Predictive Modelling of Turn-Taking in Human-Robot Interaction

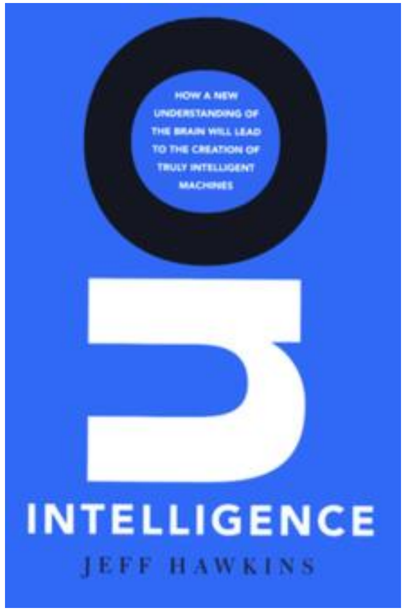
Gabriel Skantze

Professor in Speech Technology, KTH

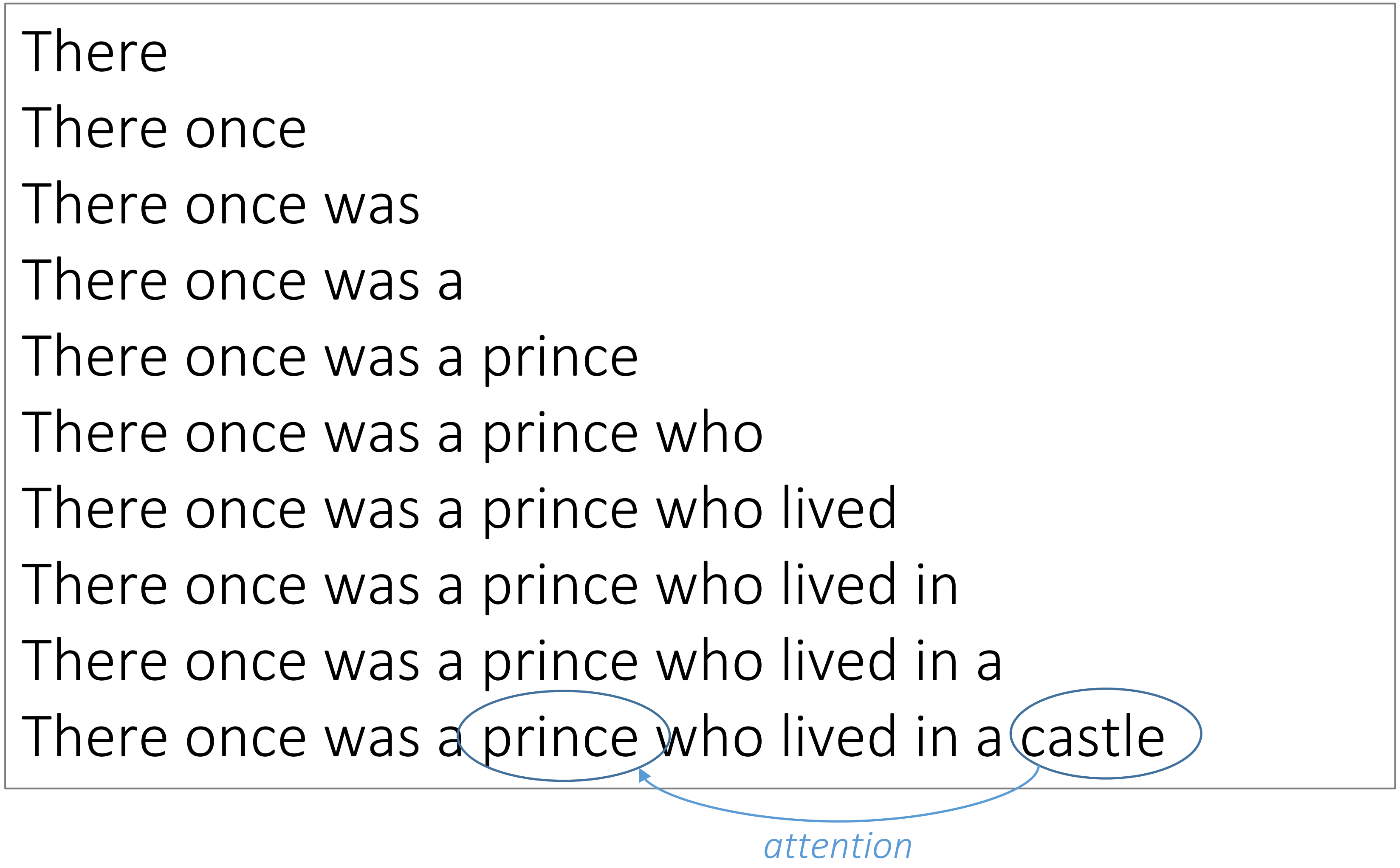
Co-founder and Chief Scientist, Furhat Robotics

Predictive modelling

- Predictive modelling = a statistical or machine learning technique used to predict future outcomes based on historical data
- Predictive modelling on spoken interaction = predict future speech activity based on historical data (the spoken interaction so far)
- Why do we want to do this?
  - Predictive modelling is useful (crucial?) for an agent/robot taking part in an interaction
  - Predictive modelling useful for an agent to learn representations of the data
  - Intelligence = Ability to predict the future? (Hawkins)
    - Bayesian Brain hypothesis


HOW A NEW UNDERSTANDING OF THE BRAIN WILL LEAD TO THE CREATION OF TRULY INTELLIGENT MACHINES
INTELLIGENCE
JEFF HAWKINS

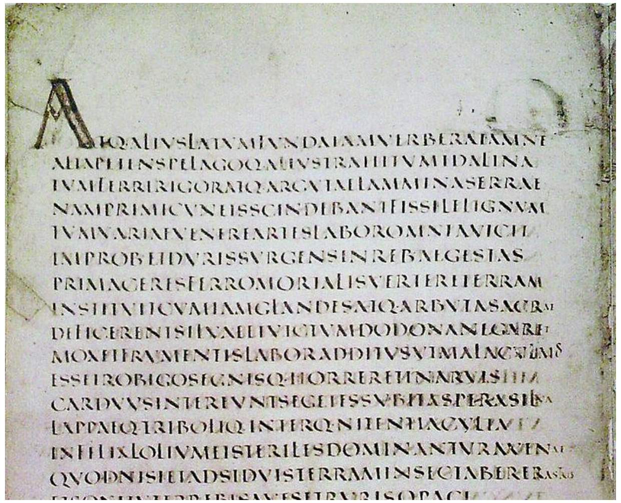
Large Language Models (ChatGPT)



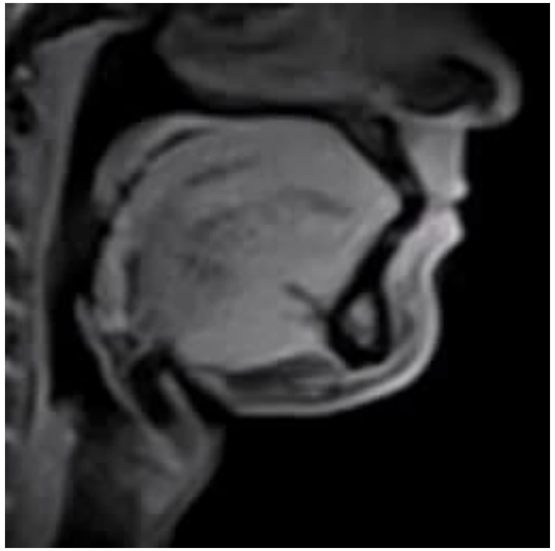
There
There once
There once was
There once was a
There once was a prince
There once was a prince who
There once was a prince who lived
There once was a prince who lived in
There once was a prince who lived in a
There once was a prince who lived in a castle

attention

Cuneiform, 3500BC



Vergilius Augusteus, 4th Century



Written Language	Spoken Language
Used since 5000 years	Used since at least 100.000 years
Words, letters, spaces, punctuation	Continuous, Highly variable, ambiguous and noisy
Asynchronous communication	**Real-time** communication
Syntactically **well-formed**	**Disfluent** (Repetitions, hesitations, truncuted words, etc)
Exclusively **symbolic & verbal** (*what* we say)	**Non-symbolic** & **non-verbal** components (*how* we talk: prosody, laughter, breathing, etc)

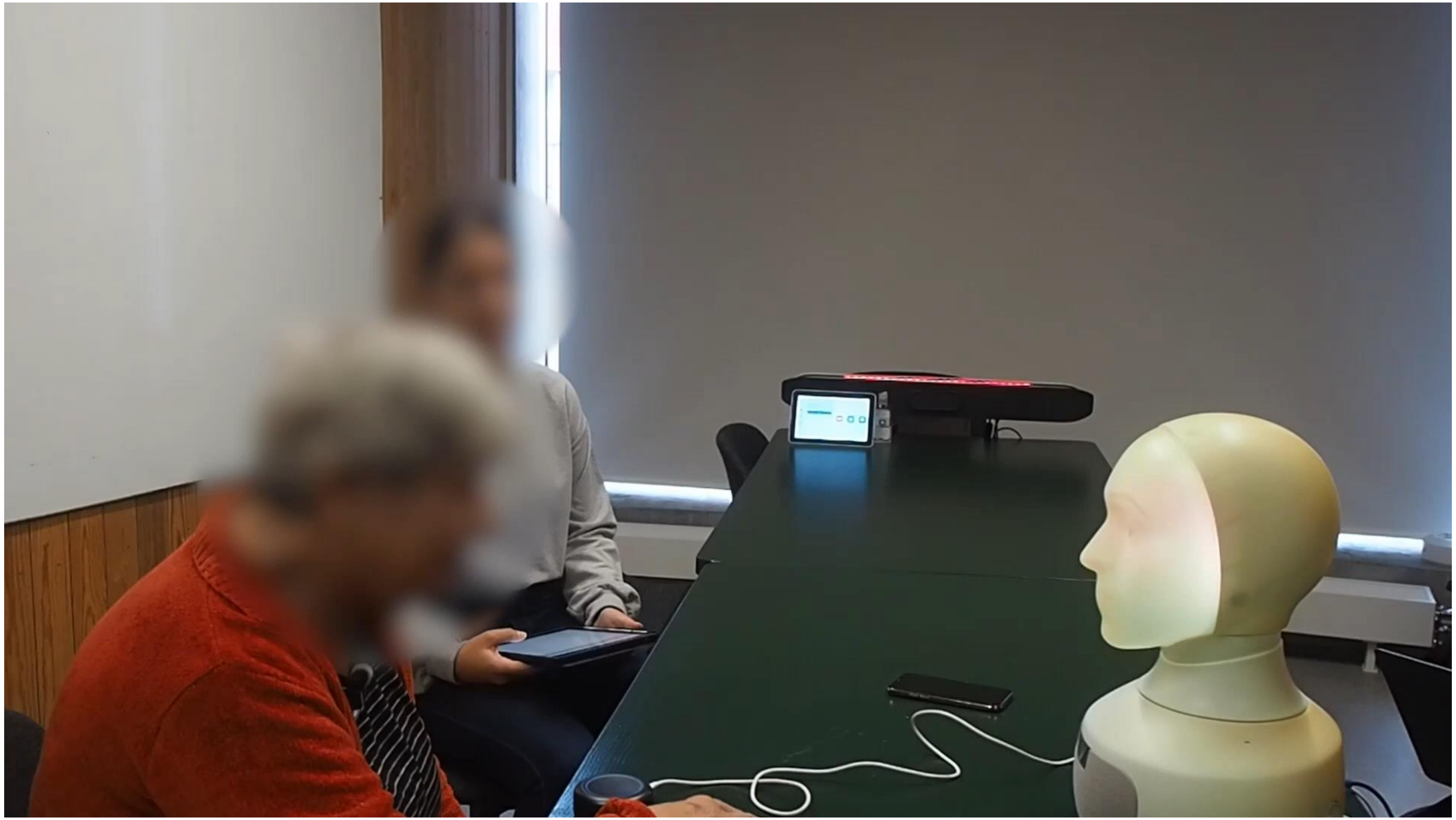
Spoken conversation is a Joint Activity happing in real time

Coordination relies on cues and signals (in the face and the voice)

Coordination requires the ability to anticipate (predict) the partner’s actions

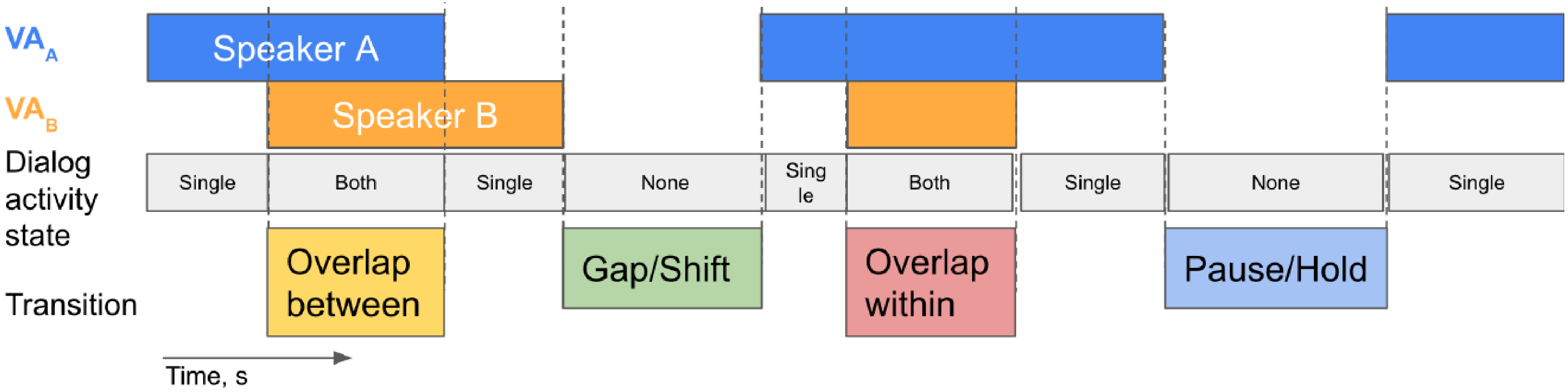
Purely reactive approaches are not enough!

Failed turn-taking

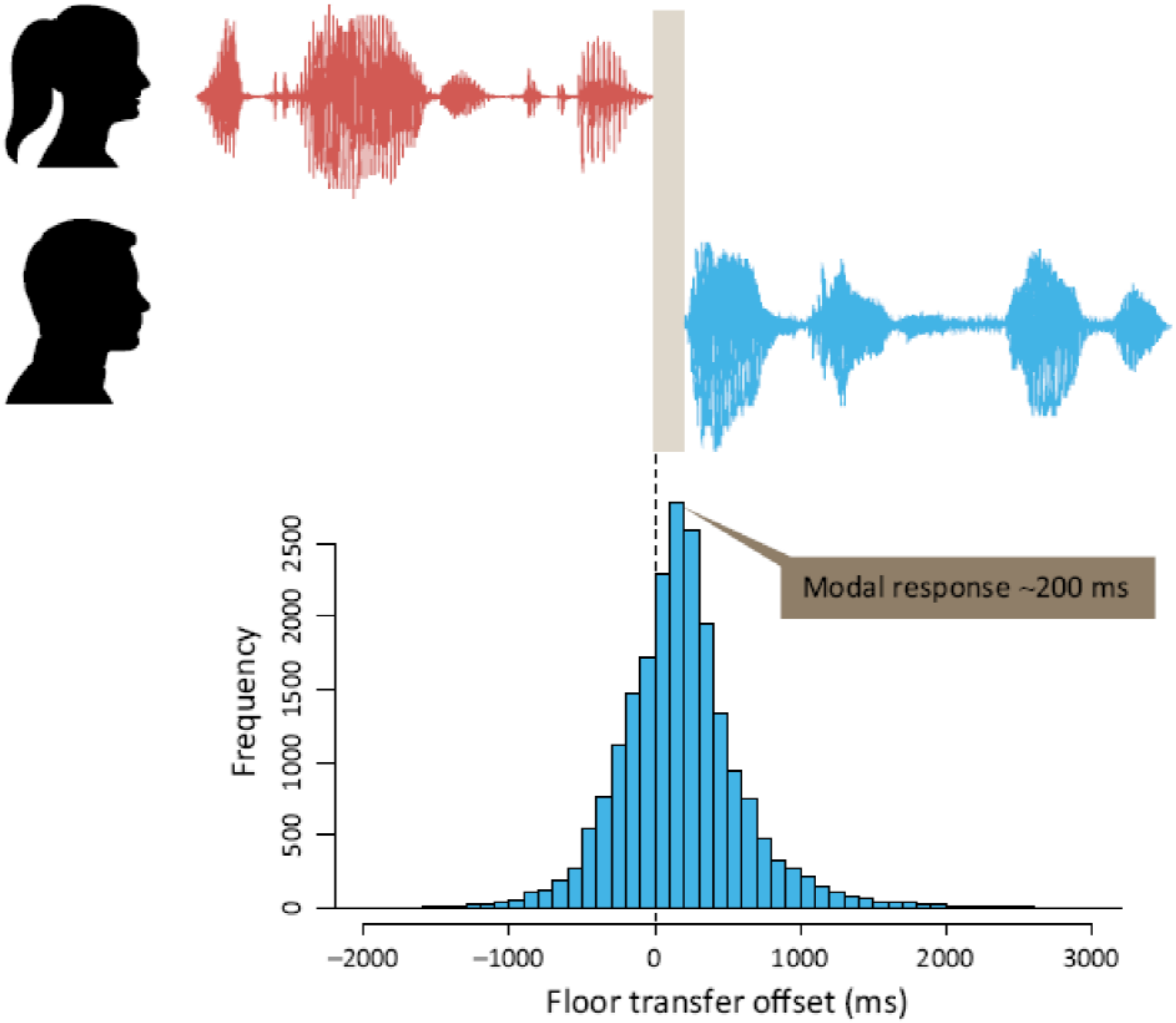


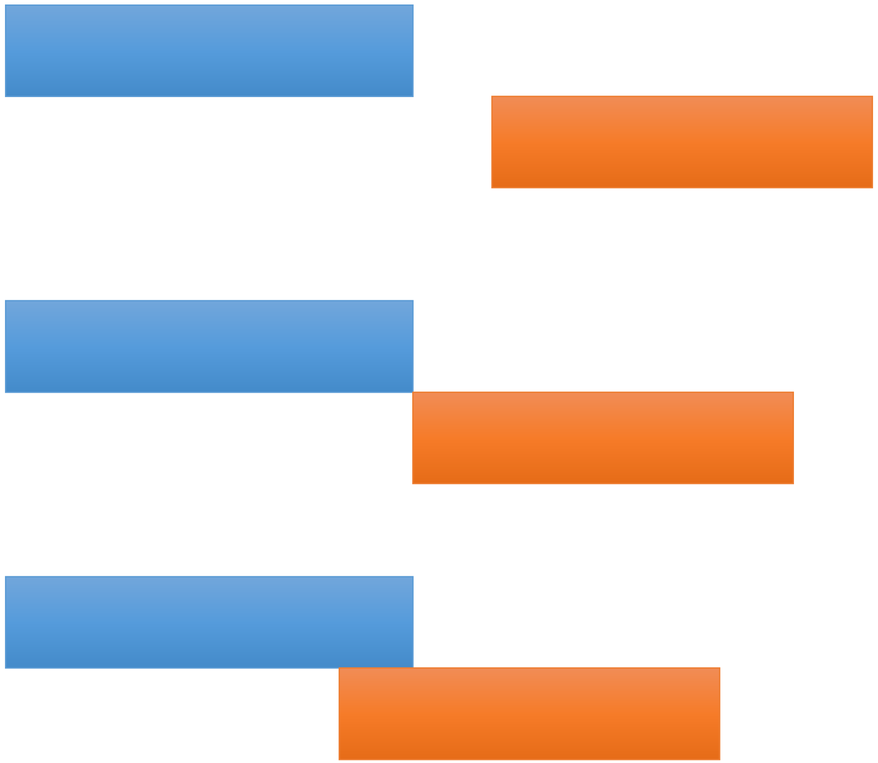
Irfan, B., Kuoppamäki, S.-M., & Skantze, G. (2023). *Between Reality and Delusion: Challenges of Applying Large Language Models to Companion Robots for Open-Domain Dialogues with Older Adults*. https://doi.org/10.21203/rs.3.rs-2884789/v1

Terminology



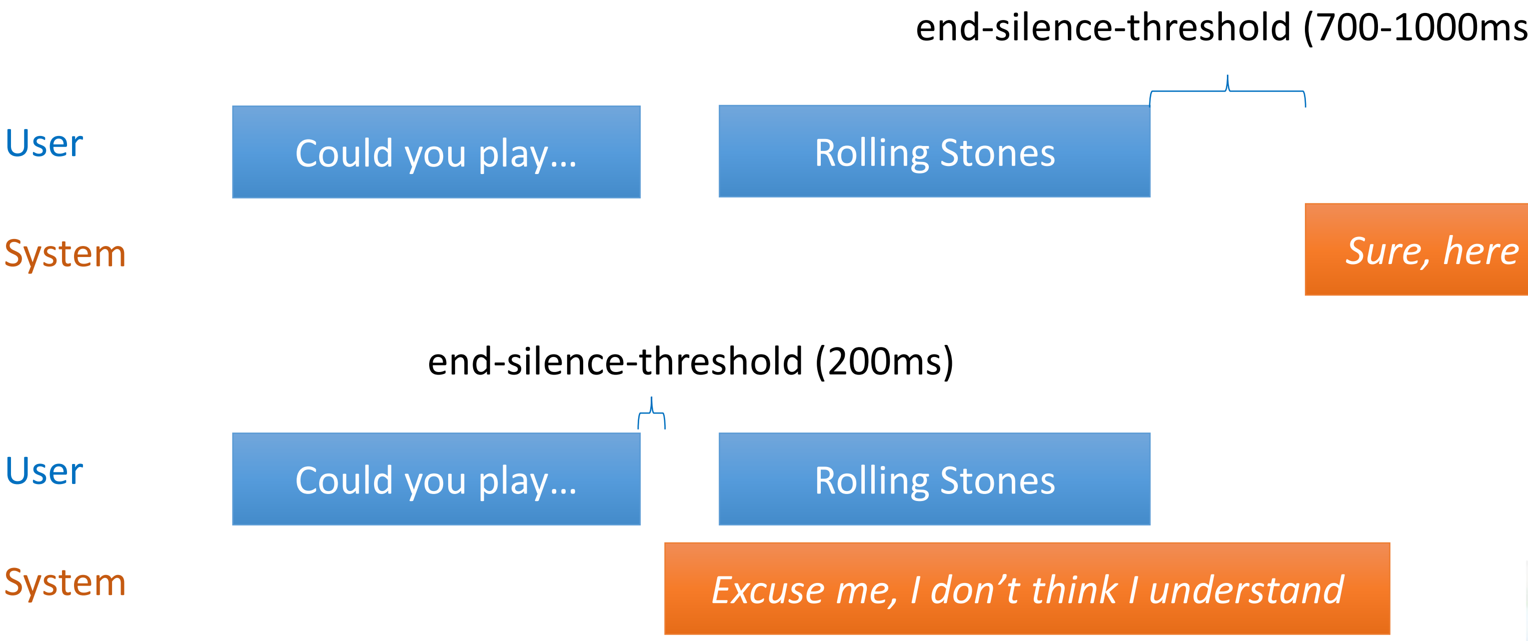
Coordination of turn-taking in spoken interaction





Levinson, S. C. (2016). Turn-taking in Human Communication—Origins and Implications for Language Processing. *Trends in Cognitive Sciences*.

Turn-taking in current systems



end-silence-threshold (700-1000ms)

User

Could you play…

Rolling Stones

System

Sure, here you go

end-silence-threshold (200ms)

User

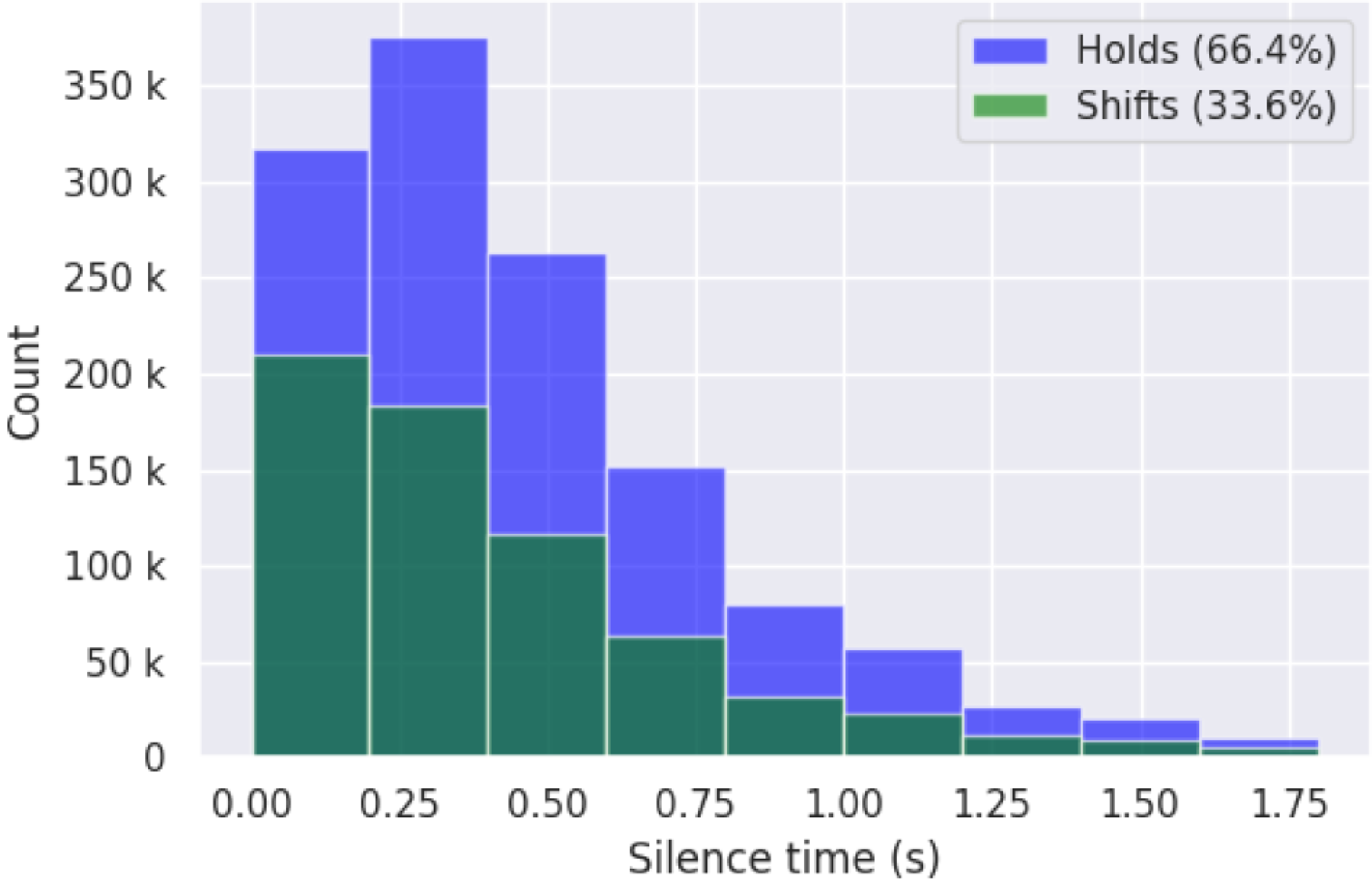
Could you play…

Rolling Stones

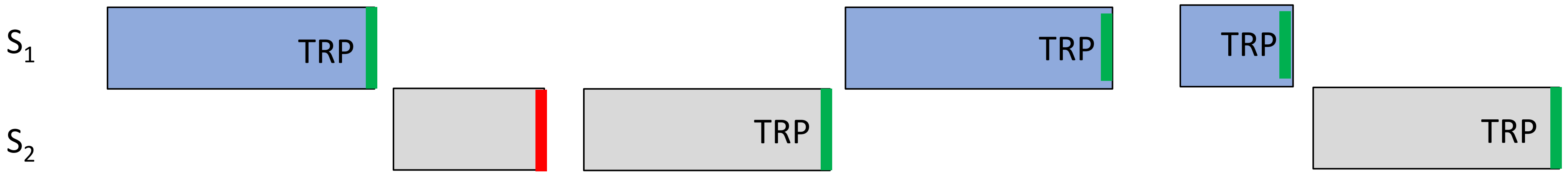
System

Excuse me, I don’t think I understand

Silence is a bad indicator of turn-taking



Coordination signals across modalities



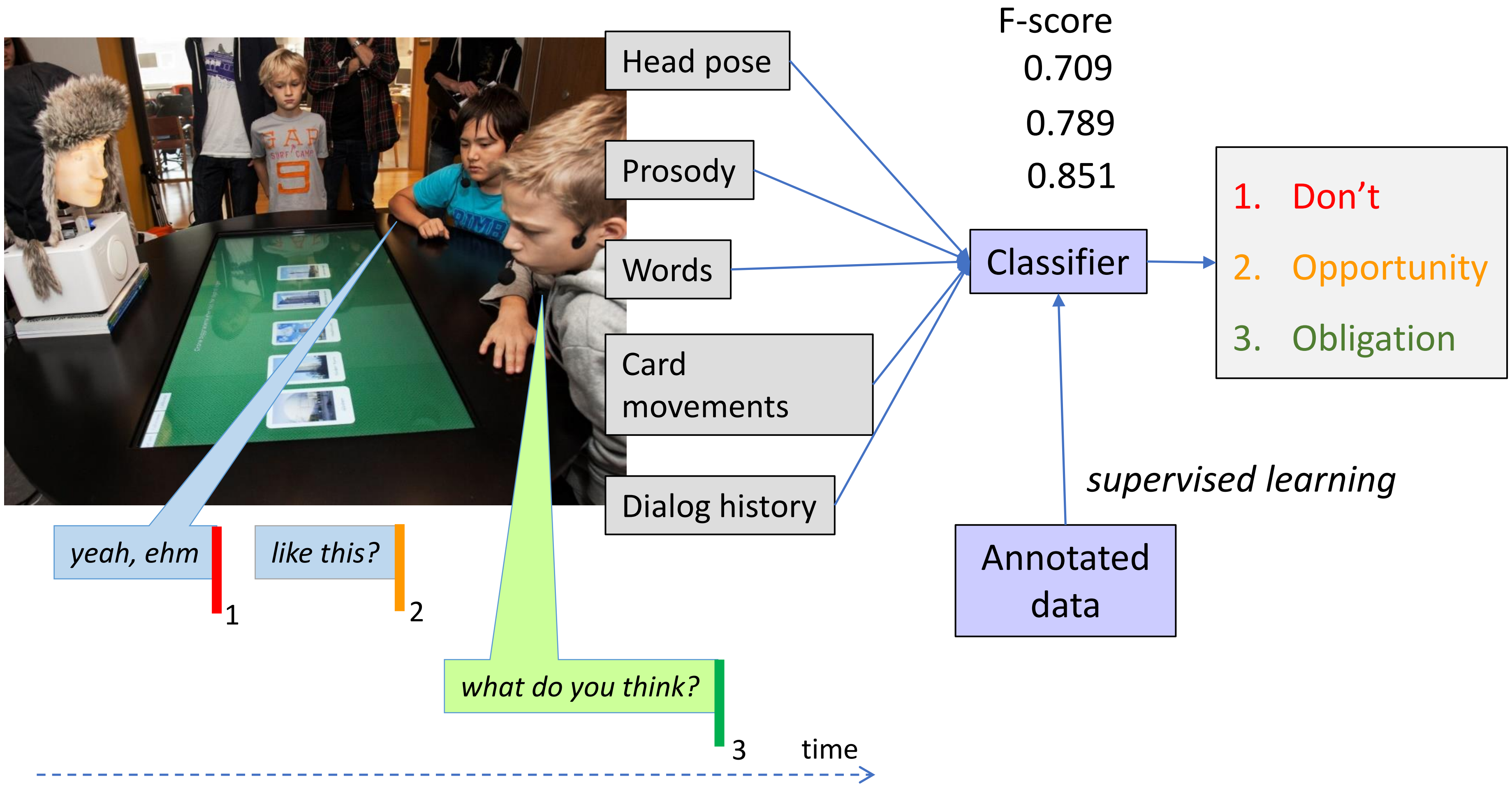
	Turn-yielding cue	Turn-holding cue
Verbal/Syntax	Complete	Incomplete, Fillers
Prosody - Pitch	Rising or Falling	Flat
Prosody - Intensity	Lower	Higher
Prosody - Duration	Shorter	Longer
Breathing	Breathe out	Breathe in
Gaze	Looking at addressee	Looking away
Gesture	Terminated	Non-terminated

TRP = Transition Relevance-Place

(Sacks et al, 1974)

The more cues, the stronger the signal! *(Duncan, 1972)*

As simple classifier for identifying turn-taking cues



Head pose

Prosody

Words

Card movements

Dialog history

F-score

0.709

0.789

0.851

Classifier

1. Don't
2. Opportunity
3. Obligation

supervised learning

Annotated data

yeah, ehm 1

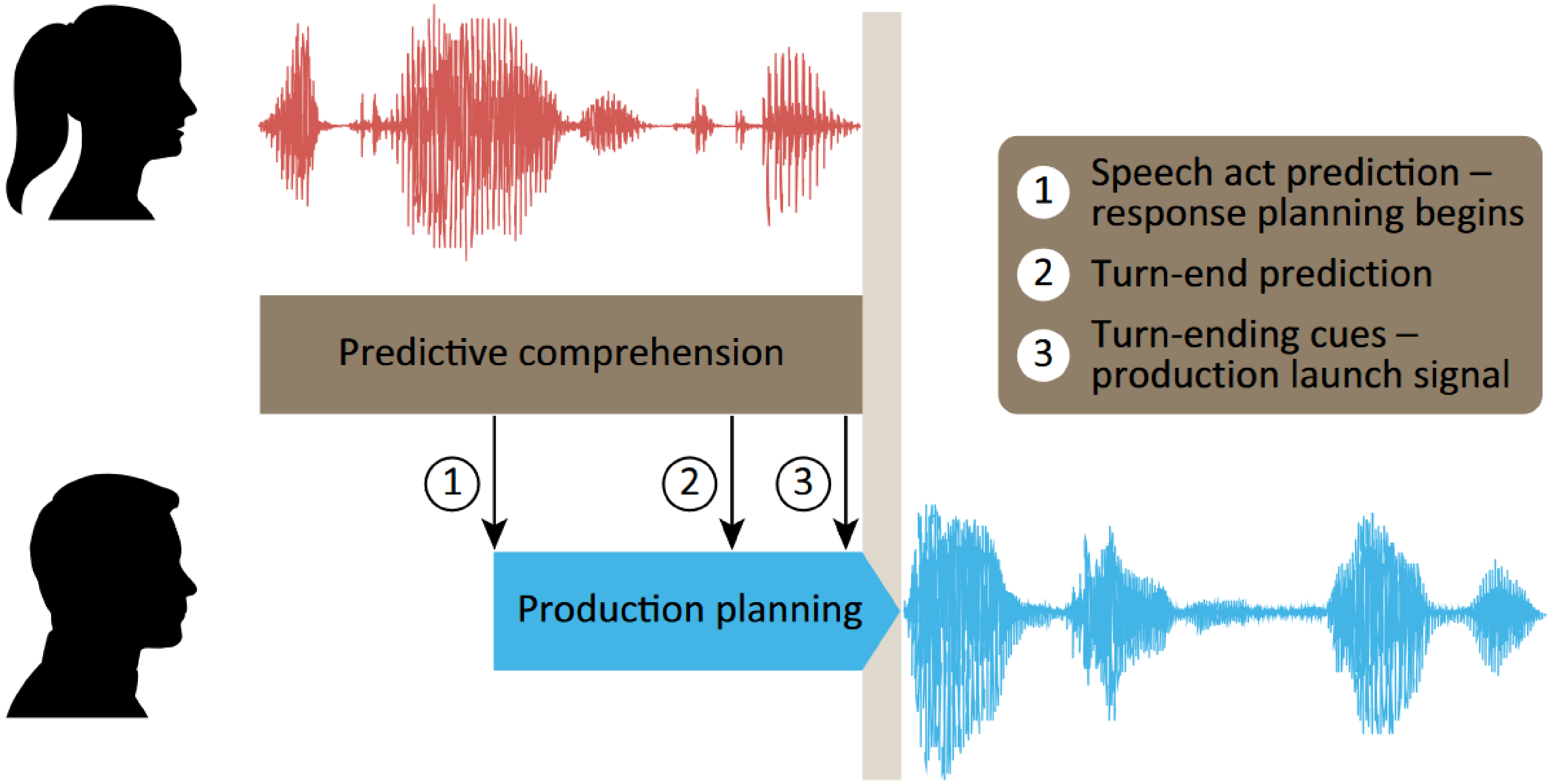
like this? 2

what do you think? 3

time

Johansson, M., & Skantze, G. (2015). Opportunities and Obligations to take turns in collaborative multi-party human-robot interaction. *Proceedings of SIGDIAL*

From reaction to prediction



Predictive comprehension

Production planning

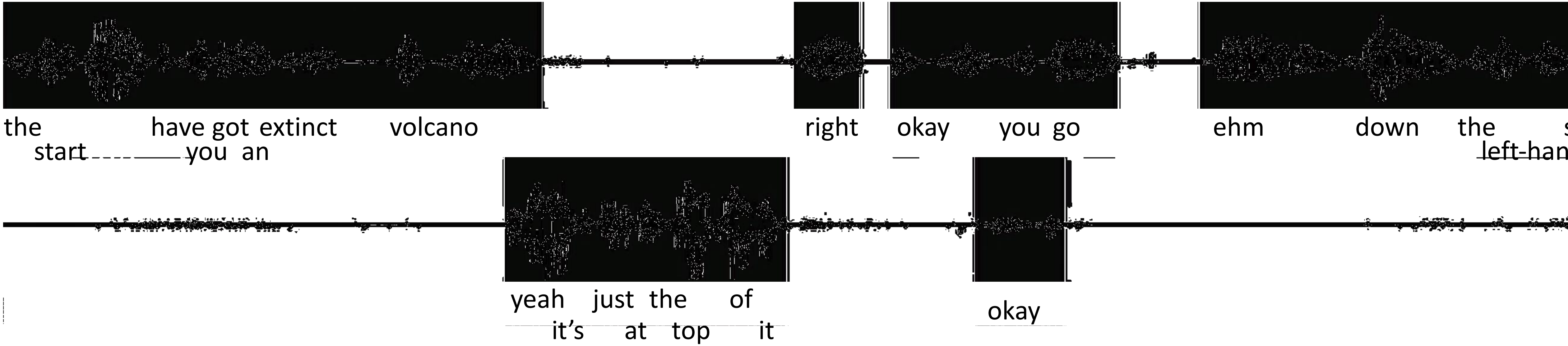
1 Speech act prediction – response planning begins

2 Turn-end prediction

3 Turn-ending cues – production launch signal

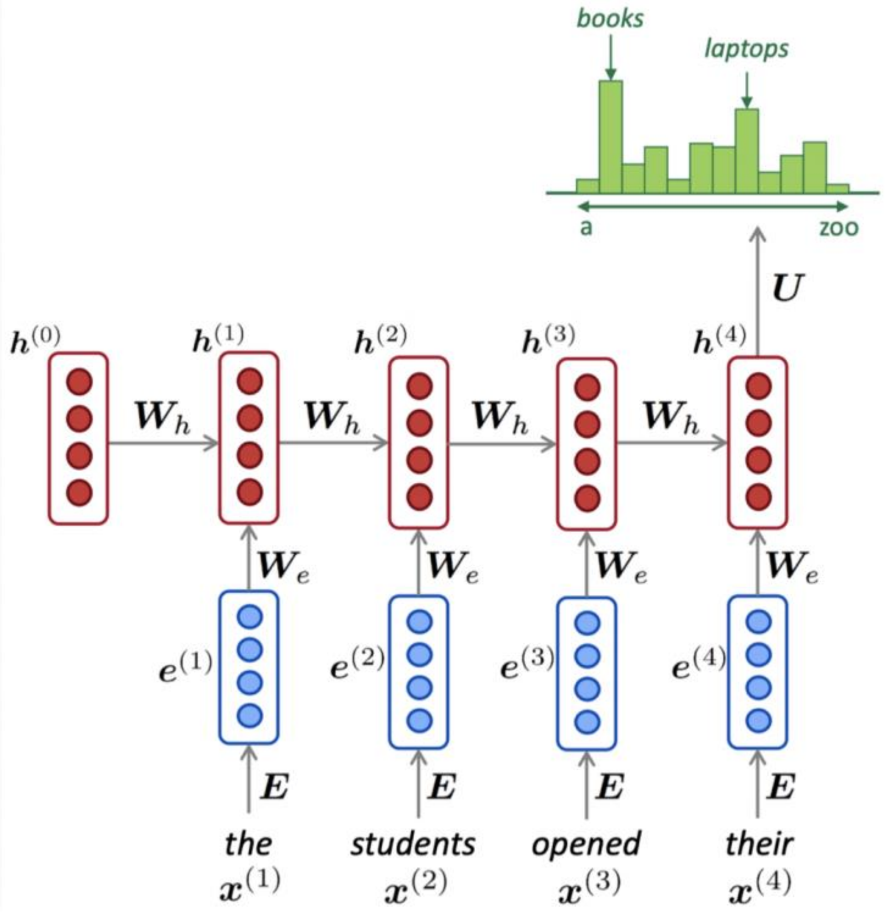
Levinson, S. C. (2016). Turn-taking in Human Communication—Origins and Implications for Language Processing. *Trends in Cognitive Sciences*.

Evidence of prediction



How can we predict the future in speech?

- Written language is made up of a sequence of discrete tokens from a fixed vocabulary
- Prediction of the next token can be formulated as a probability distribution over this vocabulary



TurnGPT: Modelling turn-taking with an LLM

- Turn completion is judged incrementally as the utterance unfolds:

What would you like <TC> to order? <TC>

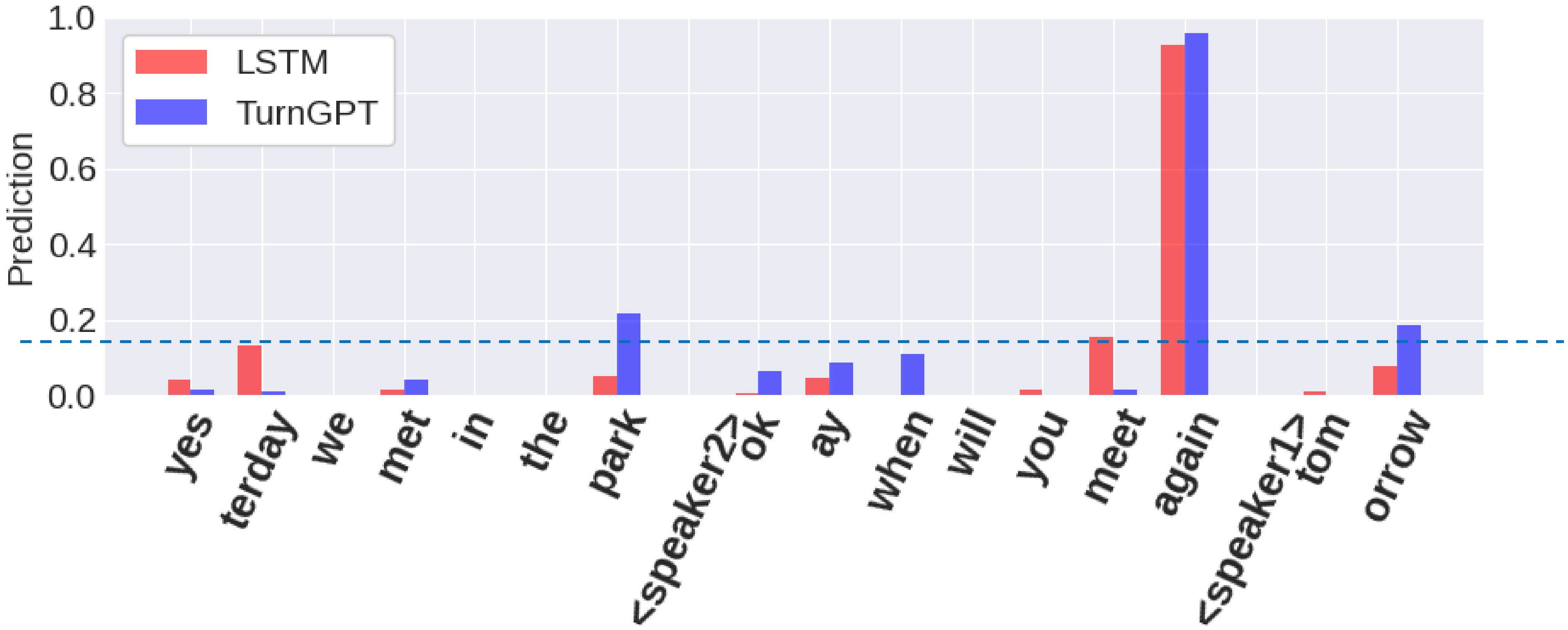
I would like a hamburger <TC> with fries <TC> and a milkshake <TC>

- Context dependence:

yesterday we met <TC> in the park <TC>
okay <TC> when <TC> will you meet again <TC>
tomorrow <TC>

Ekstedt, E. & Skantze, G. (2020). TurnGPT: a Transformer-based Language Model for Predicting Turn-taking in Spoken Dialog. *EMNLP 2020*.

TurnGPT: Probability of turn shifts

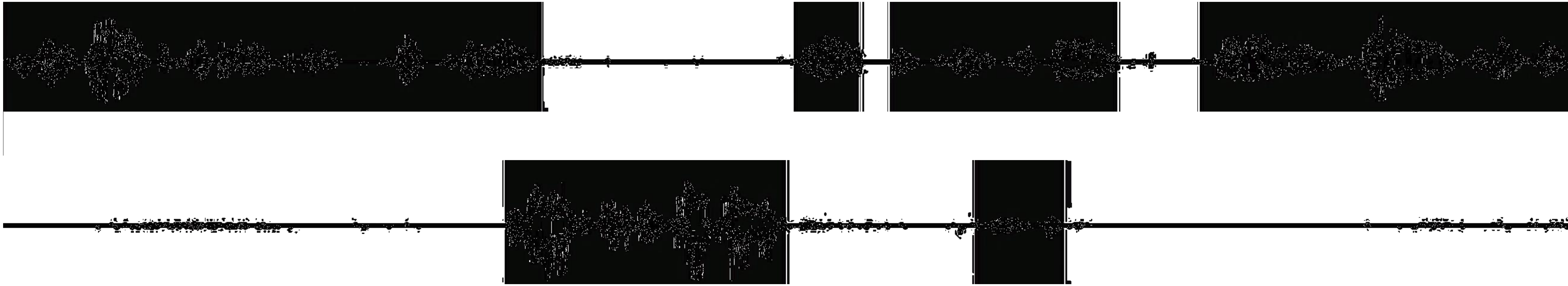


Ekstedt, E. & Skantze, G. (2020). TurnGPT: a Transformer-based Language Model for Predicting Turn-taking in Spoken Dialog. *EMNLP 2020*.

How can we predict the future in speech?

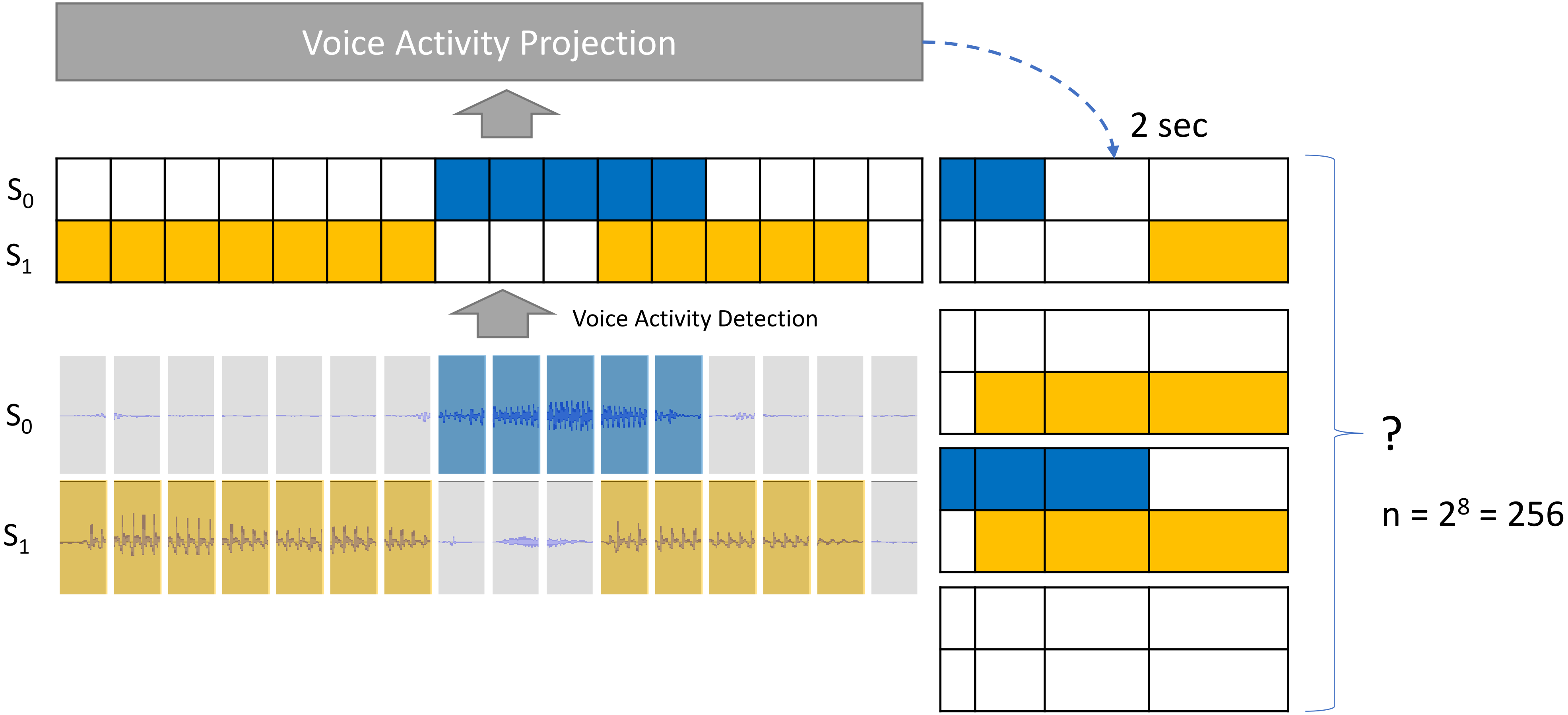
Speech is made up of a continuous sound wave.

What is a “turn”, really?

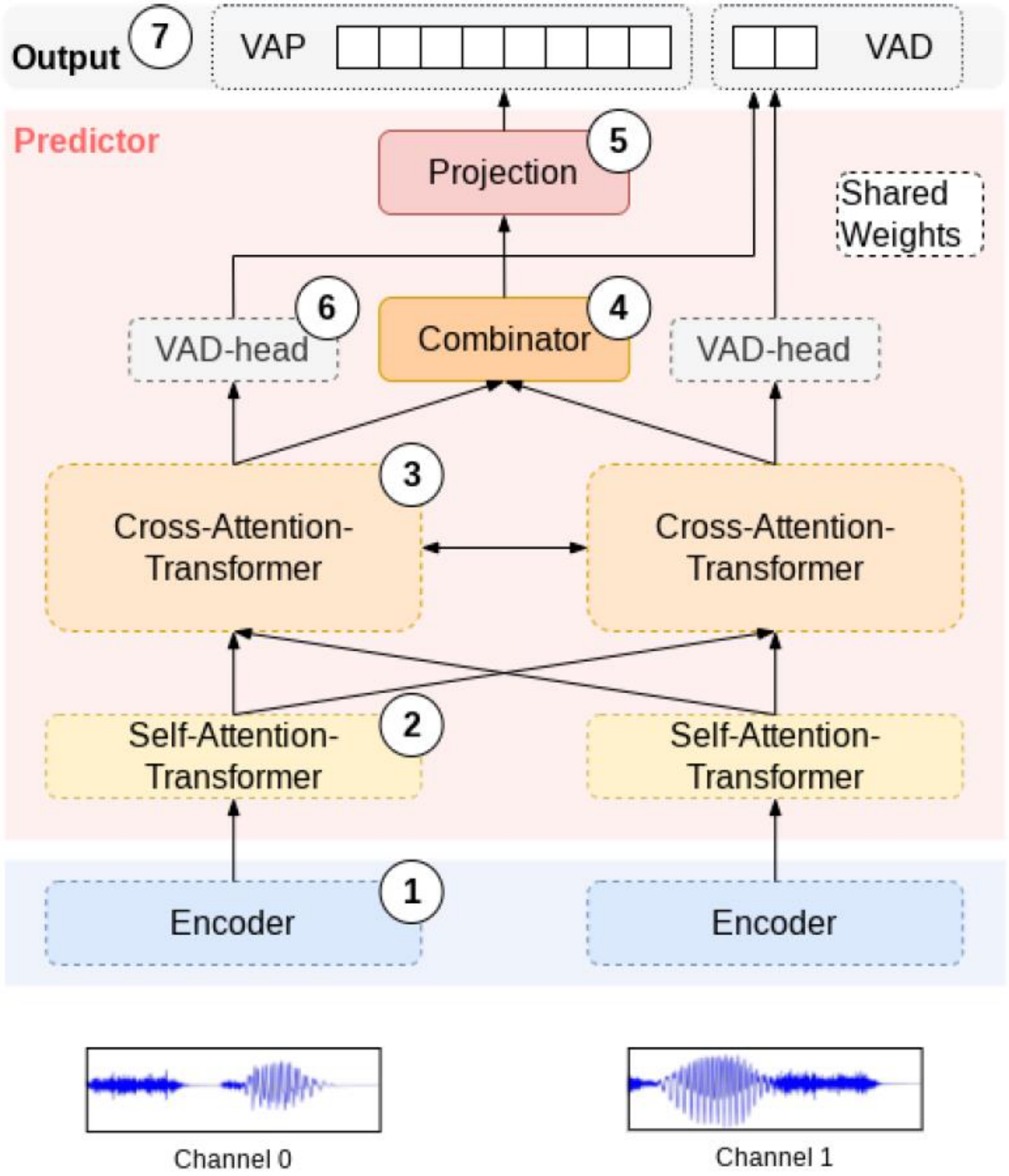


Not just words, but also prosody, timing, etc.

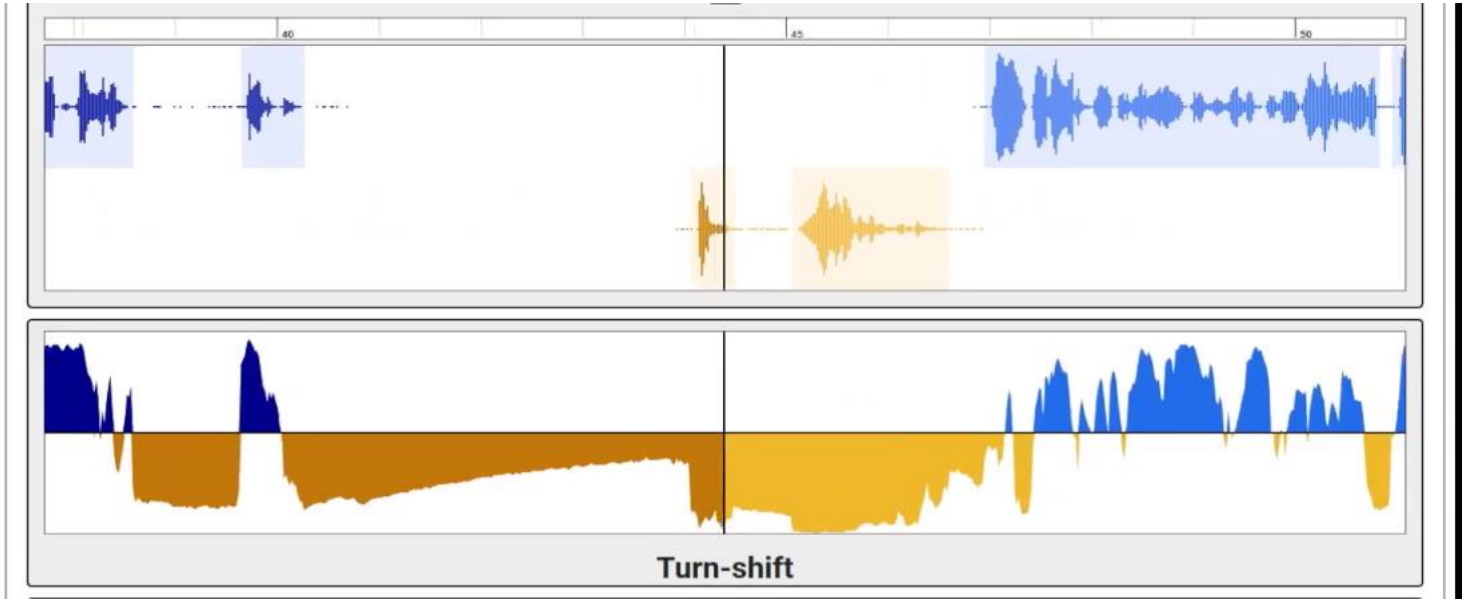
Voice Activity Projection (VAP)



Ekstedt, E., & Skantze, G. (2022). Voice Activity Projection: Self-supervised Learning of Turn-taking Events. *Interspeech 2022*


Output
7
VAP
VAD
Predictor
Projection
5
Shared Weights
6
VAD-head
Combinator
4
VAD-head
3
Cross-Attention-Transformer
Cross-Attention-Transformer
2
Self-Attention-Transformer
Self-Attention-Transformer
1
Encoder
Encoder
Channel 0
Channel 1

VAP: A turn-taking model predicting the next 2 sec of a conversation

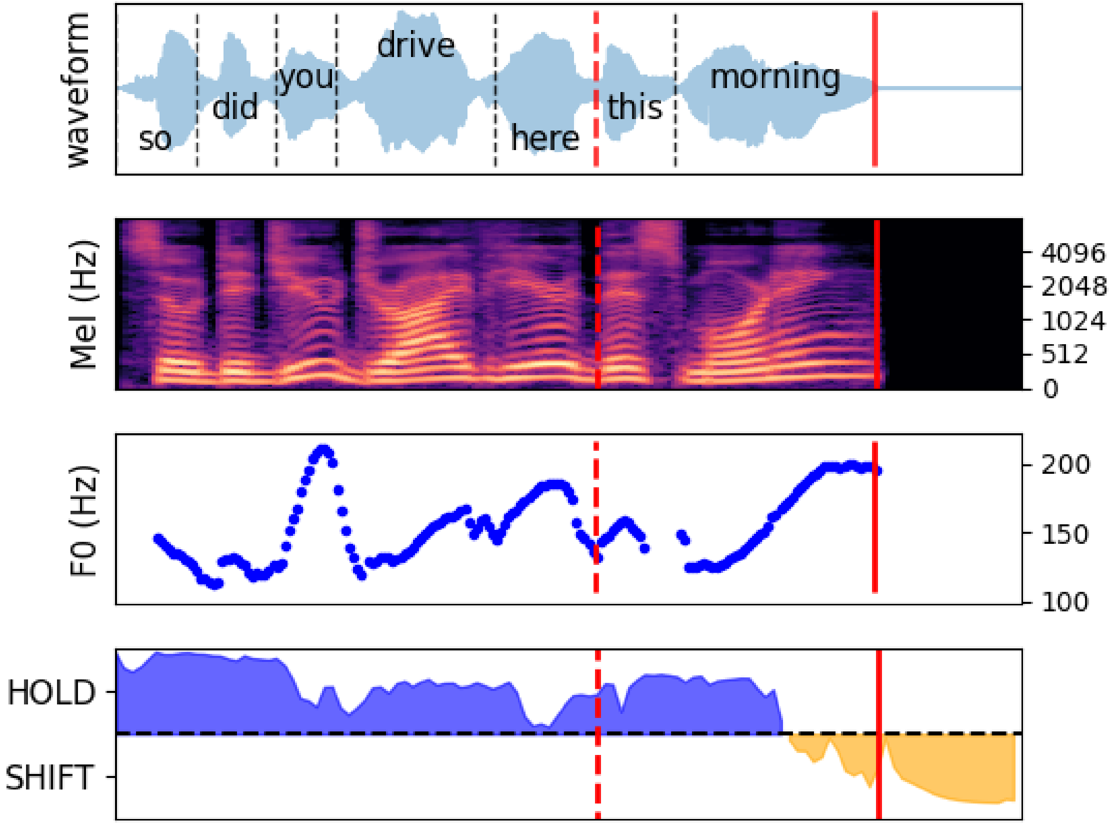


Advantages of Voice Activity Projection

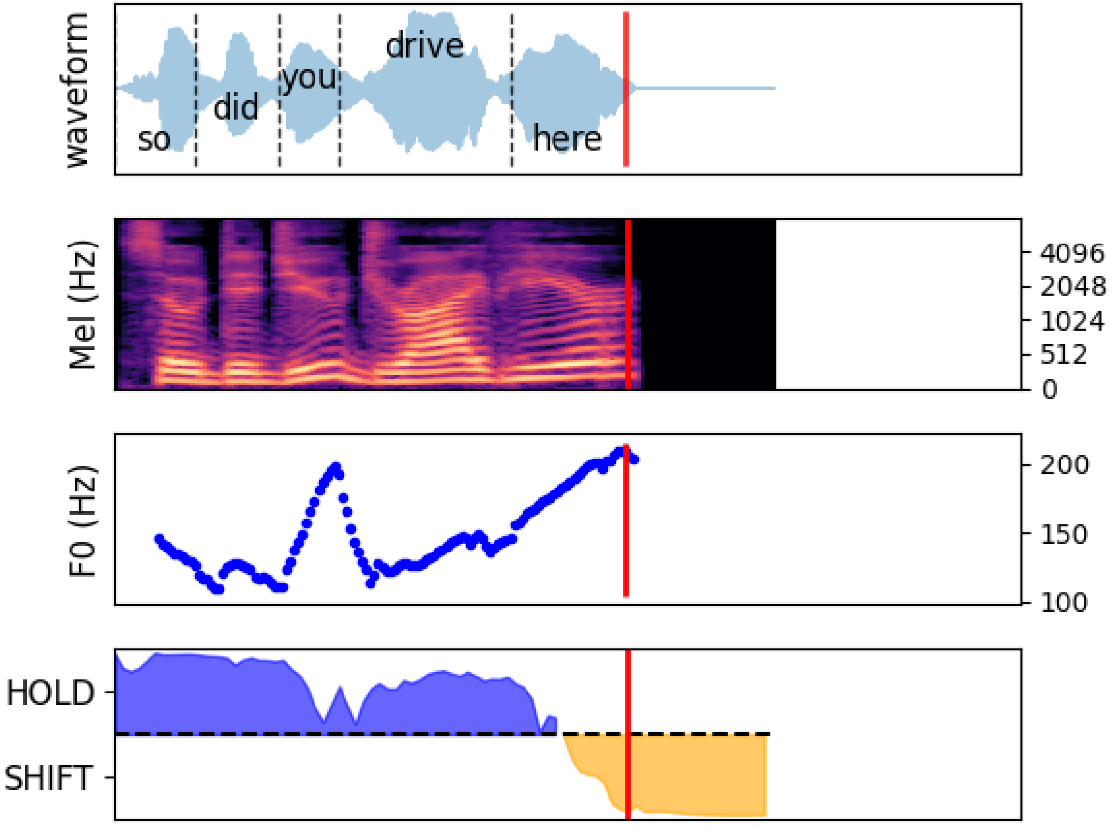
- Operates on raw audio
  - Pre-trained Constrastive Predictive Coding (CPC)
- No need for speech recognition (words) or feature extraction (prosody)
  - No need to normalize features to the speaker
  - Continuous modelling, no delay
- Only lightly annotated data needed (binary voice activity detection)
  - Can be trained on large amounts of (diverse) data

- BUT: Black-box model (trained end-to-end).
- What has it actually learned?

So, did you drive here this morning?

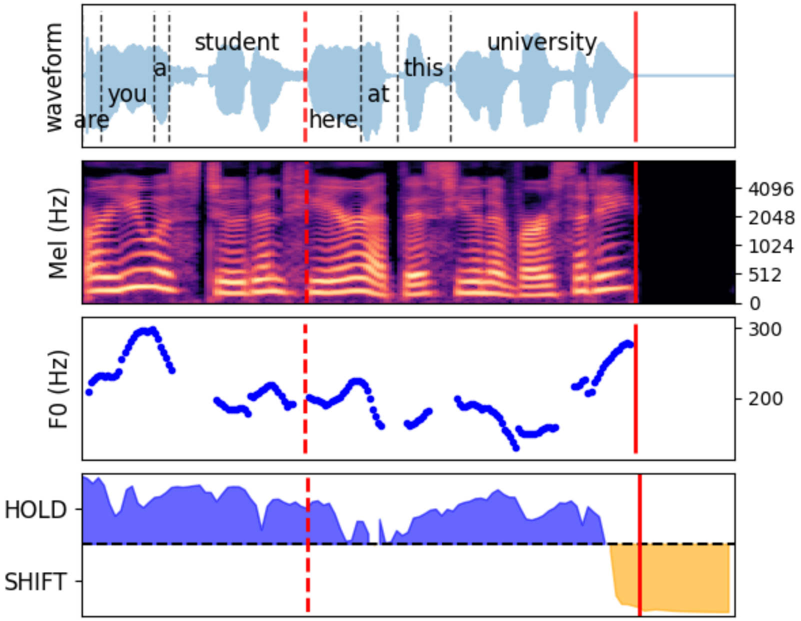


So, did you drive here?

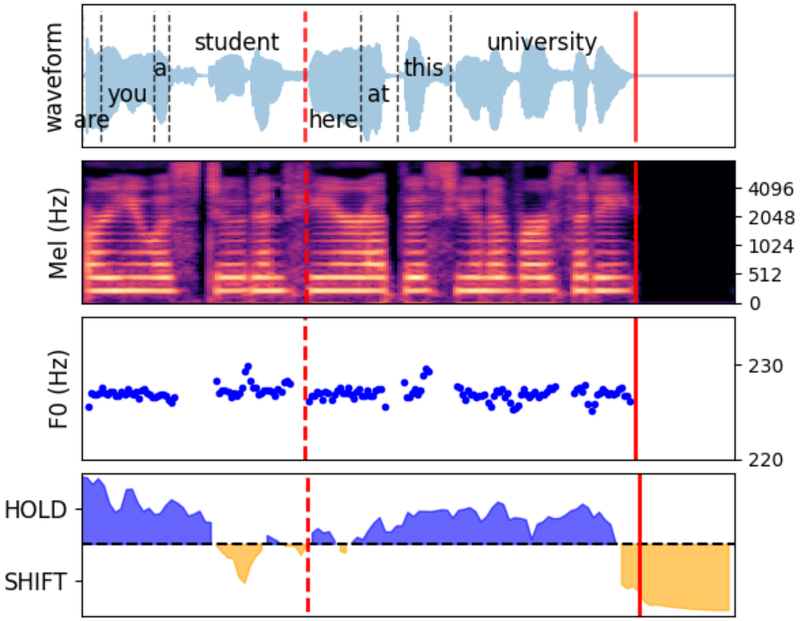


Ekstedt, E., & Skantze, G. (2022). How Much Does Prosody Help Turn-taking? Investigations using Voice Activity Projection Models. In *Proceedings of SIGDIAL*.

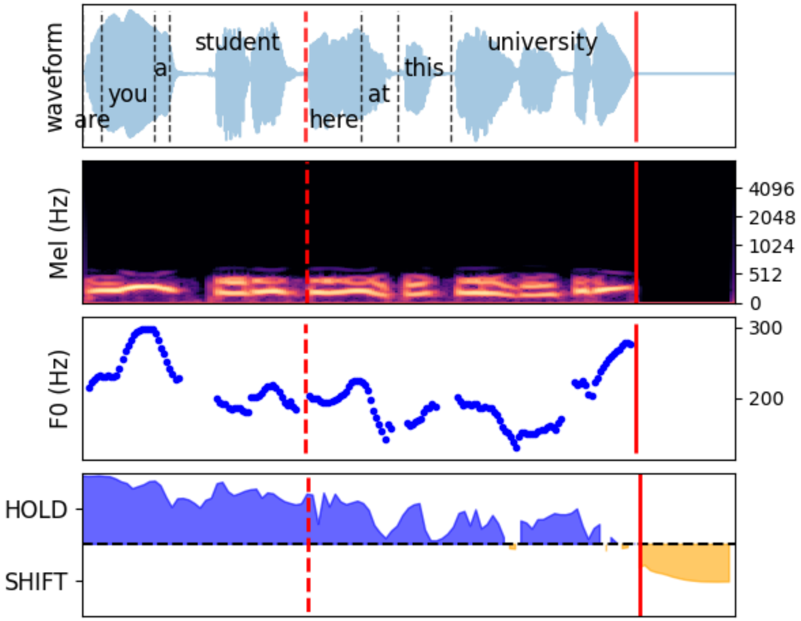
Original



Flat pitch



Low pass

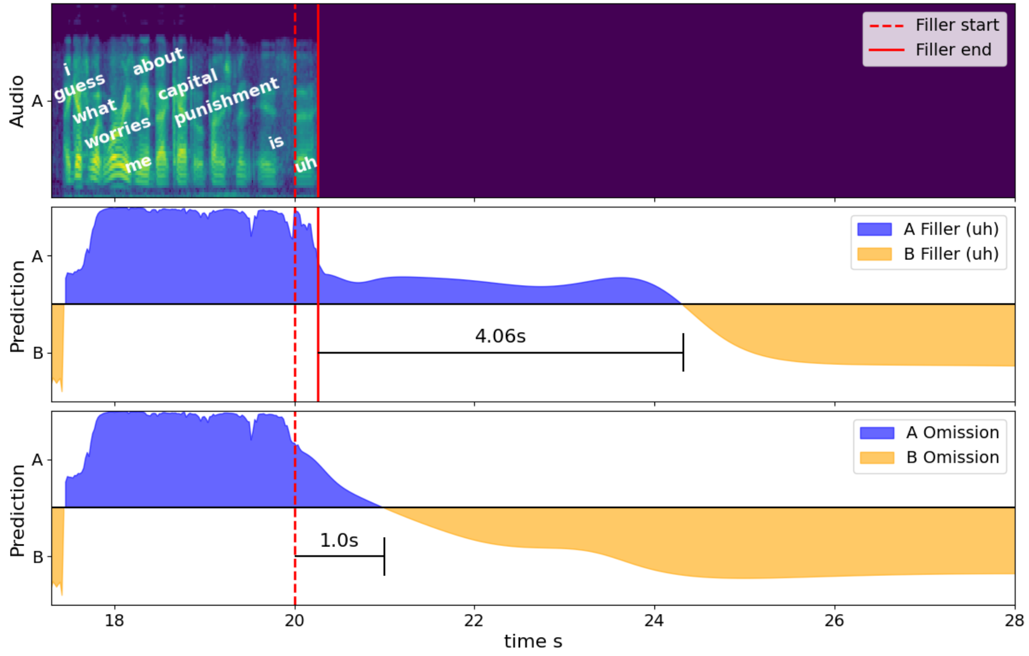


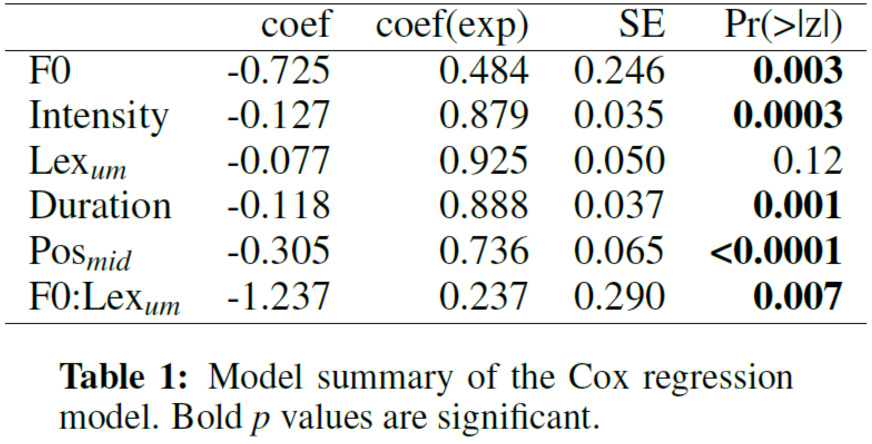
Ekstedt, E., & Skantze, G. (2022). How Much Does Prosody Help Turn-taking? Investigations using Voice Activity Projection Models. In *Proceedings of SIGDIAL*.

How much time does a filler “buy you”?

with filler

w/o filler





	coef	coef(exp)	SE	Pr(>|z|)
F0	-0.725	0.484	0.246	**0.003**
Intensity	-0.127	0.879	0.035	**0.0003**
Lex$_{um}$	-0.077	0.925	0.050	0.12
Duration	-0.118	0.888	0.037	**0.001**
Pos$_{mid}$	-0.305	0.736	0.065	**<0.0001**
F0:Lex$_{um}$	-1.237	0.237	0.290	**0.007**

Table 1: Model summary of the Cox regression model. Bold p values are significant.

What makes a good pause? Investigating the turn-holding effects of fillers. B. Jiang, E. Ekstedt & G. Skantze.
Proceedings of the 20th International Congress of Phonetic Sciences

Synthesizing turn-taking cues

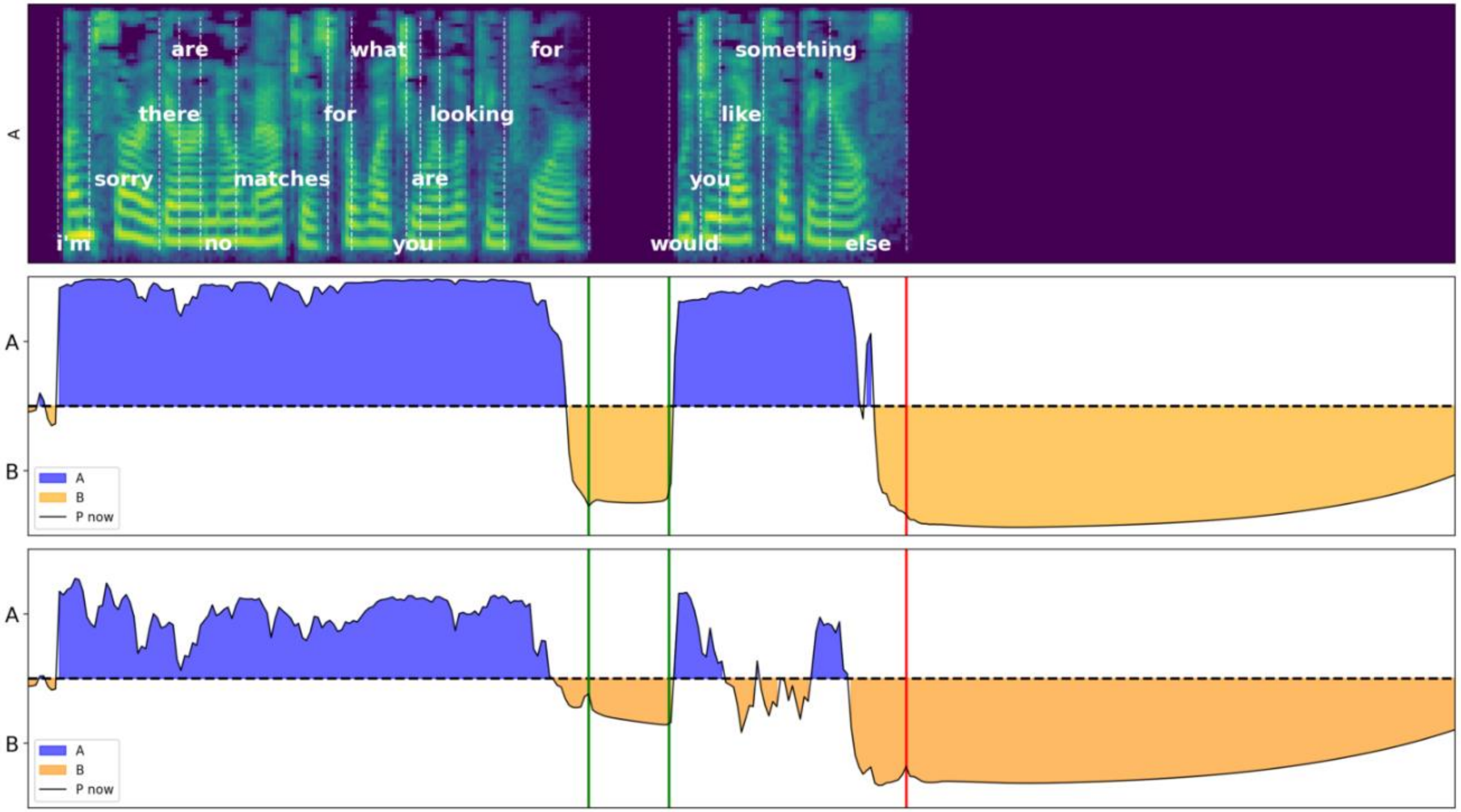
User
Do you have any ABBA compilation?

System
Yes, I have ABBA gold
Do you want me to play it for you?

User
Do you have any ABBA compilation?
Could you-

System
Yes, I have ABBA gold
Do you -

Towards a turn-taking aware TTS



now

future

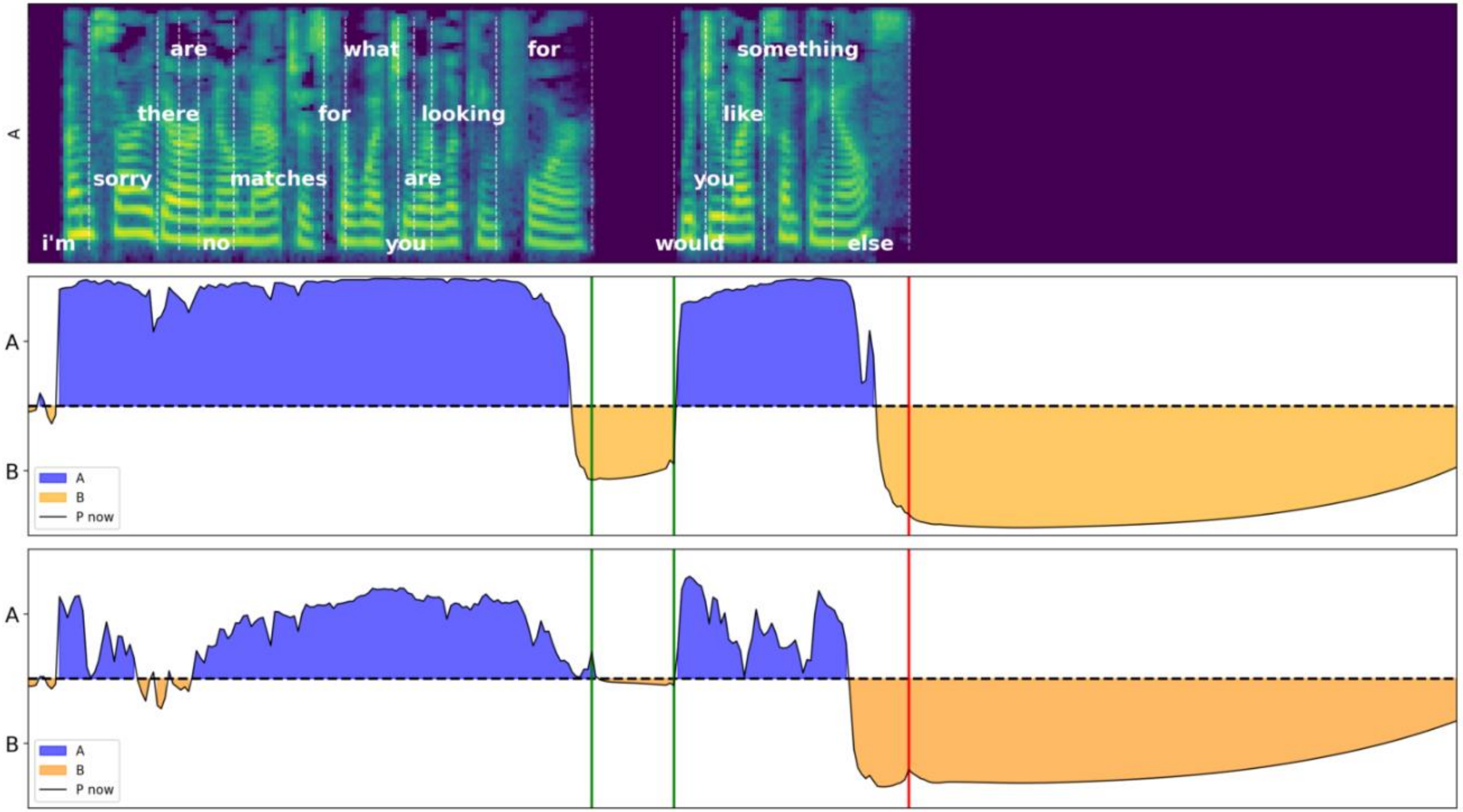
Ekstedt, E., Wang, S., Székely, É., Gustafson, J., & Skantze, G. (2023). Automatic Evaluation of Turn-taking Cues in Conversational Speech Synthesis. *INTERSPEECH 2023*, 5481–5485.

Towards a turn-taking aware TTS

Using a comma

now

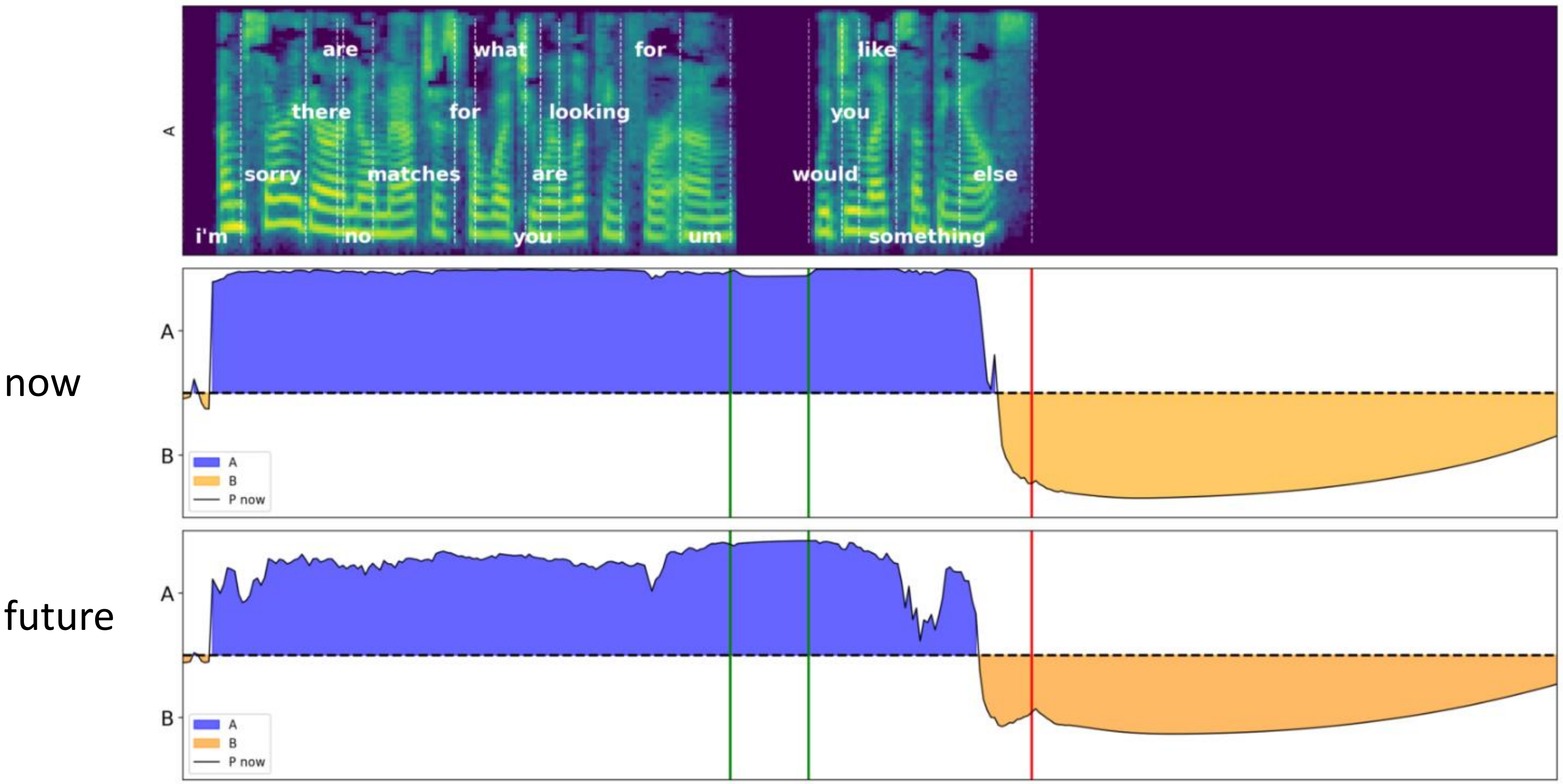
future



Ekstedt, E., Wang, S., Székely, É., Gustafson, J., & Skantze, G. (2023). Automatic Evaluation of Turn-taking Cues in Conversational Speech Synthesis. *INTERSPEECH 2023*, 5481–5485.

Towards a turn-taking aware TTS

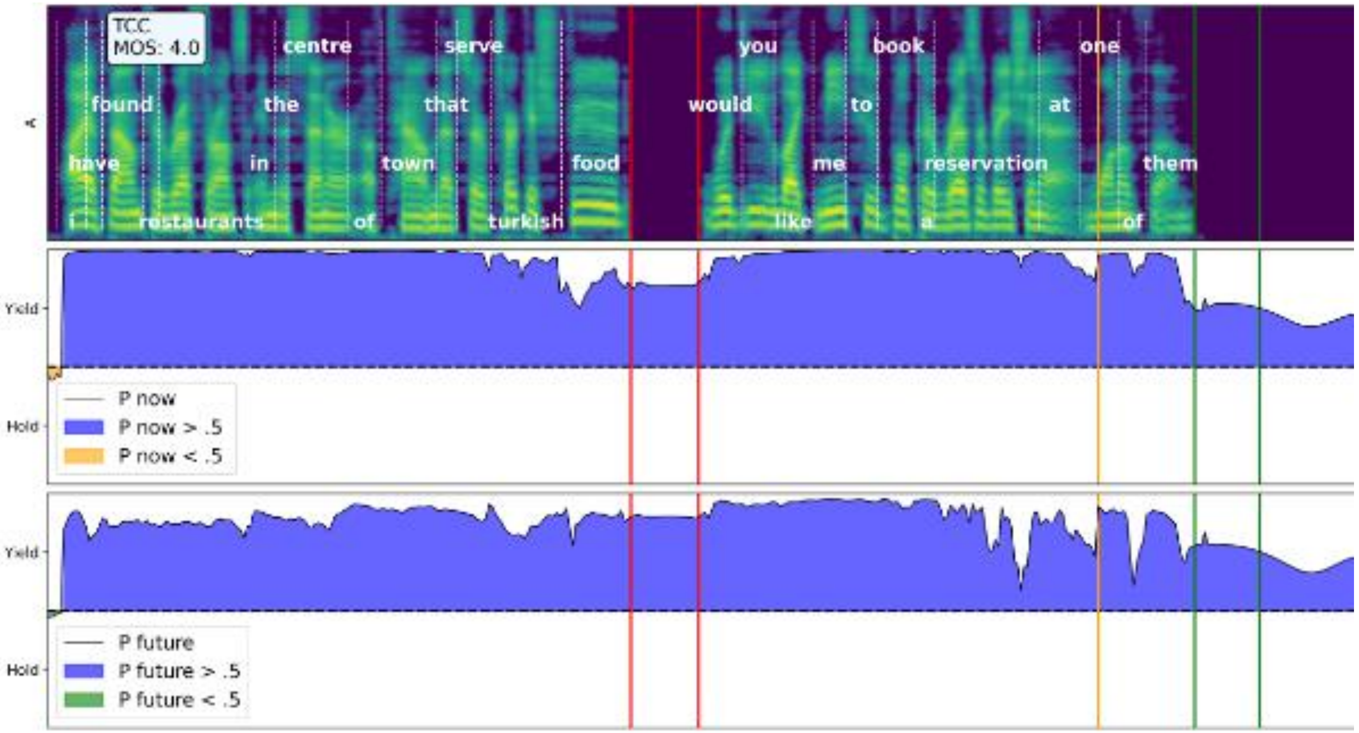
Inserting a filler



Ekstedt, E., Wang, S., Székely, É., Gustafson, J., & Skantze, G. (2023). Automatic Evaluation of Turn-taking Cues in Conversational Speech Synthesis. *INTERSPEECH 2023*, 5481–5485.

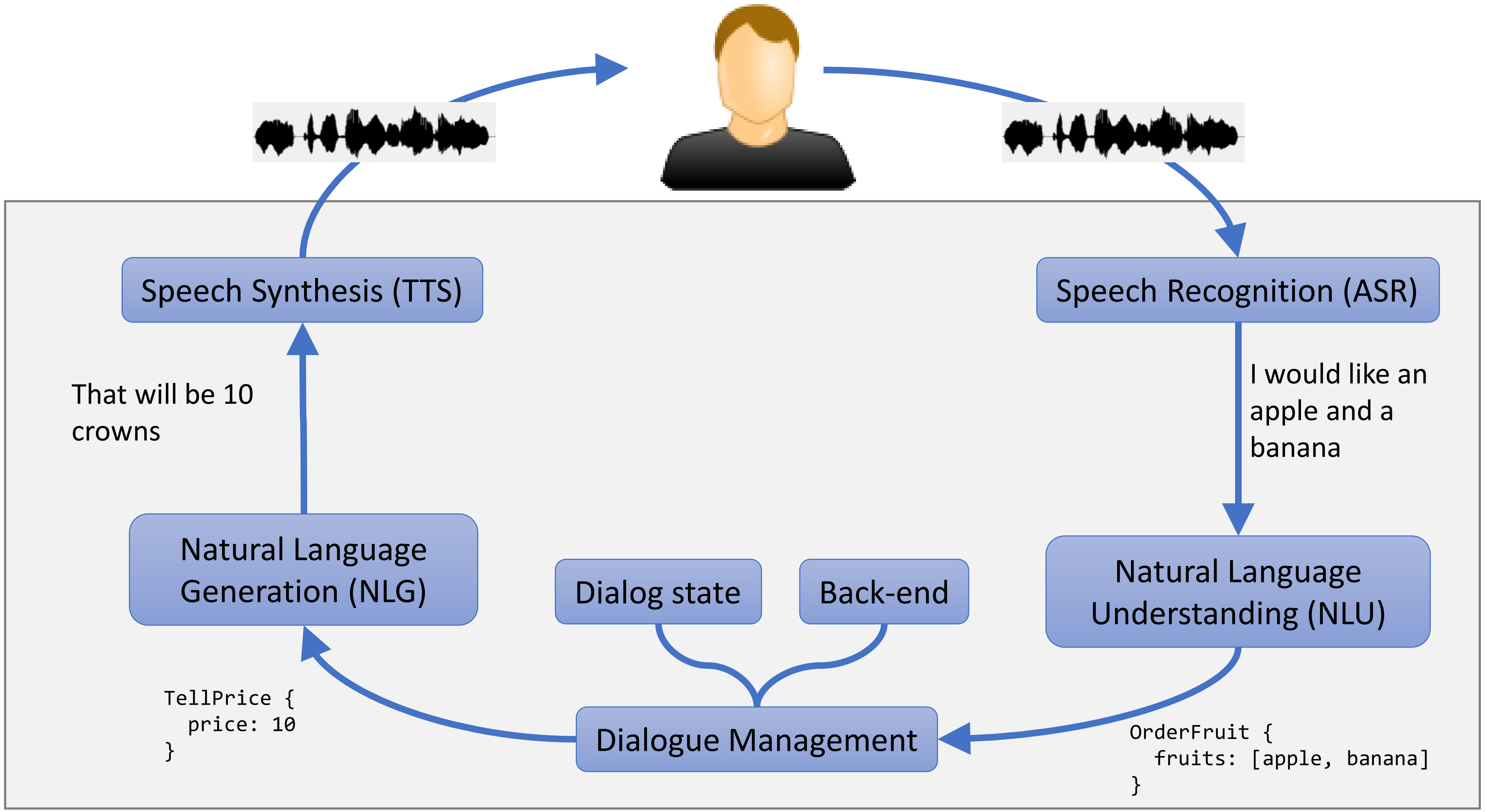
Towards a turn-taking aware TTS

A synthesizer that never yields…



now

future



Speech Synthesis (TTS)

Speech Recognition (ASR)

I would like an apple and a banana

That will be 10 crowns

Natural Language Generation (NLG)

Dialog state

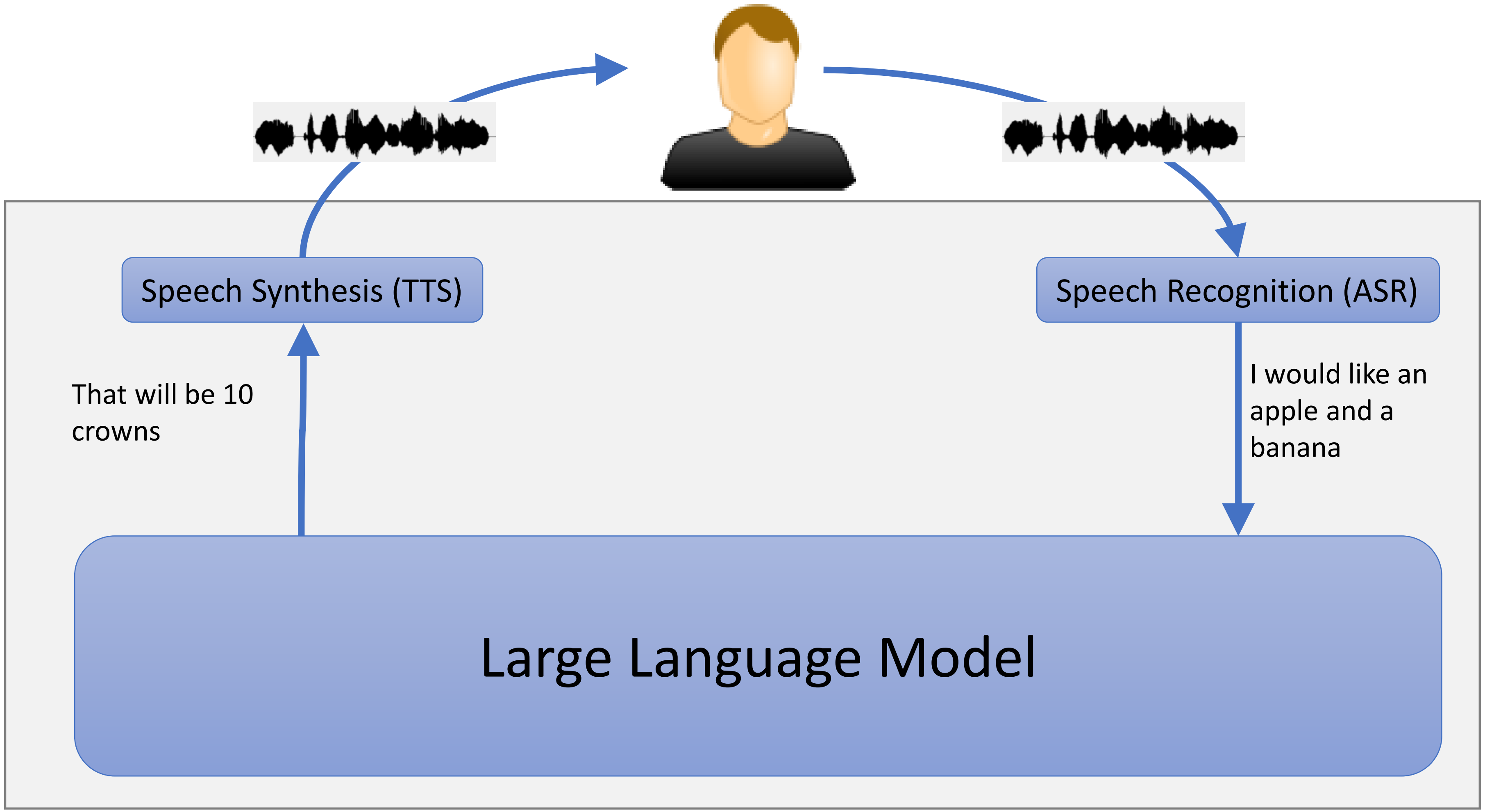
Back-end

Natural Language Understanding (NLU)

Dialogue Management

```
TellPrice {
  price: 10
}
```

```
OrderFruit {
  fruits: [apple, banana]
}
```



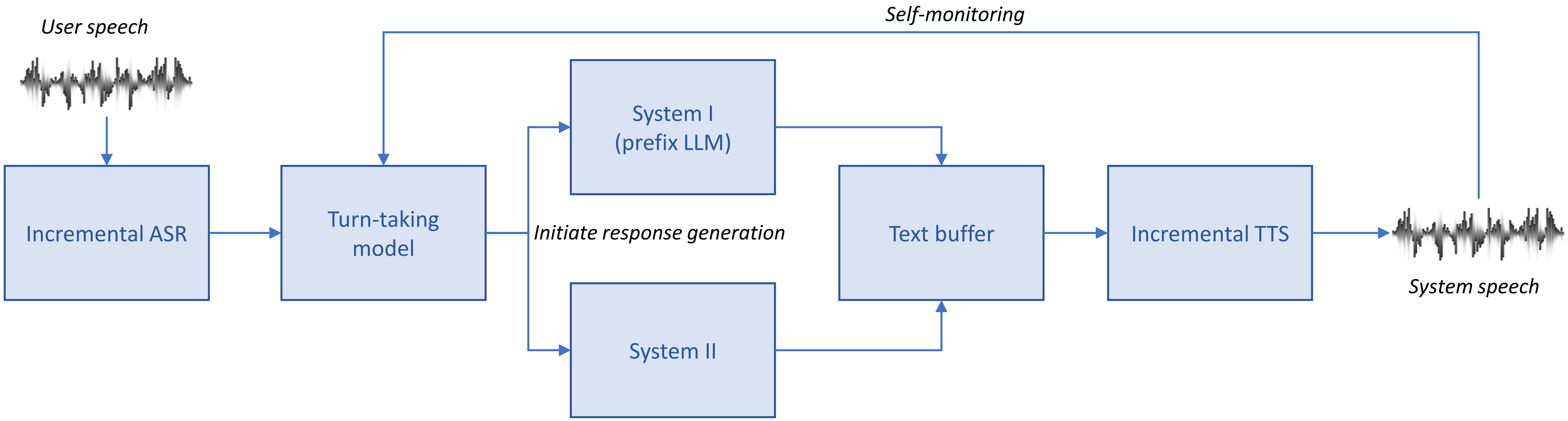
Speech Synthesis (TTS)

Speech Recognition (ASR)

That will be 10 crowns

I would like an apple and a banana

Large Language Model

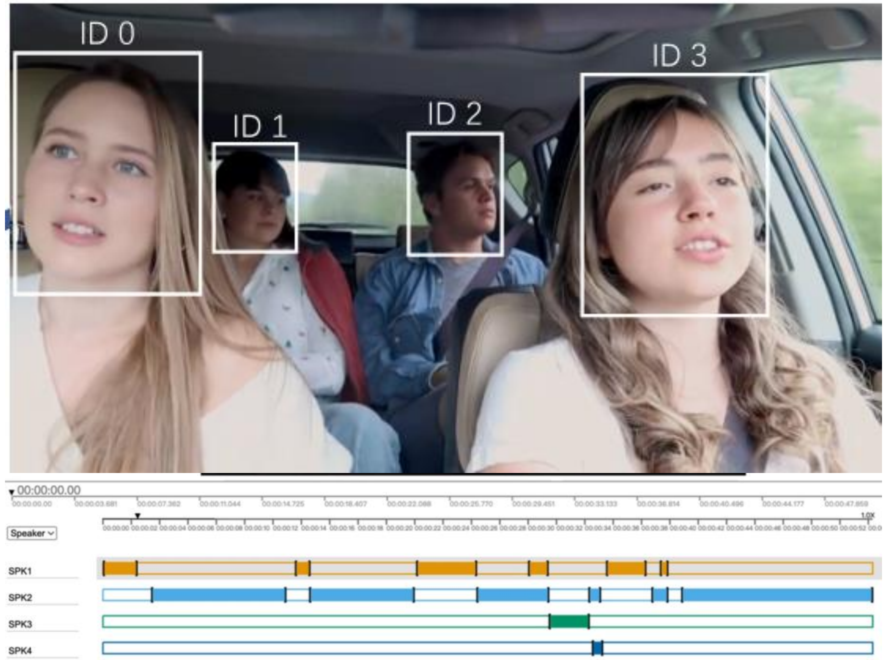

Self-monitoring
User speech
System I
(prefix LLM)
Incremental ASR
Turn-taking model
Initiate response generation
Text buffer
Incremental TTS
System speech
System II

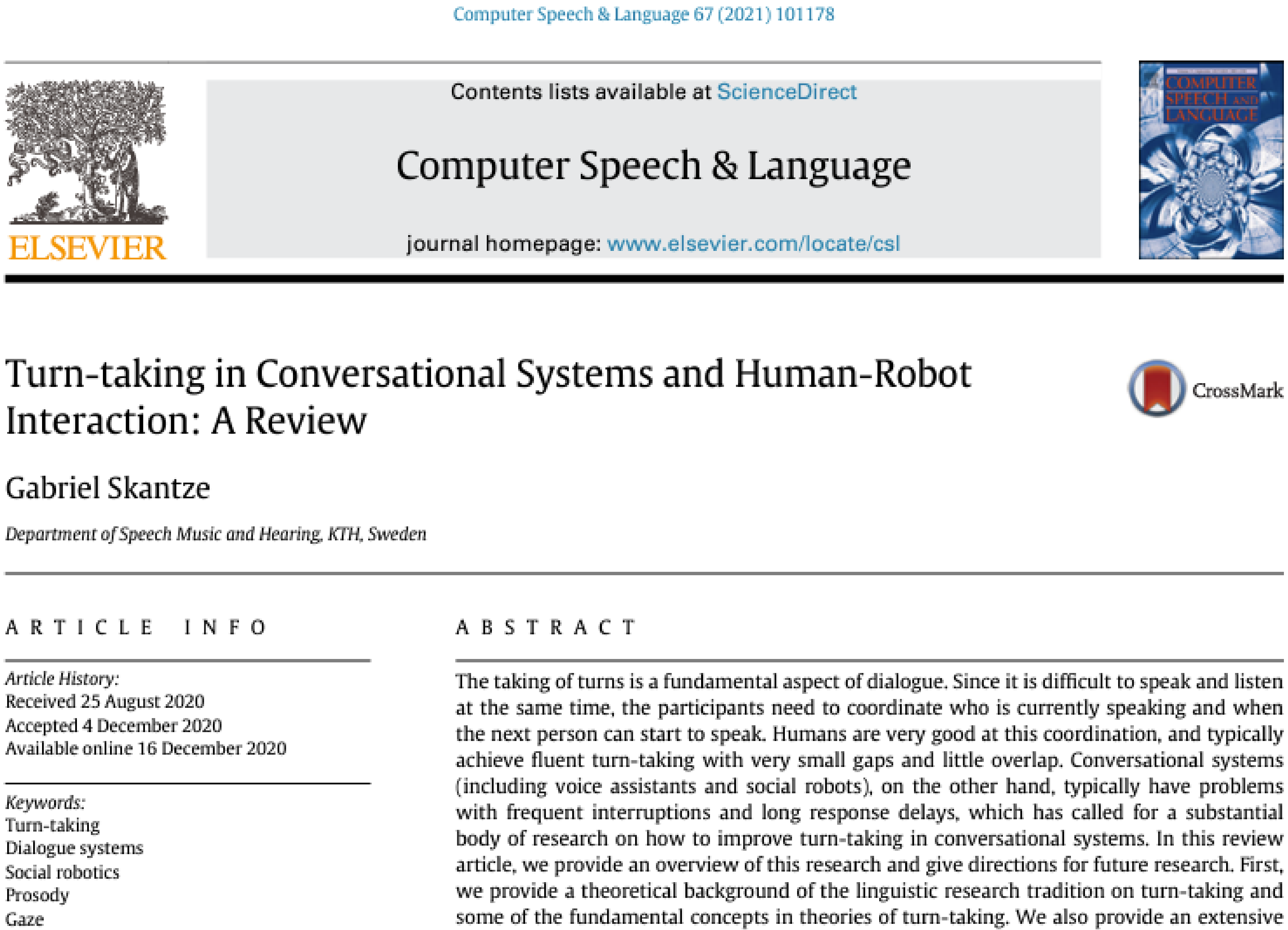
Problems with today’s systems

- System are not able to understand the user’s turn-taking cues
  - Turn-yielding vs Turn-holding pauses (”endpointing”)
  - Back-channel inviting cues
- Systems are purely reactive
  - Do not start planning responses in time
- Systems cannot start to speak before knowing what to say
- Systems cannot distinguish user interruptions from backchannels
- Systems are not aware of their own speech
  - Might accidentally yield or hold the turn in the wrong places

Current/Future work

- Implement VAP in Furhat!
  - How much can you compress the model?
- Comparing languages, multi-lingual models
- Multi-party, Multi-modal
- How can we combine audio and text?
- VAP-tuned TTS
- How can the models be used as a tool to gain insights into human-human dialogue?
  - Cues, Interaction styles, Diagnosis?





Computer Speech & Language 67 (2021) 101178

ELSEVIER

Contents lists available at ScienceDirect

Computer Speech & Language

journal homepage: www.elsevier.com/locate/csl

Turn-taking in Conversational Systems and Human-Robot Interaction: A Review

Gabriel Skantze

Department of Speech Music and Hearing, KTH, Sweden

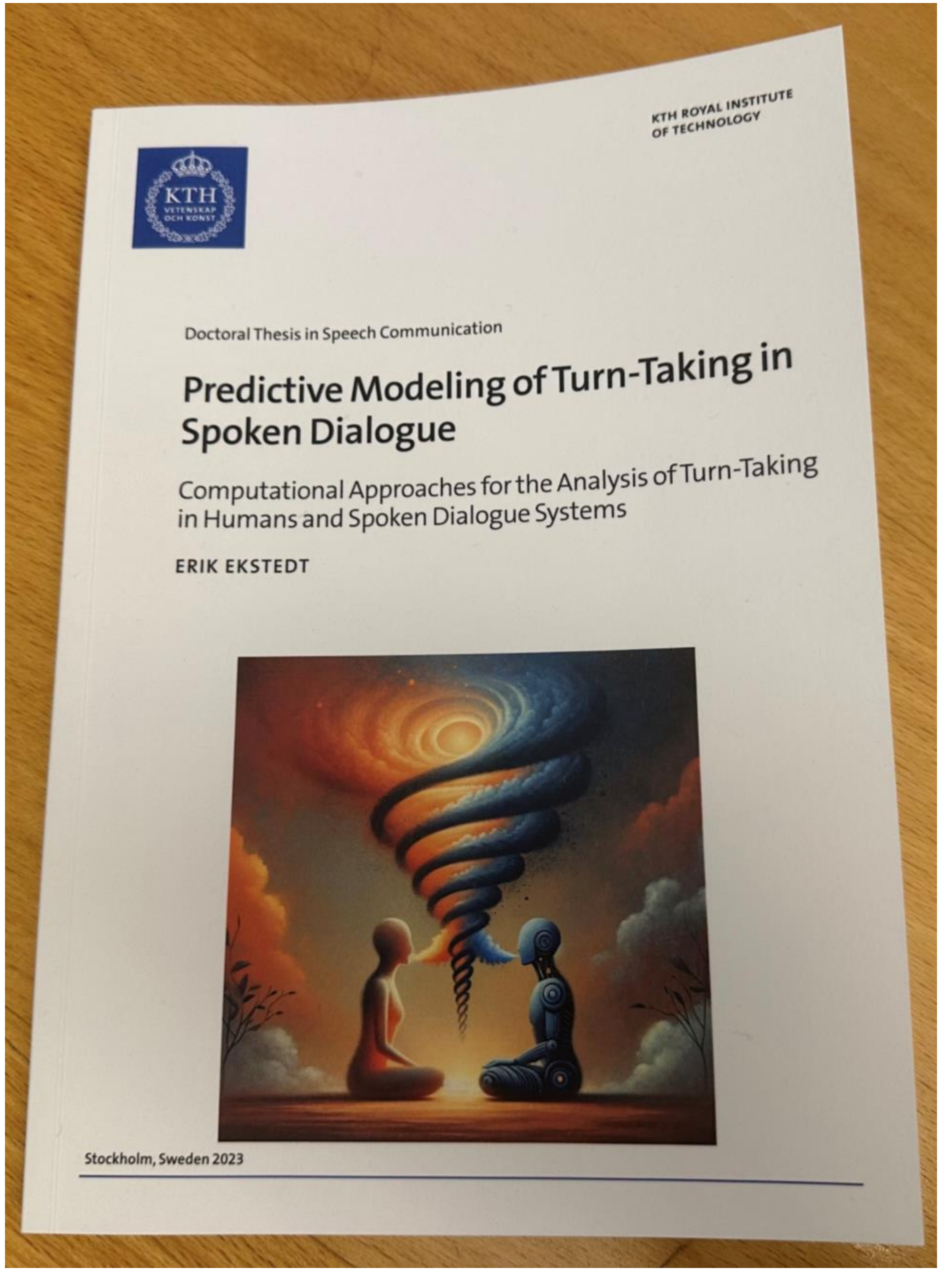
ARTICLE INFO

Article History:
Received 25 August 2020
Accepted 4 December 2020
Available online 16 December 2020

Keywords:
Turn-taking
Dialogue systems
Social robotics
Prosody
Gaze

ABSTRACT

The taking of turns is a fundamental aspect of dialogue. Since it is difficult to speak and listen at the same time, the participants need to coordinate who is currently speaking and when the next person can start to speak. Humans are very good at this coordination, and typically achieve fluent turn-taking with very small gaps and little overlap. Conversational systems (including voice assistants and social robots), on the other hand, typically have problems with frequent interruptions and long response delays, which has called for a substantial body of research on how to improve turn-taking in conversational systems. In this review article, we provide an overview of this research and give directions for future research. First, we provide a theoretical background of the linguistic research tradition on turn-taking and some of the fundamental concepts in theories of turn-taking. We also provide an extensive



KTH ROYAL INSTITUTE
OF TECHNOLOGY

Doctoral Thesis in Speech Communication

Predictive Modeling of Turn-Taking in Spoken Dialogue

Computational Approaches for the Analysis of Turn-Taking in Humans and Spoken Dialogue Systems

ERIK EKSTEDT

Stockholm, Sweden 2023

Thank you!

Vetenskapsrådet

RIKSBANKENS JUBILEUMSFOND
THE SWEDISH FOUNDATION FOR
HUMANITIES AND SOCIAL SCIENCES

WASP
WALLENBERG
AUTONOMOUS
SYSTEMS PROGRAM