Perception Experiment

September 7, 2023

Morgan Moyer

Some possible research questions:

1. Is there a reliable prosodic difference for each speech act? (Experiment 1)
2. (How much speaker variability is there?)
3. Is there a perceptible prosodic difference across the three speech acts?
4. Can hearers reliably use prosodic information to disambiguate the speaker's intended message (speech act)?
   (a) How does prosody aid a hearer in reasoning about the speaker's intended meaning?
   (b) How does a hearer use prosodic information to respond to potentially ambiguous speech acts? [mm: This is an interesting question, but it's hard to identify consistent formats for the different responses, i.e., there's too much by-item variability that might introduce external factors. Further, this kind of task relies on the (a) kind of task so we should go first with (a) before doing (b).]
5. Does prosody carry meaning?
6. Can prosody override context?

There are several ways to think about how to answer 3–6, though they may seem like the same question, they affect how we would design the experiment.

1. **How do we turn "Speech Act" into a dependant measure?**
   (a) **Paraphrase of the Speech Act**
   - Which paraphrase is the closest to what the speaker intended? What do you think the speaker intended?
   - Pro: by placing the participant in the hearer's shoes, it might more naturally get at the intition that we want, namely, how is prosody used as a cue to determine speaker meaning?
     We can quantify uncertainty in the speech signal if we have the three possible answers as a probability distribution, using a relative entropy measure (KL divergence).
   - Con: ?

(b) **Response to the Speech Act**

- Which is the best response to this statement?
- Which response do you think the speaker intended the hearer to give?
- Uses the responses from Experiment 1 as the DM.

- Pro: Puts the participant in the position of reasoning about the speaker's intention/goal
- Con: Maybe it removes the participant from the conversational situation because they're not being asked

(c) **Speech Act Categorization Task (i.e., Hellbern & Sammler 2016)**

- Participants choose the category they think the speech act is.

- Pro: Directly tests what we want?
- Con: Not the most naturalistic task, and would involve somewhat intensive training on categorization

2. **Forced Choice versus Sliders**

(a) Forced Choice between three options

- Makes trickier stats, Multi variate logistic regressions, but in Bayesian eqivalent?

(b) Sliders:

- Three options each rated on a sliding scale that adds to 1 (so we can make a probability distribution)
- This aligns with our assumptions about the ambiguity of 'Tu peux V' phrases in French:
- I.e., interpretation as the process of inference under uncertainty about the intended (speaker) meaning
- The abiguity can be modeled as a probability distribution over the three meanings (SA interpretaions)
- Prosody and Context change the probability distribution over the three meanings
- With this method we can capture changes in the probability distibution

3. **What exactly are we interested in testing?**

(a) Are we interested in quantifying the role of prosody alone?
Call this **Signal - Context**

- Yes, so we can present the utterances without the preceding contexts and see if prosody alone is sufficient to determine
- We can also quantify Accuracy (cf. Trott et al 2019) by comparing participants responses to what was meant

(b) Are we interested in quantifying the interaction between prosody and context?

i. Items with the contexts they were originally tested in Experiment 1.
Call this, **Signal + Context**

- Comparison with the results from the test without context (I did something like this with my questions database study)

- Quantify relative entropy using Kullback-Leibler Divergence from the flat distribution (Whichever task shows the least divergence
- We can also test whether participants' responses match with our a priori intuitions of what each SA context was meant (creating an Accuracy measure)

ii. Cross Items x Contexts (addressing Question 6)
Call this, **Signal x Context**

- In this way we can pit the two cues against each other
- Con: factorial explosion in the Latin Square, the minimal Latin Square is the 3 x 3, the worst (if want to quantify by-item variability) then we have 36 x 36
- Then we have to consider how to deal with by-speaker variability.

In my opinion, I think we should do the Signal - Context and the Signal + Context experiments, and conduct the comparative analysis. This is also a novel analysis that the other studies looking at similar things haven't done.

While the Signal x Context study is interesting, I think it would require much more time to design. (Plus, it would be much more expensive)

Some other questions worth looking into:

1. How reliable (unvarried) is the prosodic marker for each speech act?
Is there more than one marker for each SA? Or does there seem to be one (main)?

- This is essentially our analysis for Experiment 1, and with it we can try to understand also Question 5 (the more variation, the lower the reliability of the prosodic signal alone)
- Variation

2. How reliably can each unique prosodic signal disambiguate the speech act ?
    - For each acoustic correlate, what is it's contribution to disambiguation?
    - This would be harder to analyze because it requires further isolating the individual components (assuming there are multiple)

Some thoughts about Ruytenbeek & Trott

- They did a different acoustic analysis: mean f0, f0 slope and duration
- Perception task only looked at Assertion vs. Question as predictor
- Intent Predictor = something like the original SA context, what I was thinking would factor as Accuracy
- We might directly put acoustic values into our model? They didn't exactly directly test the effect of the signal on responses

---

Summary of meeting with Ioanna 20/06/23 to discuss handout

1. **First, What kind of task?**
    (a) **Experiment 1a**: SA Paraphrase Task +Context
        - Dependent Measure: 3 paraphrases of the target utterance, each corresponding to an SA interpretation. This is good because it seems like we can plausibly come up with general format paraphrases for all items, that differ minimally per items.
        - Participants rate each of the paraphrases on sliders with values from 0-1. Total combined slider values must be lower than 1.
        - In addition to 3 sliders, there's a checkbox option something like "None of these sound natural" which serves as criterion for excluding trials
        - 6 control items, 2 from each SA, the training items from Production Experiment
        - +Context means that participants see (hear) not only the target utterance, but the preceding context as well.
        - In some ways, this experiment serves as sanity check, because we would assume
    (b) **Experiment 1b**: SA Paraphrase Task -Context
        - Exact same method as 1a, except without the preceding context. This way we can quantify the role of prosody alone.

(c) Comparative analysis of Ex1a and Ex1b:

- Ex1b quantifies the effect of the prosodic signal alone, while Ex1a quantifies prosody + context.
- Using the sliders DM, we have for each item, a probability distribution over possible SA interpretations. We can then quantify the amount of information by comparing, per item, and per experiment, the distibution to the flat distribution (representing the uncertain state).
- we do this using Kullback-Leibler Divergence. This is a methodology (both with the sliders, +/- context, and KLD) I used in my previous post doc.

(d) **Experiment 2**: SA Response Task
NB: this is in gray to signify that we're putting it on the back burner

- More or less similar to the designs above, except rather than measuring SA paraphrase, we have participants evaluate responses that indicate SA interpretation.
- First thought would be to use the responses from Production experiemnt.
- One issue is consistency. We want the responses options to be consistant across trial, and mimnally varrying based on the particular trial. This way we minimize confounds. However, it's not straightforward that we can find such general answers for the three SAs.
- Additionally, the processes that would be involved for participants to perform this task, subsume those involved in the first set described. I.e., you have to evaluate (and therefore understand) the SA before evaluating potential behavioral responses. Therefore, we should first do the simpler task before ths one.
- Maybe we save this experiment for later.

2. Second, practical points:

- How to choose speakers to include:
  Use Assertion items from production task as criterion for including/removing speaker recordings. Only use speakers whose assertions are most cannonicallly produced. (Morgan noted that some assertions has a florish that made them sound, sassy, like the speaker was admonishing the hearer?)
- Keep all 12 Items
  - Item as between subject, so no participant hears more than one "guitar" item.
  - Participants see 12 test items total (4 items per SA)
- Paricipants hear and see the contexts
- [mm: Open Question:] Participants hear the target, do they also see it?
- post hoc analysis directly inputting acoustic values for each recording into our statistical models?

3. Production-perception link:

    - We didn't discuss running the study with the participants from the Production Experiment
    - But in any case once we get the experimental design down, it would be the same whether the participants are the old sample or a new random one.