

# Taking Turn-Taking Simulation to the Real World™

---

## Coordination Between Agents Workshop



Timo Baumann

[timo@ling.uni-potsdam.de](mailto:timo@ling.uni-potsdam.de)

<http://www.ling.uni-potsdam.de/~timo>

Acknowledgements:

David Schlangen, Michaela Atterer

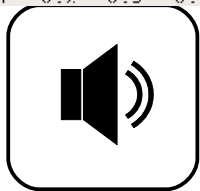
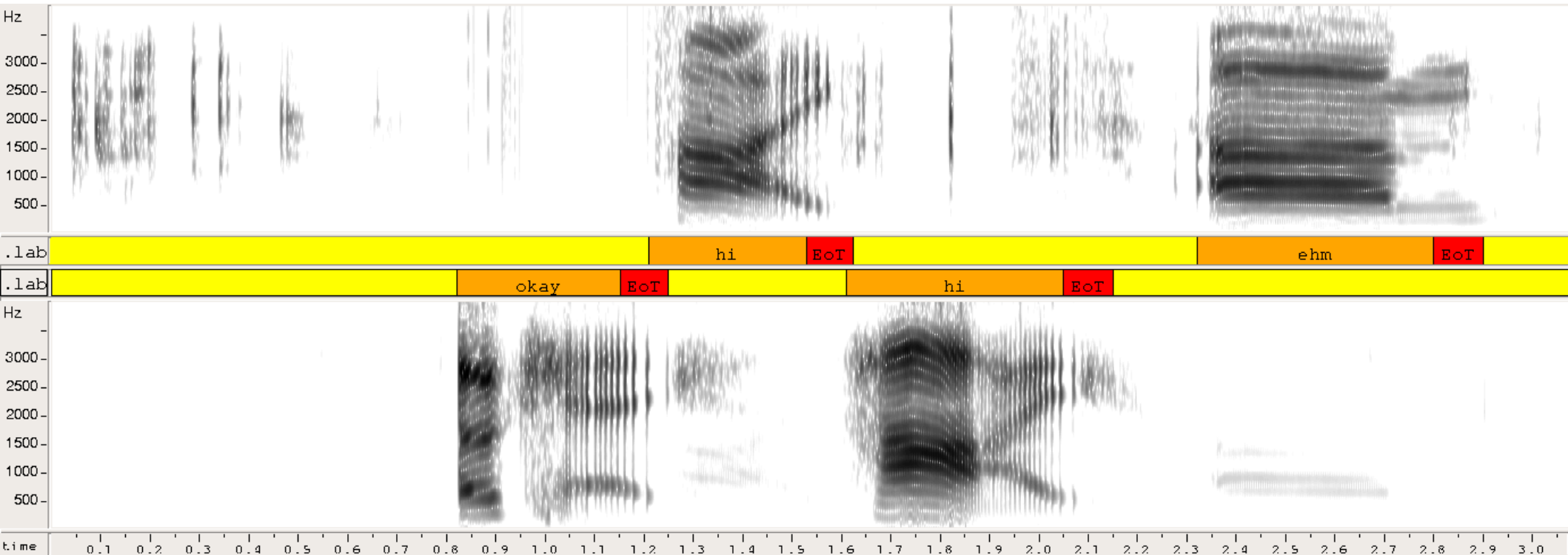
# Turn-Taking in Dialogue

---

According to Sacks, et al (1974)  
turn-taking management in dialogue is

- predictive
  - turn endings are signalled and interpreted in advance
  - syntactic, semantic, pragmatic & **prosodic** cues
- locally managed
  - relying on local context only
  - Transition Relevance Places mark possible turn-changes

# Example: Natural Turn-Taking



Example from Switchboard, telephone speech

# Our Goal:

## Simulate Realistic TT Behaviour

---

- have two agents interchange speech
  - observe „natural“ turn-taking behaviour
  - natural environment
- what rules do the agents have to follow in order to be successful?
  - what's important, what's not in turn-taking
  - use this knowledge to improve Spoken Dialogue Systems

# Taking Turn-Taking Simulation to the Real World™

---

- 1 Symbolic Turn-Taking Simulation: Padilha (2006)
- 2 Challenges for a more realistic Simulation Environment
- 3 Dialogue Simulation Architecture
- 4 Classification of Audio Into Speech States
- 5 Locally Managed Turn-Control Rules
- 6 Conclusions, Current and Future Work

# A Symbolic Approach to Turn-Taking Simulation

---

Padilha & Carletta (2002), Padilha (2006)

- agents generate TT-relevant symbols
  - *sil*, *start* (with precise timestamp), *talk*, *preTRP*, *TRP*, several kinds of back-channels
  - additional modalities: gaze, gesture, posture
- symbol exchange is synchronized on a blackboard
- rules determine behaviour
- some variables tune talkativeness, confidence, etc.

# Our Goal:

## The Real World

---

- exchange audio between dialogue participants
  - recordings allow naïve third person evaluation
- ideally, be able to interact directly with one (or more) dialogue participants
  - allows first person evaluation
- have a TT module that is useful in a general SDS
- ideally *predict* what's going to happen (like humans do, as is necessary to initiate responses)

# Our Goal:

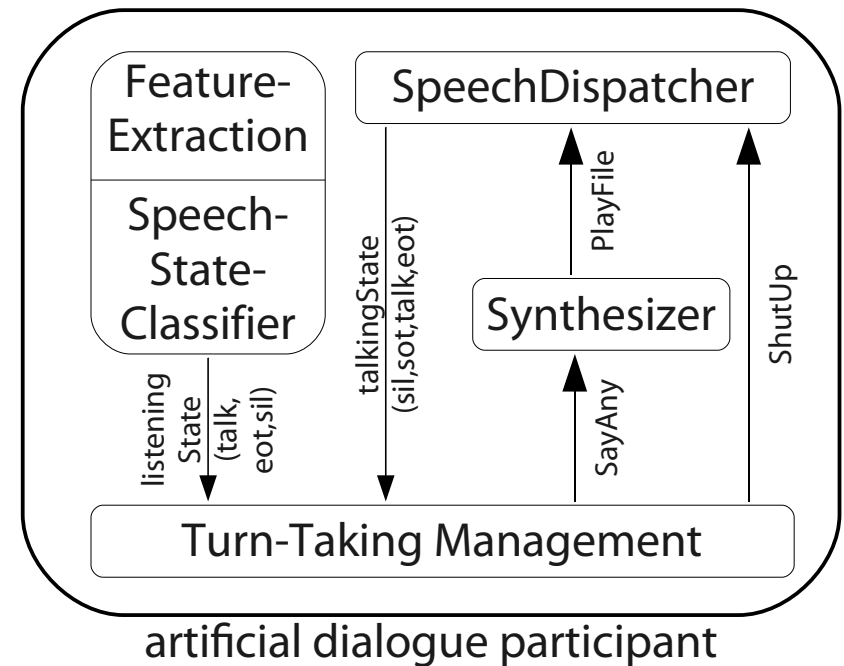
## For Now

---

- exchange audio instead of symbolic messages
- classify speech into talk, silence, end of turn (*EoT*)
- check what the turn-taking management has to know in order to show good behaviour
- abstract away from dialogue content
  - we only use prosody to find TRPs
    - ♦ even abstract away from some phonetic complexity
  - we ignore other cues (words, gaze, gesture, ...)

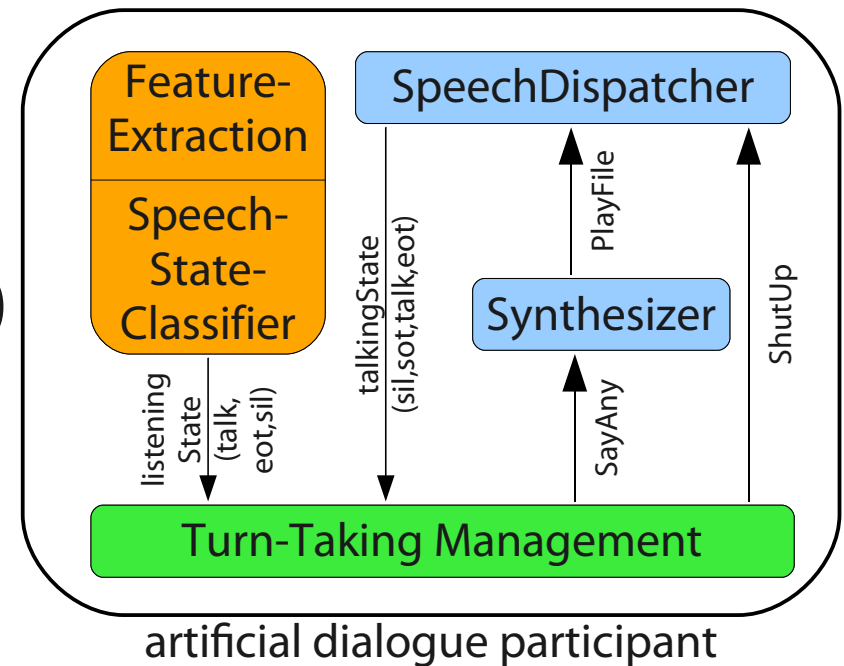
# Dialogue Simulation Architecture

- interaction through asynchronous audio streams over RTP
- headset tool, recording and monitoring tools
- artificial DPs define their own internal communication



# Simulated Dialogue Participants

- internal communication using the Open Agent Architecture (Martin, et al 1999)
  - speech generation randomly selects canned audio
  - fixed delay of 100 ms which simulates generation time (Levinson 1983 says 200 ms)
- introduces necessity for prediction



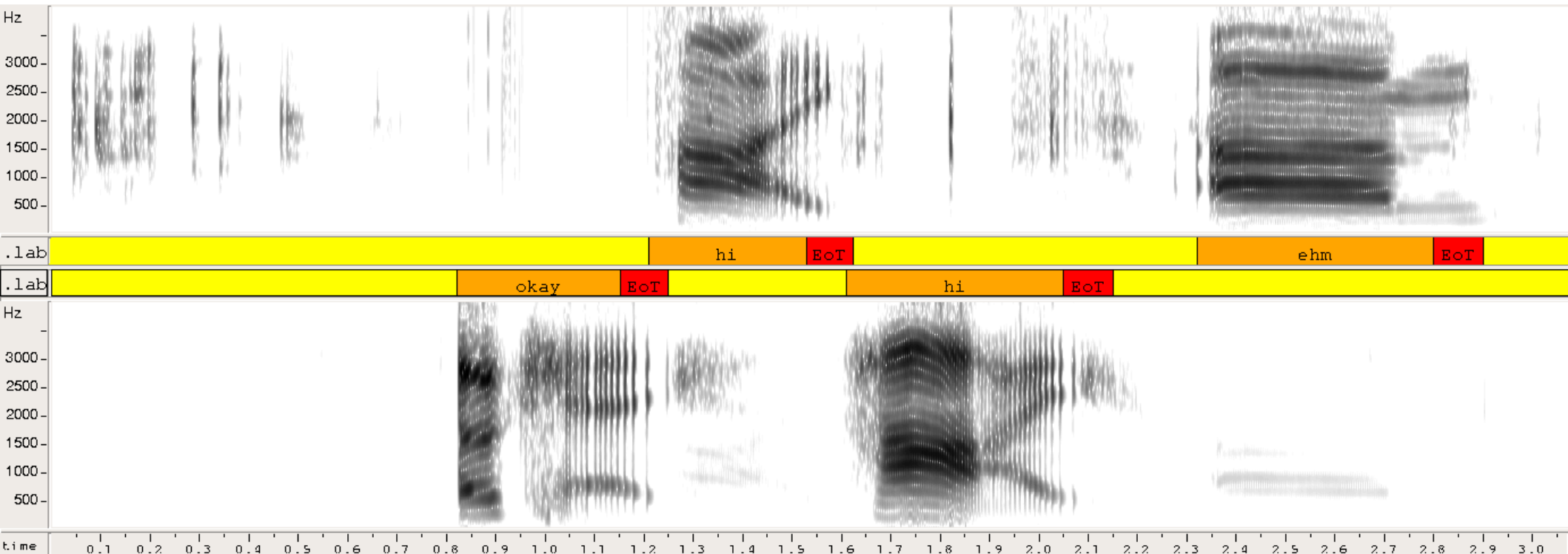
# Taking Turn-Taking Simulation to the Real World™

---

- 1 ~~Turn Taking in Dialogue~~
- 2 ~~Dialogue Simulation Architecture~~
- 3 Classification of Audio Into Speech States
  - Evaluation
- 4 Locally Managed Turn-Control Rules
  - Evaluation
- 5 Conclusions, Current and Future Work

# Classification of Audio Into Speech States

- continuously extract prosodic features from audio
- classify each frame: *silence*, *talk* or *end of turn*



# Prosodic Feature Extraction

---

very basic, not phonologically grounded features:

- we don't need syllables, phonemes, words
  - pitch and energy for each frame of 10 ms
    - no smoothing, no DP, strictly incremental
  - windows of past pitch/energy-values
  - mean, range, min, max, slope, RMSE, ...
    - short windows: remove outliers, smooth values
    - larger windows: long-term trends
-

# Corpora Used

---

- two different corpora
- each corpus contains two speakers (male, female)
- corpus of controlled pseudo-speech (next slide)
- Kiel Corpus of Read Speech (IPDS 1994)
  - about 600 utterances for each speaker
- both corpora are less complex than real dialogue
  - our results should be considered an upper bound

# Corpus of Pseudo-Speech

---

- speakers read real sentences (50 each)
- but uttered /ba/ instead of real syllables
  - „How are you today?“ → /ba ba ba baba?/
- speech is always voiced
- minimal micro-prosodic effects
- sentence intonation remains untouched
- Example: „Warum ist die Banane krumm?“



# Speech State Classification

---

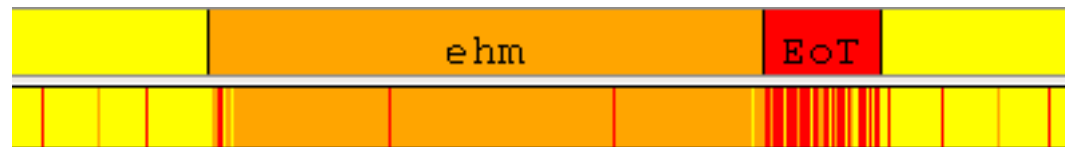
- standard machine learning classifier training
- most predictive feature (according to OneR):  
dynamic range of frame energy over 100 or 200 ms
- results for JRip, J48 (Witten & Frank 2000):

	$F_{\text{sil}}$	$F_{\text{talk}}$	$F_{\text{EoT}}$	FAR
J48	0.98	0.98	<b>0.61</b>	71.1
JRip	0.97	0.98	<b>0.73</b>	61.1

- table for Kiel-Corpus female speaker,  
other settings similar

# Smoothing of Classification Results

- data and states are sequential
- classifiers evaluate frames independently



- many false alarms (over-generated state changes)
    - many false alarms last for just one frame
- only change state after *two consecutive* classifications

	$F_{\text{sil}}$	$F_{\text{talk}}$	$F_{\text{EoT}}$	<b>FAR</b>
JRip	0.97	0.98	0.73	<b>61.1</b>
StatefulJRip	0.96	0.98	0.70	<b>31.9</b>

# Discussion of Classification Results

---

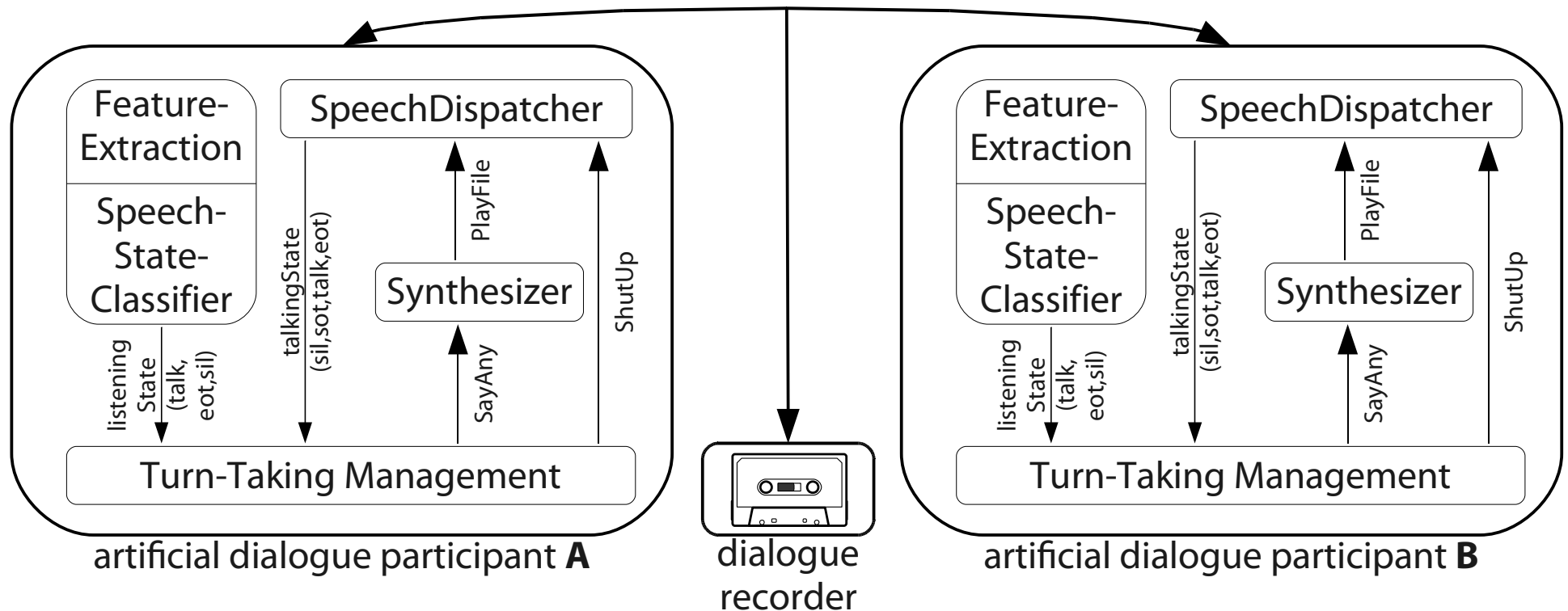
- very good distinction between speech/non-speech
  - **EoT** is of lower quality but better than Schlangen (2006)
- smoothing of classification results reduces FAR
- higher phonetic complexity of Kiel-Corpus is counterbalanced its by larger size
  - may indicate, that speech state classification for real dialogue speech would be feasible with a large corpus and speaker-normalized prosodic features

# Simulating Spoken Dialogue With a Focus on Realistic Turn-Taking

---

- 1 ~~Turn Taking in Dialogue~~
- 2 ~~Dialogue Simulation Architecture~~
- 3 ~~Classification of Audio Into Speech States~~
  - ~~Evaluation~~
- 4 Locally Managed Turn-Control Rules
  - Evaluation
- 5 Conclusions, Current and Future Work

# Setup for Turn-Taking Simulation

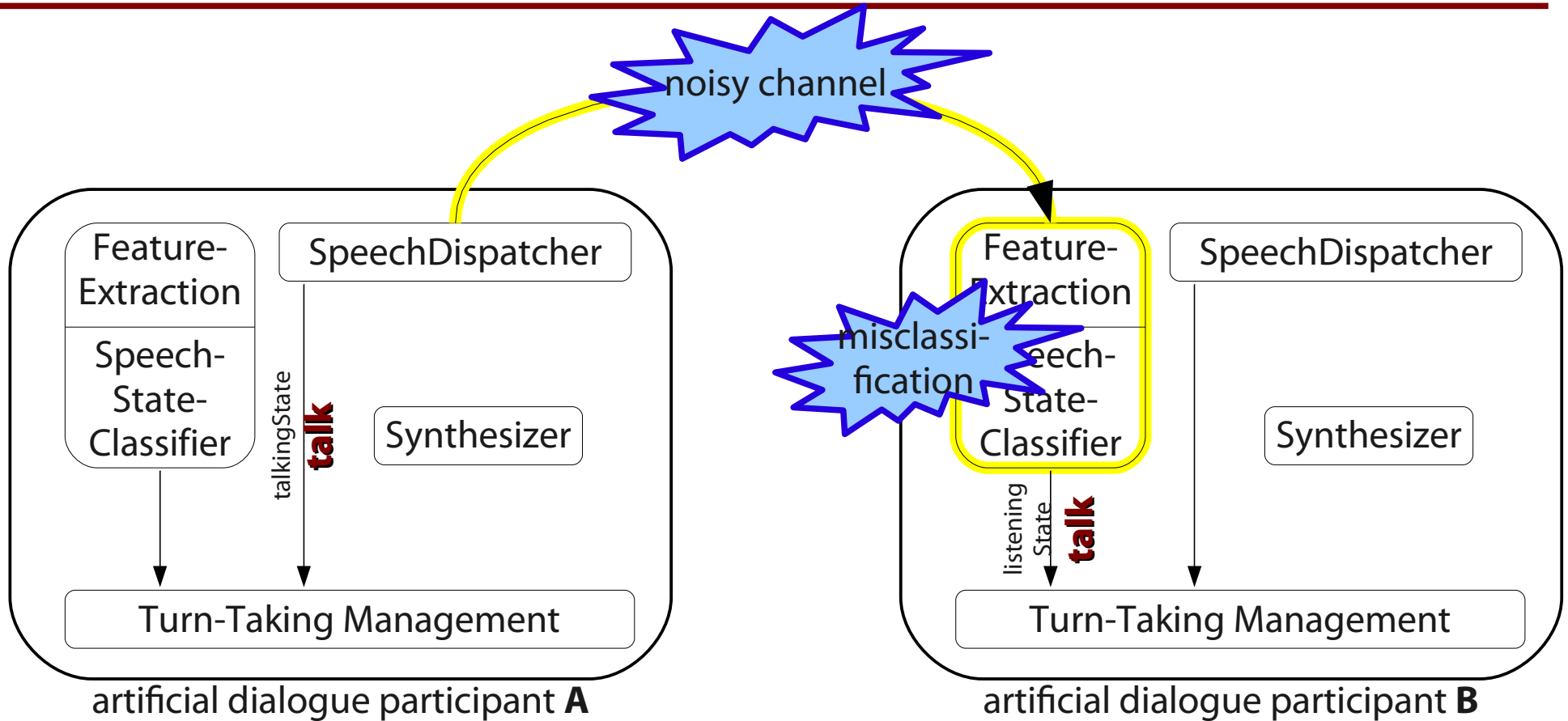


# Inside the Turn-Taking Manager

---

- knows the speech generation's *state*
  - start of turn, **talk**, **EoT**, **sil**
- the currently estimated *listening state*
  - remember? that's what we were dealing with before
  - ideally, this matches the interlocutor's speech state!
- outputs commands to the speech generation module
- always willing but never demanding to talk

when A talks, B should perceive that A is talking and vice versa



→ there will be a lag

and there will be mistakes

# The First Step:

## Do not look back!

---

- Turn-Taking Manager has no history or temporal reasoning
- very simple reflex rules, behaviour only depends on current state

# Simple Strategies for Turn-Taking

---

- 1 *Start talking when neither you nor your interlocutor is talking. Continue until your utterance is finished.*
  - 2 *as above, plus: Stop your utterance, when both you and your interlocutor are talking.*
    - results in turn truncations
  - 3 *change first rule to: Start talking, when your interlocutor is ending their turn **or** has already ended.*
    - effectively *anticipate* turn changes by exploiting the **EoT** class of speech state analysis
-

# Simple Strategies for Turn-Taking

Strategy 1

Strategy 2

Strategy 3

- add a little randomness

		listening state		
		talk	eot	sil
talking state	talk	stop talking		
	eot			
	sil		start talking	start talking

# Evaluation of Turn-Taking Success

---

- dialogue can be described by the speech states of all dialogue participants (each state either **sil** or **talk**)
  - for two-party dialogue, there are two good states
    - one's talking, the other is not (hopefully listening?)
  - and two bad states
    - *clashes*, when both participants talk simultaneously
    - and *gaps*, when neither is talking
- we choose clashes and gaps to measure turn success

# Results for Turn-Taking Strategies

---

results for Kiel-Corpus (pseudo-speech similar):

	strategy 1		strategy 2		strategy 3	
gaps	14%	528ms	21%	477ms	19%	454ms
clashes	24%	1915ms	4%	253ms	4%	243ms

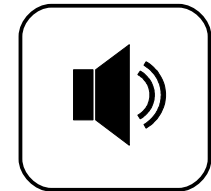
- strategy 3 is similar to Verbmobil Corpus
  - gaps: 363ms, clashes: 331ms (Weilhammer & Rabold 2003)
- predicting **EoT** improves turn-taking performance
  - despite the amount of uncertainty connected with **EoT**

# Example for strategy 3

---

this example shows:

- bad behaviour in the first half
  - many simultaneous starts (or almost simultaneous)
- smooth turn-taking in the second half
- pseudo-speech would be better for listening evaluation purposes  
(but not for fun!)



# Analysis for strategy 3

---

- false starts in the middle of an utterance
    - due to misrecognition of EoT
    - both detect crosstalk, either of them stops (while the one who started should stop more often)
  - many (simultaneous) false starts
    - Padilha resolves this with time-stamping the start, and having a rule that the lower timestamp wins
- introduce speech state timing

# Second Step:

## Time-Stamping Events

---

- time-stamp when a state change occurs
- we can now check what happened earlier
  - somewhat resolves simultaneous start dilemma
    - ♦ unfortunately, listening state always lags behind a little

# Current Extension:

## Back Channel Utterances

---

- back channel utterances (BCs) should occur during longer turns at (almost-)TRPs
- to keep it easy: also allow them at the real EoT (as a kind of turn-acquiring)
- rule: self-sil + other-EoT → give BC
- problem: back-channels *sound* like EoT, thus will be responded to by a BC and so on
- no BC when you've just finished talking yourself

# Further Extensions in a Spoken Dialogue System

---

- *Should I talk?* – Make the agents more self-aware of whether they need to communicate something
  - *Should I listen?* – Make the agents more aware of whether their interlocutor wants to communicate something
- this requires semantic and pragmatic processing
- which in turn need words and trees

# Conclusions and Future Directions

---

- speech state classification is feasible
  - in the future: speaker-normalized features
  - exploitation of higher level information (ASR, ...)
- turn-taking simulation works with real audio
  - allows new ways of evaluation of turn-taking
- turn-taking can be managed locally
  - integrate reasoning about **what** to say
- simple, local rules enough for complex behaviour

# Thanks for Listening

---

feel free to ask questions – now or later

I hope it didn't all sound like bababa to you...