UTEP Technical Report UTEP-CS-18-36

Analysis of Prosody Around Turn Starts

Gerardo Cervantes, Nigel G.Ward

Department of Computer Science
University of Texas at El Paso
500 West University Avenue
El Paso, TX 79968-0518

April 13, 2018

We are interested in enabling a robot to communicate with more natural timings: to take turns more appropriately. LSTM models have sometime been effective for this, but we found that this to be not helpful for some tasks. This technical report we look for factors that may explain this difference, by examining statistically the prosodic feature values in the vicinity of turn shift in the data. We observe that the apparent informativeness of prosodic features varies greatly from one dataset to another.

1 Motivation

In 2017 Skanze showed that a Long Short-Term Memory (LSTM) neural network could perform as well as humans, on Maptask data, and we have replicated this result [Anderson et al., 1991, Aguirre et al., 2018]. However, when we applied the same method to a Japanese dataset, the performance was very low. There are many possible explanations for this [Aguirre et al., 2018], but in this technical report, we do a low-level investigation of how the prosody compares.

2 Methods

To examine the predictability of the two data sets, we plotted the prosodic features in the vicinity of each speaker's turn starts.

A turn start is the initial point in which the turn is taken. A conversation between two speakers has many turn starts. Our plots are based on, across all the turn starts, the average of the features. Specifically we plotted the prosodic features over a window from 2.5 seconds before the turn start to 2.5 seconds after. We found the locations of the turn starts, from the human-annotated labels.

For features, we used those computed by the Midlevel Prosodic Features Toolkit [Ward, 2017]. These features are computed over 10 millisecond windows, and are designed

dataset	Maptask				Toyota data	
role of turn-start speaker	giver		follower		robot	
speaker whose features are plotted	same	other	same	other	same	other
average over all	1	2	3	4	5	6
average only when speaking	7	8	9	10	11	12

to be robust and meaningful in the face of noise and missing pitch values. Each feature is z-normalized on a per audio file basis. We then plot the average of the features in the vicinity of these turn starts.

At each offset t from the turn start, there are two possible ways to compute the average. First we can compute the average for t relative to all starts, regardless of whether or not there is speech at that time. Second, we can compute the average only when there is speech at offset t from a start. This is potentially more informative because prosodic features for pitch height, pitch range, etc may not have meaningful values when no one is speaking.

Plots using the first method are in Appendix A; those with the second method in Appendix B. Each plot in Appendix B also includes a plot of the number of turn starts in whose vicinity speech was present at offset t.

3 Corpora

We used two datasets in this analysis: the Maptask dataset and dataset from Toyota.

The Maptask dataset contains conversations between two speakers[Anderson et al., 1991]. In each conversation, one speaker has a map with a route; the other speaker has only a map. One speaker then has the task of explaining the route to the other. In these conversations one person thus naturally comes to lead the discussion. We refer to these as the giver and follower respectively.

The Toyota dataset contains conversations between a user and a robot. The robot is controlled using Wizard-of-Oz (WOZ) to decide when it should take the turn. Thus the user has a conversation with the robot.

The table summarizes all the plots. Note that we generally expect the “other” features to be more informative, as the speaker starting a new turn will generally have been silent up to that point.

4 Observations on the Maptask Data

1. Figures 1-4: Intensity, unsurprisingly, dips starting around 500ms before the turn start. This happens for both speakers. The tendencies for the other features are much weaker.

2. Figures 2 and 8: A predictive feature for the giver’s turn starts could be the decrease in lengthening by the follower starting around 300 ms before the turn start, as seen in Figure

2. However this feature is not robust, so this may also reflect creaky voice. Furthermore, Figure 8 suggests that there is actually an increase, that is, turn-final lengthening.

5 Observations on the Toyota Data

1. Figures 6 and 12: There is a peak disalignment bump for the user. In addition the lengthening of the speaker decreases around 500 ms before the robot turn start. Again this may be due to creaky voice.
2. Figure 11: Pitch peak disalignment by the robot also seems to be a predictive feature, around 600 ms before it starts a new turn.
3. Figure 12: The user exhbits low pitch height from around 2000 ms to around 600 ms before the robot's turn start.

6 Observations Across Both Datasets

1. Figures 7-11: Lengthening is a predictive feature, occuring around 300 ms before the turn start.
2. Figure 7, 9, and 11: There are apparent sudden dips in all prosodic features right before the turn start happens. However this is not meaningful, and simply reflects the fact that there are no instances of speech at those times. Specifically, in these datasets there are few overlaps, and both speakers are typically silent before a turn start.

Acknowledgment

We thank the Toyota Motor Corporation for their support.

References

[Aguirre et al., 2018] Aguirre, D., Ward, N., Cervantes, G., and Fuentes, O. (2018). An improved deep-learning model of turn-taking in spoken dialogue. In *Sigdial, submitted*.

[Anderson et al., 1991] Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The HCRC map task corpus. *Language and Speech*, 34:351–366.

[Ward, 2017] Ward, N. G. (2017). Midlevel prosodic features toolkit. https://github.com/nigelgward/midlevel.

7 Appendix A

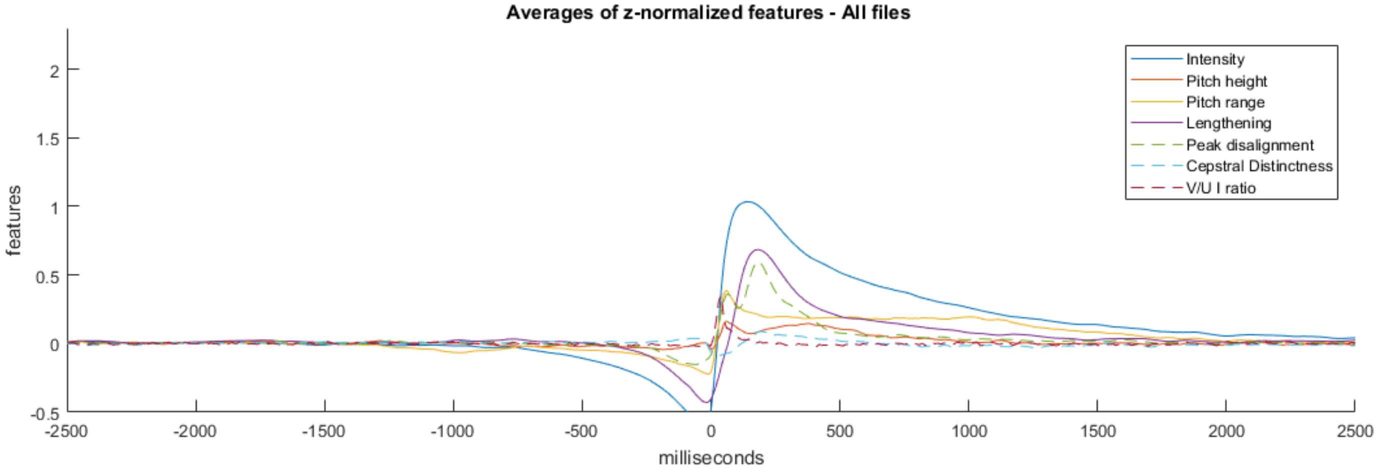

Figure 1: Maptask, giver. Values for features of the giver around his turn starts, at time 0.

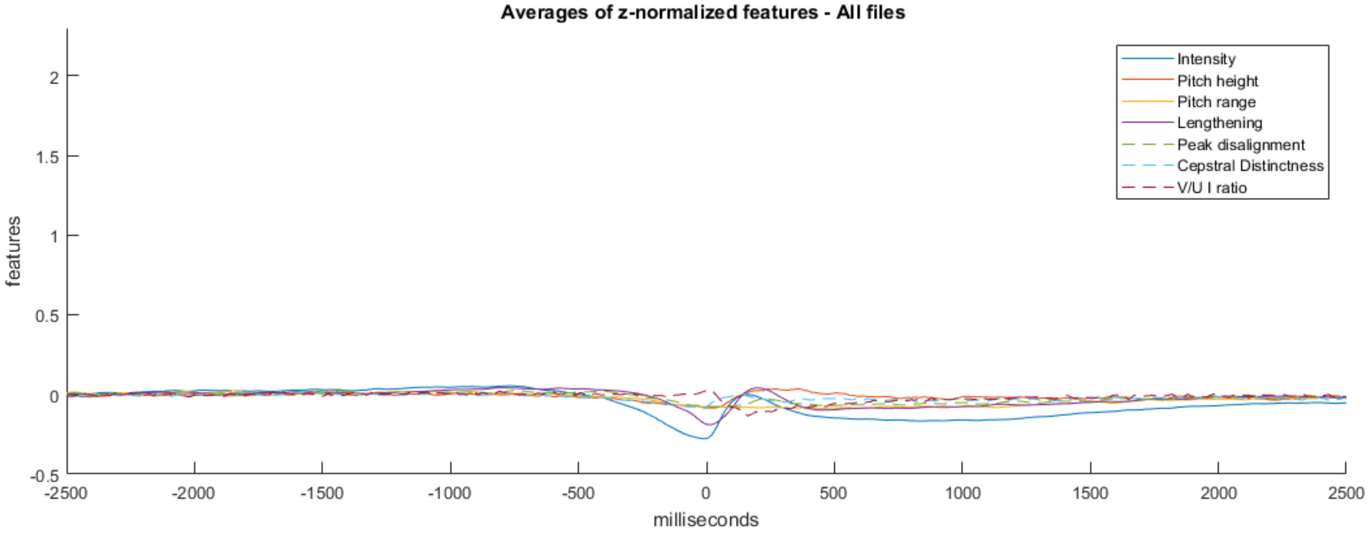

Figure 2: Maptask, giver. Values for features of the follower in the vicinity of the giver's turn starts.

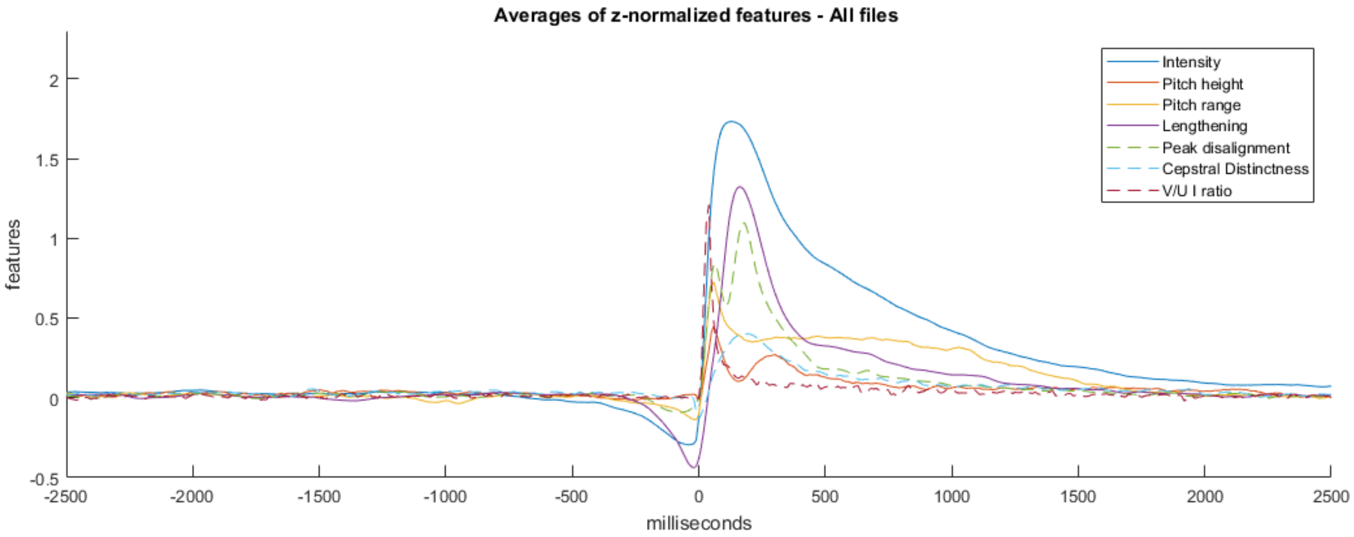

Figure 3: Maptask, follower. Values for features of the follower around his turn starts.

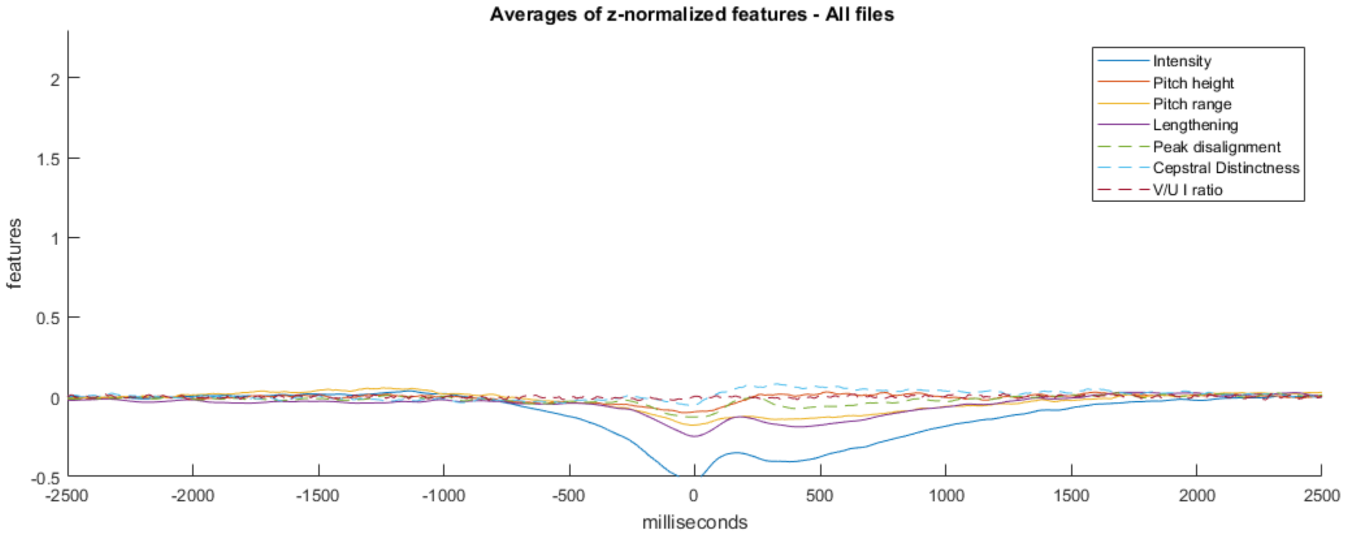

Figure 4: Maptask, follower. Values for features of the giver, in the vicinity of turn starts by the follower.

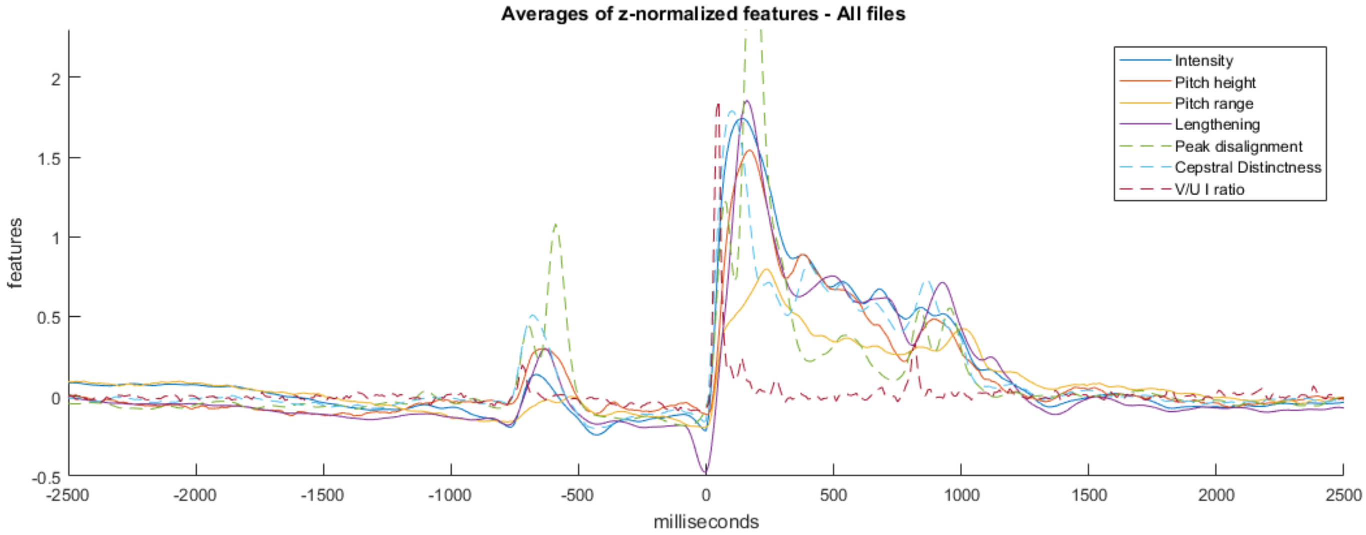

Figure 5: Toyota data. Values for features of the robot in the vicinity of robot turn starts.

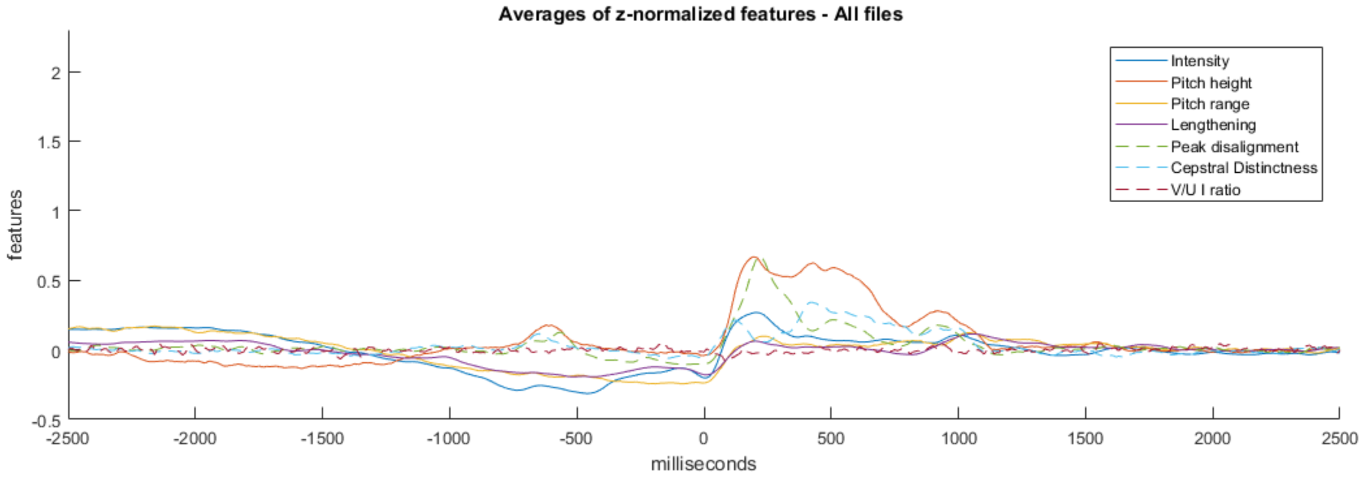

Figure 6: Toyota data. Values for the features of the user in the vicinity of robot turn starts.

8 Appendix B

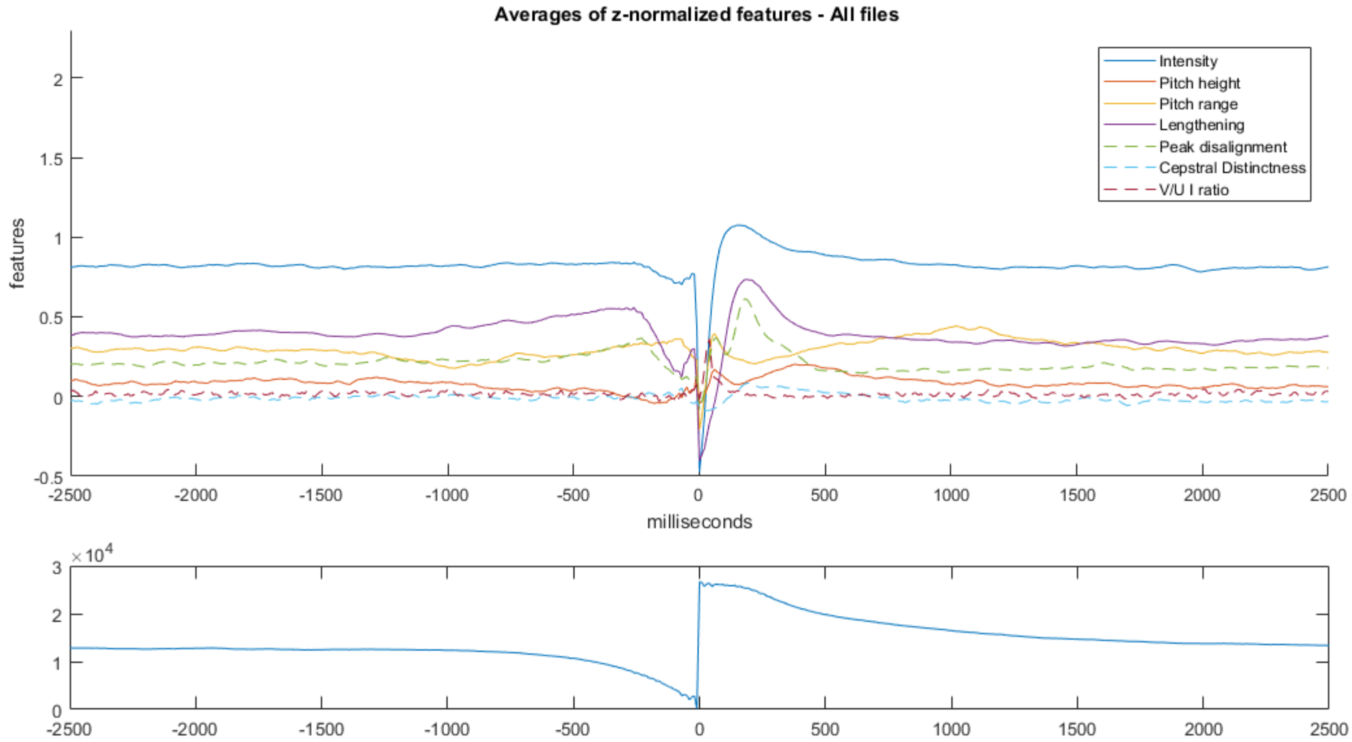

Figure 7: Maptask, giver. Values for features of the giver around his turn starts, at time 0, averaged only for times when speaking.

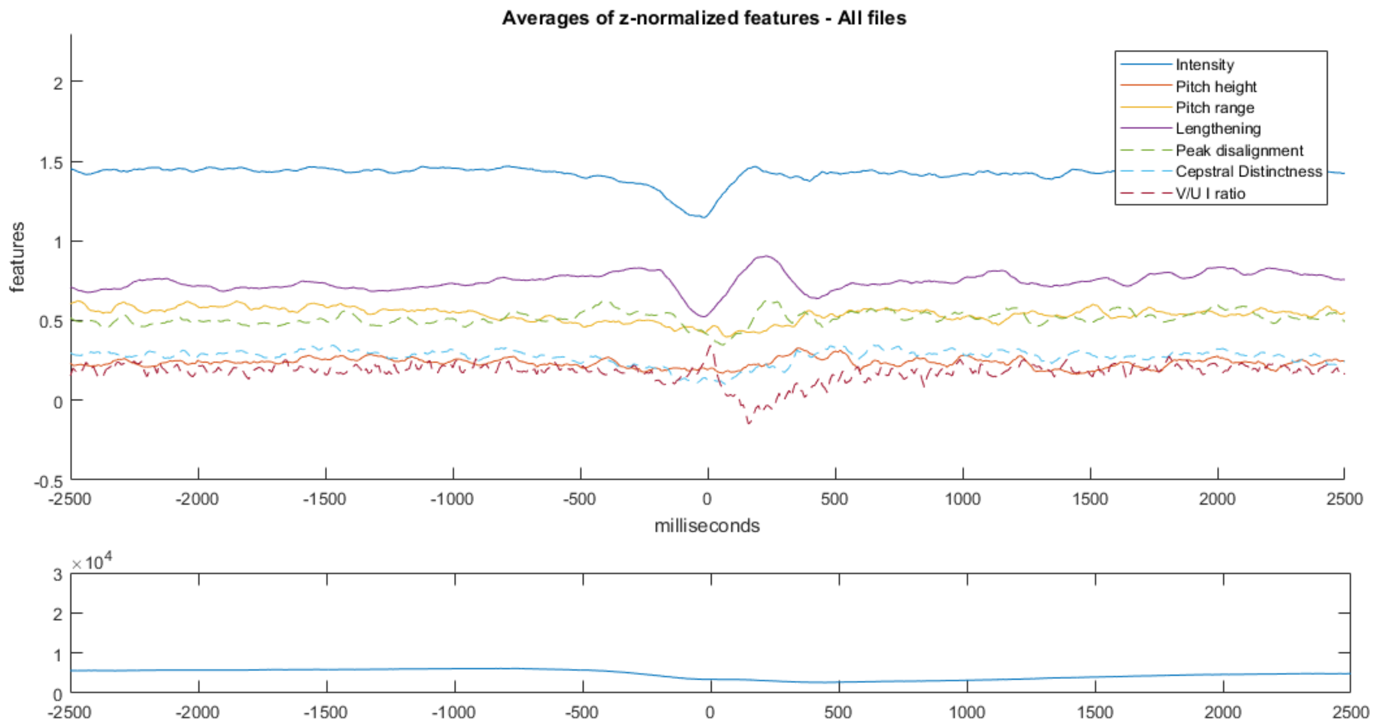

Figure 8: Maptask, giver. Values for features of the follower, as above.

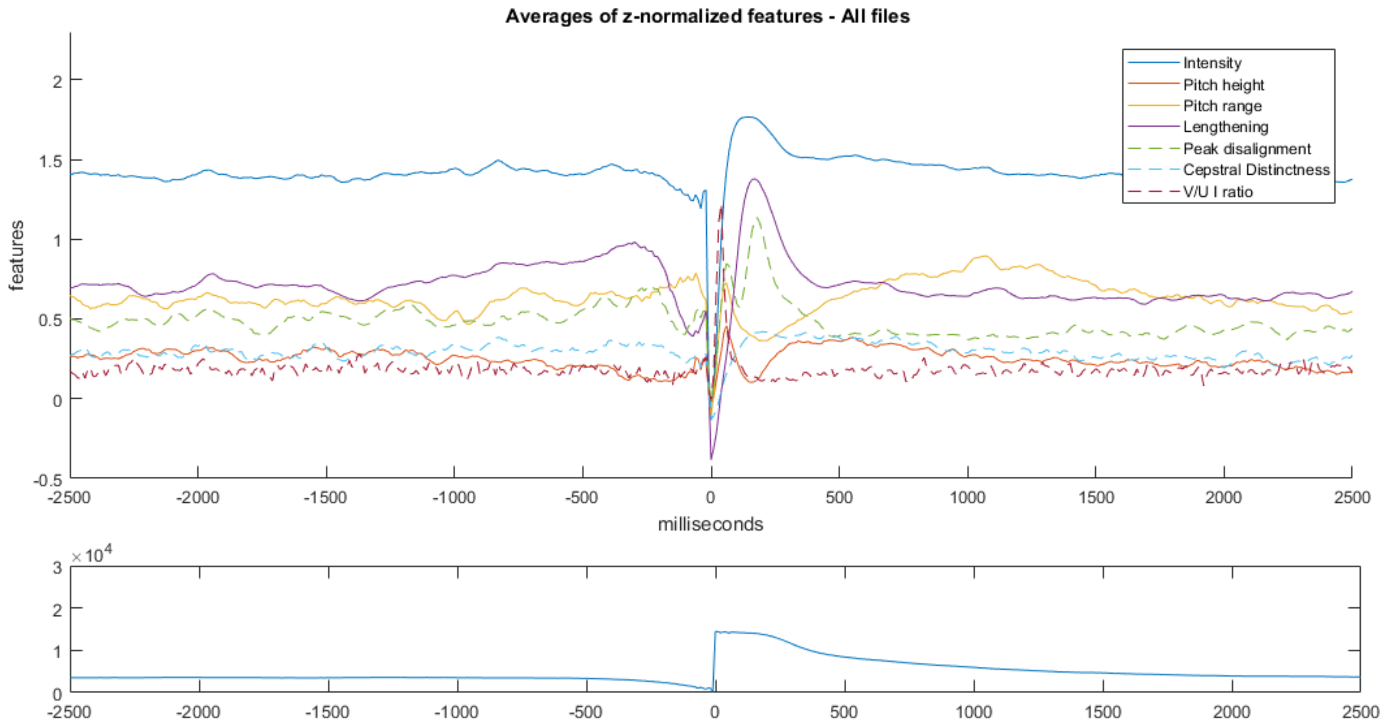

Figure 9: Maptask follower. Values for features of the follower in the vicinity of their turn starts, as above.

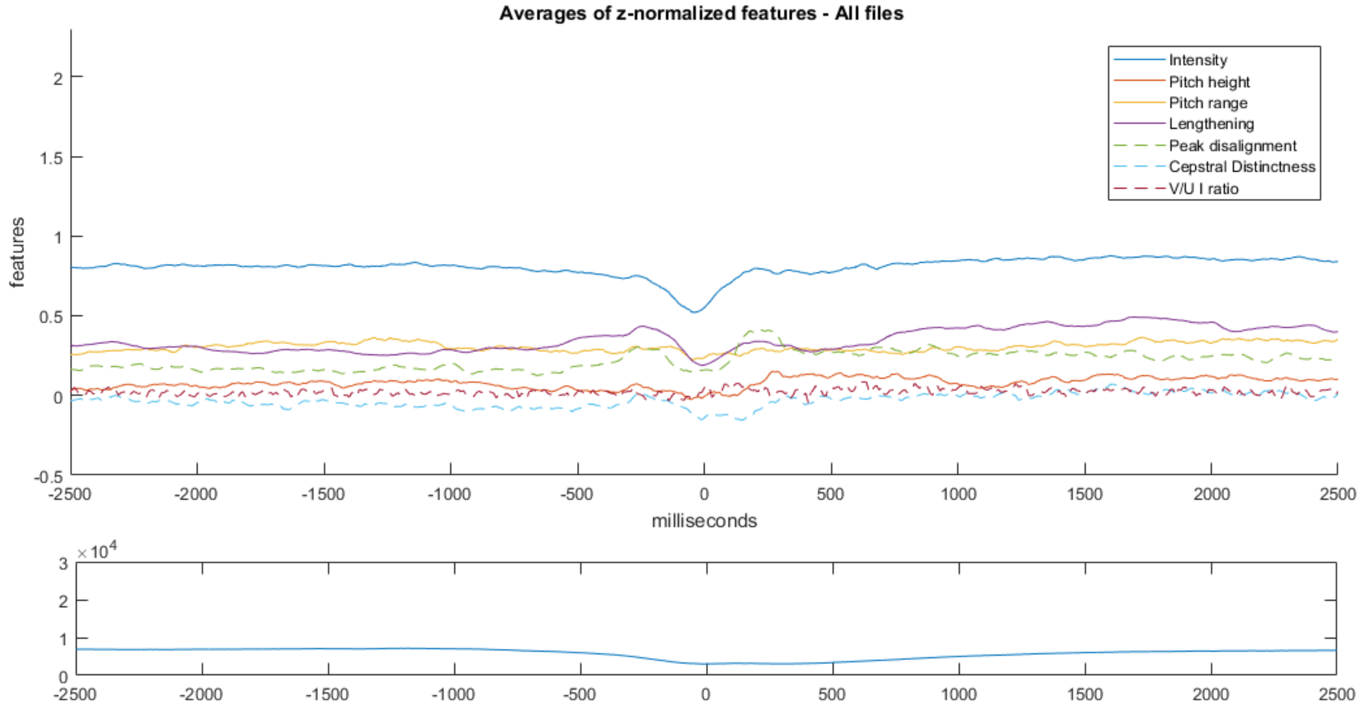

Figure 10: Maptask follower. Values for features of the giver, as above.

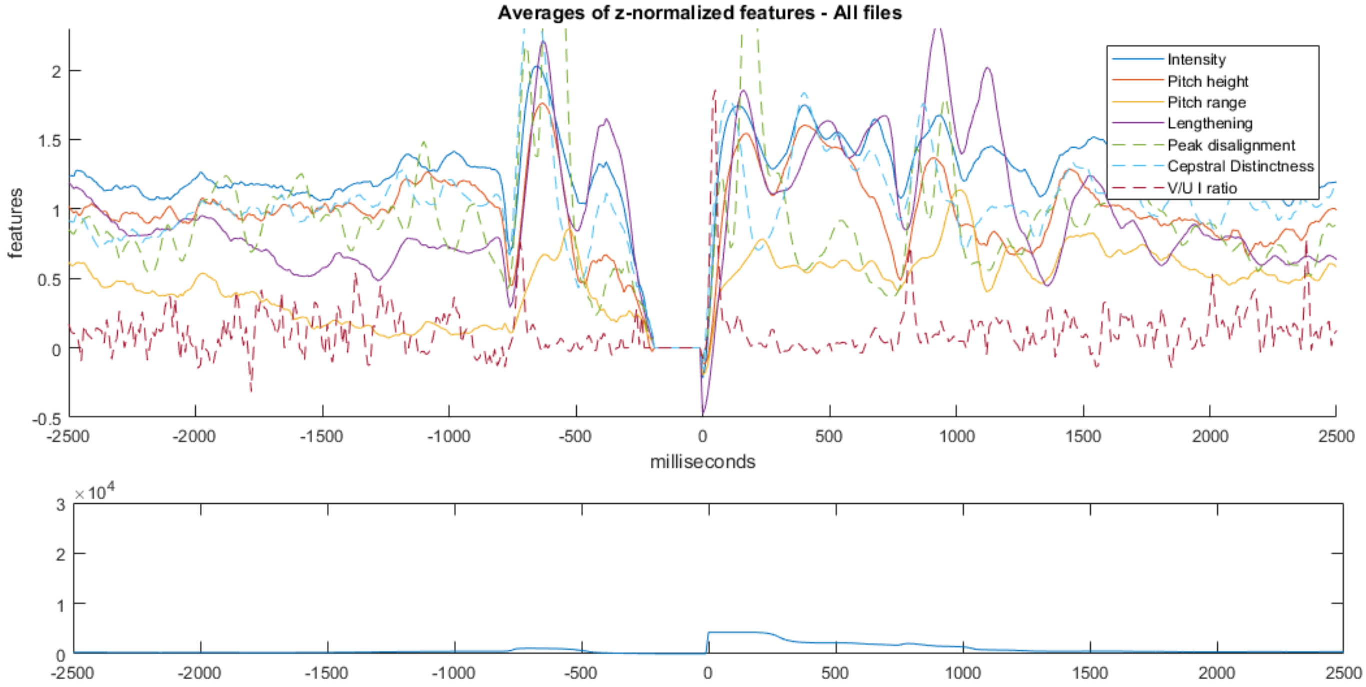

Figure 11: Toyota data. Features from the robot in the vicinity of robot turn starts, averaged only over times when the robot is speaking.

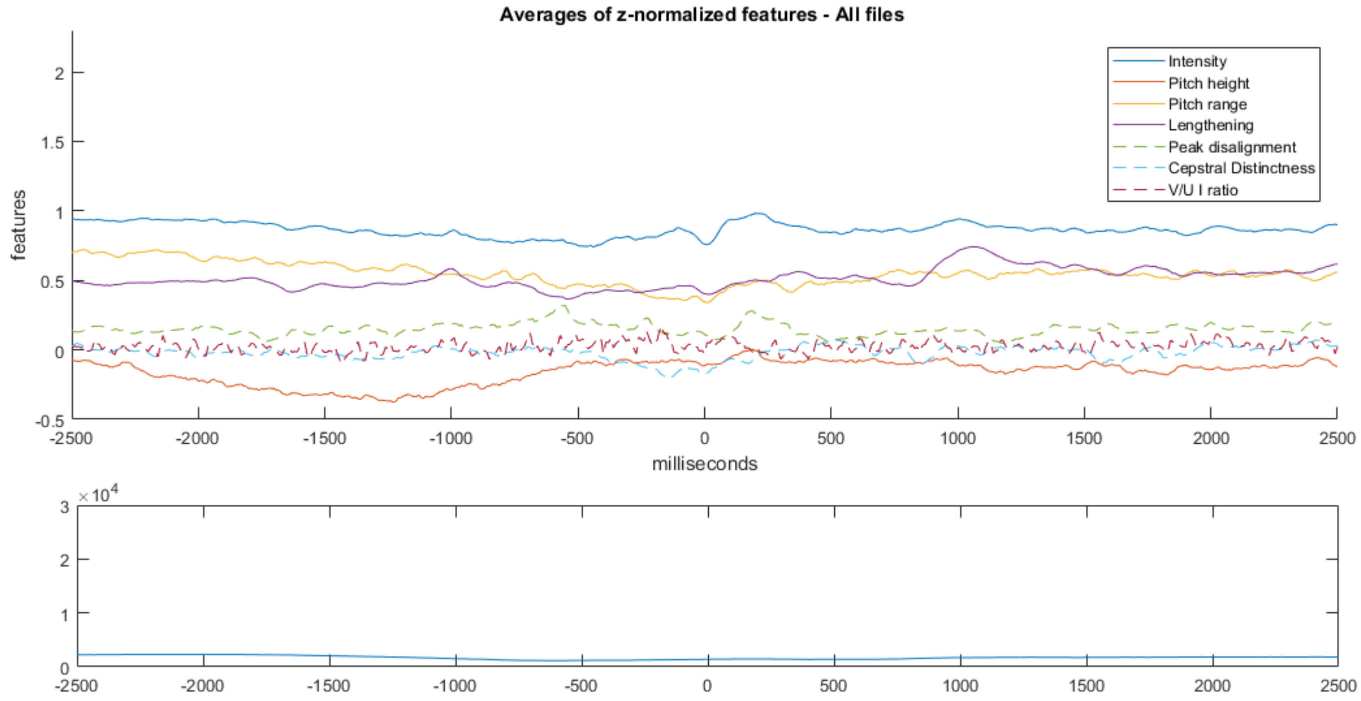

Figure 12: Toyota data. Features from the user robot in the vicinity of robot turn starts, averaged only over times when the user is speaking.