Multimodal Continuous Turn-Taking Prediction Using Multiscale RNNs

Matthew Roddy, Gabriel Skantze, Naomi Harte

Sept 24, 2018

Introduction

- **Spoken Dialogue Systems (SDSs)**: A computer system that is able to converse with a human with voice.
  - e.g. Alexa, Cortana, Google Duplex, conversational robots.
- **Turn-taking Modeling**: Modeling the decisions as to when a dialogue system should start or stop speaking.

Research Objective

- **Research Objective**: Design conversational turn-taking models that can aid SDSs in producing naturalistic interactions.

Example: Microsoft Directions Robot

- Heuristic turn-taking model: Silence thresholding (500ms)
- Problem: 500ms silence is sometimes too long, sometimes too short.
  - Human-human modal turn-switch time: 200ms (very fast)
  - User could be taking a mid-turn pause
- Error analysis: 28% of utterances had a timing problem (e.g. interruptions, long pauses)

Microsoft directions robot, Andrist et al. [2017], Bohus and Horvitz [2011]

Turn-Taking Cues

- **Turn-taking cues**: Cues used during conversations to coordinate turn exchanges.
  - Used by speakers to signal either relinquishing the turn, or holding on to the turn.
  - Used by listeners to **predict** when an appropriate turn-transition point is coming up.
- **Prediction** is how we are able to achieve fast turn-switch times of 200ms or less.

Turn-Taking Cues: Acoustic

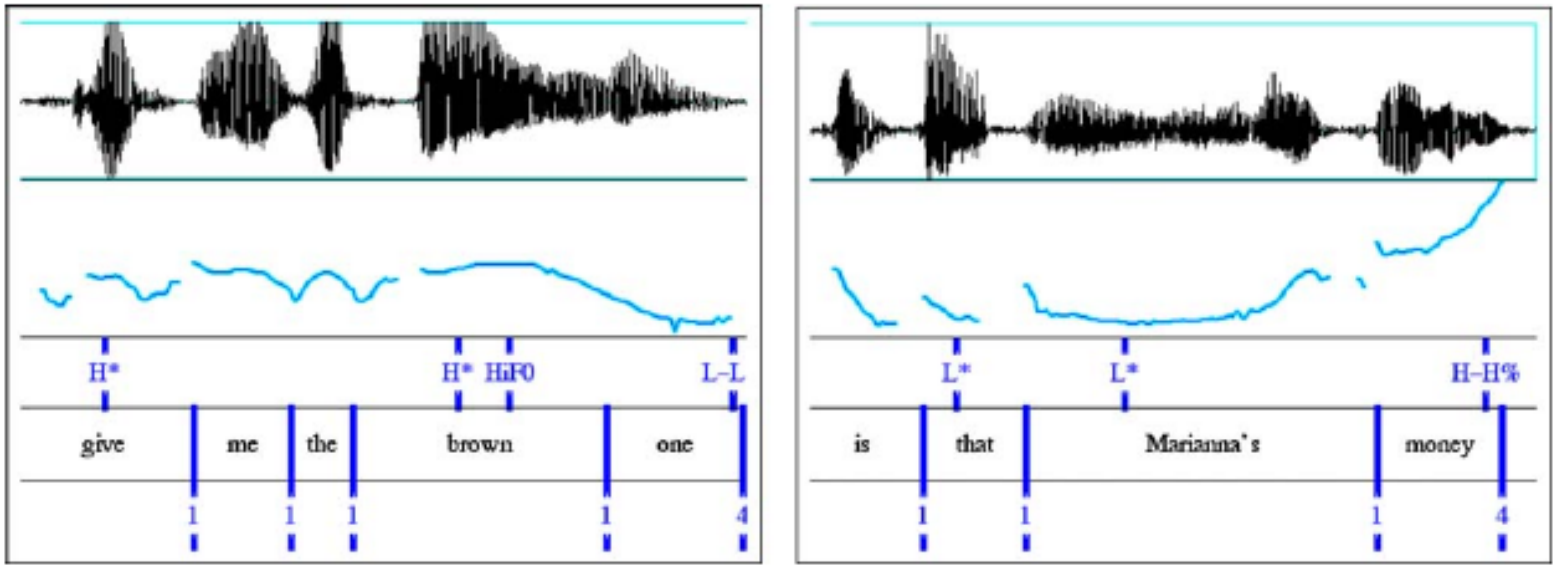


Figure: Pitch contour turn-taking cues. Gravano and Hirschberg [2011]

- Falling/rising pitch
- Lowering of intensity
- Final lengthening

Turn-taking Cues: Linguistic

- **Cue phrases**: E.g. "Well...", "I mean..."
- **Semantic completeness**: Whether or not an utterance is the complete response to the previous turn.
  - This requires the context of the previous phrase.

Example

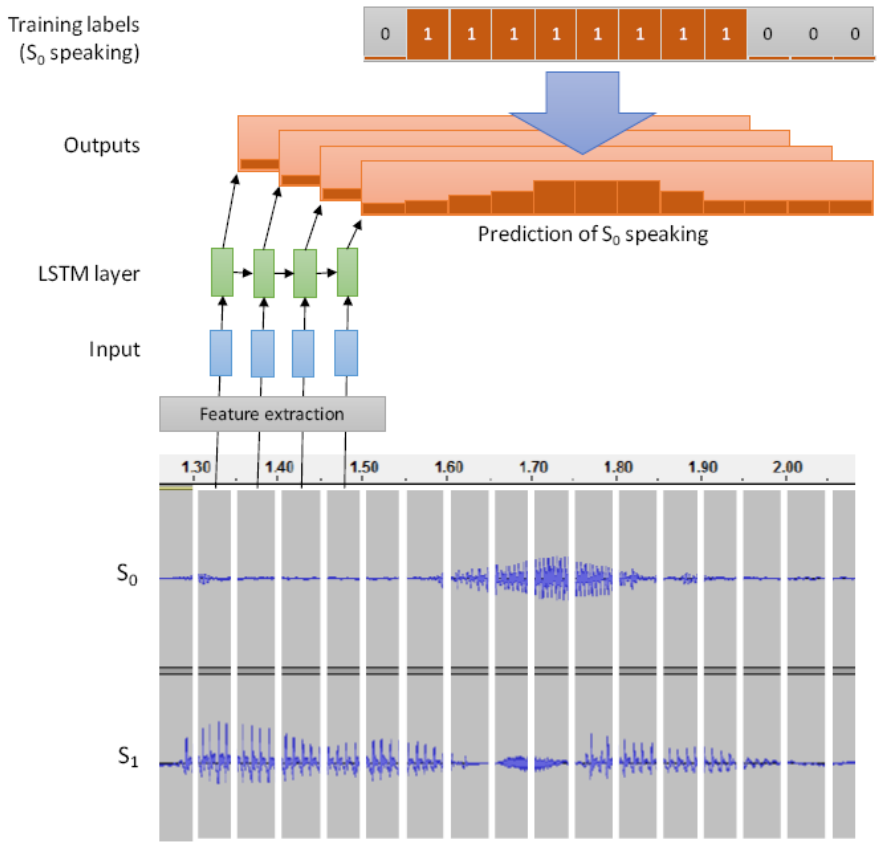
"Would you like an apple, an orange, or a banana?"
"An apple."

Traditional models based on turn-taking cues

- Many proposed models that exploit turn-taking cues.
  - Mostly posed as a classification problem: whether a period of silence in the conversation is a turn-switch point or a pause.
- Limitations:
  - Limited to making binary turn-switch decisions after a pre-defined silence threshold
  - The question of how much context is to be included in the decision-making is open ended.

Continuous turn-taking modeling (1)

- **Model**: LSTM that forms predictions at intervals of 50ms.
- **Objective**: Model probability of speaker voice activity at discrete points in the immediate future.
- **Output**: A vector of 60 predictions (3 seconds) for speech activity at points in the future.
- **Input**: Acoustic/linguistic features extracted from both speakers.



Speech activity prediction using LSTMs Skantze [2017]

Continuous turn-taking modeling (2)

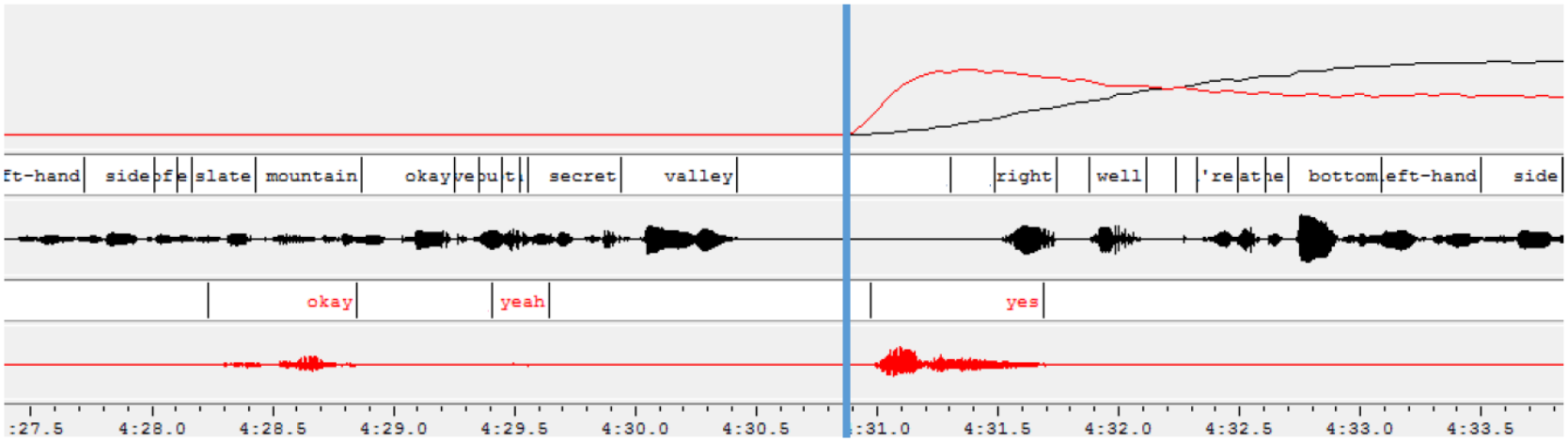
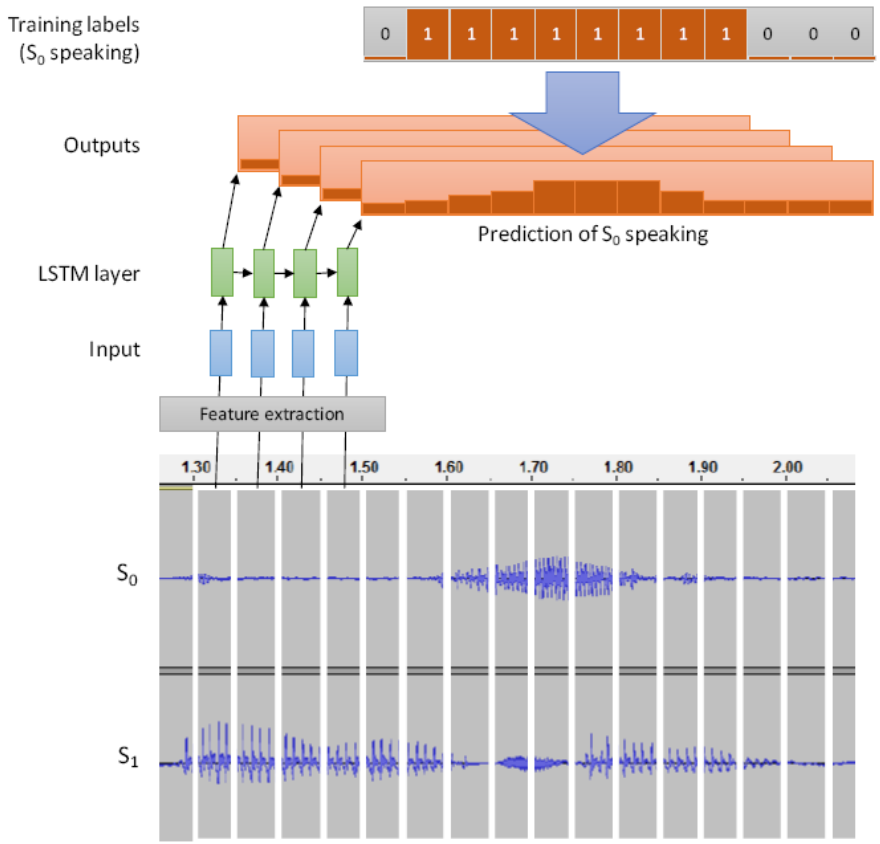


Figure: The model predicts that the red speaker will give a (short) response, and the black speaker will continue later on. Skantze [2017]

- Captures general information about turn-taking in the data.
- When applied to standard turn-taking decisions outperforms previous non-continuous approaches.
- Even out-performs humans.

Drawbacks of using single LSTM

- Features need to be all processed at the same temporal rate.
- Rate of information for linguistic features is variable. Results in a sparse linguistic feature representation.
- Semantic completeness difficult to learn because of long-term dependencies (vanishing gradients).



Speech activity prediction using LSTMs Skantze [2017]

Multiscale approach

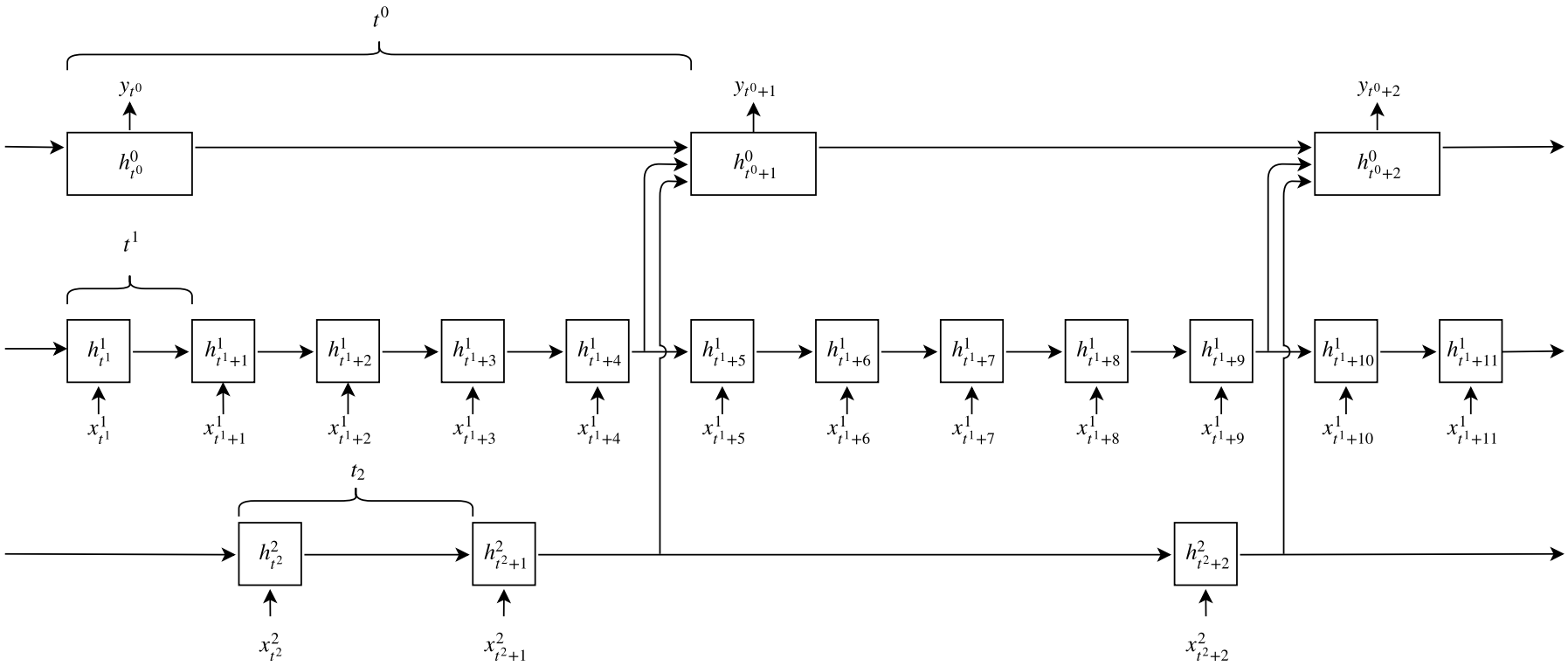


Figure: Multiscale RNN Architecture, Roddy [2018]

Results on HCRC Maptask

Majority-class baseline	0.421
Human performance	0.709
Skantze, Sigdial, 2017	0.762
Roddy et. al., Interspeech, 2018	0.813
Roddy et. al., ICMI, 2018	0.8553

Table: F-score comparisons for predicting turn-shifts at 500ms pauses.

Main Points

- Continuously predicting future speech activity using LSTMs is a promising direction for designing naturalistic turn-taking models.
- Multiscale approach (separate LSTMs for modalities) facilitates the learning of long-term dependencies.
- Papers:
  - Skantze. Towards a General, *Continuous Model of Turn-taking in Spoken Dialogue using LSTM Recurrent Neural Networks*, SIGDIAL, 2017.
  - Roddy, Skantze, Harte. *Investigating Speech Features for Continuous Turn-Taking Prediction Using LSTMs*, Interspeech, 2018.
  - Roddy, Skantze, Harte. *Multimodal Continuous Turn-Taking Prediction Using Multiscale RNNs*, ICMI, 2018.
- Code: www.github.com/mattroddy

Bibliography

Sean Andrist, Dan Bohus, Ece Kamar, and Eric Horvitz. What Went Wrong and Why? Diagnosing Situated Interaction Failures in the Wild. In *International Conference on Social Robotics*, pages 293–303. Springer, 2017.

Dan Bohus and Eric Horvitz. Multiparty turn taking in situated dialog: Study, lessons, and directions. In *Proceedings of the SIGDIAL 2011 Conference*, pages 98–109. Association for Computational Linguistics, 2011.

Agustín Gravano and Julia Hirschberg. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634, July 2011. ISSN 08852308.

Gabriel Skantze. Predicting and Regulating Participation Equality in Human-robot Conversations: Effects of Age and Gender. pages 196–204. ACM Press, 2017. ISBN 978-1-4503-4336-7.