A FRAMEWORK FOR THE AUTOMATIC INFERENCE OF STOCHASTIC TURN-TAKING STYLES

Kornel Laskowski

Carnegie Mellon University, Pittsburgh PA, USA
Voci Technologies, Inc., Pittsburgh PA, USA

Goals

A **stochastic turn-taking model** predicts a speaker's speech activity at instant ***t***, given that speaker's and their interlocutors' speech activity at preceding instants.

It is known that ...

For conversant-independent models, training with more data helps **on average**.

For conversant-dependent models, **within-conversation** adaptation helps.

But it is not known ...

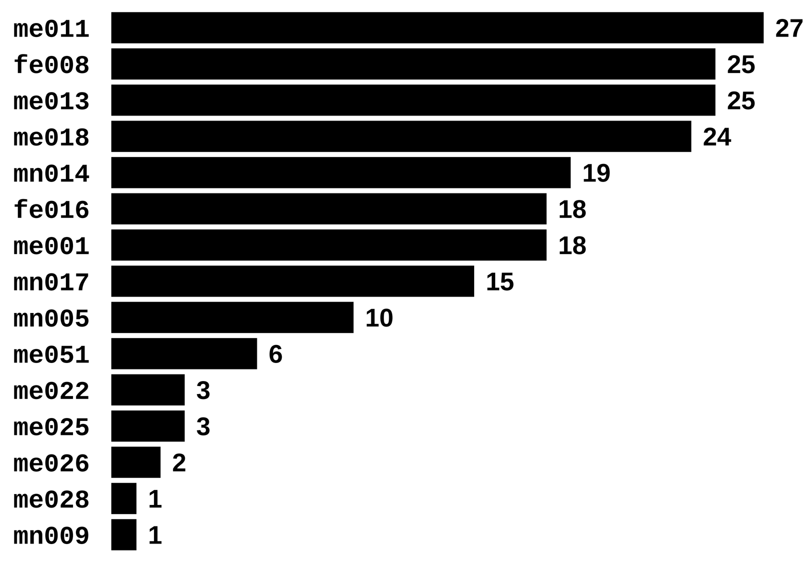
1. Whether conversants are self-consistent **across conversations**, implying quasi-stationary turn-taking styles?

2. What may account for the variability observed in models?

Longitudinal, Conversational Dataset

"`Bmr`" Subset of ICSI Meeting Corpus

- recorded over the course of a year
- allegedly **natural** turn-taking: "would have been held even if they were not recorded"
- total 29 conversations
- average 48.4 minutes per conversation
- total 15 participants
- average 6.8 participants per conversation
- total 197 conversation sides



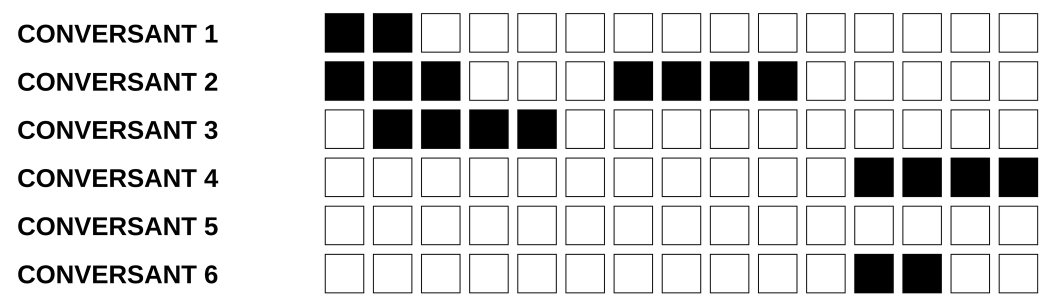
Findings from Dataset

1. Intra-person (within-class) variability is smaller than inter-person (between-class) variability.
2. People are surprisingly self-consistent, even in the extremely impoverished representation of speech/non-speech over 500 ms.
3. Longer-conversation observations exhibit smaller intra-person (within-class) variability.
4. Greater "talkativity" exhibits larger inter-person (between-class) variability.
5. Stochastic turn-taking models appear to be correlated with conversational-group role.
6. Person-discriminative aspects of stochastic turn-taking models appear to lie on a low-dimensionality manifold.

1. Inference of Turn-Taking Models

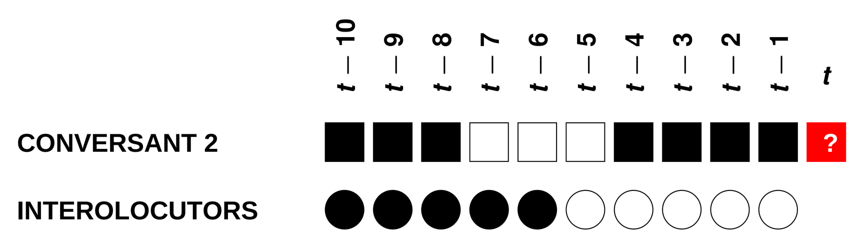
Construct Speech/Non-speech Chronogram

a. Throw away voice, words, prosody, etc.
b. Discretize using 100-ms frames



Infer an STT Model for Each Conversant

c. Compute exclusive-OR of interlocutors
d. Train n-gram stochastic turn-taking model



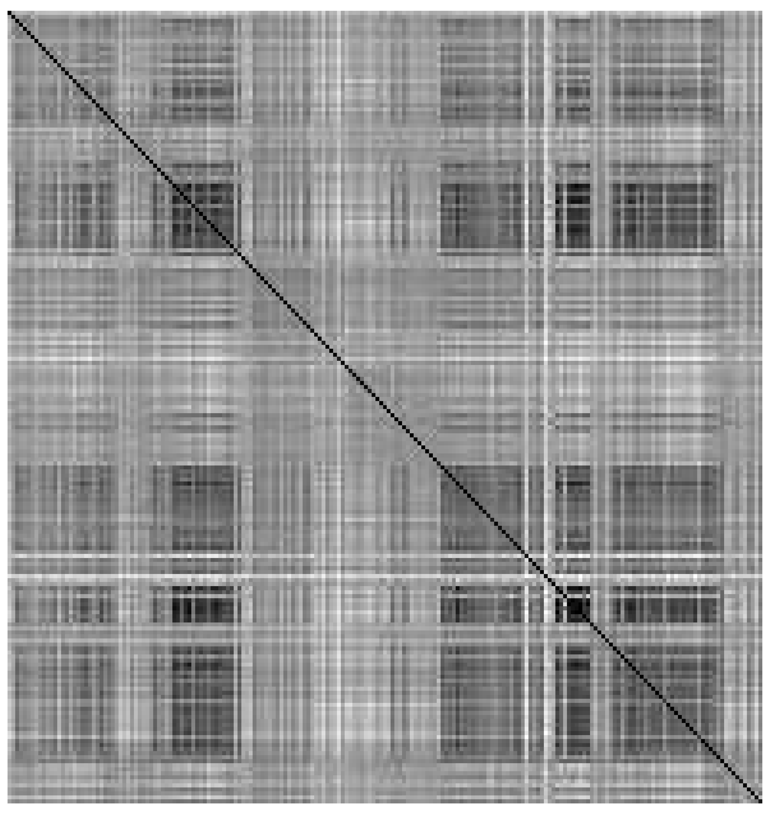
Unsmoothed model characterizes participant's contextualized speech deployment timing in one conversation (ie. one "side").

2. Computation of Inter-Model Distances

Compute Jensen-Shannon (JS) Distance

a. n-grams are conditional probability models
b. JS distance is symmetric and bounded

Form Distance Matrix Over All Model Pairs



Can perform distance-based clustering and/or classification of conversational sides directly from complete matrix.

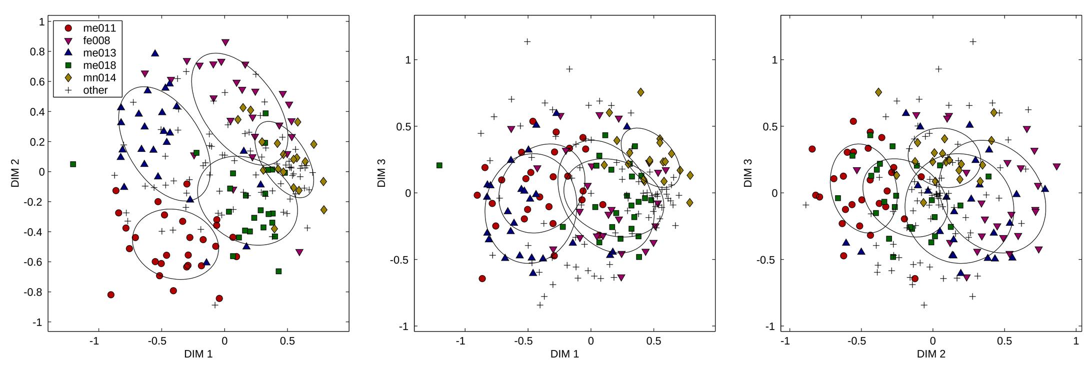
3. Nearest-Neighbor (NN) Classification

Classify Participants Directly from Matrix

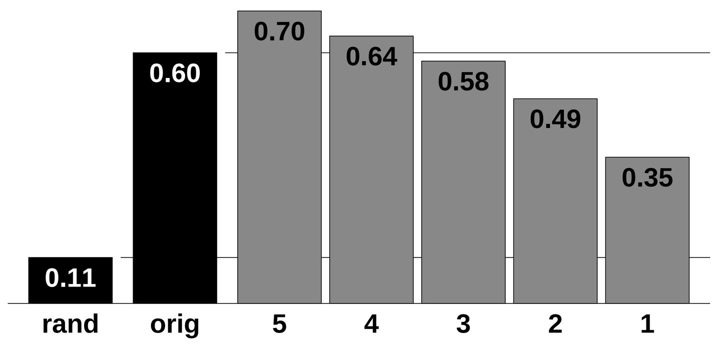
a. Accuracy of 60% is achieved

Classify After Multi-Dimensional Scaling

b. Apply MDS to ***N*** dimensions (***N*** small)



c. Recompute distances
d. Repeat NN classification

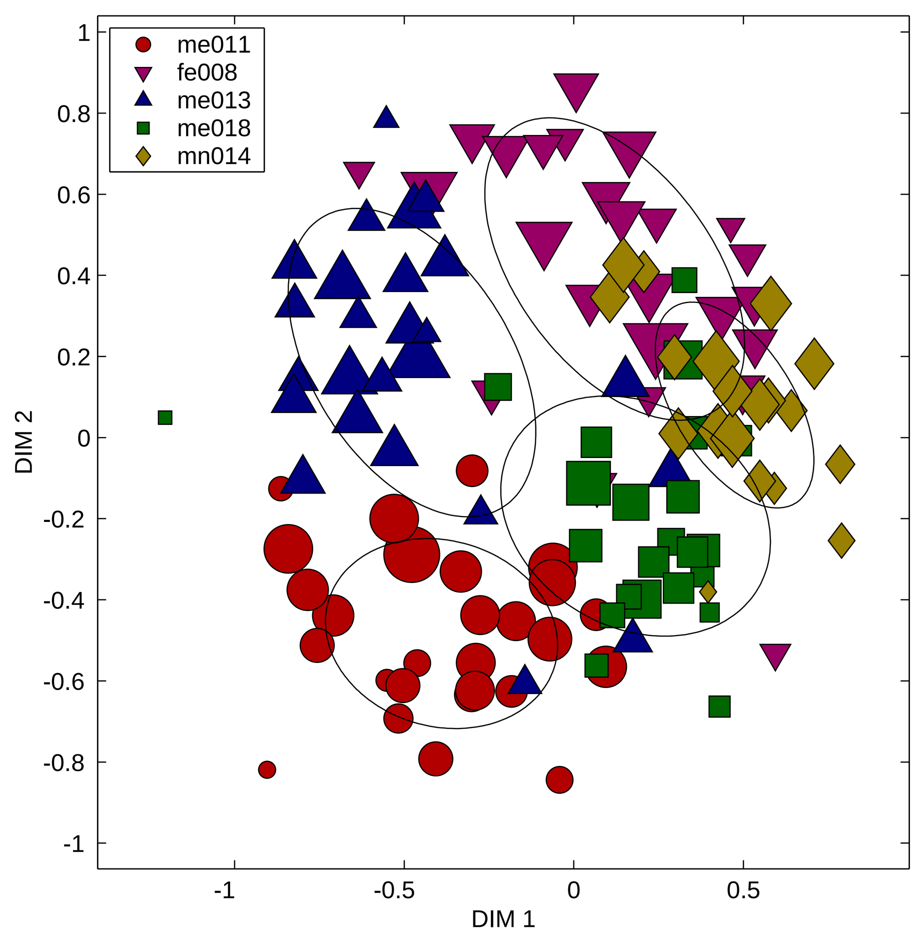


The space of models appears to lie on a low-dimensional manifold.

Intra-Person: Duration of Observation

Variability due to duration of sides

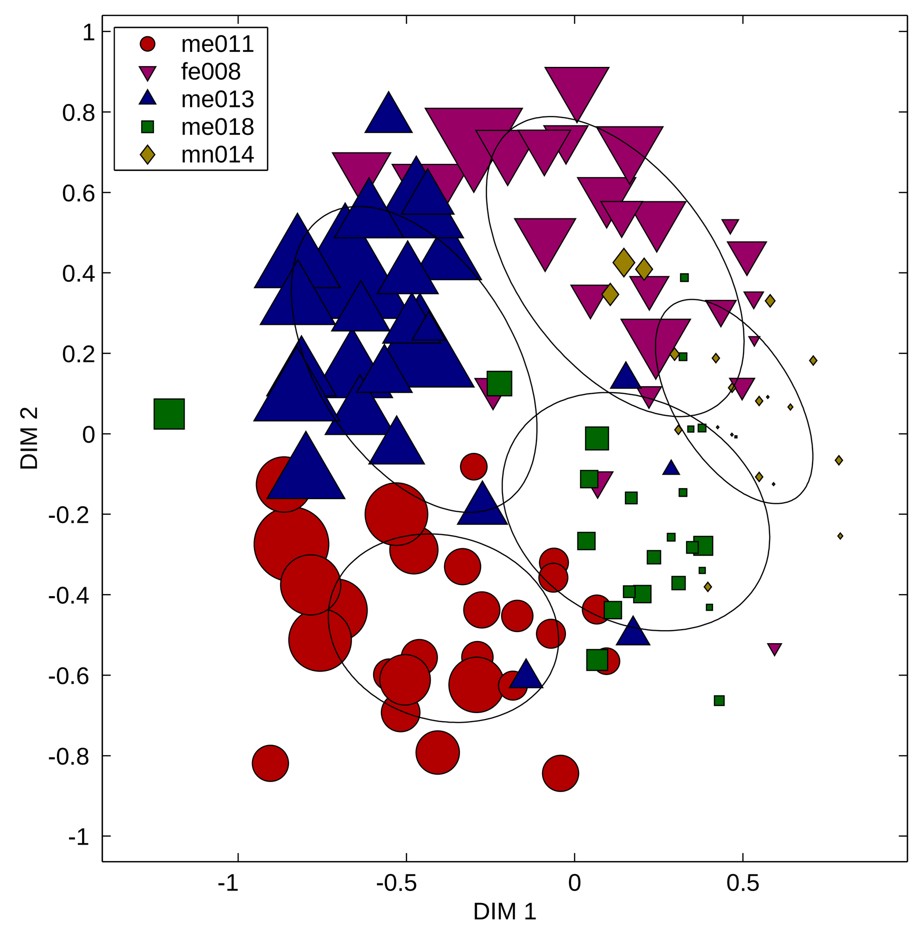
bigger markers ⟶ longer sides



Intra-Person: Unconditional "Talkativity"

Variability due to the proportion of speech

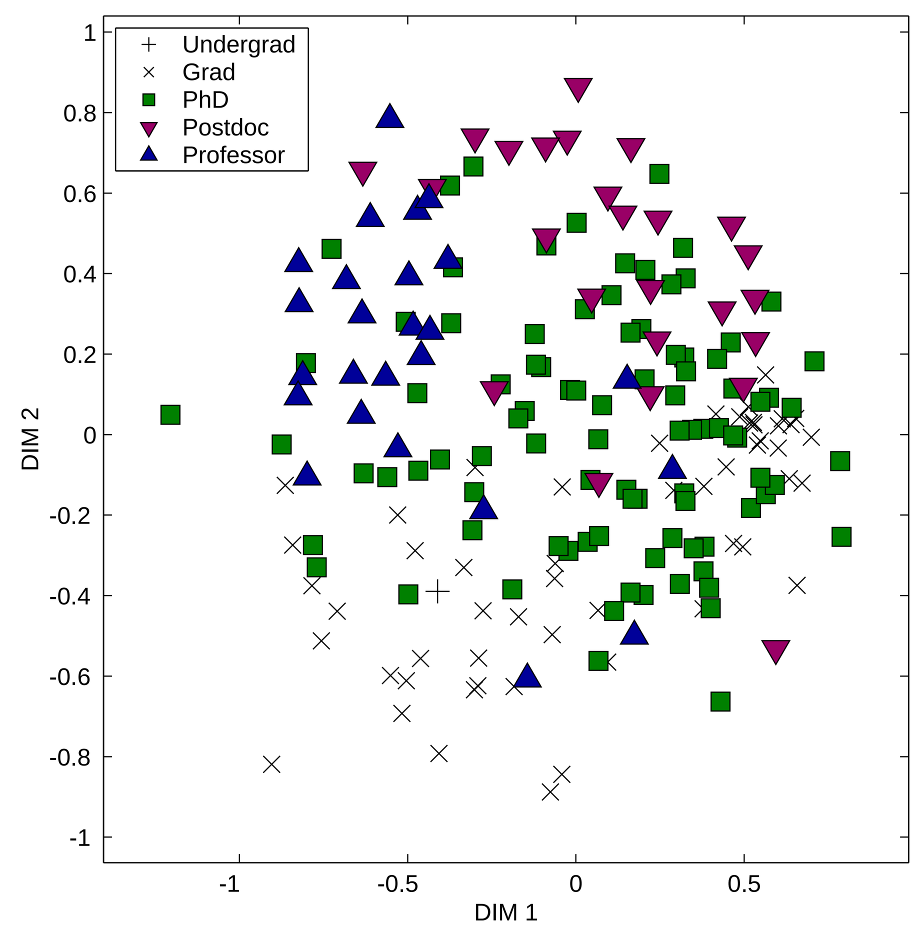
bigger markers ⟶ talkative sides



Inter-Person: Organization Seniority

Variability due to organizational role

self-reported seniority as proxy



kornel@cs.cmu.edu