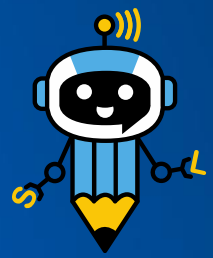
A Collection of Pragmatic-Similarity Judgments over Spoken Dialog Utterances

Nigel Ward, Divette Marco

University of Texas at El Paso

Linguistic Resources and Evaluation Conference

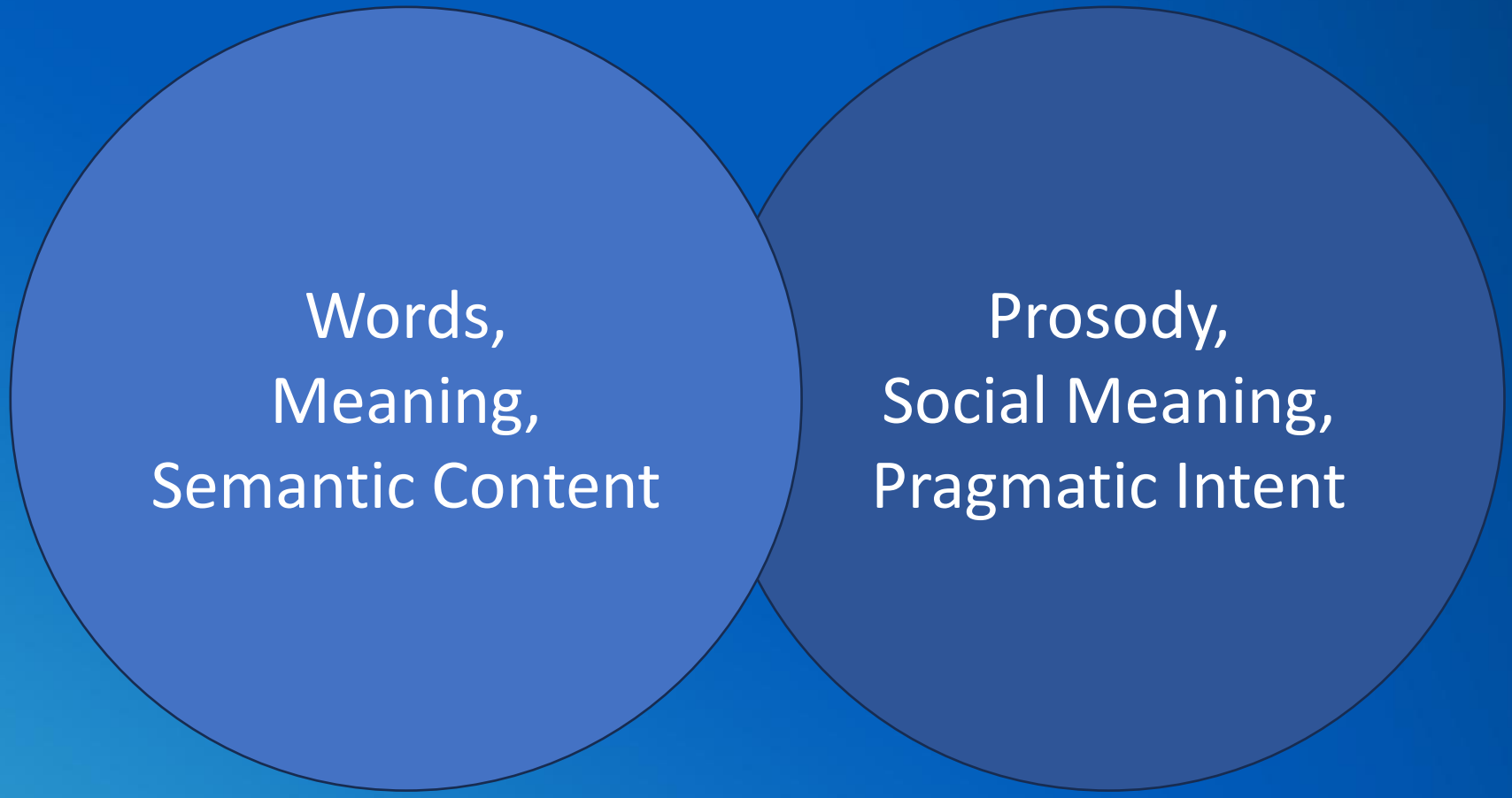
(LREC) 2024



NATIONAL AI INSTITUTE for Exceptional Education

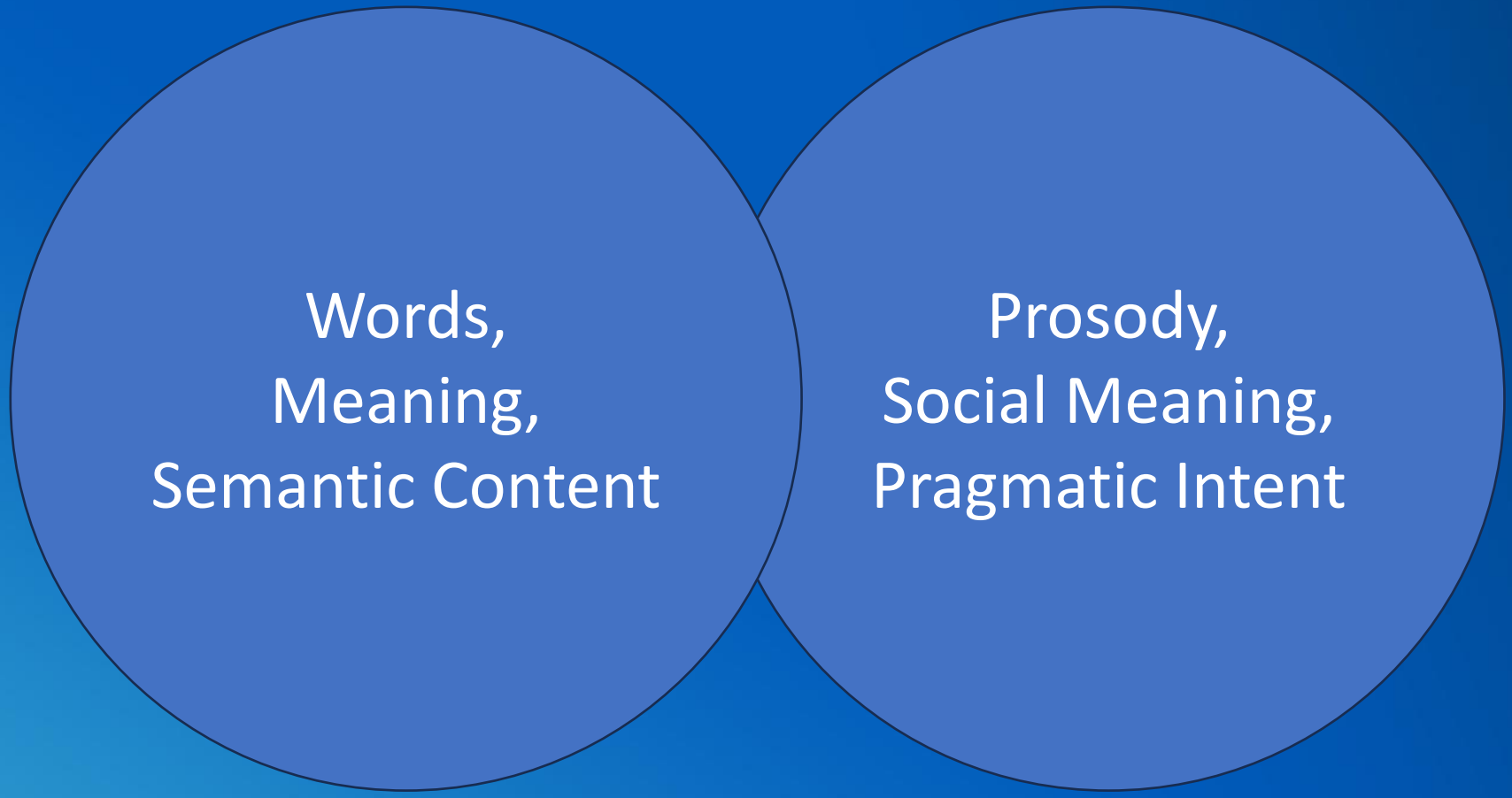
Human Spoken Language is Multifaceted

NATIONAL AI INSTITUTE for Exceptional Education



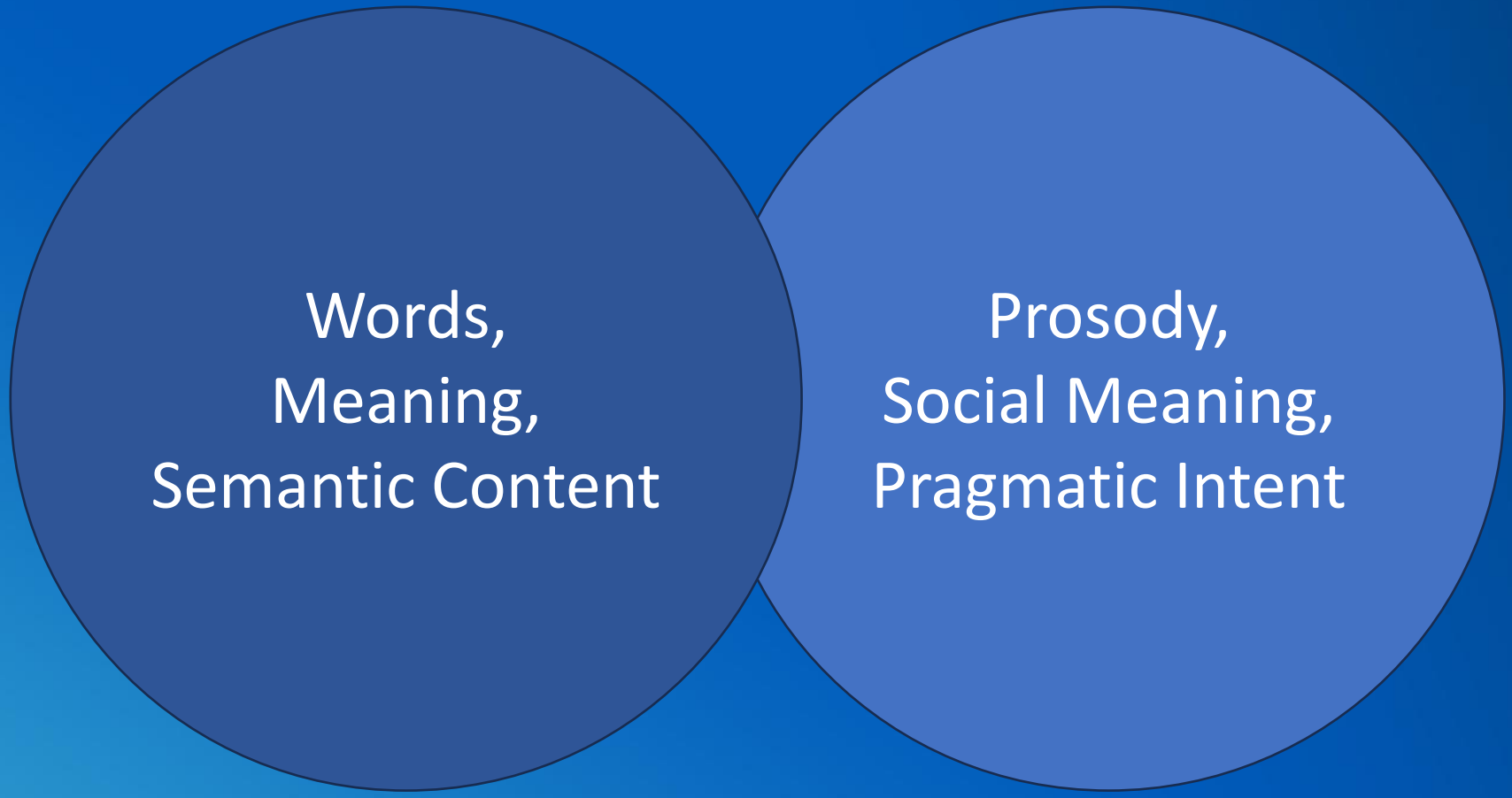
Human Spoken Language is Multifaceted

NATIONAL AI INSTITUTE for Exceptional Education



Human Spoken Language is Multifaceted

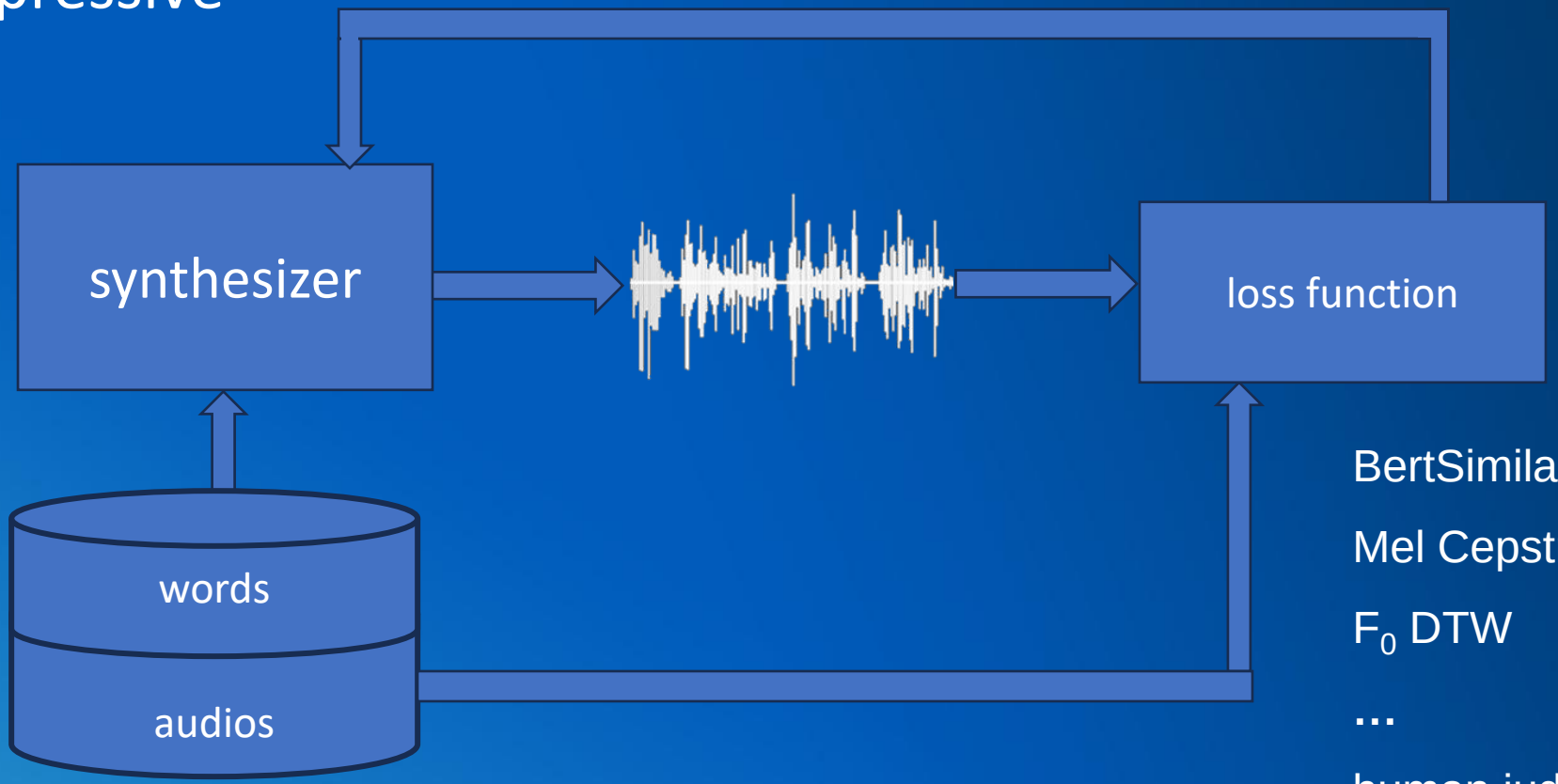
NATIONAL AI INSTITUTE for Exceptional Education



Speech Synthesis

NATIONAL AI INSTITUTE for Exceptional Education

- Already highly intelligible
- Not very expressive



Uses for a Pragmatic Similarity Measure

NATIONAL AI INSTITUTE for Exceptional Education

- For Speech Synthesis:

  How close is an utterance to the target?

- For Second-Language Training:

  How close is a learner utterance to a target?



- For Diagnosis:

  Are the two utterances close enough to infer that the speakers have the same medical condition?

- For Retrieval-based Chatbots:

  . . .

Related Work

NATIONAL AI INSTITUTE for Exceptional Education

- Semantic-similarity models
  - address a different problem
- Prosodic-similarity models
  - designed only for read speech
- Same-speech-act models (Pragst 2022)
  - inadequate for nuanced or multifaceted utterances

A Collection of Pragmatic-Similarity Judgments over Spoken Dialog Utterances

Nigel Ward, Divette Marco

University of Texas at El Paso

A Collection of Pragmatic-Similarity Judgments over Spoken Dialog Utterances

Nigel Ward, Divette Marco

University of Texas at El Paso

A Collection of Pragmatic-Similarity Judgments over Spoken Dialog Utterances

Data is needed

- To train models
- To evaluate models

A Collection of Pragmatic-Similarity Judgments over Spoken Dialog Utterances

Data is needed

- To train models
- To evaluate models

Outline

NATIONAL AI INSTITUTE for Exceptional Education

- The need
- The data we collected
- Cleverness and weakness in the data collection
- A model trained using this data

Session Statistics

NATIONAL AI INSTITUTE for Exceptional Education

	English 1	English 2	Spanish
Stimuli (clip pairs)	220	233	235
Judges	9	9	6
Total judgments	1980	2098	1410
Agreement*	0.45	0.72	0.66

*average inter-judge correlation, Pearson's

https://github.com/divettemarco/PragSim

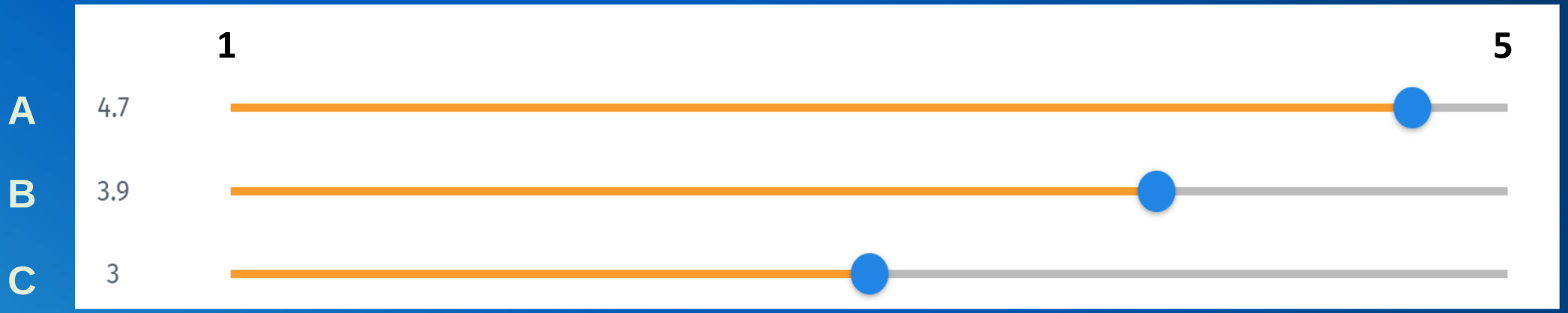
Judgment Examples

NATIONAL AI INSTITUTE for Exceptional Education

A

B

C



"How pragmatically similar are these, in terms of the overall feeling, tone, and intent?"

Judges, Procedure

NATIONAL AI INSTITUTE for Exceptional Education

2nd Session, October 7, 2023

Other Design Choices

NATIONAL AI INSTITUTE for Exceptional Education

- Dialog data

- Rating (vs ranking, ABX, etc.)
- Continuous rating 1 – 5 (vs discrete)
- Minimal delay between presentation
- Context-free presentation

Design Choices: The Instrument

NATIONAL AI INSTITUTE for Exceptional Education

"How pragmatically similar are the two clips, in terms of the overall feeling, tone, and intent.

- Try to ignore:
  - speaker differences,
  - differences in the words said
  - insignificant differences in pitch, rate, pausing, etc.

- Maybe consider: • Similarity in the contexts where they would likely appear • Similarity in how a listener would likely respond • Similarity in how the speaker may have felt (confident, positive, offended, enthusiastic, etc.) • Similarity in the dialog activity (correcting a misconception, teasing, holding the floor, asking a question, implying something, etc.)"

Stimulus Creation

NATIONAL AI INSTITUTE for Exceptional Education

Each pair has

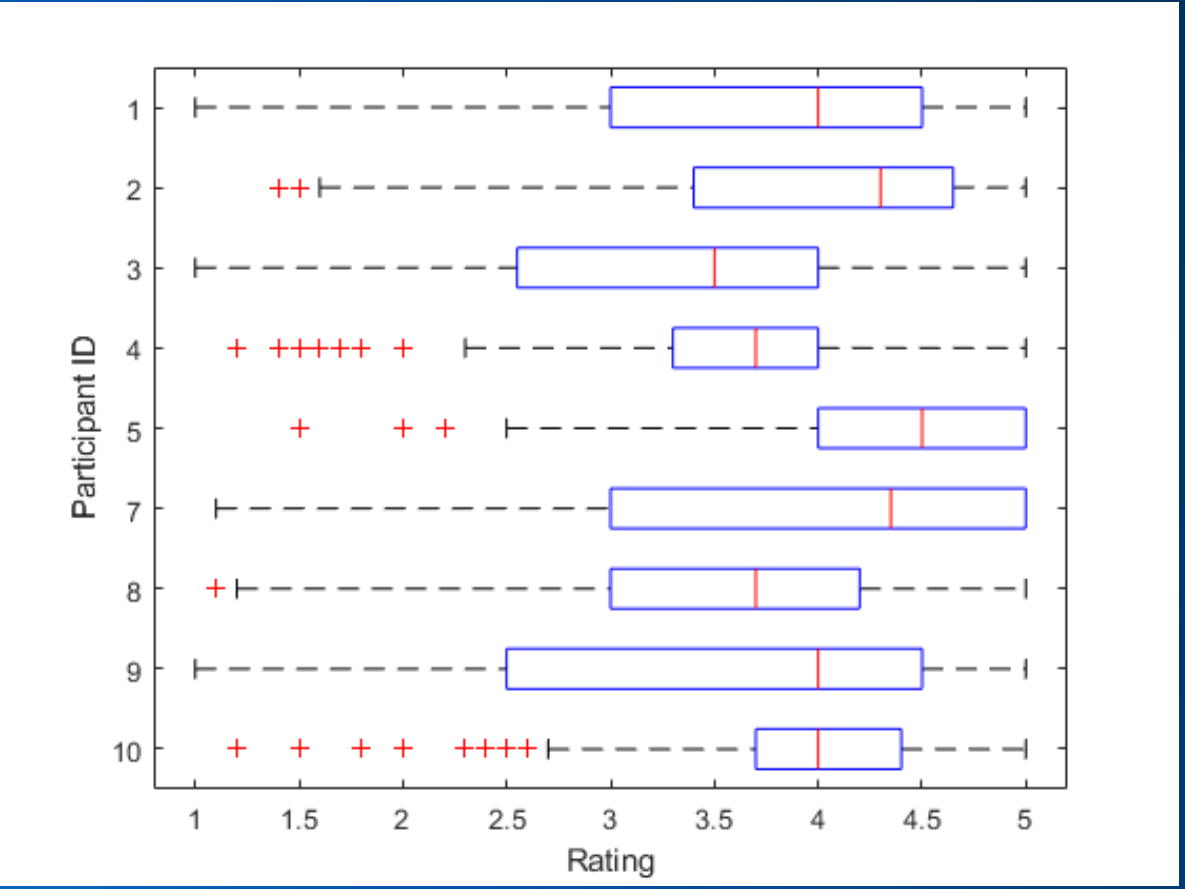
- An utterance from a real dialog, chosen for interestingness
- A re-enactment, done under various conditions:
  - Mimic the audio
  - See the words
  - Reproduce audio with different words
  - See the words and the context
  - Hear only the context
  - Speech synthesizer

very similar

moderately similar

Factors Affecting Ratings

- Judges varied
- Judges got slightly more generous over time
- Judges learned to use more of the scale



Factors Affecting Agreement

- Judge identity

Inter-Annotator Agreement (correlations), Session 1

judge	1	2	3	4	5	7	8	9	10
1									
2	0.40								
3	0.38	0.61							
4	0.37	0.59	0.59						
5	0.19	0.30	0.31	0.49					
7	0.41	0.67	0.66	0.54	0.33				
8	0.39	0.64	0.60	0.80	0.40	0.54			
9	0.21	0.40	0.36	0.18	0.19	0.51	0.20		
10	0.42	0.62	0.50	0.59	0.29	0.52	0.63	0.27	
Per-Judge Means									
	0.34	0.53	0.50	0.52	0.31	0.52	0.52	0.29	0.48

Session 2
average agreement: 0.72

Other Factors Affecting Agreement

NATIONAL AI INSTITUTE for Exceptional Education

Poor agreement

Generally better agreement

- For blandly-spoken pairs

  i.e., without laughter, ingressive fillers, breathiness, falsetto …

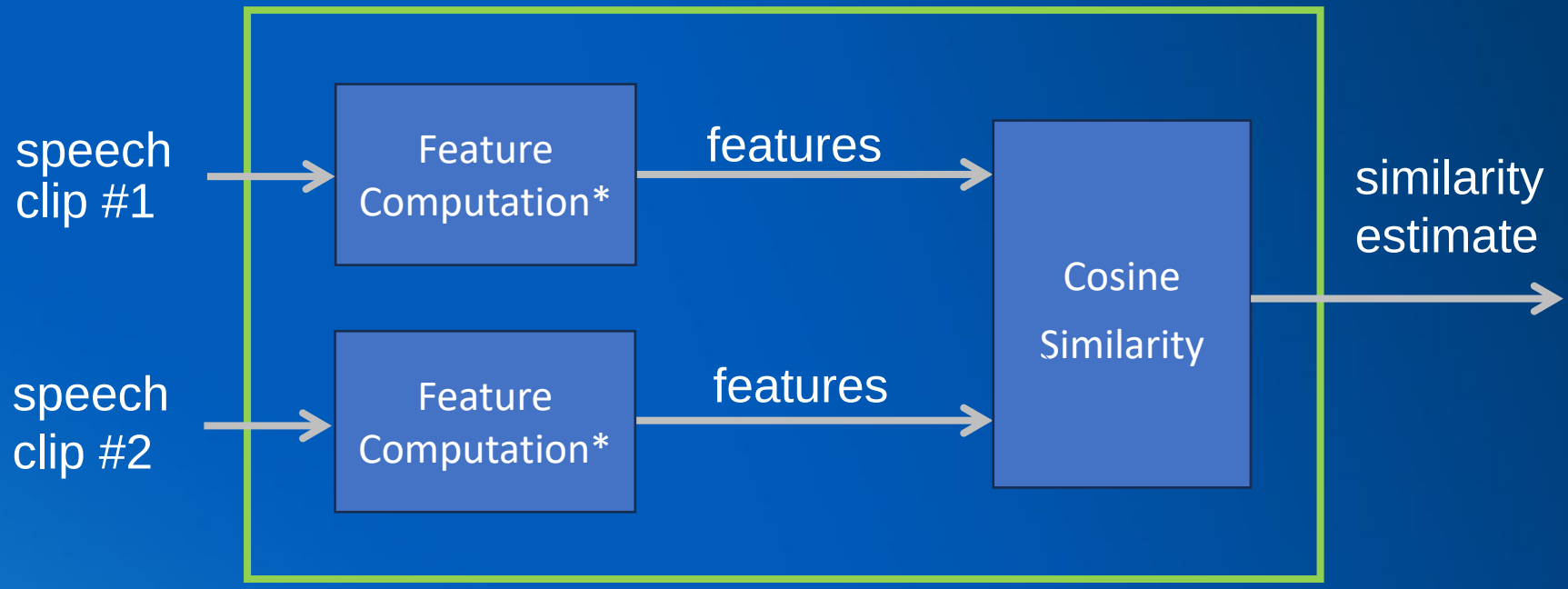
- For similar-personality speakers

- For judges with more experience
- Near the top of the scale
- For pairs with same lexical content
- For pairs similar in duration

NATIONAL AI INSTITUTE for Exceptional Education

Bonus Topic

A Similarity-Prediction Model



*Features:

- 103 features from the HuBert pretrained model
- selected to optimize performance on a training set of 1980 human judgments of similarity
- averaged over each entire clip

Comparison to human agreement

NATIONAL AI INSTITUTE for Exceptional Education

Average of Correlations* with Every Human Judge

	English 1	English 2	Spanish 1
Wav2Vec 2.0	.31	.41	.24
HuBert	.45	.41	.40
Selected HuBert	.50	.64	.45
Worst Human	.29	.68	.62
Average Human	.45	.72	.66
Best Human	.53	.78	.70

* Not correlations with the human average, like before

Utility for Finding Most-Similar Utterances

NATIONAL AI INSTITUTE for Exceptional Education

- An utterance from a conversation last week
  I drive a Hyundae Elantra, it's a gray color. Um, I chose it

- The most similar utterance out of 5000+ Switchboard utterances
  I use, 1-2-3, a lot. It's a Lotus product. It has a spreadsheet and I have, I use a

Notes:

- Talking about a product choice
- Early in the conversation
- Surprised by the question, disfluent
- Unsure whether the listener will recognize the name
- Satisfied with the product
- Intending to explain why they chose it

Pragmatic Similarity Demo

Contributions

NATIONAL AI INSTITUTE for Exceptional Education

- A protocol to collect pragmatic-similarity perceptions
- Observations of factors affecting ratings and agreement
- A set of 5000+ ratings of pragmatic similarity, for use in:
  - Speech-to-speech translation
  - Assessment of speaking skills
  - Dialog systems
  - Diagnosis

UTEP

NATIONAL AI INSTITUTE for Exceptional Education

Common Pragmatic Functions

NATIONAL AI INSTITUTE for Exceptional Education

- Positive assessment
- Cueing action
- Marking a shift in activity
- Showing empathy
- Yielding the turn vs Holding the floor

and many more, often nuanced, often in combination

A Conversation

NATIONAL AI INSTITUTE for Exceptional Education

Utility for Classifying Speech Disorders

NATIONAL AI INSTITUTE for Exceptional Education

Challenge: given data from a new, unknown speaker, is he/she autistic or not?

UTEP

Utility for Classifying Speech Disorders

NATIONAL AI INSTITUTE for Exceptional Education

Challenge: given data from a new, unknown speaker, is he/she autistic or not?

We used the NMSU ASD-NT dataset (thanks to Dr. Lehnert-LeHouillier)

- 28 Participants
  - 14 Neurotypical
  - 14 Autism spectrum disorder

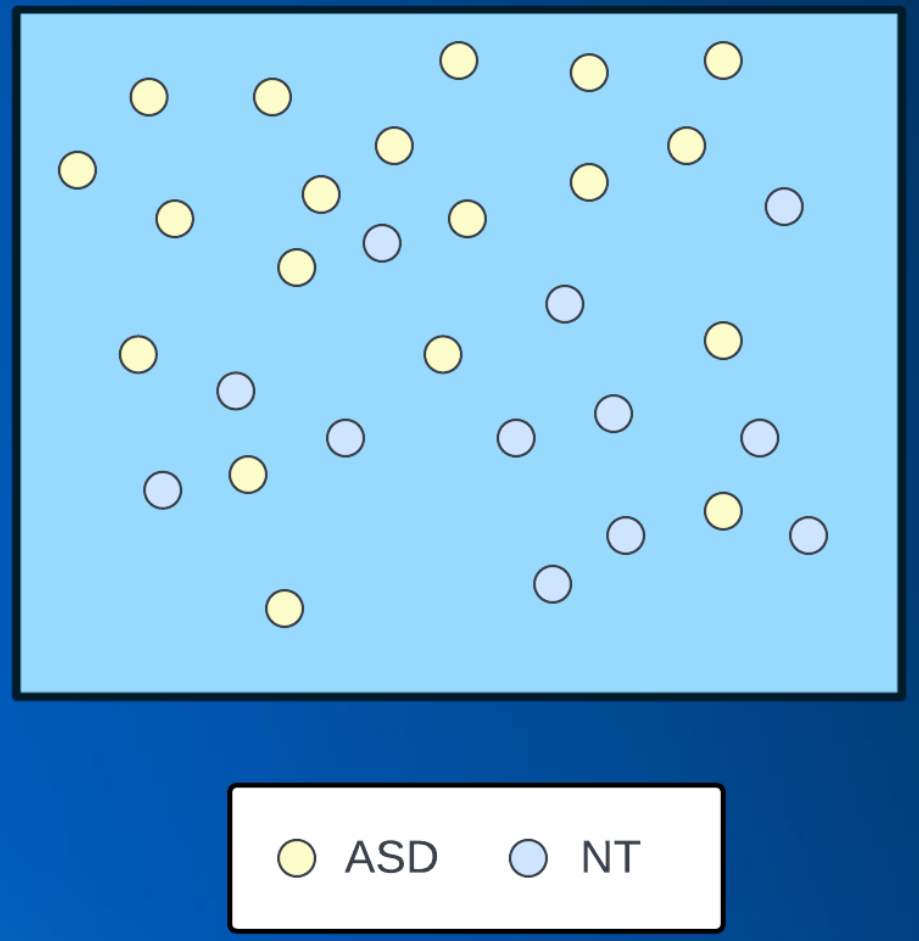
- 789 ASD audio clips
- 702 NT audio clips

UTEP

The Problem

NATIONAL AI INSTITUTE for Exceptional Education

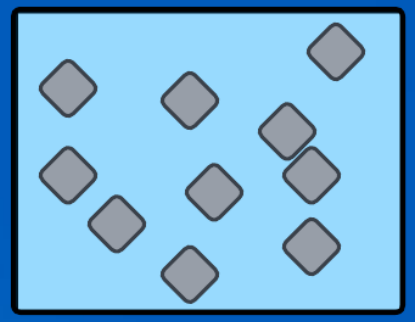
Known-Clip Representations



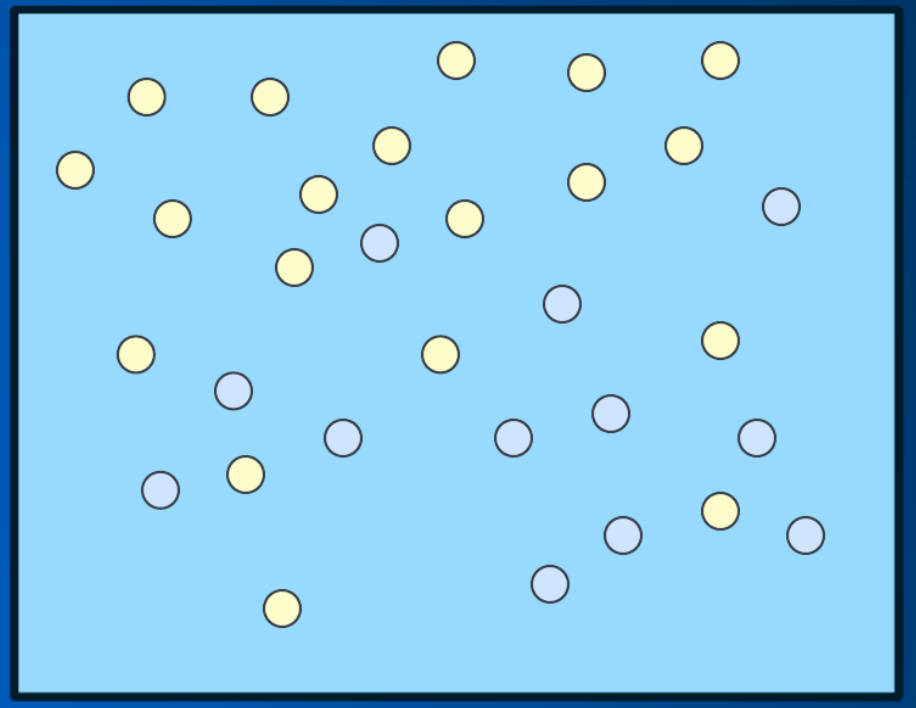
The Problem

NATIONAL AI INSTITUTE for Exceptional Education

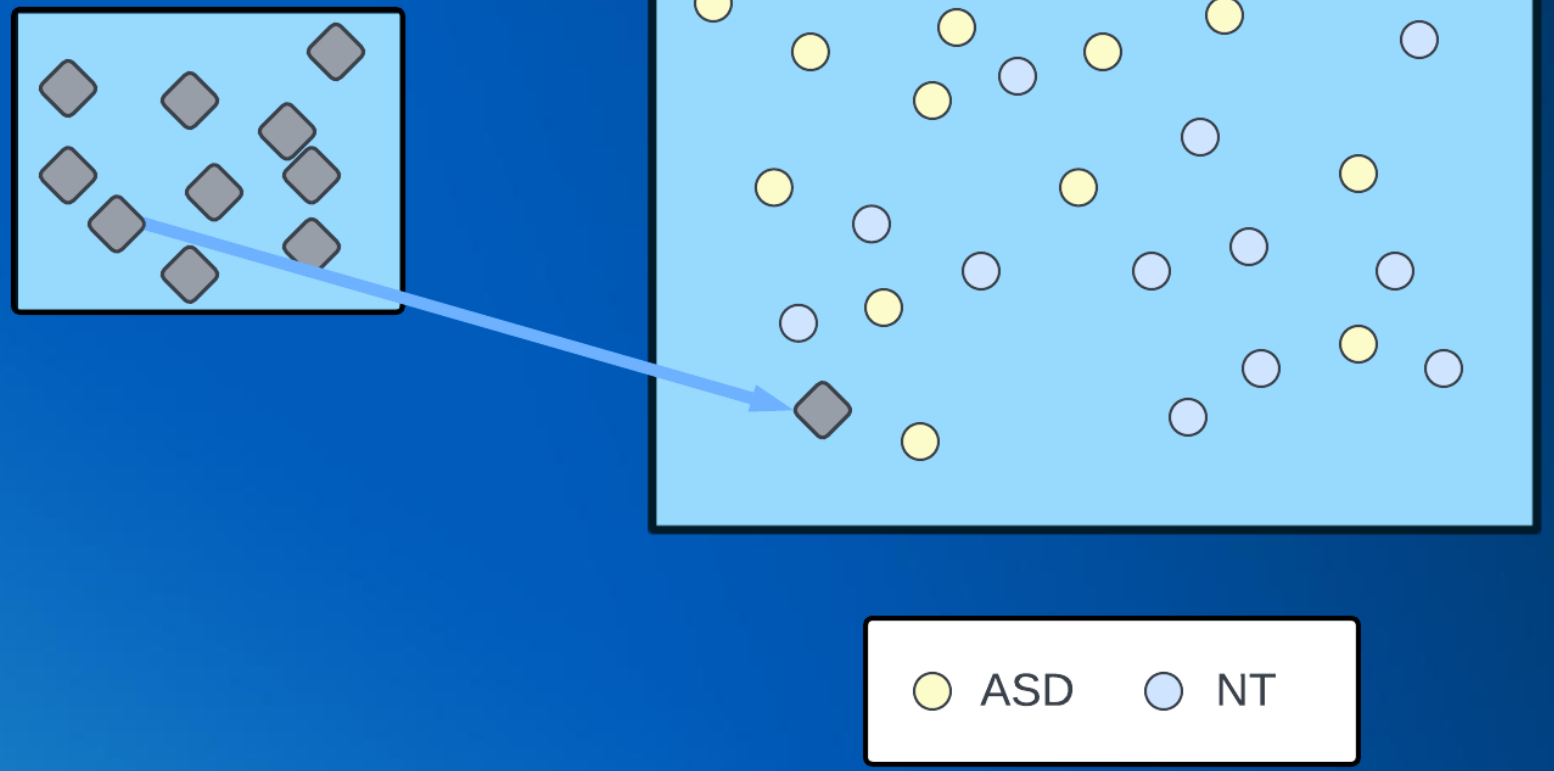
Child X and some of his Speech Clips



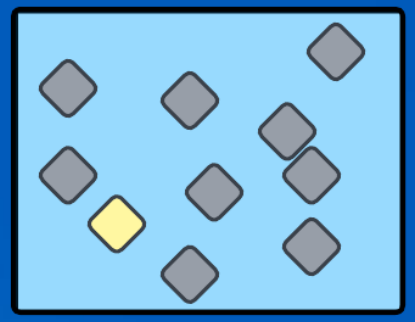
Known-Clip Representations



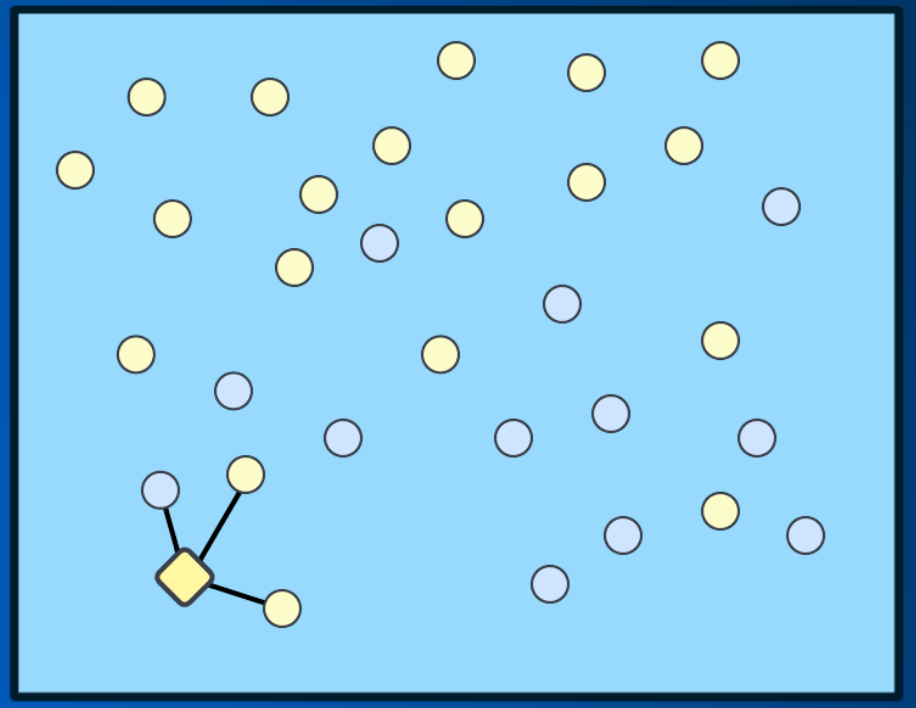
Classification by kNN



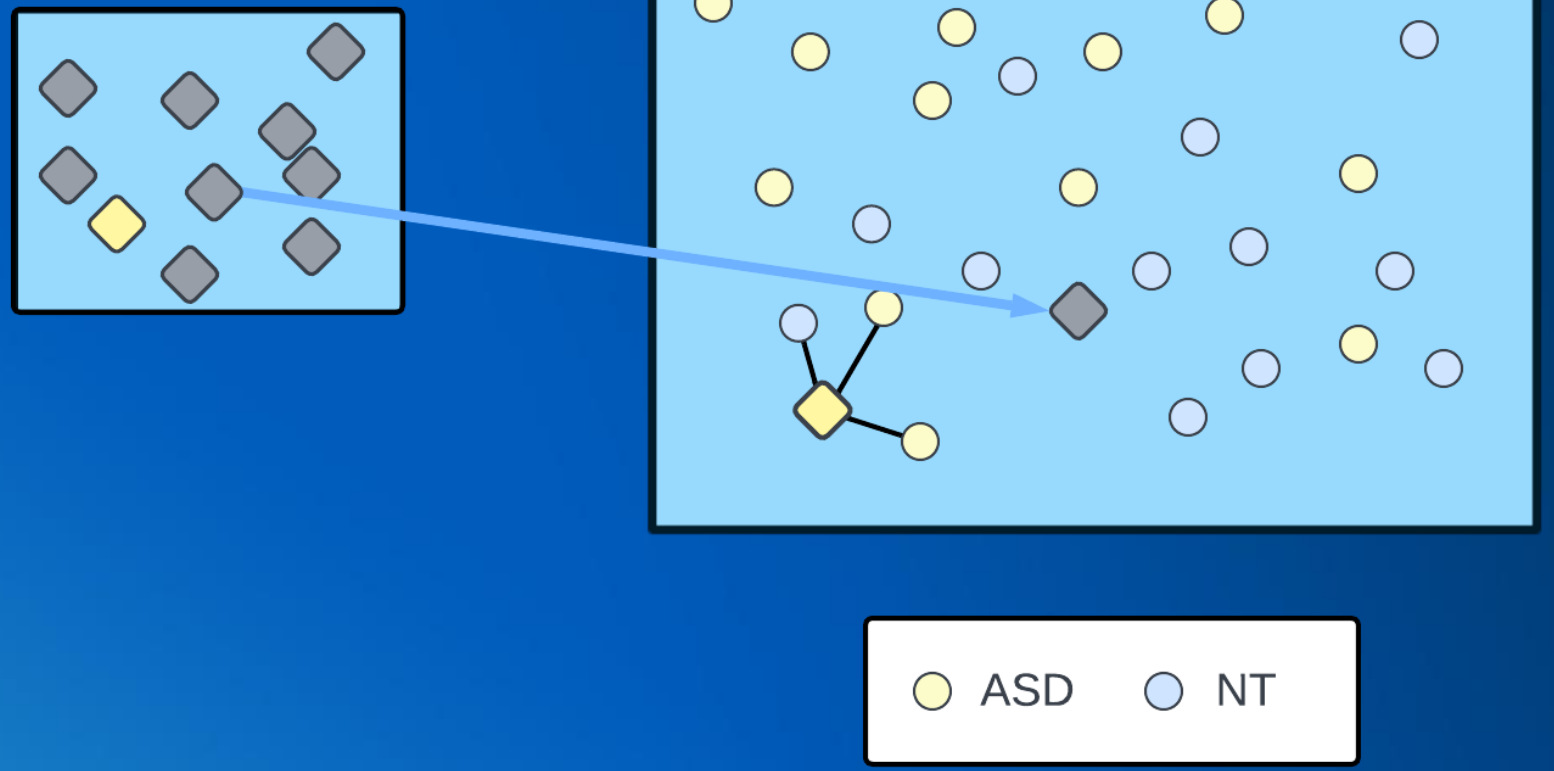
Classification by kNN



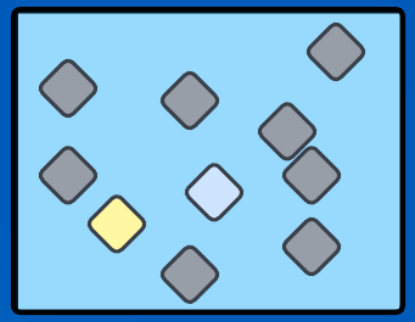
Known-Clip Representations



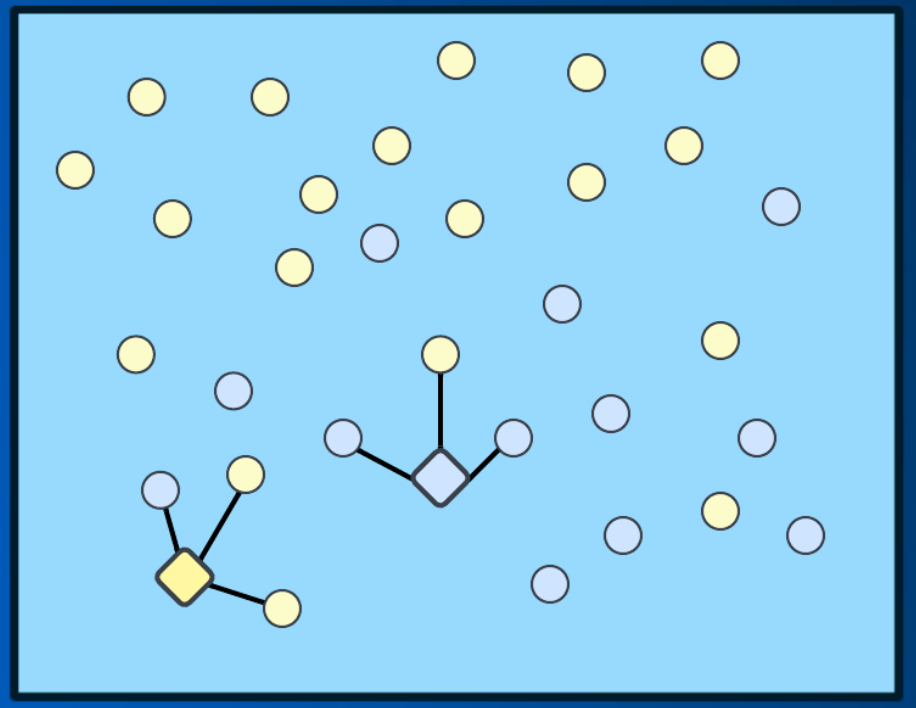
Classification by kNN



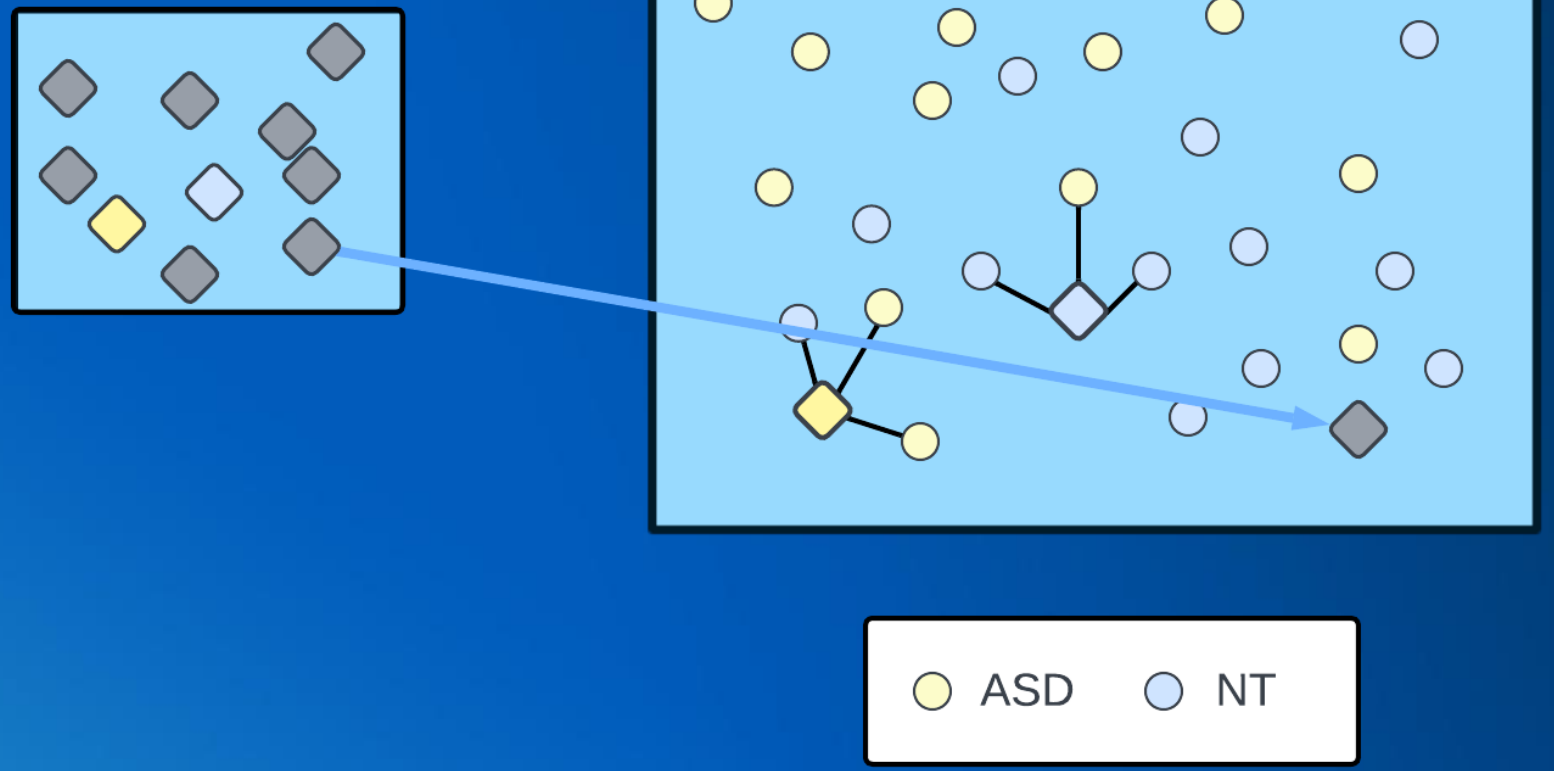
Classification by kNN



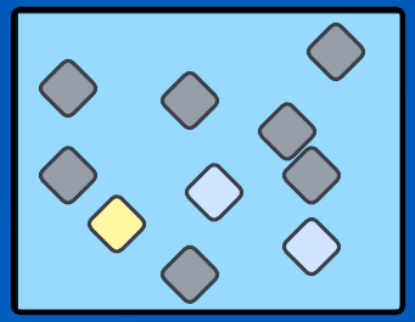
Known-Clip Representations



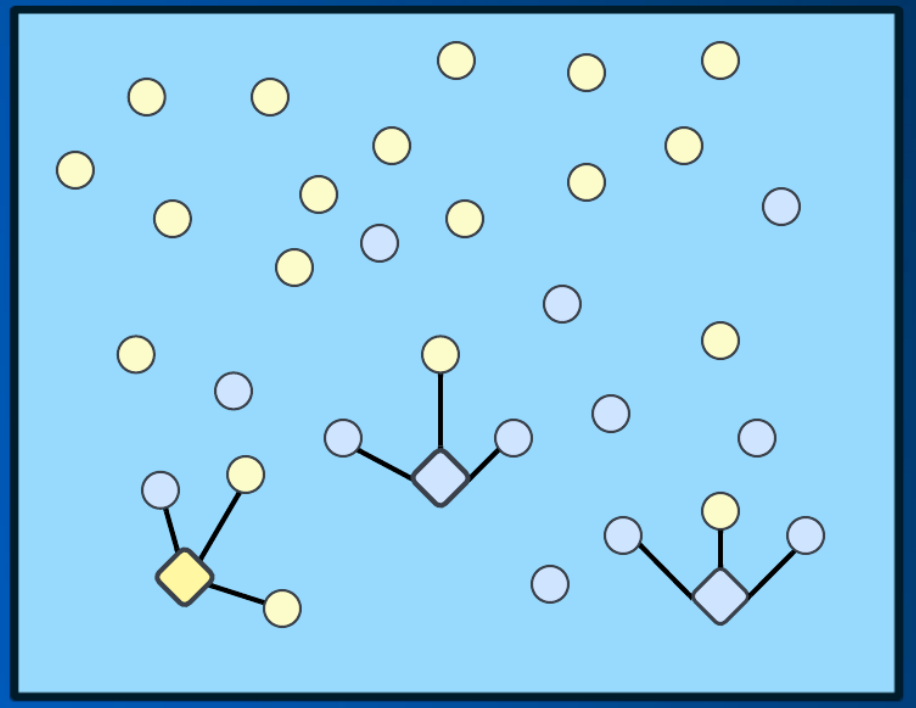
Classification by kNN



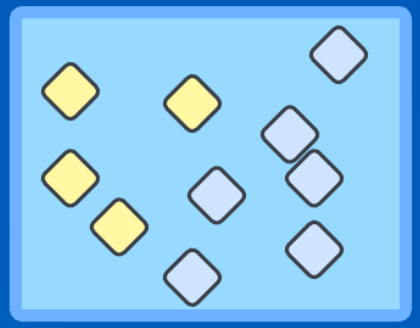
Classification by kNN



Known-Clip Representations



Classification by kNN



We classify the child by their most frequent clip label

Results

	Autistic	not
Predicted Autistic	10	1
Predicted not	4	13

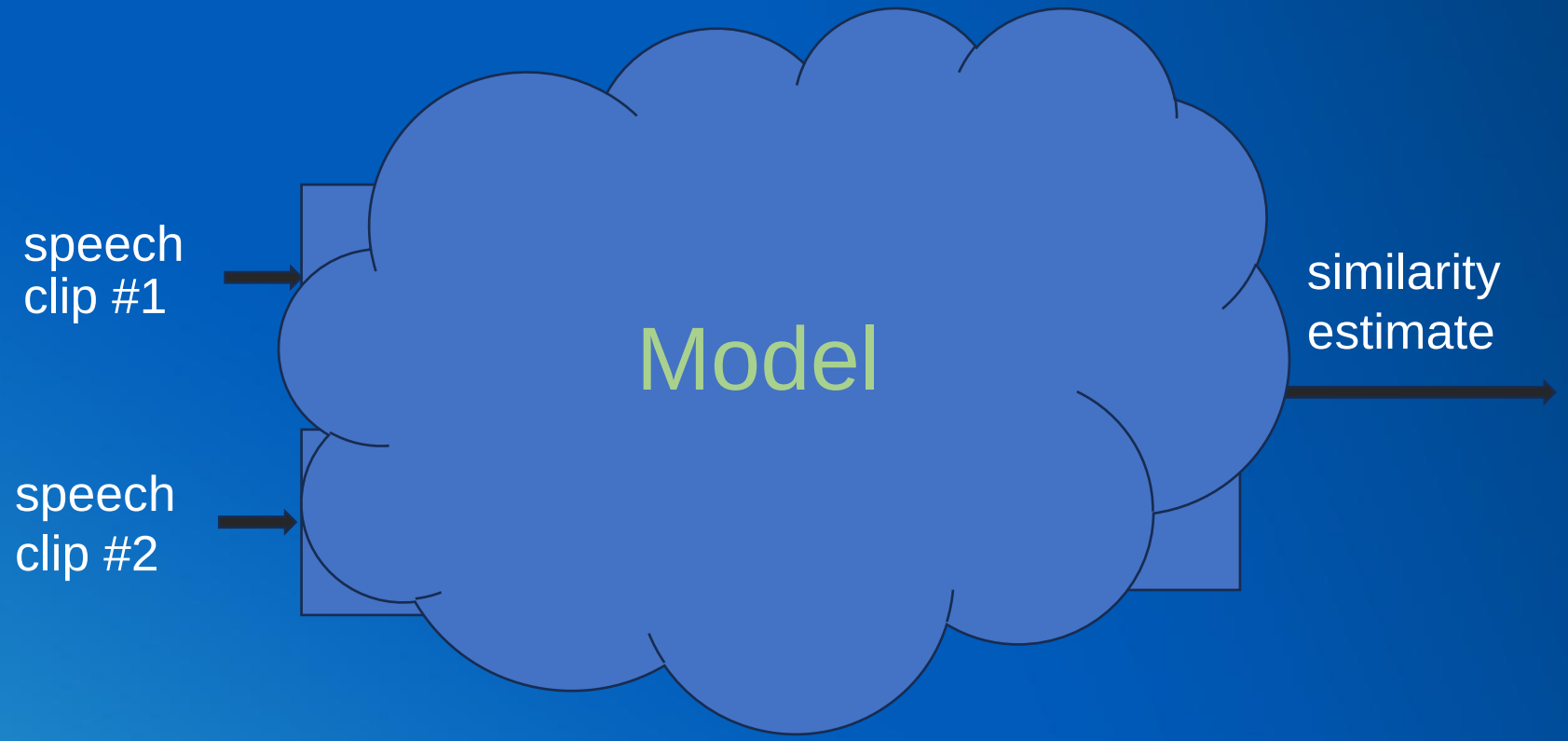
81% accuracy

Exculpatory factors

- the misclassified NT speaker was one of the youngest
- 3 of the misclassified autistic speakers had lower ADOS scores
- 2 of them had very few audio clips to go on

Note: best performance with k=7, but not highly sensitive

Problem (restated)



Results

NATIONAL AI INSTITUTE for Exceptional Education

Correlation with Human Judges' Averages

	English 1	English 2	Spanish 1
Cepstral distance	.09	.24	.22
F0 DTW	.08	.11	.07
Mel-cepstral DTW	.16	.23	.22
Duration	.24	.05	.20
WavLM	.12	.17	.06
Wav2Vec 2.0	.31	.41	.24
HuBert	.45	.41	.40

Language Dependencies

NATIONAL AI INSTITUTE for Exceptional Education

Correlation with human judgment averages

	English 1	English 2	Spanish 1
Original HuBert	.45	.41	.40
English-tuned HuBert	.69	**.74**	**.53**
Spanish-tuned HuBert	.59	**.63**	**.72**

- Feature selection helps
- Language-specific features selection helps more

Comparison to BERTSimilarity

NATIONAL AI INSTITUTE for Exceptional Education

Correlation with average human judgments on the lexically-distinct subset*

	English 1	English 2	Spanish
selected HuBert	0.31	0.20	0.38
duration	0.49	0.11	0.20
BertSimilarity	0.57	0.50	0.38

* for the rest, BertSimilarity performance is of course 0.0

Demo Procedure

NATIONAL AI INSTITUTE for Exceptional Education

- An undergraduate, native English speaker volunteers
- He/she has a short conversation with Andy
- The system extracts their utterances.
- For each, it finds utterances in the corpus that it thinks are very similar, less similar, etc.
- We listen and see if we agree

UTEP

NATIONAL AI INSTITUTE for Exceptional Education

After this point is just spare slides

Other Use Cases, with Healthcare Utility

NATIONAL AI INSTITUTE for Exceptional Education

A similarity metric can support

- Detecting atypical speakers
- Finding similar speakers
- Finding representative utterances
- Finding atypical/outlier utterances
- Finding comparable utterances (as in the demo)

What are Prosody and Pragmatics?

NATIONAL AI INSTITUTE for Exceptional Education

- Prosody is the patterns of rhythm, stress, and intonation in speech.
- Pragmatics is the study of how context contributes to meaning.
- Prosodic features convey pragmatic meaning.
- Pragmatic Similarity defines how closely the meaning of two utterances are to each other.

Shallow Modeling Options

NATIONAL AI INSTITUTE for Exceptional Education

- Supervised learning (requires labeled data)
- Unsupervised learning
- Self-supervised learning

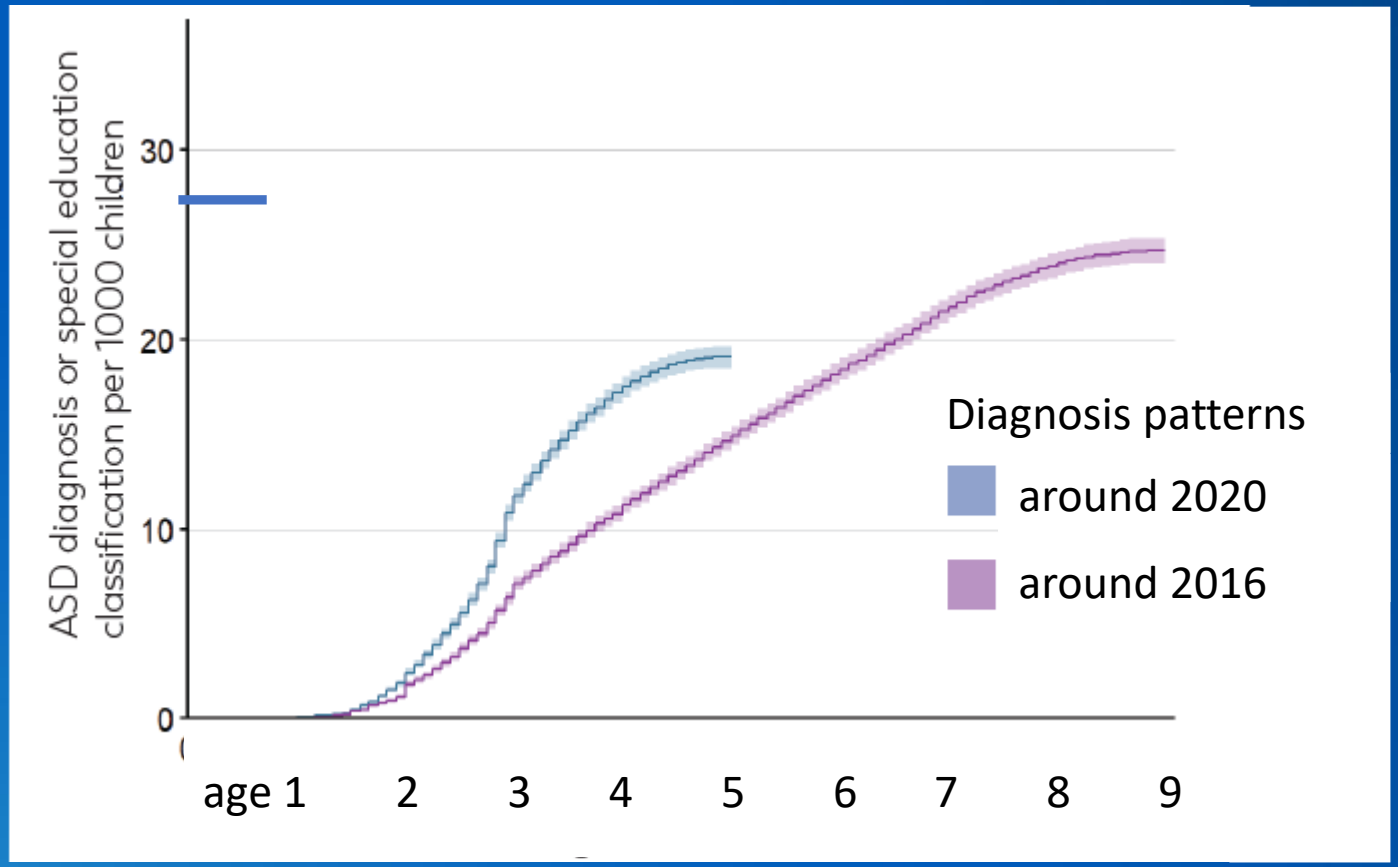
Childhood Communication Disorders

- Apraxia
- Dysarthria
- Articulation disorders
- Stuttering
- Specific language impairment
- Autism (1 in 36 children)

  etc.

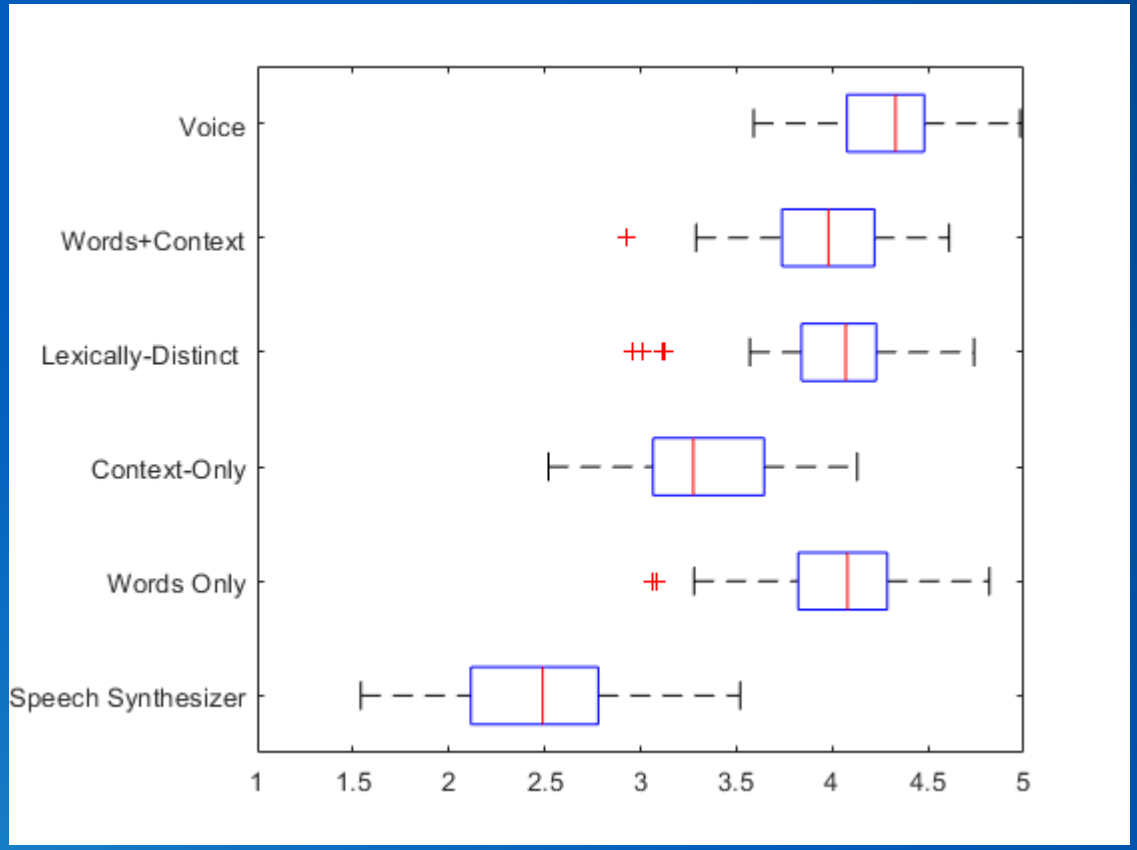
Early intervention can help … but this requires early screening

Early Diagnosis is Hard



https://www.cdc.gov/ncbddd/autism/addm-community-report/spotlight-on-COVID-disruption.html

NATIONAL AI INSTITUTE for Exceptional Education



Common Pragmatic Functions

NATIONAL AI INSTITUTE for Exceptional Education

- Cueing action
- Positive assessment
- Marking a shift in activity
- Showing empathy
- Yielding the turn vs Holding the floor



All of these are mostly conveyed with prosody non-trivially