MTurk Pupillometry Analysis (Exposure)

Zach Burchill

February 21, 2018

Contents


```
[1] "Does the data pass the sanity check?"
```

```
[1] "Sanity check ok!"
```

```
[1] "Sanity check ok!"
```

1 Overview of study

1.1 Verbal summary

1.2 Experiment 1

At its current stage, is basically a type of over-the-web near-replication of Clarke & Garrett (2004), but with slightly more trials (instead of **4/6 trials per block**, there are now **8**), and featuring an Indian-accented talker. This experiment was meant to make sure we could find an adaptation effect with our stimuli. For a more detailed analysis, see the other results report.

1.3 Experiment 2

The current report focuses on the findings of Experiment 2. Experiment 2 was very similar to Experiment 1 in both task and stimuli, but the visual-probe blocks were interleaved with approximately 7.5 minutes total of the exposure talker reading sections from a short story. We did this so that participants would have approximately twice as much exposure time to the talkers' voices, and because in the final pupillometric experiment we want to be able to track pupil size while participants are passively listening to accented speech.

1.3.1 Runs

Due to findings in our comparison between Exp2 and Exp1 (detailed in the overview results report) we ran Exp2 again with 64 subjects, but with different positions for the items. We'll therefore refer to the original run as "Exp2v1" and the second as "Exp2v2". This results report will not only give the details of Experiment 2, but allow us to compare these two runs and see the effect of the position of the items.

Currently, the exclusion criteria for Experiment 2 were applied on both runs lumped together.

1.4 Total number of subjects

As shown in the tables below, we start with 64 participants per condition, the vast majority of whom were monolingual English speakers who used (in-ear or over-ear) headphones to complete the task.

Table 1: Total number of subjects per condition and Run

Condition	n
accented	64
unaccented	64

Table 2: Total number of subjects by language background and audio equipment type.

LgBackground	AudioType	n
monolingual	in-ear	49
monolingual	over-ear	67
monolingual	external	1
other	in-ear	4
other	over-ear	4
	in-ear	2
chinese	in-ear	1

Figure 1 shows the distribution of subjects' mean raw response time by Block and Condition. Note that for the majority of subjects in each condition, the mean RT in each block was (considerably) less than 2.5 seconds. However, a one or two subjects registered extremely slow RTs (e.g., mean RT ~4 seconds in one or more blocks), which was likely due to divided attention (e.g., multi-tasking during the experiment) or to technical issues affecting the recording of RTs.

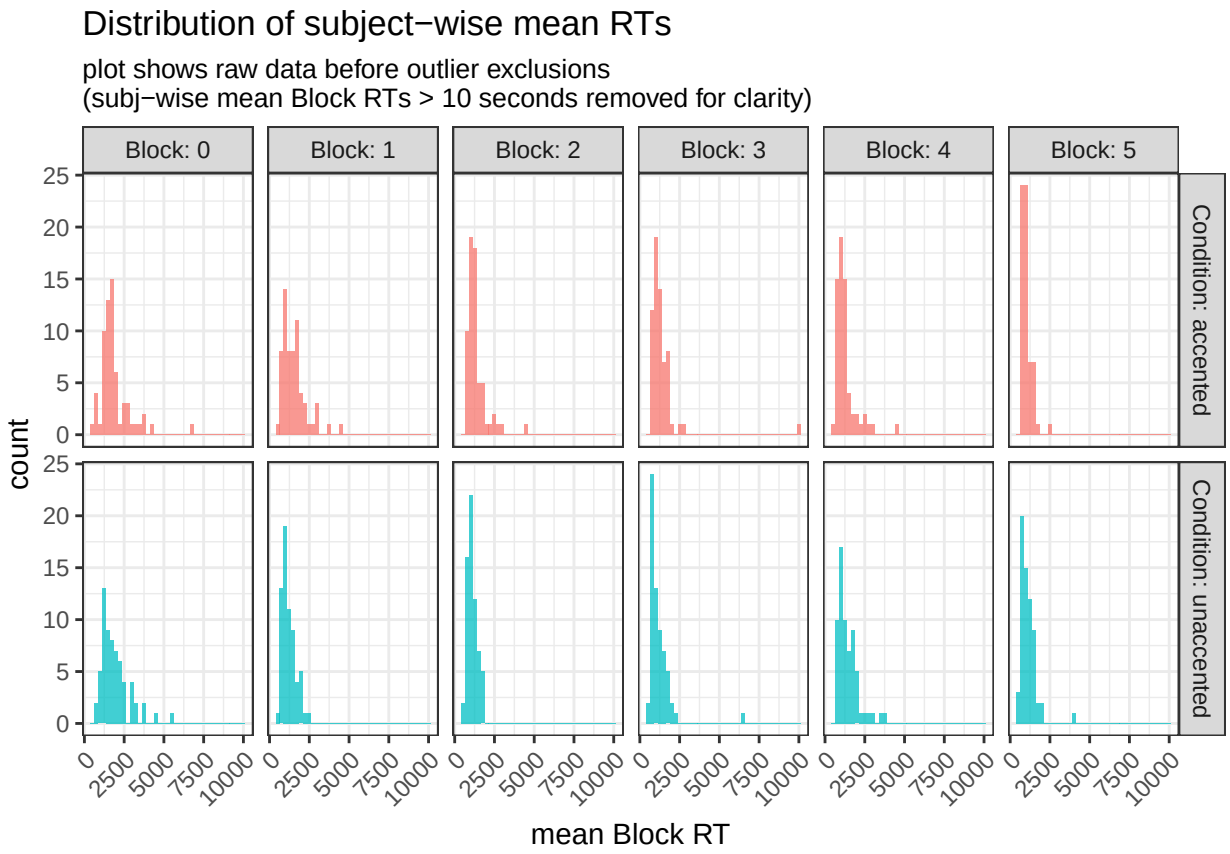


Figure 1: Distribution of subjects mean RTs by Block and Condition, prior to outlier exclusions.

2 Data preparation

2.1 Outlier Exclusion

Outlier exclusion was a multi-step process. The full set of exclusion criteria and the % of data lost for each criterion at each step are summarized in 2.4.

2.1.1 Step 1: Ineligible participants

The first step of outlier exclusions was to exclude participants who did not meet participation criteria. We excluded the following:

1. **Language background:** non-monolingual English speakers
2. **Audio equipment:** participants who did not use (in-ear or over-ear) headphones
3. **Accent familiarity:** participants who reported high familiarity with Indian-accented speech
    - subjective report of hearing an accent like the one in this experiment regularly or "all the time"

The number of exclusions based on these eligibility criteria was similar across conditions (see Table 3).

We additionally implemented an exclusion criterion based on task performance:

4. **"Cheating":** participants with any block mean RT < 200ms
5. **Task performance:** participants with mean RT in non-practice block > 3 SDs from Condition mean

NOTE: We did *not* exclude participants based on mean RTs in the practice block, whereas our previous work had. Additionally, our previous work *did not exclude* participants with block means < 200ms

This fourth criterion was an attempt to identify and remove subjects who consistently registered slow response times because they did not perform the task faithfully (e.g., multi-tasking) or because their computer equipment did not provide reliable recording of RTs over the web. **Again, exclusion rates are relatively similar across conditions** (see Table 4).

Figure 2 shows the distribution of RTs by condition and block *after* removing ineligible participants.

Table 3: Number of subjects excluded per condition based on language background, audio equipment usage, and accent familiarity.

Condition	Freq
accented	15
unaccented	13
TOTAL	28

Table 4: Number of participants excluded based on mean RTs

Condition	Freq
accented	3
unaccented	3
TOTAL	6

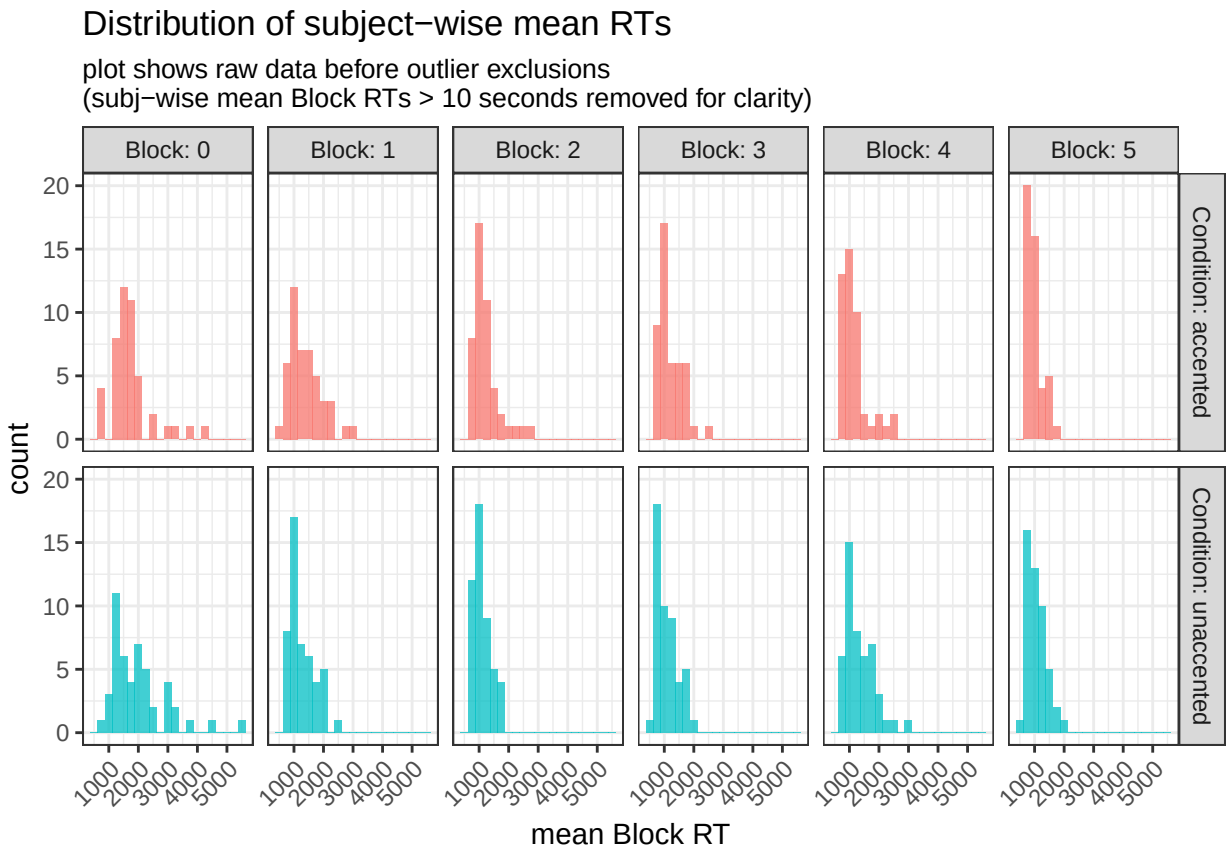


Figure 2: plot shows all data after outlier exclusion Step 1

2.1.2 Step 2: Trials with extreme RTs

The second step of outlier removal was to exclude **trials** with atypical RTs. We omitted trials based on the following criteria:

- RTs less than 200ms
  - based on the assumption that it takes approx. 200ms to program a motor response; hence RTs less than 200ms from the onset of the target stimulus reflect processing of earlier information
- RTs greater than 3 SDs from subject's mean

The proportion of trials excluded based on these criteria was similar across conditions (see Table 5).

Table 5: Proportion of trials excluded per condition.

Condition	n_subjs	n_useableTrials	n_excludedTrials	prct_excludedTrials
accented	46	2064	52	2.46
unaccented	48	2150	58	2.63
TOTAL:	94	4214	110	2.54

Figure 3 shows the distribution of raw RTs after both subject-wise and trial-wise outlier exclusion (i.e., outlier exclusions steps 1 and 2). **There are still a few slow RTs. We could consider adding an upper bound (e.g., exclude RTs > 4 or 5 seconds)?**

Figure 4 shows the difference between subjects' mean baseline (Block 5) RT when calculated before vs. after exclusion of trial-wise outliers. There are several points to make here:

- For the vast majority of subjects, trial-wise outlier exclusion doesn't affect estimation of baseline RTs
  - i.e., difference btw baseline calculation methods ~0ms
- However, **when trial-wise outlier exclusion *does* matter, it matters a lot**!
  - i.e., for subjects with a non-zero difference on these two baseline RT measures, the mean size of the difference is nearly 400ms.
  - For perspective, 400ms is several orders of magnitude larger than the expected main effect of accent in this experiment (e.g., in C&G 2004, the difference between accent and control conditions in Block 1 is ~100-150ms across experiments).
- Thus, if we don't exclude trial-wise outliers, we not only massively mis-estimate the baseline RT for a subset of subjects, we also propogate this estimation error into the rest of the data via the RT normaization procedure (experimental RTs - subject's mean baseline RT).

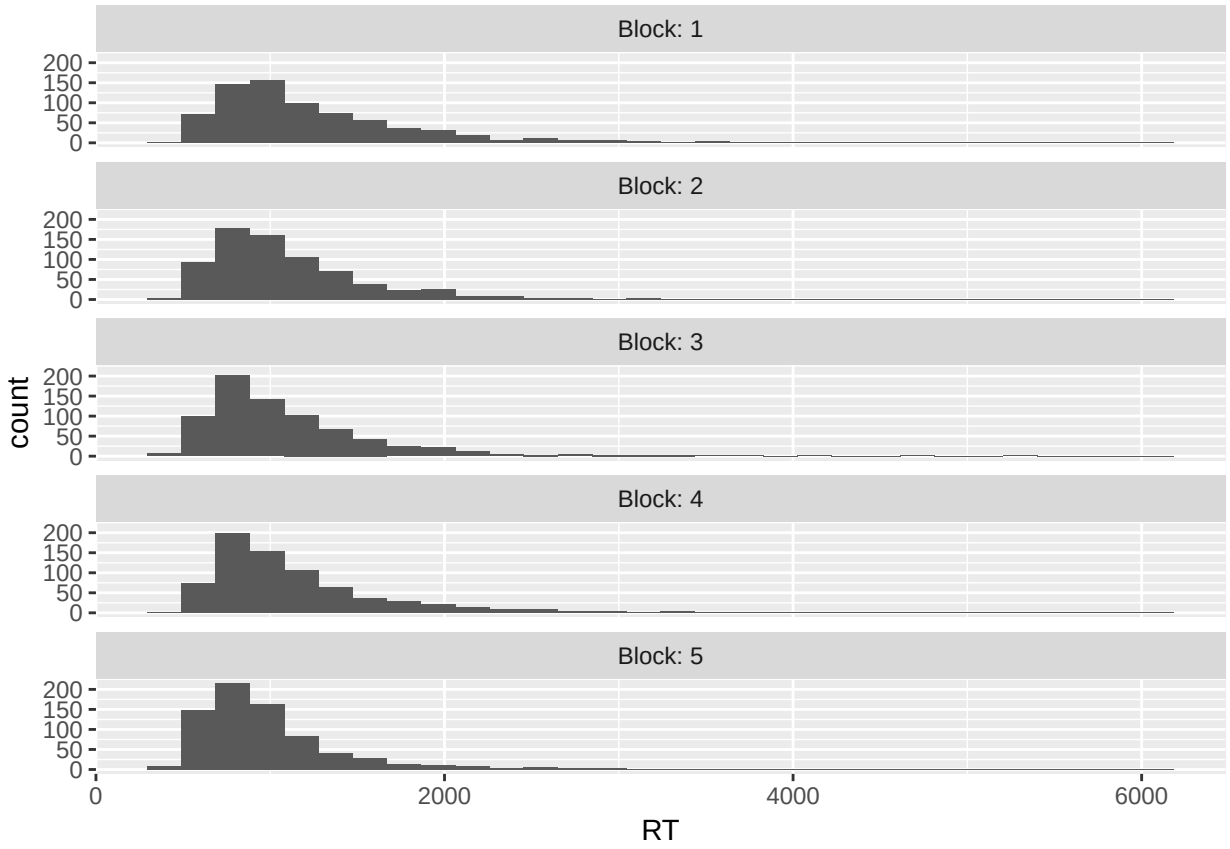


Figure 3: The distribution of raw RTs after both subject-wise and trial-wise outlier exclusion (i.e., outlier exclusions steps 1 and 2). Some very slow RTs persist.

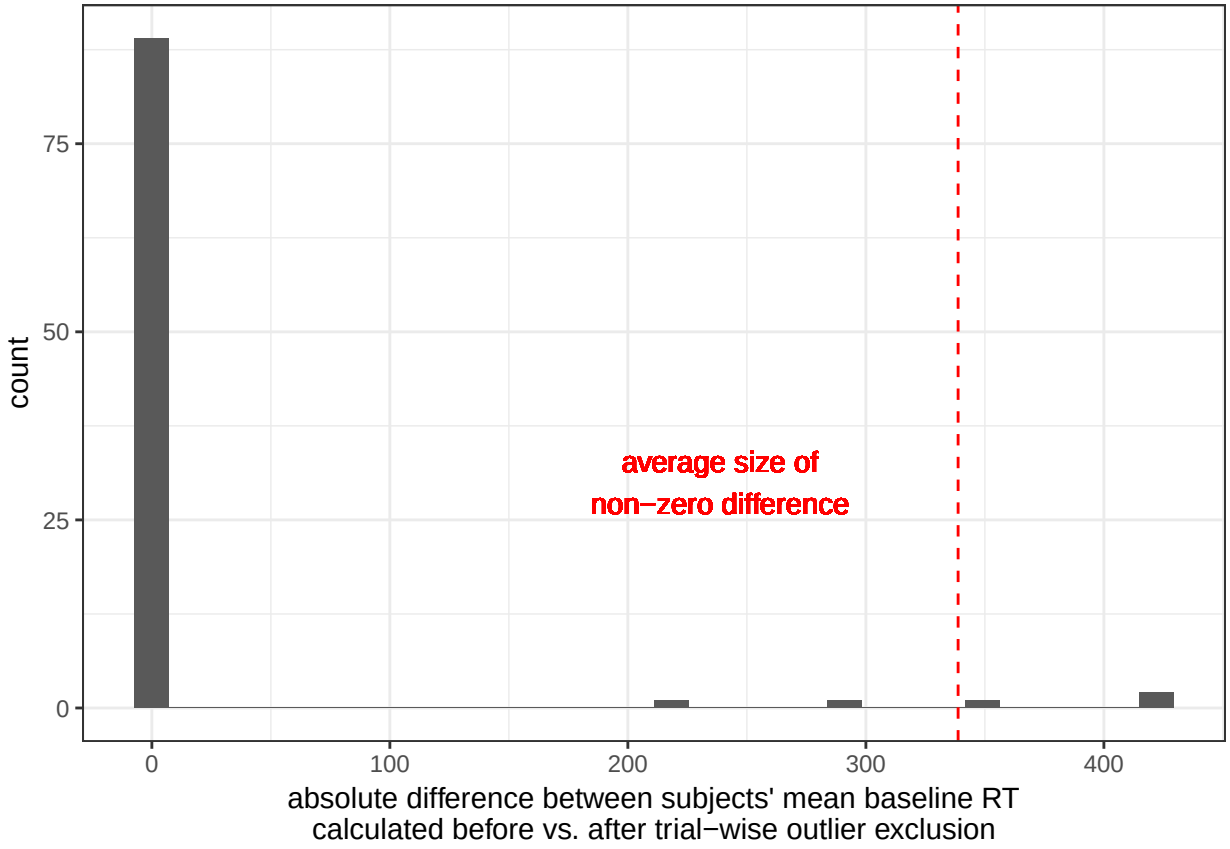


Figure 4: The absolute differences of subjects' mean RT on their baseline block with and without trial-wise exclusions. The average change for mean RTs that ARE different is marked with a dotted line. Without trial-wise exclusions, some participants' adjusted RTs would be quite skewed.

2.2 Examine RTs and Accuracy during practice and baseline (after exclusion steps 1 and 2)

Now that we've excluded extreme subject and trial outliers, we can look at the practice and baseline data to assess our high-level predictions about how participants should perform on this web-based task.

1. **One data pattern that we expect to find is that performance (both RTs and accuracy) in the practice and baseline blocks is comparable across experimental conditions.** We expect this because these blocks of the experiment were identical across conditions (i.e., native-accented stimuli presented in the clear).
    - ... *if performance in the* ***practice block*** *differs substantially across conditions*, we would need to consider whether the subjects in each condition were sampled from the same underlying population (e.g., did we run all conditions at approximately the sme time of day?).
    - ... *if performance in the* ***baseline block*** *differs substantially across conditions*, we would need to consider whether exposure to different types of speech during the main block of the experiment induced overall differences in task performance (in which case the baseline block doesn't provide a reliable condition-independent "baseline" for normalization purposes).
2. **A second data pattern that we expect to find is evidence of improvement (adaptation) over the course of the task.** One way this would manifest is faster RTs and increased accuracy in the post-experiment baseline block, relative to the practice phase.

Figure 5 shows the distribution of subject-wise mean RTs during the practice and baseline blocks as a function of exposure condition.

1. The distributions are similar across exposure conditions. Thus, listening to foreign-accented speech or speech in noise during the exposure phase did not induce weird response behavior.
2. As expected, RTs are consistently faster and less variable in the baseline block, relative to the practice block, across conditions. Thus, participants are adapting to the task.

Figure 6 shows the distribution of subject-wise mean Accuracy during the practice and baseline blocks as a function of exposure condition.

1. There's a bit of variability between conditions during the practice block – but not enough to be troubling. Performance in the baseline task is comparable across conditions.
2. Accuracy is *higher* in practice task than during the baseline task. This is the opposite of what we expected, but the decrease in accuracy may be insignificant. More likely, participants are shifting how much they weigh accuracy in the speed-accuracy trade-off as they get familiarized with the task, preferring quick responses to accuracy.

NOTE REGARDING OUTLIER EXCLUSION. So far, we haven't implemented any accuracy-based exclusion criteria. Figure 6 shows that all subjects are above chance-level accuracy in the baseline phase (except for one subject). Hence, I don't think we need to implement accuracy-based exclusions.

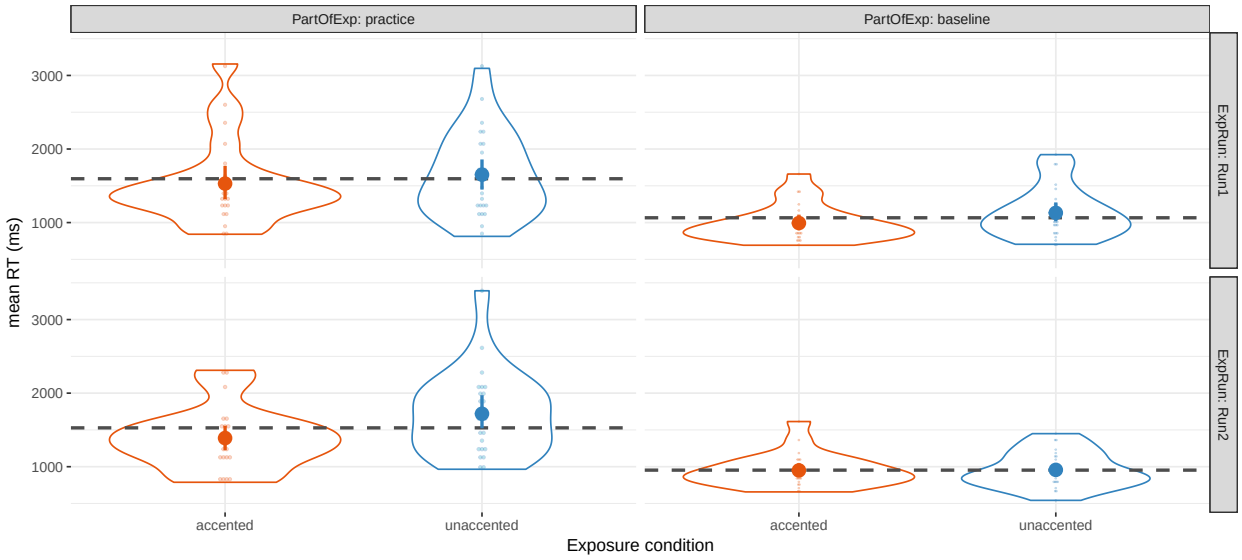


Figure 5: Distribution of subject-wise RTs during practice and baseline blocks. Each small dot indicates one subject's mean. Dashed lines indicate block means across conditions. As predicted, participants are quicker at responding at the end of the experiment.

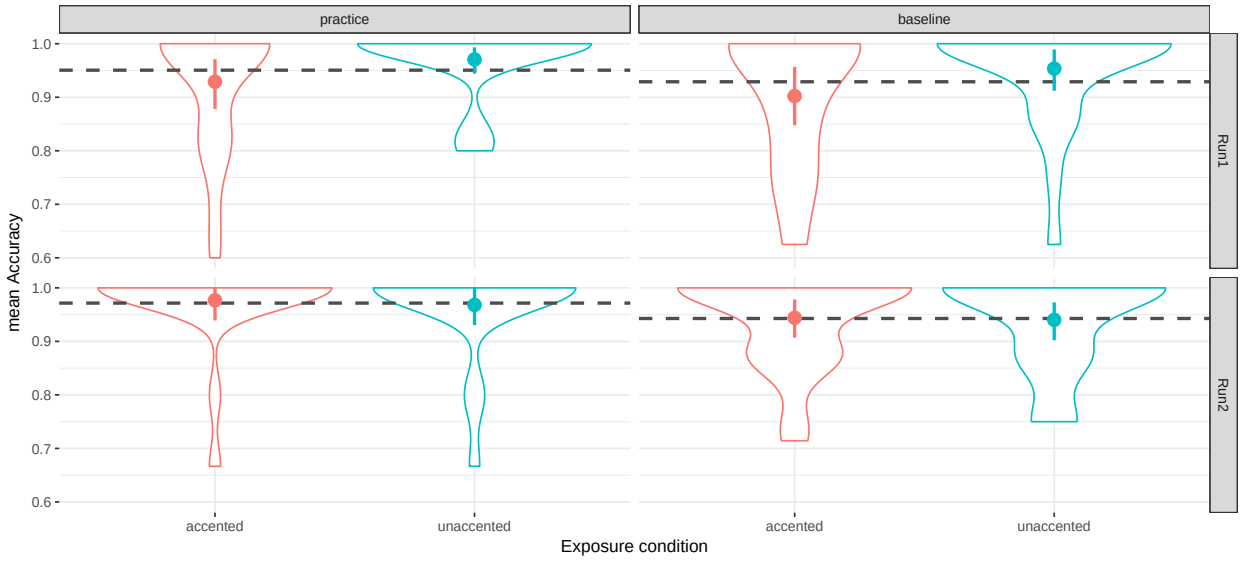


Figure 6: Distribution of subject-wise accuracy during practice and baseline blocks. Dashed lines indicate block means across conditions. Against our prediction, participants are slightly more accurate in practice. This appears insignificant, but may be due to participants changing their weighing of their speed-accuracy trade-offs.

2.3 Normalize experimental RTs relative to baseline

Now that we've completed all trial-wise RT exclusions, we can calculate *normalized* RTs that take into account each subject's baseline speed on this task. For this procedure, we adjust the RTs on each trial by subtracting out the corresponding subject's mean RT during the baseline phase. We refer to the resulting measure as *adjusted RTs*.

Now we want to check the distribution of adjuted RTs to make sure it seems reasonable, given our expectations about task performance.

Note that we expect baseline RTs to be faster on average than RTs during the experimental block, regardless of exposure condition. We expect this for two reasons. First, the baseline task occurred at the end of the experiment, after participants had adapted to the task. Second, *all* participants heard native accented speech during the baseline phase; hence, there was no need for accent adaptation during this phase.

If raw baseline RTs are, indeed, faster on average than raw RTs on experimental trials, then we expect each subject's mean *adjusted* RT (experiment RTs - baseline) to be greater than 0.

Figure 7 shows the distribution of subject-wise mean adjusted RTs during the experimental block, plotted by exposure condition. **Note that there are several subjects with a mean Adjusted RT of less than 0** (i.e., subjects who were much *slower* during baseline than during the main block).

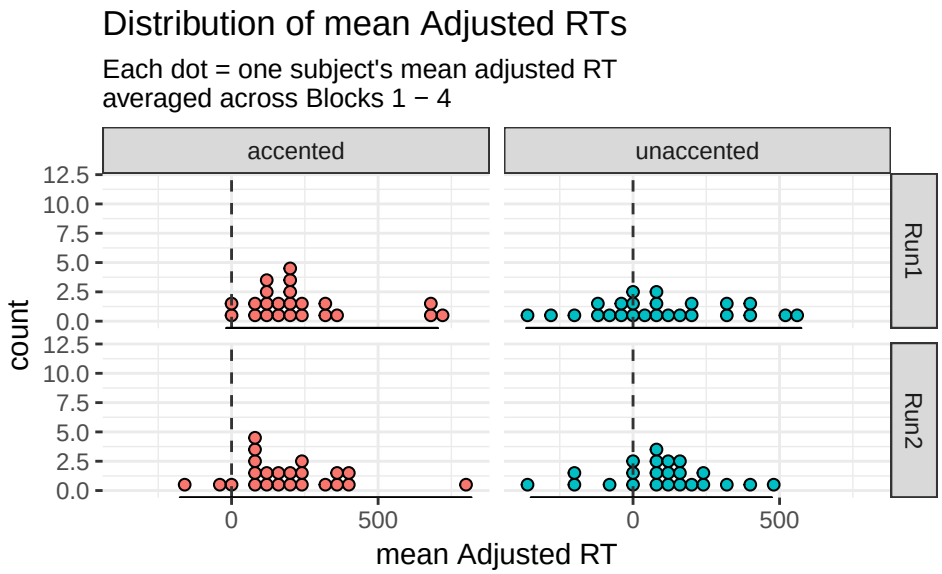

Distribution of mean Adjusted RTs
Each dot = one subject's mean adjusted RT
averaged across Blocks 1 − 4
accented
unaccented
Run1
Run2
count
12.5
10.0
7.5
5.0
2.5
0.0
0
500
mean Adjusted RT

Figure 7: ...

2.4 Summary of exclusion criteria:

- Participant-level exclusions:
  - non-monolingual English speakers
  - subjects who used audio equipment other than (in-ear or over-ear) headphones
  - subjects with high familiarity with Indian-accented English (based on subjective self report)
  - subjects whose mean RT in any block was $>$ 3 SDs from Condition mean
  - subjects who had a mean RT in any block $<$ 200 ms
- Trial-level exclusions:
  - raw RTs $<$ 200ms
  - RTs $>$ 3 SDs from subject's mean RT in each block

3 Experiment 2

3.1 Visual analysis

In this section, we'll plot the two runs separately. To see the visual analysis of them combined, see Section 5.2 in the Appendix.

Let's plot the adjusted RTs by trial (Figure 8) and by block (Figure 9).

And the mean accuracy (Figure 10).

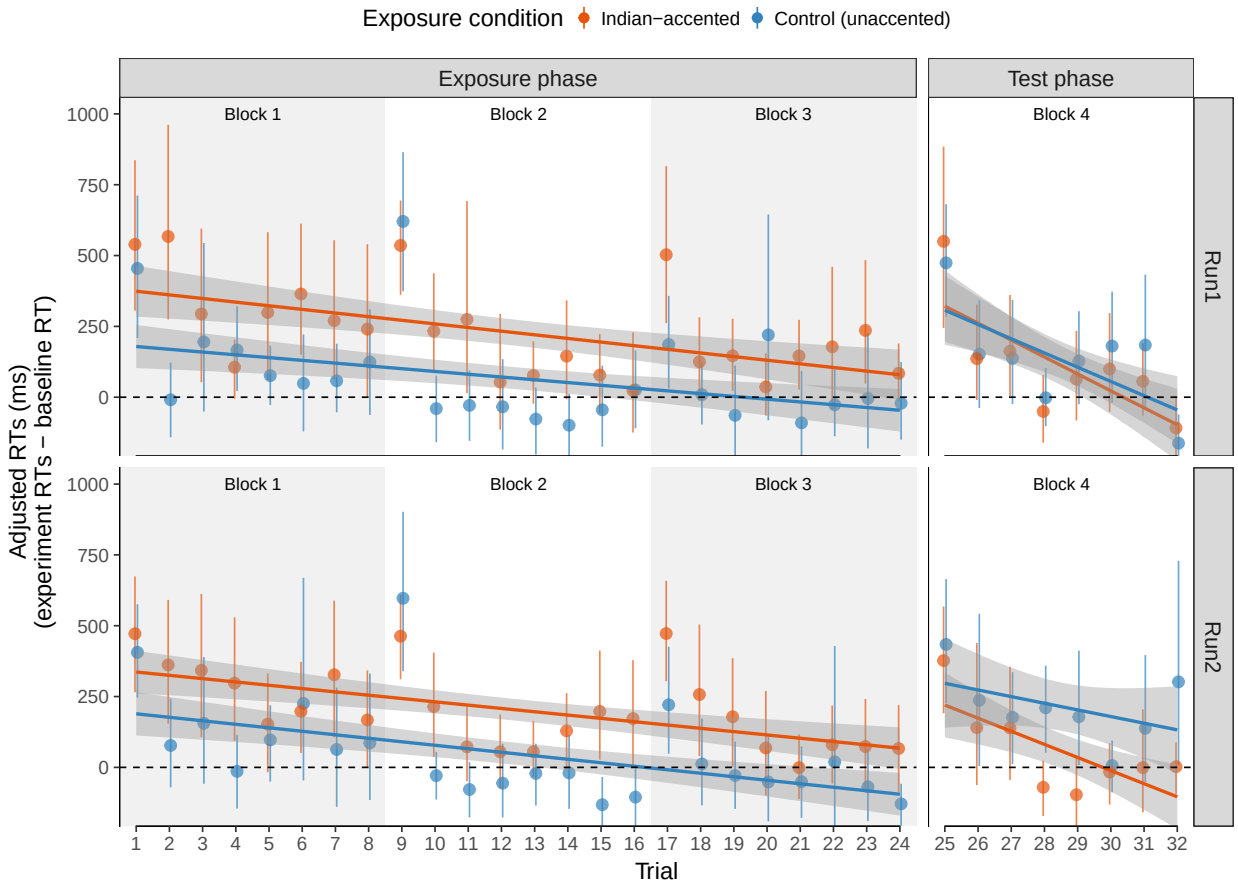


Figure 8: The adjusted RTs for correct trials in the main experiment. Note the seeming adaptation to the accented talker in test by the participants in BOTH exposure conditions, unlike what we see in Experiment 1, where adaptation seems to be confined to those in the unaccented exposure condition.

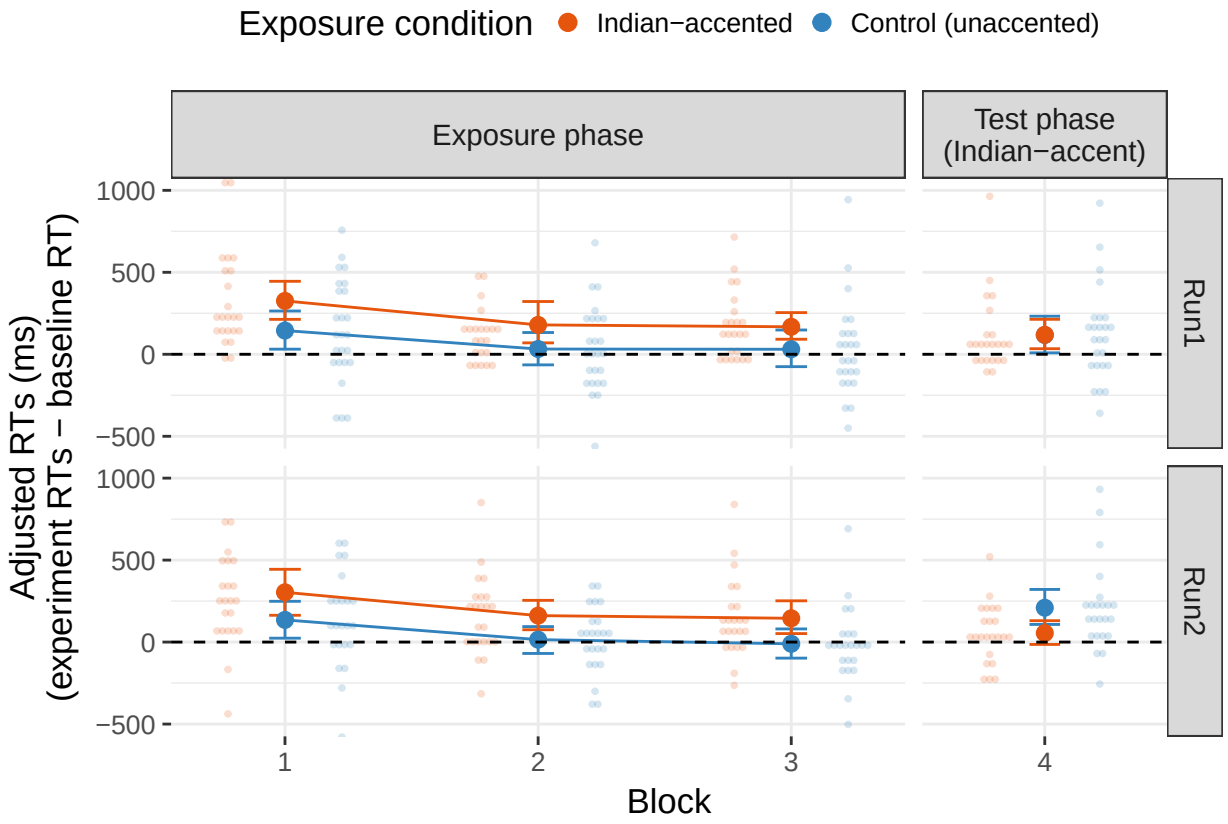


Figure 9: The adjusted mean RTs for each block of the main experiment. You can see that it follows the pattern we expected pretty much, with the unaccented participants doing poorly on the accented test block (block 4). However, unlike in Experiment 1, they seem to be doing equally poorly compared to the accented condition.

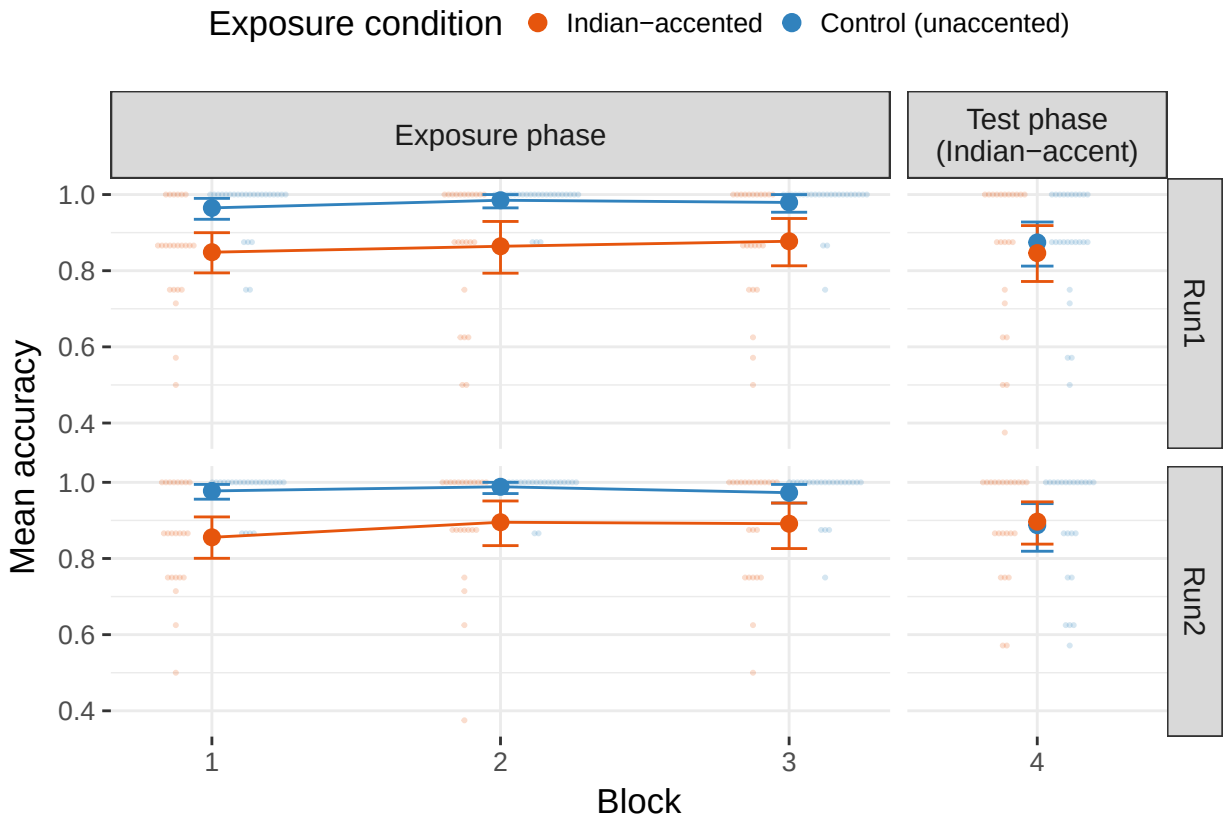


Figure 10: Each dot is a participant's mean accuracy for a particular block.

As in the overview report, we also plot the progression of adaptation within each block (Figure 11) and adaptation within each block, ignoring the first trial of every block (Figure 12). Similarly, we look at the unadjusted RTs in log-log space in Figure 13.

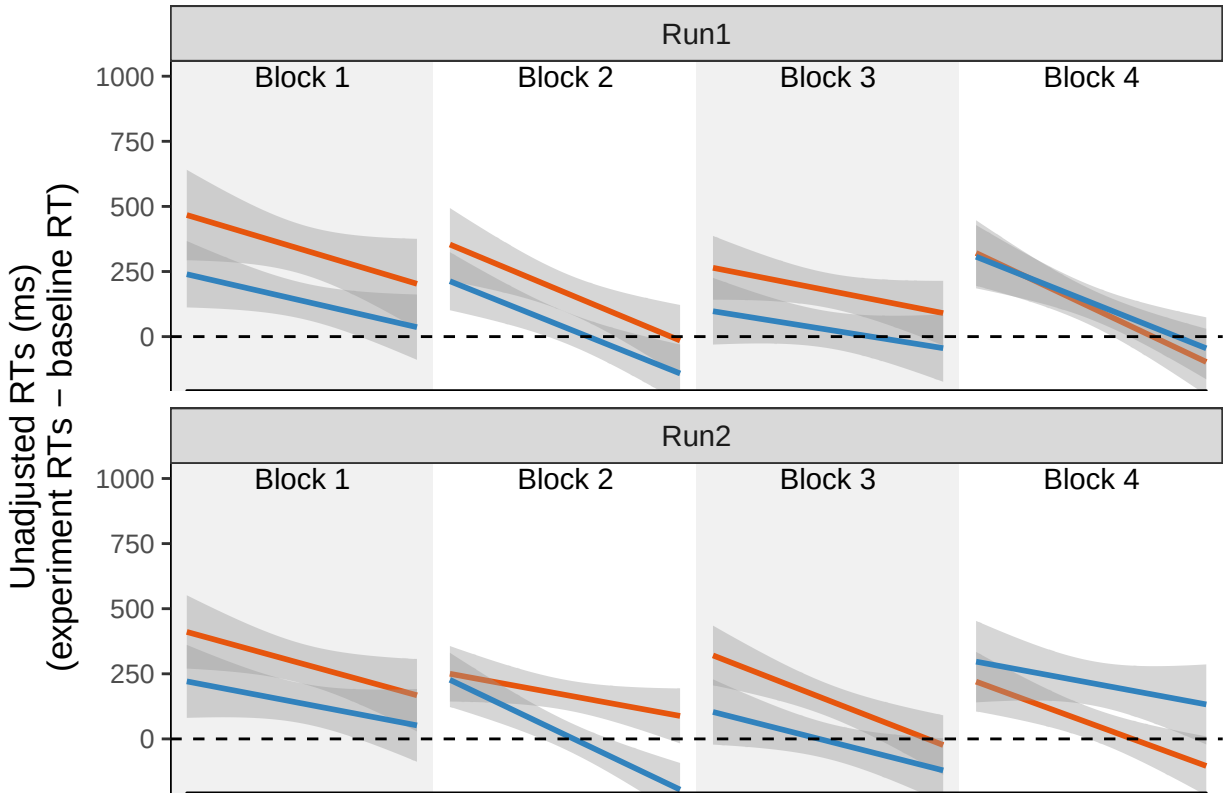


Figure 11: It seems as if Exp2 participants are resetting their adaptation each time they return to the task. There were no audio problems or glitches in how the exposure material was presented, nor were there special instructions before Block 4.

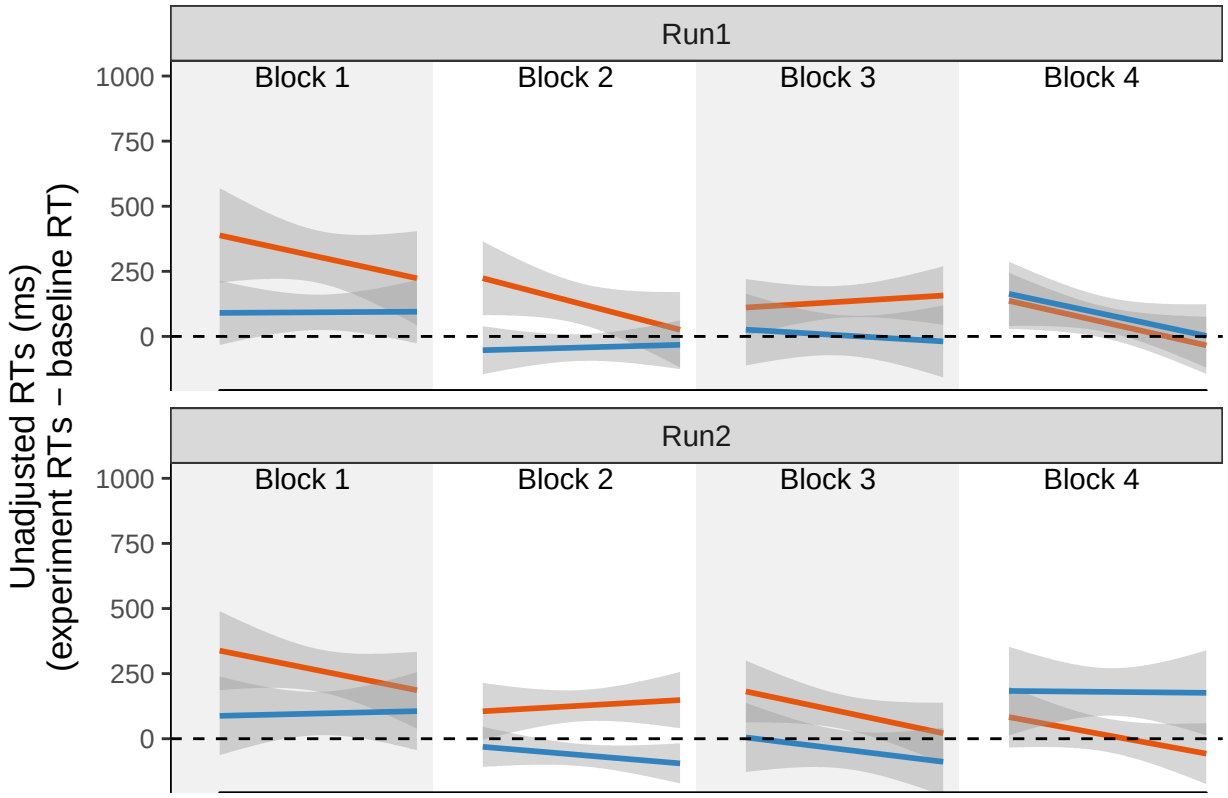


Figure 12: The same as Figure 11, but with the first trial of each block excluded from the data.

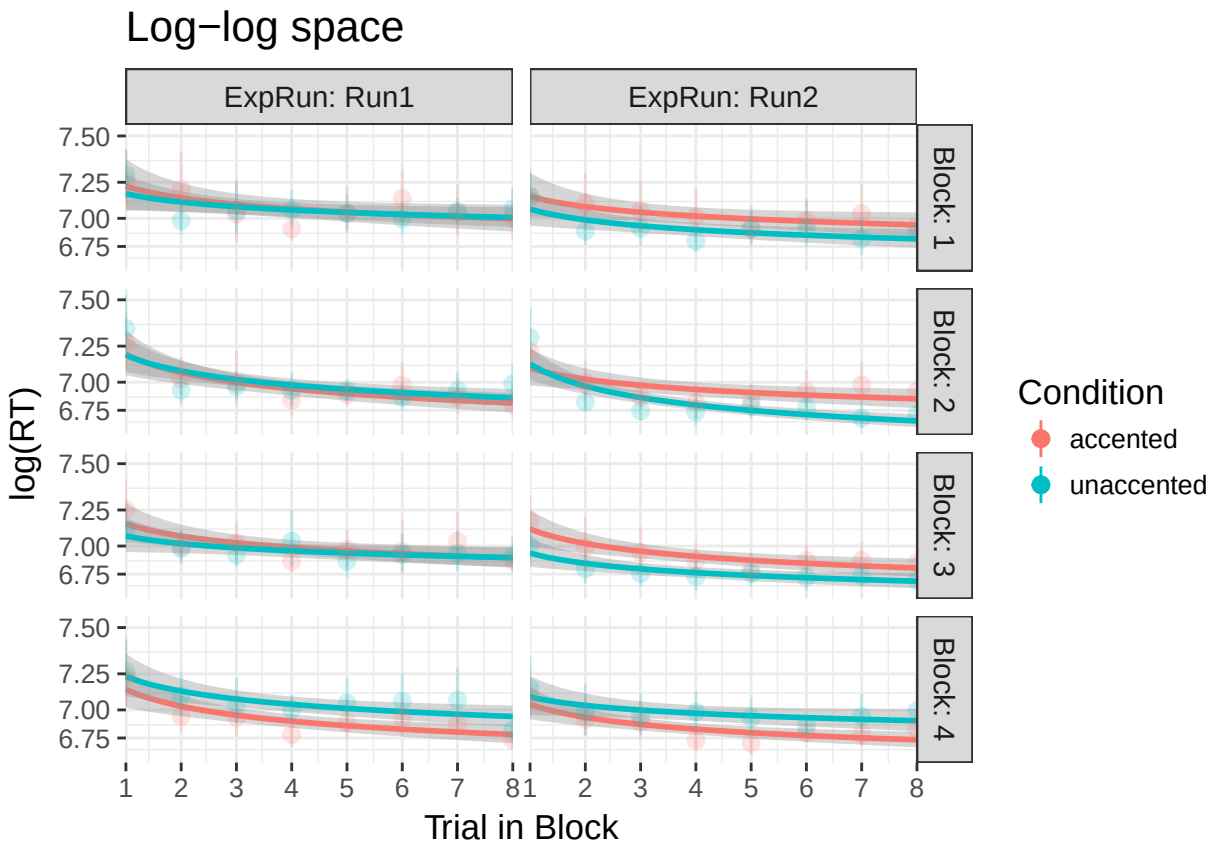


Figure 13: Plotting the UNADJUSTED RTs in log-log space by block.

4 By-stimulus analysis

We discussed the fact that due to the counter-balancing methods, each particular trial index only had 8 different audio clips per condition in exposure. Therefore, individual sentences and trial numbers had lots of potentially dangerous collinearity.

4.1 Accuracy-Trial collinearity

Figure 14 shows the mean accuracy for every audio clip by each speaker. It seems there are six sentences that have means below 0.8 for the accented speaker that seem like they might be significantly bringing down the mean. Figure 15 shows that this is somewhat true.

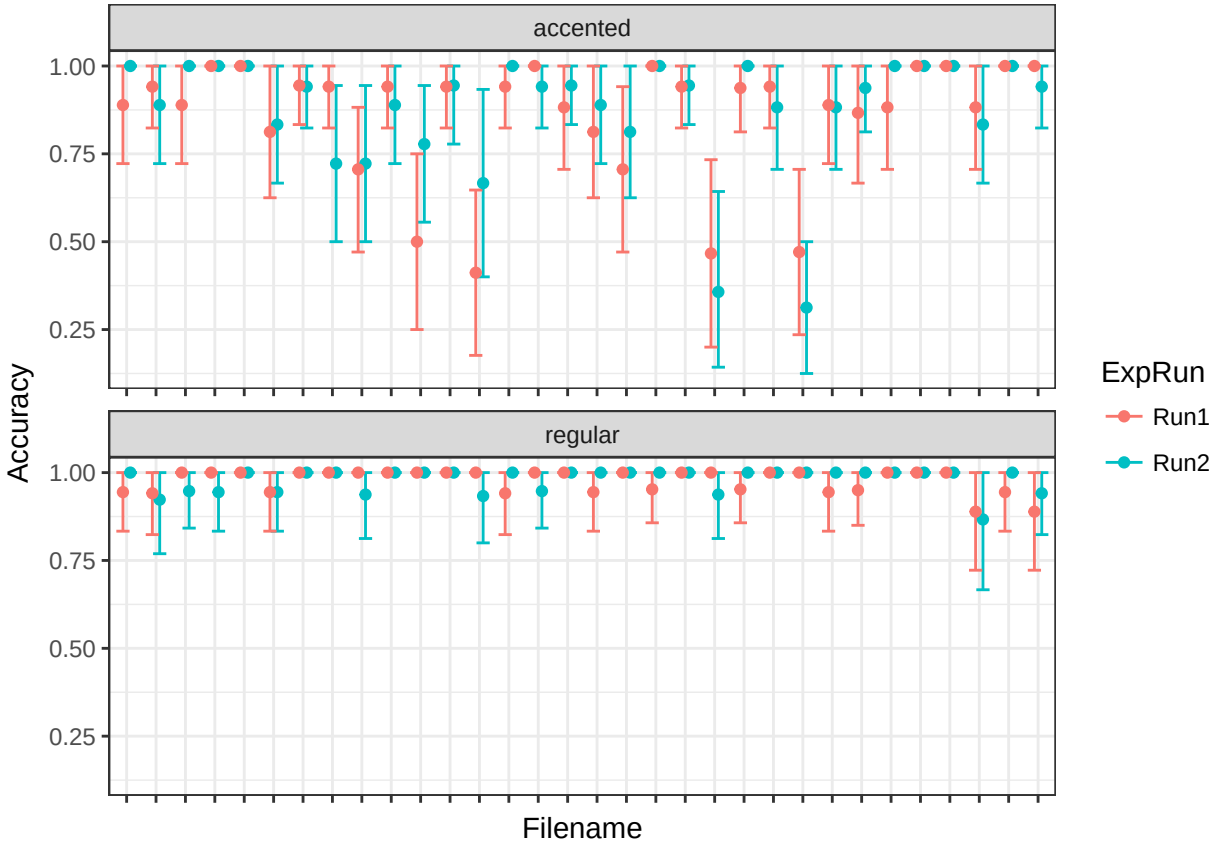


Figure 14: The average sentence accuracy of both speakers in the exposure blocks (i.e. 1-3) collapsed across trial index. It seems like a lot of the inaccuracy can be attributed to a few sentences.

4.2 RT-Trial collinearity

Figure 16 shows RTs for each audio clip. Nothing seems striking.

4.3 Speed-Accuracy by Trial

Figure 17 shows speed and accuracy plotted together.

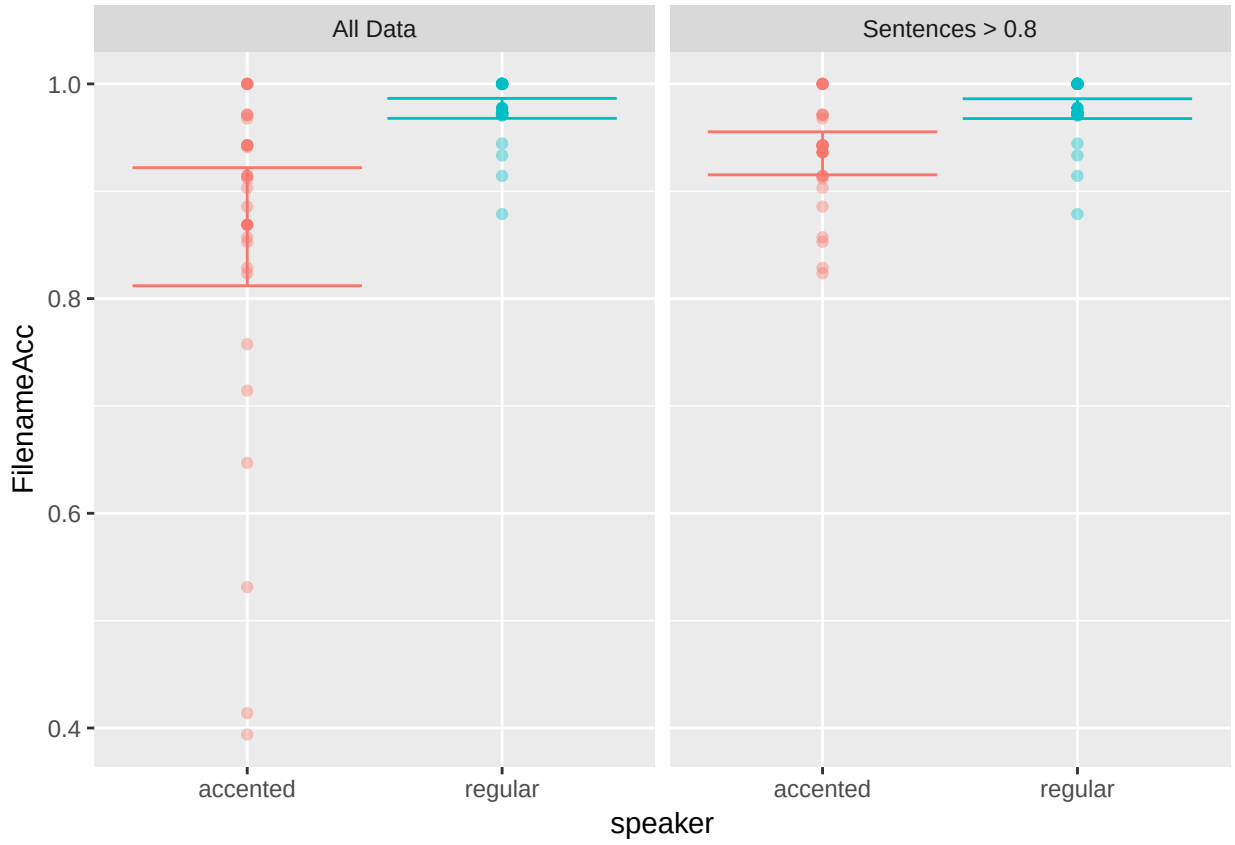


Figure 15: The mean accuracy for all sentence stimuli for each speaker, with all the data and without any audio clips with mean accuracies below 0.8. Each point represents an audio clip's mean accuracy. Removing the six low-scoring audio clips makes the accuracies very comparable.

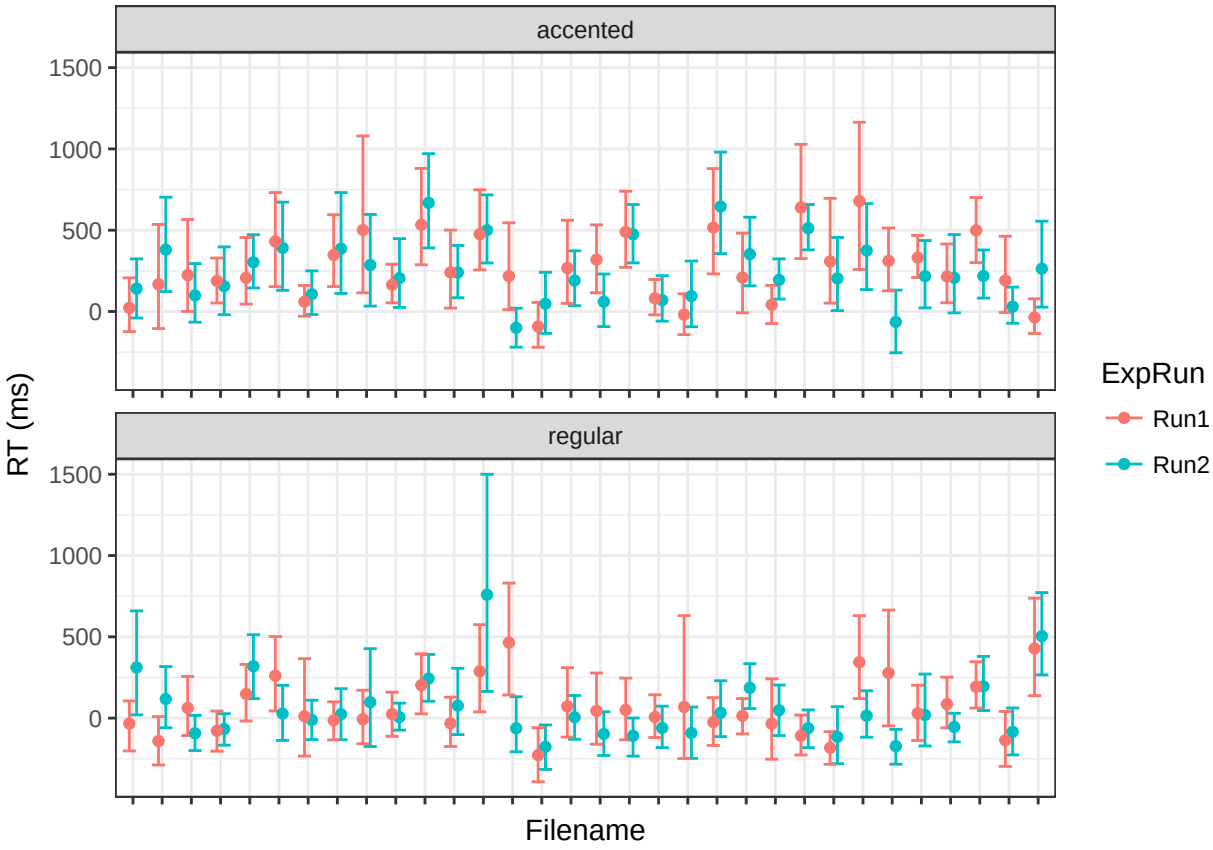


Figure 16: RTs for each audio clip in the exposure phase. Nothing looks too weird here.

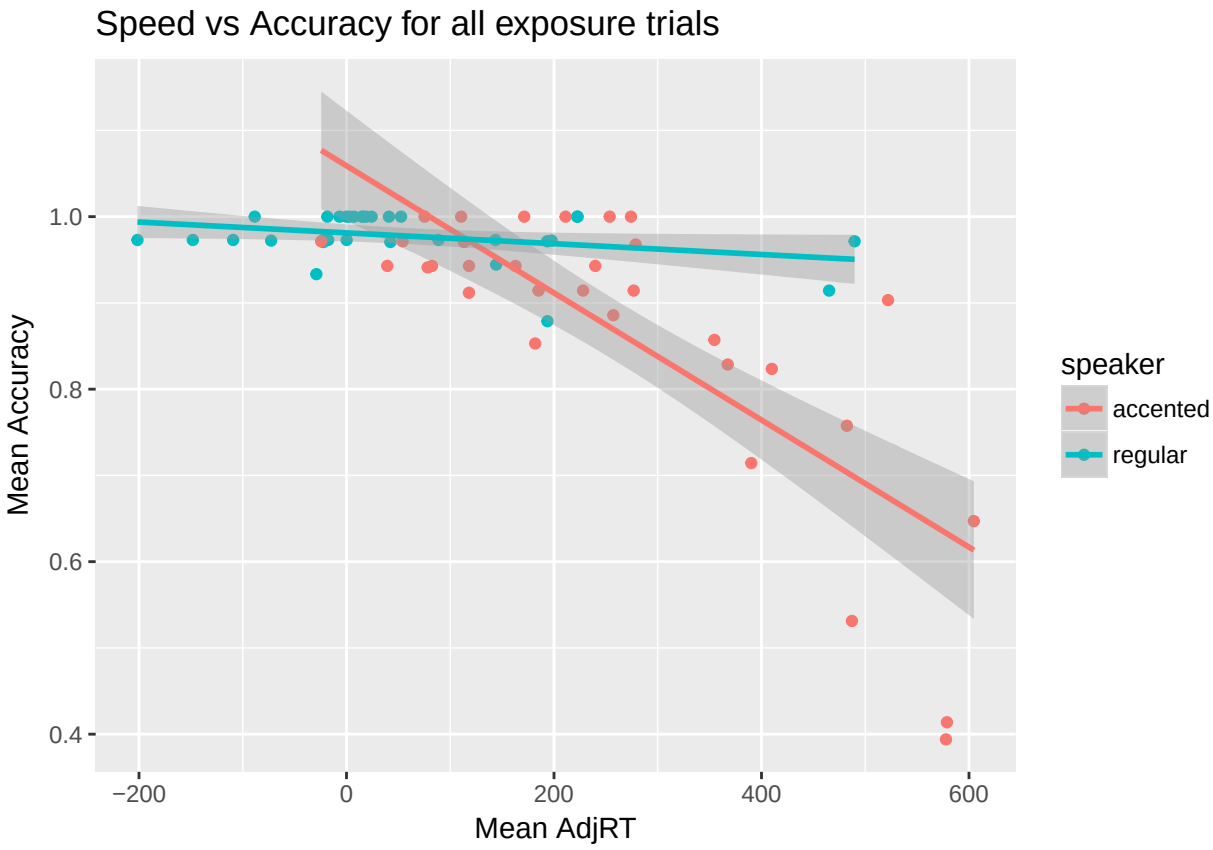
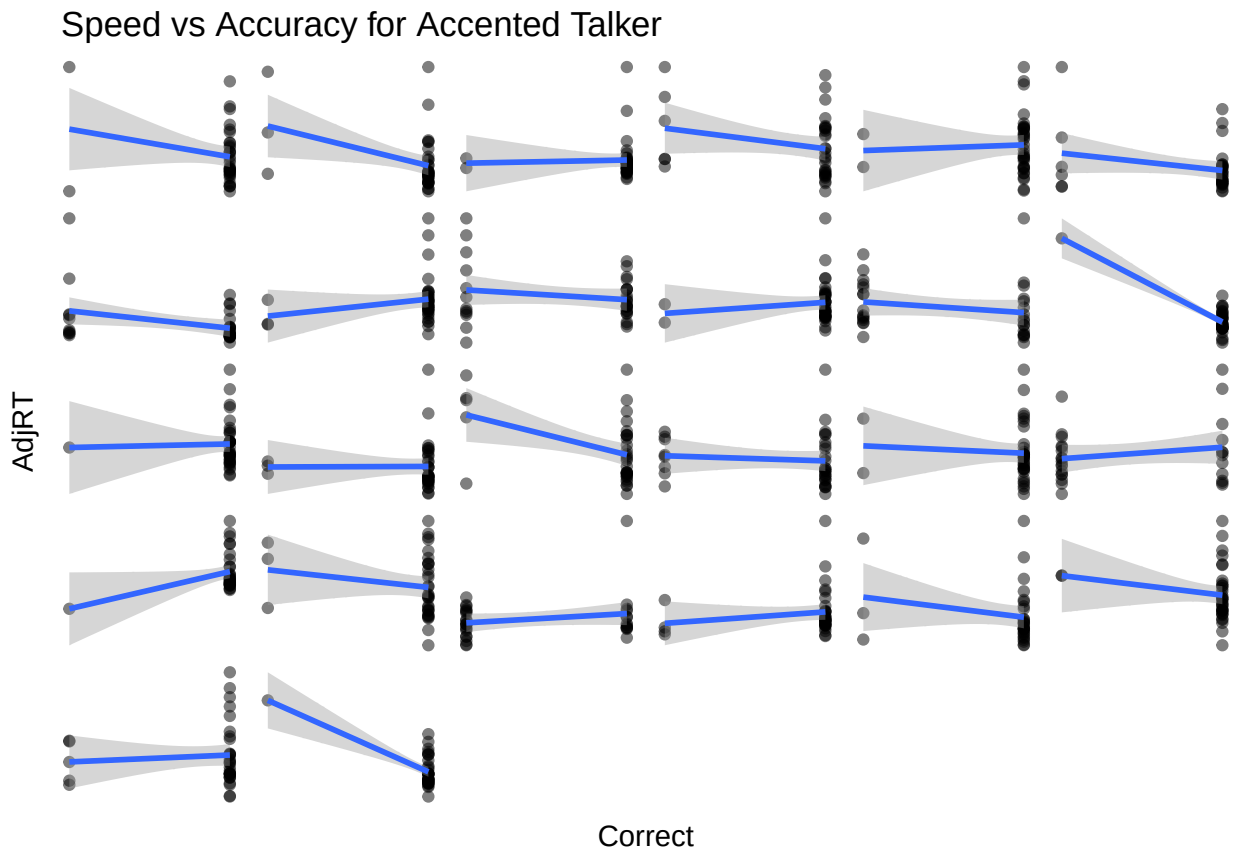


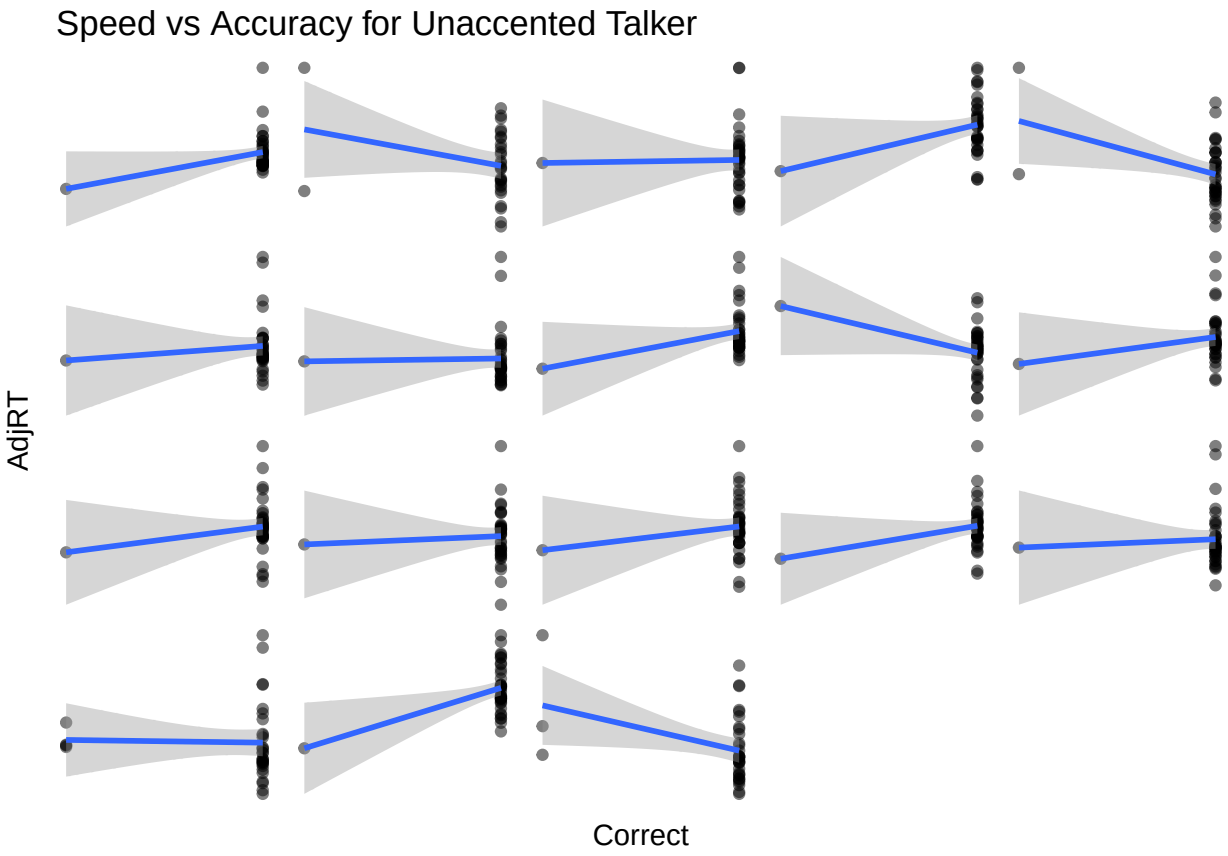
Figure 17: Each point represents an audio clip. Those low-accuracy audio clips look a little weird.

5 Appendix

Everything after this point is part of the appendix!

5.1 Speed-accuracy for individuals audio clips





5.2 Visual analysis of runs combined

Here we plot all the data from Experiment 2 as one dataset.

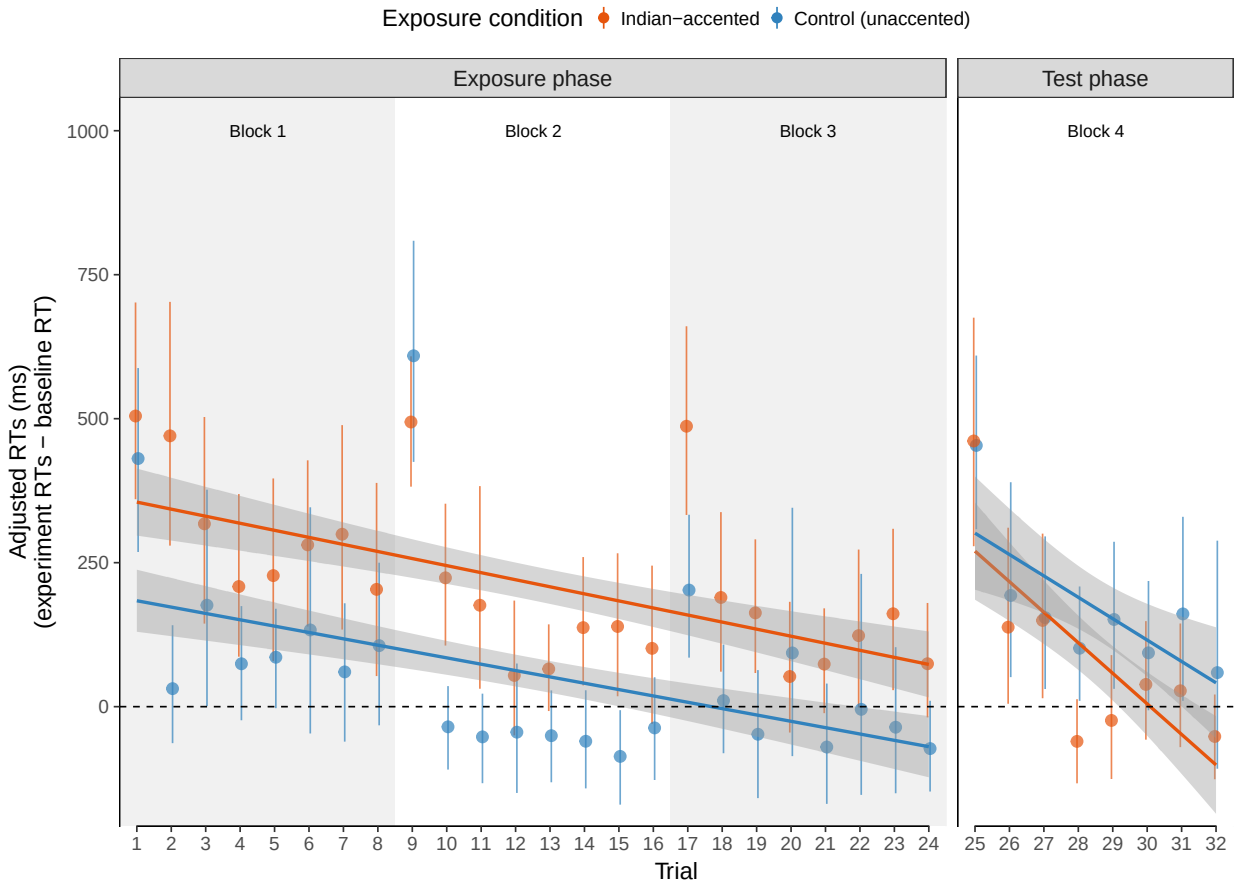


Figure 18: The adjusted RTs for correct trials in the main experiment. Note the seeming adaptation to the accented talker in test by the participants in BOTH exposure conditions, unlike what we see in Experiment 1, where adaptation seems to be confined to those in the unaccented exposure condition.

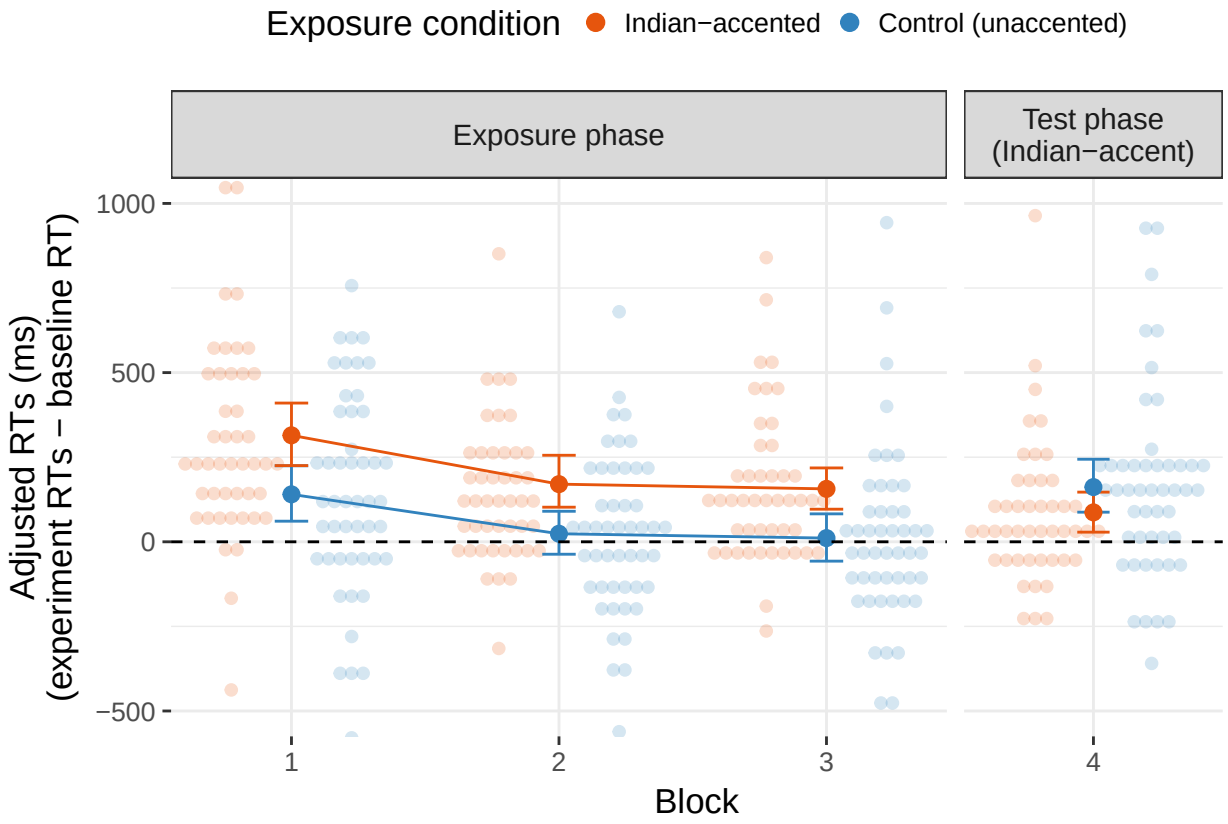


Figure 19: The adjusted mean RTs for each block of the main experiment. You can see that it follows the pattern we expected pretty much, with the unaccented participants doing poorly on the accented test block (block 4). However, unlike in Experiment 1, they seem to be doing equally poorly compared to the accented condition.

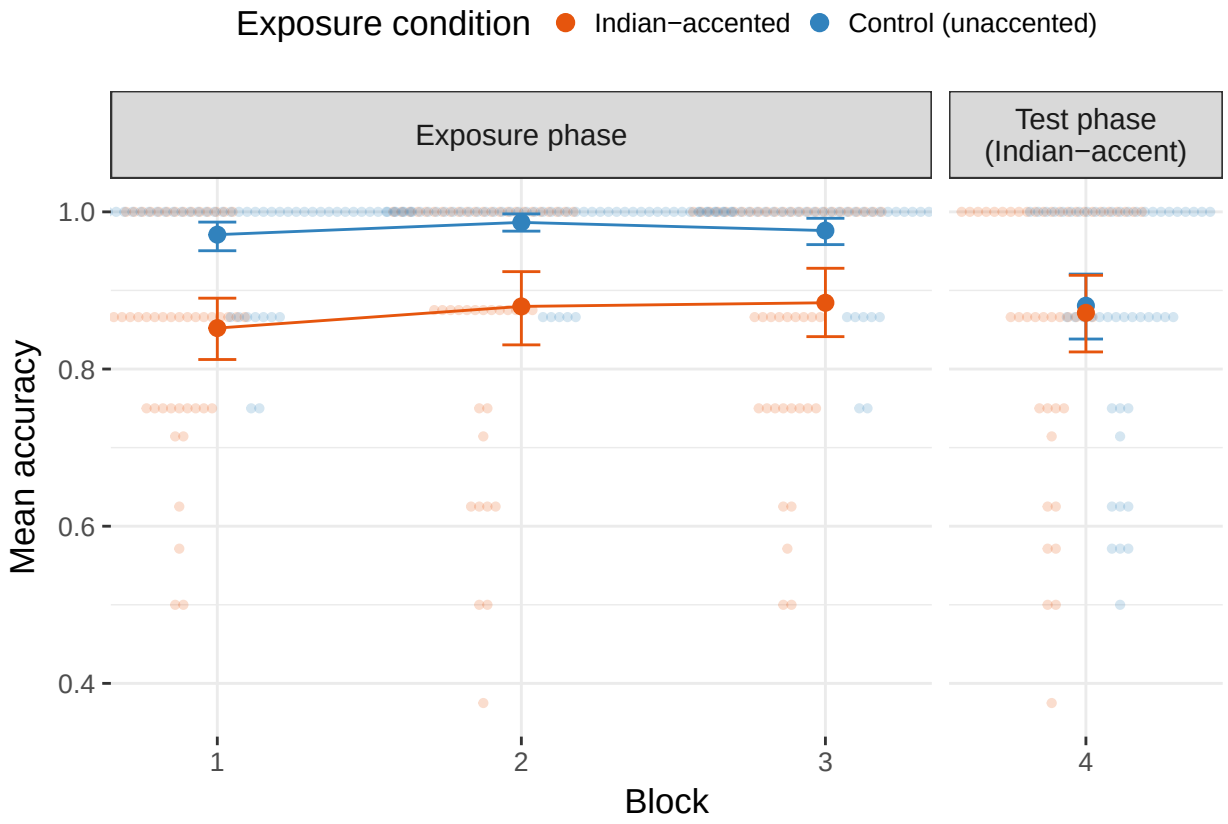


Figure 20: Each dot is a participant's mean accuracy for a particular block.

We also plot the progression of adaptation within each block (Figure 21) and adaptation within each block, ignoring the first trial of every block (Figure 22). Similarly, we look at the unadjusted RTs in log-log space in Figure 23.

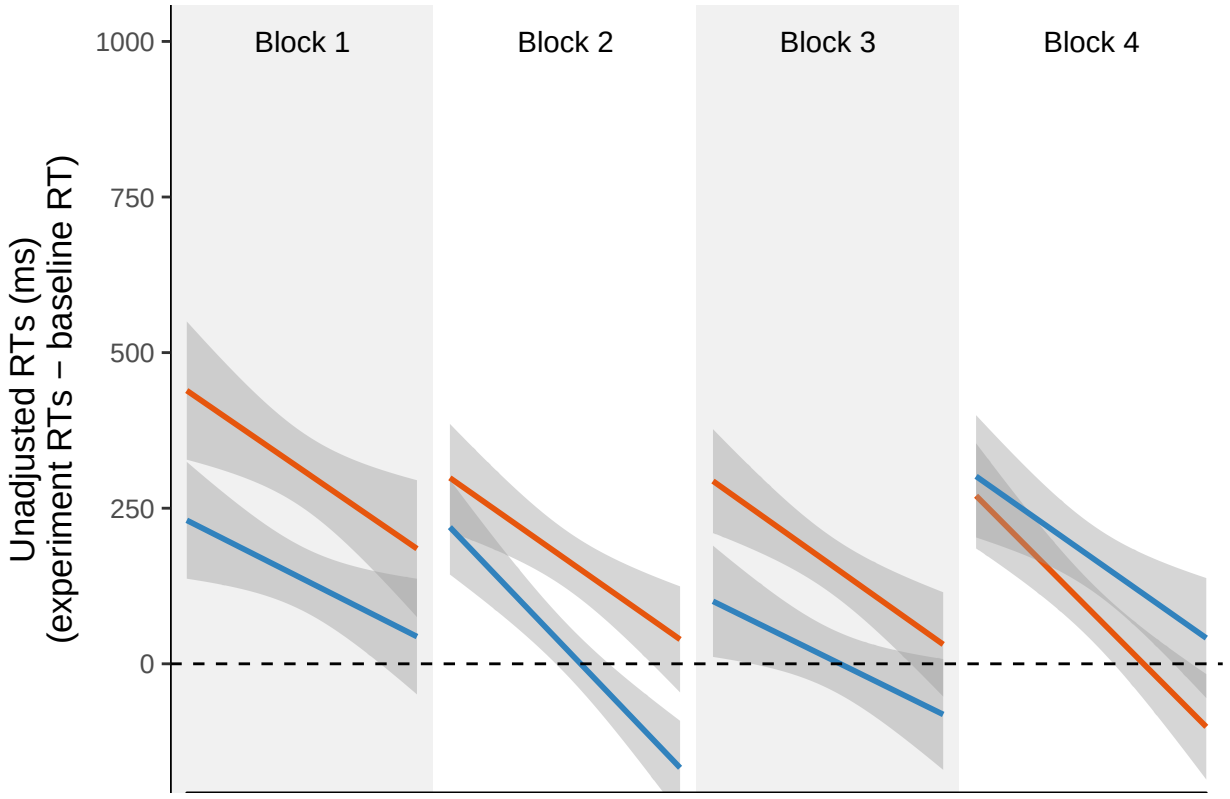


Figure 21: It seems as if Exp2 participants are resetting their adaptation each time they return to the task. There were no audio problems or glitches in how the exposure material was presented, nor were there special instructions before Block 4.

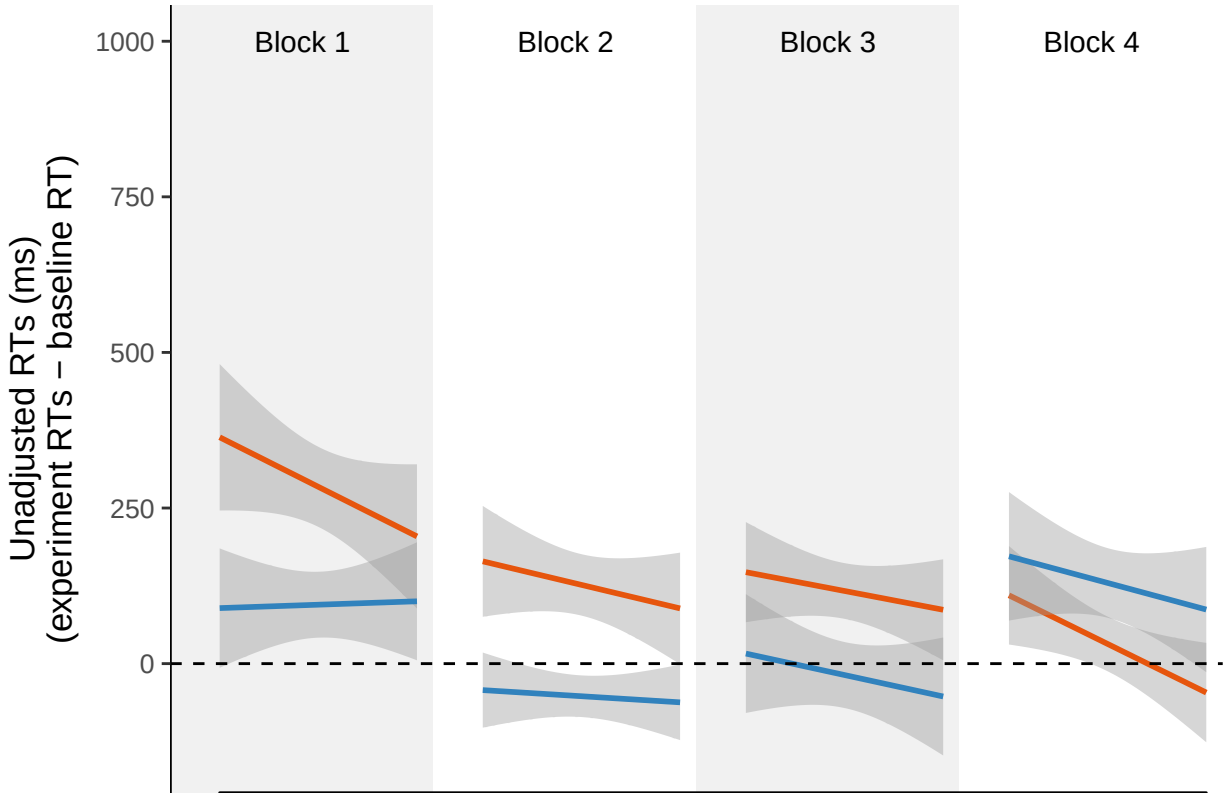


Figure 22: The same as Figure 11, but with the first trial of each block excluded from the data.

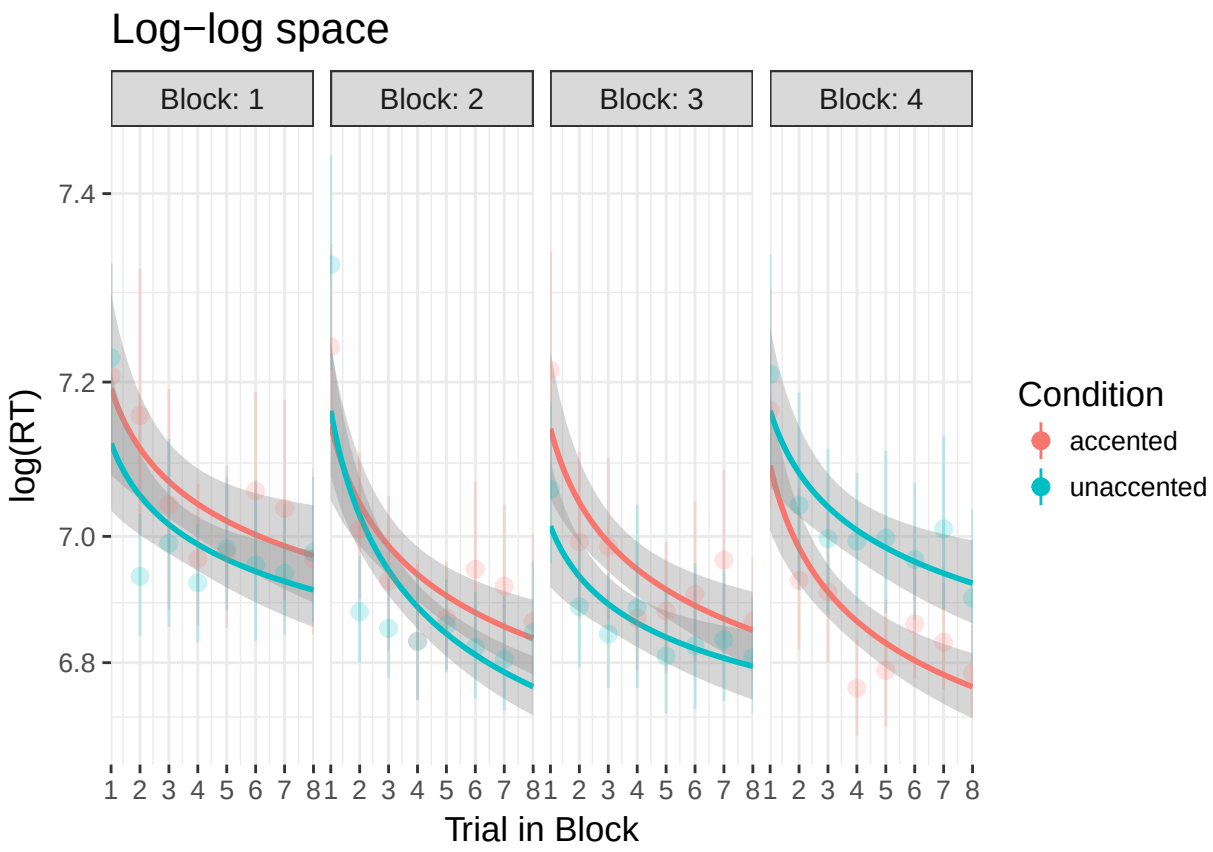


Figure 23: Plotting the UNADJUSTED RTs in log-log space by block.