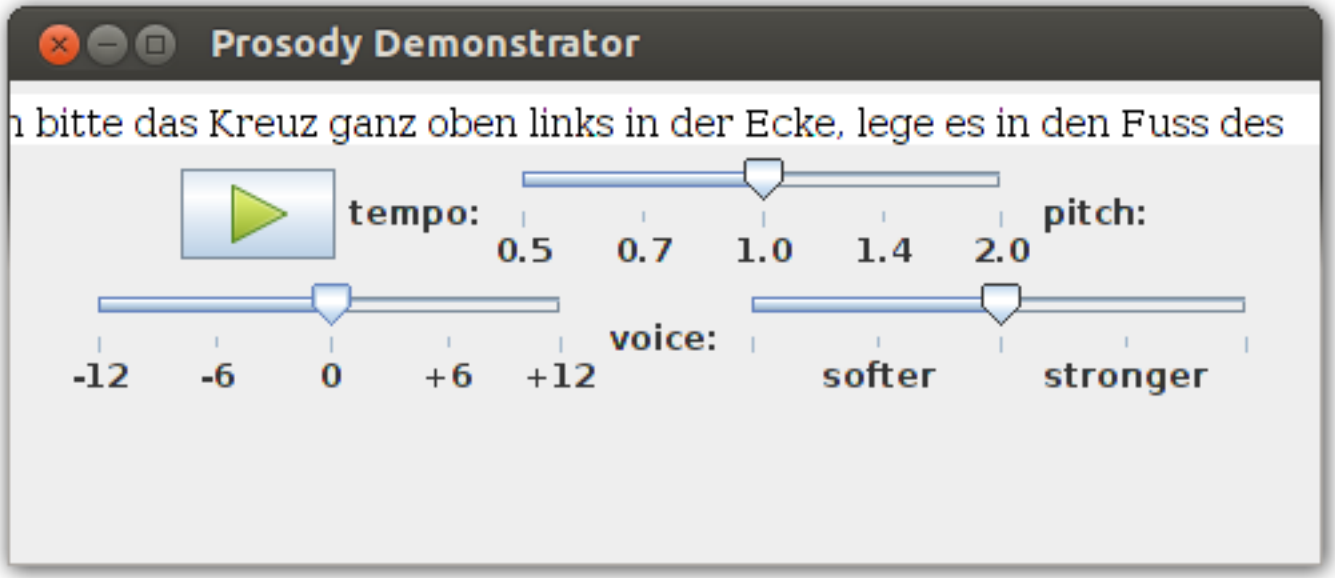


- t_1: Auto fährt durch die Hauptstraße
- t_2: Auto wird sicherlich abbiegen (bloß wohin?)
- t_3: Auto biegt nach rechts ab

Inkrementelle Sprachsynthese Beispielvideo



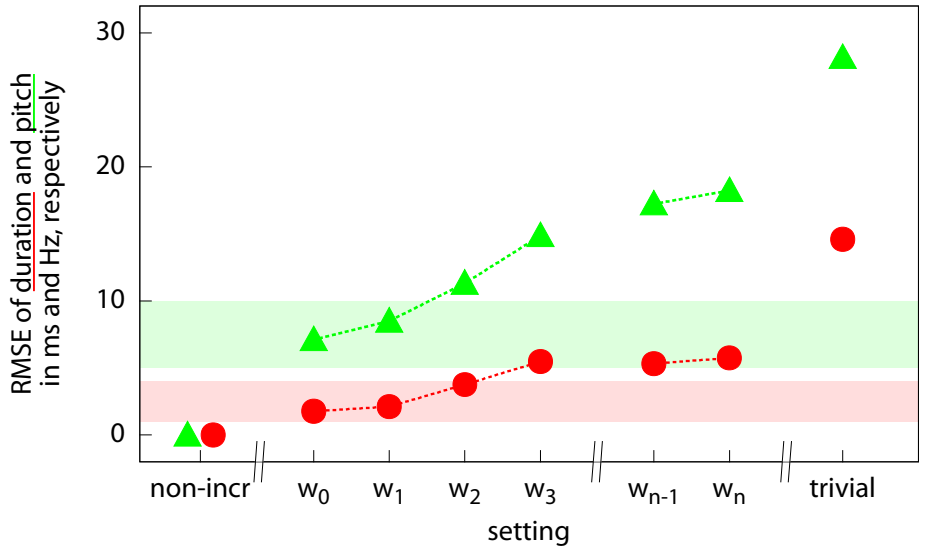
Interaktive Prosodieadaptierung



- Steuerung unterhalb der HMM-Optimierungsebene
  - minimale Verzögerung
- bisher keine gute Anbindung an IU-basiertes Modell

Prosodiequalität inkrementeller Sprachsynthese

- praktisch unverändert solange eine *Phrase* als Kontext in die Zukunft zur Verfügung steht



- vollständig *inkrementelles Prosodiemodell* ist ein zukünftiger Forschungsgegenstand

Inkrementelle Prosodiemodellierung

- Incremental Units für alle relevanten Einheiten
  (insb. für durch Mary bestimmte Intonationsphrasen)
- unterspezifizierte Phrasen und Wörter
  - oft ist bekannt, *dass* eine Äußerung weitergeht, bloß noch nicht *wie*
- just-in-time-Berechnung von Grundfrequenz aus Phrasenakzent+Wortakzent+affect state+lautinhärente Eigenschaften
  - erlaubt die unabhängige Anpassung noch zur Laufzeit
    (z.B. Änderung des Phrasenakzents führt *automatisch* zu verändertem Intonationsverlauf)

Future Work (Left-overs)

- Architektur zur inkrementellen Verarbeitung:
  - *n-best* Hypothesen bzw. *Lattice Processing*; Einfluss aufs Datenmodell
  - Fusion paralleler Eingabe- und Verarbeitungsströme (Gestik, Mimik, …)
  - Analyse inkrementeller Systeme und Synthese von *Entwurfsmustern*, sowie Bestimmung von *Best Practices*
- inkrementelle Spracherkennung:
  - Untersuchung der inkrementellen Qualitäten von A*-Suche anstatt Viterbi
  - *Top-Down-Feedback* höherer Verarbeitungsstufen
  - Verhalten grammatikbasierter LMs, morphembasierte Erkennung
  - inkrementelle Konfidenzschätzung
  - einfach: frühes Commitment für unstrittige Hypothesen

Future Work (Left-overs) II

- inkrementelle Sprachsynthese
  - Integration aller Syntheseschritte in die inkr. Architektur
    - HMM-Zustandsauswahl (MSc-Arbeit)
    - Tokenisierung, etc.
  - inkrementelles Prosodiemodell für reaktivere Ausgabe
- Interaktionssteuerung / Dialogfluss-Vorhersage
  - inkrementelle Prosodie/Dauer-Modellierung
  - Nutzung, zum Beispiel für Backchannel-Feedback

Bisher kein vollständiges iSDS für angemessen komplexe Domäne

- ➔ daher auch kein Nachweis, dass IV dort nützlich ist
- übersteigt den Rahmen der Arbeit
  - insb. Dialogmanagement war hier nicht das Ziel
- stattdessen Proof-of-Concept-Systeme
  - Evaluation eines Systems inklusive Dialogmanagement und Floor-Tracking durch Overhearer-Studie
  - alle Schritte (außer DM) in Teilsystemen getestet
  - vgl. auch (Baumann et al. 2013) für Hybridsystem aus inkrementeller und nicht-inkr. Verarbeitung

stubborn/yielding vs. monoton/nicht-monoton

- monotone inkrementelle Verarbeitung:
  Hypothesen können nicht nachträglich verändert werden
  (in der Arbeit wird nicht-monoton verarbeitet)
- nachgiebige (yielding) Verarbeitung:
  nutzt Nicht-Monotonität um zu finalen Ergebnissen zu gelangen wie sie unter Berücksichtigung der gesamten Eingabe möglich sind

  yielding → non-monotonic

  ¬(non-monotonic → yielding)