Measuring Naturalistic Speech Comprehension

Irmak Ergin

2024-02-22 14:49:40.34904

Instructions

The project proposal is due on **Thursday, February 22nd at 8pm**. It should not be longer than **700 words**. It may contain code (code doesn't count toward the word limit).

Research question (3 pt)

What is your main research question? (1 pt)

Research question: How can we measure naturalistic speech comprehension in real-time and in a time-resolved manner?

We will introduce a measure allowing participants to provide continuous comprehension ratings while listening to naturalistic speech segments. This measure will be validated against traditional post-hoc assessments.

The Big Goal: The research proposed here represents the first experiment in a series aimed at answering the question: "What are the causal neural mechanisms of speech comprehension?"

Background

Speech comprehension has been described as an effortless, robust, and automatic process (Shannon et al., 1995; Gwilliams & Davis, 2022). Yet, in real-world contexts, it is common for a listener to understand something different from what was said, or to fail to derive meaning entirely. Beyond merely perceiving speech, how humans understand speech is still largely unknown. Furthermore, the processes engaged during perception versus comprehension are not always clearly distinguished in the literature (Hickok & Poeppel, 2007). While there is a considerable amount of research on speech (mis-)comprehension in neuro-atypical individuals (Gaillard et al., 2007; Rohde et al., 2018), when and why speech comprehension breaks down in neuro-typical adults is understudied. Tackling this question will give us insights into the causal mechanisms involved in speech comprehension. In other words, by understanding the reasons for its failure, we can discern what is required for it to function successfully.

Current methods assess comprehension post-hoc, posing limitations: (1) Responses rely on memory, conflating comprehension with retention (Emmorey et al., 2017; Tun et al., 1991). (2) Static post-hoc comprehension measures do not allow us to relate behavioral reports of speech comprehension with dynamic processes happening in the brain while listening. Therefore, there is a need for a measure that can capture comprehension as it is happening, including its moment-by-moment fluctuations.

What hypotheses are you trying to test? (2 pt)

Enumerate 1-3 hypotheses in terms of directional relationships. For example, you might write something like "X is predicted to increase as Y increases", or "We predict that performance in Group 2 will be significantly better than performance in Group 1".

Real-time comprehension measure: The novel real-time comrehension measure scores will be predicted by the comprehension scores of the traditional post-hoc tests. i.e, the scores of our novel real-time comprehension measure can be explained by comprehension, above and beyond the contribution of individual differences in working memory and speech perception in noise capacities.

Behavioral performance: Comprehension will be worse for higher speed speech segments compared to lower ones.

Methods (4 pt)

Study Design (1 pt)

Describe the study design and data. How were the data collected? From whom? Are the data from an experiment, or will you work with existing data? If the former, what were the experimental conditions? Is the design cross-sectional or longitudinal, and if the latter, what was the time period over which the data was collected? You can use the acronym QALMRI to guide your write-up (https://www.cnlm.uci.edu/files/2015/01/QALMRI.pdf).

Stimuli

An audiobook, Someday, Someday, Maybe by Lauren Graham (2013), will serve as the source for our original recording, from which we will extract 30-second segments that begin and end with sentences beginning and endings. These segments will be used to create stimuli with six different speech rates, ranging from a minimum speeding up factor of 1 (the same as the original recording rate) to a maximum factor of 4 (four times faster than the original recording rate). The six speech rates will be selected to be 1, 2, 2.5, 3, 3.5, and 4 times faster than the original recording rate.Due to different speech rates, the final duration of segments will vary from 30 seconds to 7.5 for the slowest and fastest rates, respectively. We decided to not match segments based on compressed duration because the effects of speech rate vs. information per second would become hard to disentangle as segments with faster rates would inevitably have more content. To avoid potential differences in context or word salience, and to ensure uniformity in the difficulty of post-hoc comprehension questions, we will randomize segment-to-speed allocation across participants. In total, there will be 180 segments, consisting of 30 instances for each of the six speech rates.

Measures

Control Measures

Digit Span: We will measure working memory capacity using the forward and backward Digit Span Test (Richardson, 2007; Wechsler, 1981). The forward and backward tasks will consist of different digit sets, and the forward test will be applied before the backward test. Each of them will include seven levels (3 to 9 digits) with 2 different spans (sequences) for each length level. The procedure will start with the shortest digit spans (3 digits) and will stop as soon as

participants fail to correctly repeat both spans belonging to one level or when the two longest digit spans (9 digits) are successfully completed. Digit spans will be presented auditorily and participants will verbally report their responses. Digit-In-Noise Test: We will measure speech comprehension ability in noise using "hearWho", a mobile and web-based software application for hearing screening developed by the World Health Organization (WHO). HearWho utilizes a Digit-In-Noise Test by presenting 23 spoken sets of US English digit triplets over progressively increasing white noise (World Health Organization, 2022).

Speech Comprehension Measures

Real-time Speech Comprehension Measure: Participants will report continuous comprehension ratings while listening to the segments on a slider. They will express their comprehension level by adjusting the distance between their index finger and thumb of their dominant hand — maximizing the distance indicates complete comprehension, while minimizing it signifies no comprehension (Pelli & Vale, 2014). This reporting approach allows participants to convey continuous feedback without needing to visually engage with the slider, relying on their innate sense of distance between their index finger and thumb of the same hand. These features are particularly important for the proposed research program, as the measure can be applied in the scanner, facilitating potential follow-up studies employing brain imaging methods.

Post-hoc Comprehension Rating: After each segment is finished, participants will rate their comprehension on a 10 point scale, with 0 indicating no comprehension and 10 indicating full comprehension. The purpose of this task is to have a post-hoc self-report comprehension measure in addition to the real-time self-report comprehension measure.

Summary Task: Participants will be asked to give a verbal summary of the segment. The semantic similarity of the segment and the summary will be evaluated using a Large Language Model by creating a cross-correlation matrix. Order of recall is not important and would not be taken into account.

Multiple Choice Question: A four option multiple choice question will be administered after each segment, to assess whether participants understood the content delivered in the segments.

Design and Procedure

Firstly, participants will complete the control tasks which are the Digit Span and Digit-In-Noise tests. Then, prior to the experimental trials, participants will have 2 training blocks with the slowest and fastest speech rates. The experiment will employ a block design. Each block will start with the speech segment, the real-time comprehension measure parallel to listening to the segment, and will end with post-hoc comprehension measures. Each speech segment will be on a single speech rate. Participants will provide real-time comprehension measures while listening to the story. When the story ends, they will complete three post-hoc comprehension measures, which are:the 10-point comprehension scale, the summary task, and the multiple choice question. The blocks will be presented randomly.

Planned sample (2 pt)

What is the sample size? What level of power will you have for testing your main hypothesis? Make sure to include output that shows your power analysis, e.g., code with relevant output and/or a clearly labeled graph.

I wanted to run a power analysis based on pilot data but I coulnd't have the pilot data in time. There are no similar studies that I can use the effect size of. Therefore, I am adding a formulation for the power simulation but not running the code. I will perform the following power analysis using data for my final report.

What is the sample size I would need for 80% power with alpha level of 0.05? Formulation:

```
# library(simr)
# model <- lmer(real_time_comprehension_scores ~ multiple_choice_questions +
#scale_comprehension_ratings +
#verbal_summaries +
#(1 | Digit_Span) + (1 | Digit_In_Noise), data = data)
#power_analysis <- powerSim(model, nsim = 10, n=n)
#nsim- number of simlutaions, n= sample size
#effect_size <- r.squaredGLMM(model)
#sample_size <- pwr.f2.test(u = NULL, v = NULL,
#                           f2 = effect_size,
#                           sig.level = 0.05,
#                           power = 0.80)$n
```

What would be the power for different sample sizes? Formulation:

```
#sample_sizes <- seq(10, 50, by = 5)
##vector to store results
#power_results <- numeric(length(sample_sizes))
# loop through each sample size
#for (i in seq_along(sample_sizes)) {
## calculate power for current sample size
# power_results[i] <- pwr.f2.test(u = NULL, v = NULL, f2 = effect_size, sig.level = 0.05,
# plot
#plot(sample_sizes, power_results, type = "b", # line plot with points
#     xlab = "Sample Size", ylab = "Power",
#     main = "Power for Each Sample Size",
#     ylim = c(0, 1), pch = 16, col = "blue")
# abline(h = 0.8, lty = 2, col = "red")  # add line for 80% power
```

Variables (1 pt)

List your variables of interest (including dependent and independent variables) and describe how they are operationalized.

I mentioned the variables and how they are operationalized in the methods- measures section, but the main IV for the behavioral task is various speech rates (to create a comprehenison challange) and the DV is the comprehension (comprehension scores on various measures).

Analysis (5 pt)

Primary analyses (3 pt)

What statistical analyses are you planning to use to test the hypotheses described above? If possible, formalize your hypotheses as linear models in R, e.g: `response ~ 1 + group + age`. *It is also fine to talk about statistical techniques here that we haven't and/or won't cover in class.*

A Mixed Effects Linear Regression Analysis to model the median real-time comprehension scores. Comprehension scores of the post-hoc tests, which are the multiple-choice questions, 10-scale comprehension ratings, and verbal summaries, will be utilized as factors in the regression. Digit Span and Digit-In-Noise scores will be assigned as random effects, accounting for individual subject variations. This approach allows us to explore to what degree the scores of our novel real-time comprehension measure can be explained by comprehension, above and beyond the contribution of individual differences in working memory and speech perception in noise capacities. Formulation:

```
#lmer(real_time_comprehension_scores ~ multiple_choice_questions +
#scale_comprehension_ratings +
#verbal_summaries +
#(1 | Digit_Span) + (1 | Digit_In_Noise), data = data)
```

Other analytic choices (2 pt)

Are there any other analytic choices a reader of your project proposal should know about? For instance, will you check model assumptions and, if so, how? Will you do any data transformations? If you have missing data, what will you do to address this?

Criteria for participants: -Native English speakers -Normal hearing

Data transformations: -The scores of the physical slider (the novel real-time comrehension measure) and the scores of the post-hoc comprehension rating will be rescaled to values between 0-1. Currently the range of values is 0-250 for the physical slider and 0-10 for the post-hoc rating.

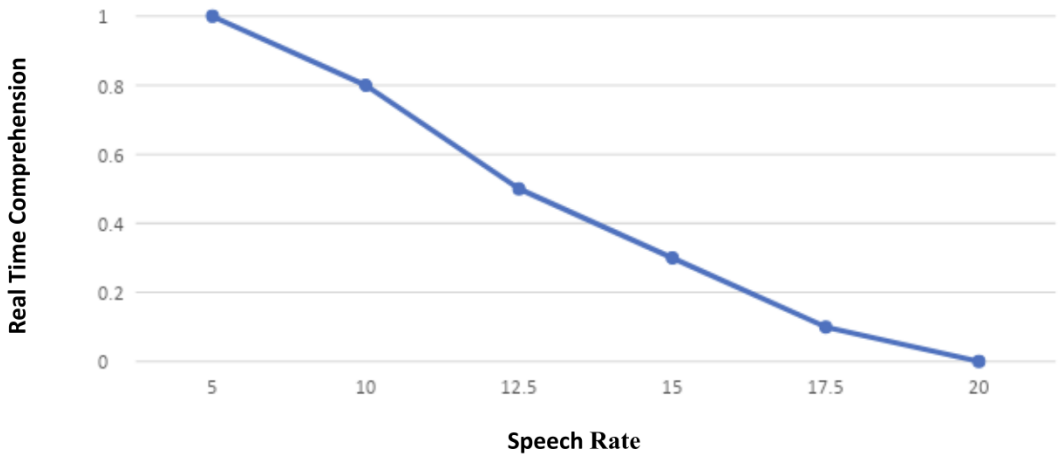
Exclusion criteria For the post-hoc comprehension measures, to ensure that we are validating the novel measure against actual comprehension: -Absolute threshold for comprehension: 75% correctness on the multiple choice questions for the slowest (easiest to comprehend) speech rate. -Ratings on the 10-point scale should be significantly different for the slowest and fastest speech rates. For the physical slider to ensure that participants actually used it actively: - Distribution of the amount of slider movements across trials and participants- If a participant's slider movements exceeds ± 3.5 standard deviations from the mean, we will remove those participants' data.

Key Visualizations (2 pt)

What figures and/or tables are you planning on showing? For figures, what will be on the X and Y axes, what `geom(s)` will you use (e.g. point, bar, line)? No need to include code here, but you can if you think it would be helpful. For tables, what will be in the columns and rows?

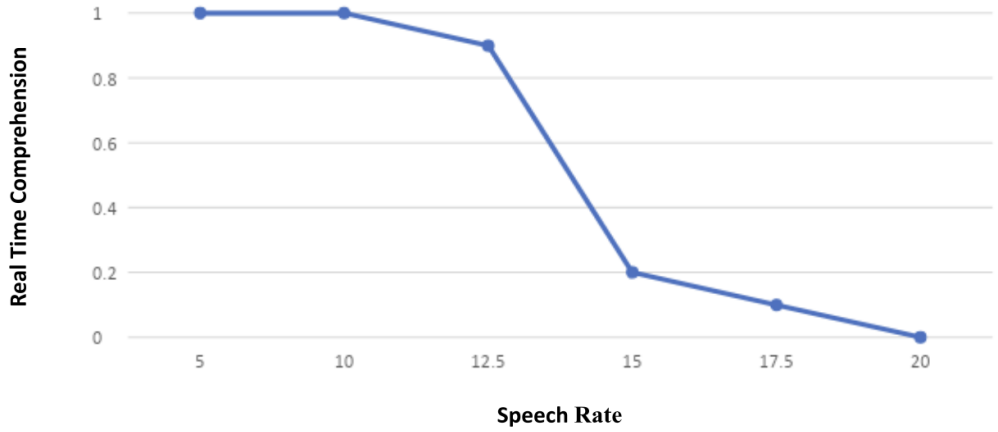
I will plot the patern of comprehension (real-time comrehension scores- y axis) across speech rates (x axis) to see whether there is a categorical or linear decline in comprehension across different speech rates (e.g., Figure 1 and 2, respectively).

Figure 1. Prediction Plot for Linear Decline in Comprehension



Figures:

Figure 2. Prediction Plot for Categorical Decline in Comprehension



I will also plot the regression model (real time comprehension scores-posthoc measures) Formulation:

```
#ggplot(data, aes(y = real_time_comprehension_scores)) +
#  geom_point(aes(x = multiple_choice_questions),
#  color = "blue") +
#  geom_smooth(aes(x = multiple_choice_questions),
#  method = "lm",
#  se = TRUE,
#  color = "blue") +
#  geom_point(aes(x = scale_comprehension_ratings),
#  color = "red") +
#  geom_smooth(aes(x = scale_comprehension_ratings),
#  method = "lm",
```

```
#  se = TRUE,
#  color = "red") +
#  geom_point(aes(x = verbal_summaries),
#  color = "green") +
#  geom_smooth(aes(x = verbal_summaries),
#  method = "lm", se = TRUE,
#  color = "green") +
#  labs(x = "Predictors",
#  y = "Real Time Comprehension Scores") +
#  ggtitle("Regression Plot of Real Time Comprehension Scores")
```

References

Emmorey, K., Giezen, M. R., Petrich, J. A. F., Spurgeon, E., & O'Grady Farnady, L. (2017). The relation between working memory and language comprehension in signers and speakers. Acta Psychologica, 177, 69–77. https://doi.org/10.1016/j.actpsy.2017.04.014 Gaillard, W. D., Berl, M. M., Moore, E. N., Ritzl, E. K., Rosenberger, L. R., Weinstein, S. L., Conry, J. A., Pearl, P. L., Ritter, F. F., Sato, S., Vezina, L. G., Vaidya, C. J., Wiggs, E., Fratalli, C., Risse, G., Ratner, N. B., Gioia, G., & Theodore, W. H. (2007). Atypical language in lesional and nonlesional complex partial epilepsy. Gwilliams, L., & Davis, M. H. (2022). Extracting Language Content from Speech Sounds: The Information Theoretic Approach. In L. L. Holt, J. E. Peelle, A. B. Coffin, A. N. Popper, & R. R. Fay (Eds.), Speech Perception (Vol. 74, pp. 113–139). Springer International Publishing. https://doi.org/10.1007/978-3-030-81542-4_5 Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. Nature Reviews Neuroscience, 8(5), 393–402. https://doi.org/10.1038/nrn2113 Pelli, D. G. and Vale, L. (2014). Lingering pleasure in the experience of beauty. International Association for Empirical Aesthetics, New York, New York, August 22-24, 2014. Pulvermüller, F., Shtyrov, Y., Ilmoniemi, R. J., & Marslen-Wilson, W. D. (2006). Tracking speech comprehension in space and time. NeuroImage, 31(3), 1297–1305. https://doi.org/10.1016/j.neuroimage.2006.01.030 Richardson, J. T. E. (2007). Measures of Short-Term Memory: A Historical Review. Cortex, 43(5), 635–650. https://doi.org/10.1016/S0010-9452(08)70493-3 Rohde, A., Worrall, L., Godecke, E., O'Halloran, R., Farrell, A., & Massey, M. (2018). Diagnosis of aphasia in stroke populations: A systematic review of language tests. PLOS ONE, 13(3), e0194143. https://doi.org/10.1371/journal.pone.0194143 Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. Science, 270(5234), 303–304. https://doi.org/10.1126/science.270.5234.303 Tun, P. A., Wingfield, A., & Stine, E. A. (1991). Speech-processing capacity in young and older adults: A dual-task study. Psychology and Aging, 6(1), 3–9. https://doi.org/10.1037/0882-7974.6.1.3 Wechsler, D. (1981). Wechsler Adult Intelligence Scale-Revised. Psychological Corporation, New York.