Setting the Stage for Disfluency: Implications of Contextual Task Framing Effects for the Design of Listening Tasks

Ambika Kirkland, Jens Edlund
KTH Royal Institute of Technology
Stockholm, Sweden
kirkland@kth.se, edlund@speech.kth.se

Abstract

Speech disfluencies have been shown to impact both judgments about a speaker's competence and decisions about which source of information to rely on. However, fluency effects more broadly are highly sensitive to context: they are strongest when there is little other information available to inform judgments and decisions, and can be attenuated or even reversed by metacognitive processes. Speech is generally experienced in the context of interactions, where listeners have access to a plethora of information about the speaker and other parameters relevant to decision-making. It is hence crucial to consider how the outcomes of studies on speech disfluencies might be impacted by the framing of experimental tasks and the information available to participants. We carried out a decision-making task where participants had to choose which of two speakers, one fluent and one disfluent, had answered a trivia question correctly. The task was presented in the context of three scenarios which provided different information about the task and speakers. We replicated previous findings that listeners preferred fluent answers in only one of these three contexts, demonstrating the importance of task framing.

Keywords: speech disfluencies, speech perception, metacognition, spontaneous speech, decision-making, prosody, paralinguistics

1. Introduction

Speech disfluencies, such as filled pauses and repetitions, are a hallmark of spontaneous speech and serve myriad important functions in human interaction. They can help a speaker maintain the smooth flow of conversation (Clark and Wasow, 1998; Clark and Fox Tree, 2002), hold the floor (Jiang et al., 2023), orient a listener's attention to unexpected information later in an utterance (MacGregor et al., 2009), or provide information about cognitive load (Clark and Wasow, 1998), metacognitive (Brennan and Williams, 1995) or emotional states (Harrigan et al., 1994). Despite this prominent role in spontaneous conversation, there is robust evidence that disfluencies can negatively impact judgments about speakers. In particular, filled pauses and repetitions are associated with lower ratings of trustworthiness (Lay and Burron, 1968), competence (Kirkland et al., 2023c,a) and confidence (Gustafson et al., 2021; Kirkland et al., 2022, 2023c,a).

These findings may reflect a general phenomenon that has been observed in the cognitive science literature, namely that fluency or the lack thereof impacts judgments about sources of information (Bornstein and D'Agostino, 1994; Alter and Oppenheimer, 2009; Unkelbach and Greifeneder, 2013). Even disfluency irrelevant to the information being evaluated, such as hard-to-read text (Oppenheimer, 2006) or white noise (Dragojevic and Giles, 2016), can result in the information or its source seeming less reliable.

These effects do come with an important caveat: they are highly sensitive to context. Contextual information guides metacognitive processes (thinking about thinking) which in turn determine how fluency information is used (Alter and Oppenheimer, 2009). Fluency is a stronger cue in low-information contexts and may not be used as a basis for judgments or decisions at all if other, better or more explicit information is available (Schwarz and Clore, 1996; Alter and Oppenheimer, 2009). Another illustration of the crucial role of context is the so-called discounting effect. People tend to accept the single most salient explanation for disfluency rather than attributing it to multiple causes (Einhorn and Hogarth, 1986), and as a result, will discount its relevance to making judgments if given a good reason to do so (Bornstein and D'Agostino, 1994; Alter and Oppenheimer, 2009). In some contexts, the typical consequences of disfluency can be completely reversed. When disfluency is assumed to reflect the amount of effort put into a task or the skill required to complete it, it can make the source seem more competent (Song and Schwarz, 2008; Thompson and Ince, 2013). Finally, the effects of context can be further mediated by internal states. For example, people who feel more powerful and in control are more strongly influenced by brief changes in their subjective experience and are hence more responsive to fluency manipulations (Thompson and Ince, 2013).

This may present an especially thorny dilemma when it comes to investigating disfluency effects in speech. Laboratory investigations of cognitive

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8491–8497
11–16 May 2026.

phenomena always entail a tradeoff between experimental control and external validity (Bracht and Glass, 1968), but this is particularly apparent with speech phenomena, since speech is typically experienced in the messy, chaotic and information-rich context of human interaction (Verbeke, 2024). This makes it all the more pressing to examine how the context in which speech stimuli are evaluated might affect listener behavior in speech disfluency research and research on spontaneous speech more generally.

Some evidence already exists for a role of context in mediating the effects of speech disfluencies. Recently, Kirkland et al. (2023a) found that disfluent speech was no longer viewed as sounding less confident and competent when a speaker confessed to nervousness about public speaking: an example of discounting. But given the dramatic ways in which even subtle differences in contextual framing have been shown to alter fluency effects in other domains (Oppenheimer, 2008; Alter and Oppenheimer, 2009; Thompson and Ince, 2013; Schwarz et al., 2021), and the evaluation of speech stimuli more generally (Edlund et al., 2024; Lameris et al., 2023; O'Mahony et al., 2021; Kirkland et al., 2023b), this bears considerably more investigation.

Here, we adapt a decision-making task which was previously used by Kirkland and Edlund (2025) to examine how listeners use fluency information to make decisions in the face of conflicting information. They found that listeners were less likely to choose disfluent responses to guesstimation-style trivia questions (e.g., "What was the 10th most popular women's name in 1901?") compared to fluent ones. We chose this task for a number of reasons. First, it uses a more behavioral measure instead of Likert-style rating scales. At the same time, unlike more complex in-person study designs such as the bluffing games used by De Keersmaecker et al. (2024), this procedure could be easily implemented in several variations with many participants on a crowdsourcing platform. This allows us to compare the outcome of different context manipulations to a previous finding under similar conditions.

We used two types of contextual task framing manipulations that might affect how listeners use disfluency information based on previous research. One scenario implied that the speakers were under intense time pressure. Emphasis on task complexity and effort can in some cases reverse the direction of typical disfluency effects, making a source seem more competent, skillful or diligent (Song and Schwarz, 2008; Thompson and Ince, 2013). We also included a scenario in which the speakers were described as struggling to navigate a social situation. This provided listeners with an opportunity to discount disfluency as a cue for decision-making by attributing it to something not relevant to the speaker's trivia knowledge. This sort of manipulation has been previously shown to attenuate the effects of disfluent speech on judgments (Kirkland et al., 2023a). The third scenario was intended to provide a more neutral comparison to the others while still involving the general context of a quiz show, and described the speakers as practicing their answers with a teammate. These scenarios are described more extensively in section 2.1.1.

2. Method

2.1. Materials

Our stimuli were adapted from those used by Kirkland and Edlund (2025), which consisted of short two-party conversations synthesized with the ElevenLabs Eleven Flash v2.5 text-to-speech model[1]. In these exchanges, the speakers debated the answers to guesstimation-style trivia questions similar those used by Al Moubayed et al. (2013): one speaker stated what he or she believed to be the correct answer and the other speaker contested this and provided his or her own answer. The question itself was not stated, but was implicit in the first speaker's utterance. One of the two speakers produced disfluent speech. The type and locations of disfluencies were the same across utterances.

The design of the original stimuli addressed several potential pitfalls and confounds. The questions have objective, true or false answers (the identity of the nth item on a ranked list), but listeners cannot easily rely on prior knowledge, search engines, or generative AI tools to choose the correct answer since both alternatives are incorrect to the same degree (i.e., one alternative is a rank higher than the correct answer, and the other is a rank lower). The alternatives are hence equally reasonable "near misses". Furthermore, neither the correct answer nor either of the two alternatives were returned by search engine or chatbot queries. The stimuli were also vetted through piloting and with post-hoc confound checks to rule out answer bias, recency and ordering effects, and preferences for specific voices.

We modified these stimuli to create more a more immersive quiz show scenario with three variants, described in the following section.

2.1.1. Quiz show scenarios

We created three different types of quiz show contexts by generating audio clips to introduce a scenario listeners were told to imagine themselves observing, and appending a short context cue to the beginning of each stimulus item. The introduction clips were 20-27 seconds in length. In each clip, a

[1] https://elevenlabs.io/

Table 1: Scenario introductions for each of the three contexts

Scenario	Host monologue
Neutral	Welcome back, ladies and gentlemen! The contest will begin shortly, but in the meantime, let's join the participants backstage as they gear up for the preliminary round. Each team has been given a fixed set of questions that might be asked during the qualifiers and the players are reviewing them with their teammates.
New teammate	Welcome back, ladies and gentlemen! We're now in the final round of the contest. The participants will have a few minutes to discuss and lock in their answers. There's been a lot of preparation leading up to this moment but all of the teams have been reshuffled at the last minute, and it's becoming apparent from their awkward interactions that these contestants are still warming up to their new team assignments.
Time pressure	Welcome back, ladies and gentlemen! We're now in the final round and things are starting to get tense. The time penalty from the last round has left the contestants with very little time to discuss and lock in their answers. There's been a lot of preparation leading up to this moment, but the ticking clock is enough to test anyone's nerves.

short snippet of music played over the sound of an applauding crowd. The "quiz show host" then described one of three scenarios. Transcripts of the scenarios (not shown to participants) are shown in Table 1

In the **time pressure** scenario, the host stated that the quiz show participants had incurred a time penalty and felt pressured by the ticking clock. In the **new teammates** scenario, the host mentioned that all of the quiz show teams had just been reshuffled, and implied that the speakers felt awkward interacting with their new teammates. In the **neutral** scenario, the host stated that the speakers were preparing for the qualifying round of the show by reviewing a fixed set of questions. We included this scenario instead of simply replicating the task without any context, to ensure that the effects of our context manipulations did not result from simply adding *any* additional context.

In the shorter context cues, the host made a comment referring back to the quiz show scenario (e.g., "let's see how the teams are coping with the ticking clock!" in the time pressure condition) over background applause. These clips ranged from 3-4 seconds in length. To create the final stimulus items for the listening task, we appended the short context clips to the beginning of the two-party conversations between quiz show participants.

The quiz show host's utterances were synthesized with the Eleven Flash v2.5 text-to-speech model, using the following prompt: "A British male quiz show host in his 50s. Professional, cool, impartial, intelligent. TV or radio voice." The music and crowd noise were generated with a commercial text-to-audio model, also from ElevenLabs.

2.2. Procedure

The procedure, including the process of counterbalancing stimulus items, largely followed that of Kirkland and Edlund (2025), aside from the context manipulation. We included the same filler items and attention checks used in their experiment, but modified them in the same manner as the test items by appending them to the context cues described in 2.1.1.

2.2.1. Participants

We recruited 144 participants via the crowdsourcing platform Prolific[2]. Participants were from the United States, UK, Ireland, Canada and Australia, and were native speakers of English. We pre-screened potential participants so that only those with a high submission approval rate on Prolific (at least 90%) could take part in the study. Participants were assigned in equal numbers to each of the three contextual framing conditions detailed in 2.1.1 (48 participants per condition). The number of participants per condition was decided upon beforehand and was the same number as in Kirkland and Edlund (2025) so that our results would be more comparable to theirs.

If a participant failed one or more of the attention checks, their responses were not included in the analysis and we recruited a new participant to ensure complete counterbalancing and to reach the predetermined sample size.

[2] https://www.prolific.com/

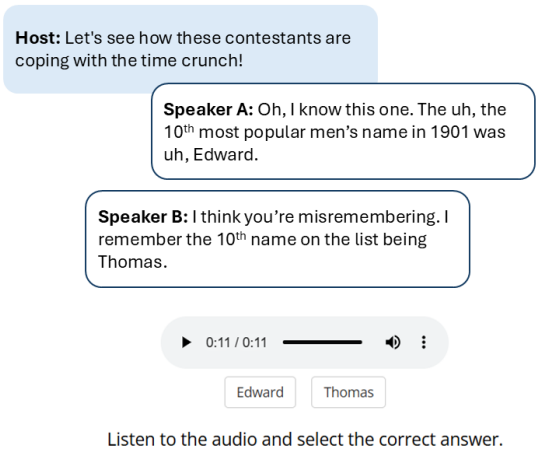

Figure 1: An example experimental trial. Transcripts are included here to illustrate what participants heard but were not shown during the experiment.

2.2.2. Listening task

Participants took part in an online listening task hosted on the experiment presentation platform cognition.run[3]. They first read an overall description of the task they would perform, which informed them that the voices were computer generated and described the task procedure. Performance was incentivized in the same manner as in (Kirkland and Edlund, 2025), by offering a 33% bonus payment for 'good performance'. We did not provide specific criteria for good performance but encouraged participants to do their best to respond both promptly and accurately (in reality, all participants who successfully completed the experiment and passed both attention checks would receive the bonus payment).

After viewing the task instructions, participants listened to one of the three introductions to the quiz show scenario, depending on which condition they were assigned to. They could not continue with the rest of the task until this clip had finished playing. They then completed a short practice round to familiarize themselves with the test procedure before beginning the trials.

During the trial phase, participants responded to a total of 14 items (8 test items, 4 filler items and 2 attention checks) by listening to the audio and clicking on the answer they thought was the correct choice out of the two alternatives. The stimulus played once automatically during each trial and participants could not make their choice or advance to the next trial until the audio was done playing. They were able to play the audio more than once before responding, though the instructions encouraged them to only do so if necessary. An example of an experimental trial is shown in Figure 1.

[3] https://www.cognition.run/

3. Results

Z-tests were carried out to assess whether participants chose the fluent response as the correct one at a rate that significantly differed from chance (i.e., 50%). The proportions of fluent responses chosen as correct for each condition are shown in figure 2.

The proportion of fluent answers significantly differed from chance in the *new teammate* condition ($Z = 2.01$, $p < 0.05$) but not in the *neutral* ($Z = 1.32$, $p = 0.19$) or *time pressure* condition ($Z = -0.09$, $p = 0.93$).

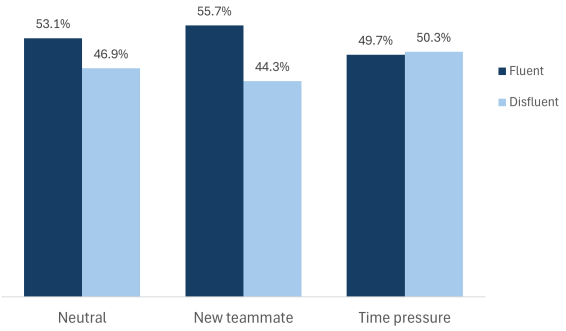

Figure 2: Percentage of answers chosen as correct by fluency and scenario

4. Discussion

Our results show that different types of contextual task framing can produce different outcomes in terms of how speech disfluencies affect decision-making. In the time pressure scenario, we did not find the preference for fluent answers observed by Kirkland and Edlund (2025). This scenario emphasized the difficulty of the task for the quiz show participants, and prior research has shown that highlighting task difficulty can cause disfluency to be interpreted as a sign of effort and skill, rather than a lack of competence (Song and Schwarz, 2008; Thompson and Ince, 2013). We did not, however, find a preference for the disfluent responses in this case.

We also failed to find any preference for the fluent answers in the neutral scenario. This may indicate nothing more than the fact that the effect was rather small, but it may also mean that the "neutral" context was not entirely neutral. This scenario included information about the speakers and setting that was not present in the original study: for example, the fact that the speakers were participating in

a quiz show in front of a large audience and had spent time carefully reviewing their answers with a teammate. Even the greater degree of immersion potentially provided by the contextual audio might have affected the outcome.

The new teammate scenario was the only one in which participants chose fluent answers as correct at a higher rate than disfluent answers. This is somewhat inconsistent with previous findings that attributing speech disfluencies to anxiety or discomfort can attenuate their impact on competence judgments (Kirkland et al., 2023a). However, there are a few important considerations. First of all, the relationship between evaluative judgments (like how competent a speaker seems) and decisions made on the basis of those judgments (like which answer to choose), is not always straightforward (Tittle and Hill, 1967). Furthermore, potential anxiety about working with a new teammate is not extraneous to task performance in this case. A speaker who is particularly flustered about the team assignments (more so than his or her teammate who seems to be taking things in stride) might reasonably be expected to choke up and make mistakes.

The takeaway here may be that we should carefully consider the sorts of implicit and explicit cues contained in any sort of task framing when we set out to design more realistic settings for studying spontaneous speech perception. This is even more important when researching phenomena that are particularly sensitive to the information and assumptions that might be carried by context. Our results also imply that findings from experiments with little or no contextual framing might not generalize very well.

Further research is needed to better understand the nuances of contextual framing effects and develop best practices for designing more realistic listening tasks. Although it is important to study spontaneous speech in settings that more closely resemble how we experience speech "in the wild" (Verbeke, 2024), we need a clearer picture of how characteristics of these settings affect evaluations and behaviors. A good starting point is to consider some of the metacognitive processes documented in the cognitive science literature (e.g., (Alter and Oppenheimer, 2009)), which could be triggered by different types of task framing or even by seemingly superfluous information provided to participants. Understanding how more general cognitive heuristics are leveraged during speech perception may help with the daunting task of designing more realistic experiments.

5. Conclusions

Our findings showed that the impact of disfluent speech on decision-making behavior varies according to the contextual framing of the task. Previous findings that disfluent answers to trivia questions were more likely to be judged as incorrect were not reproduced in two out of our three contextual framing conditions. This highlights the need to carefully consider which kind of information is provided to participants, either implicitly or explicitly, by the context in which the task is presented. It also emphasizes that caution should be exercised when generalizing highly controlled experiments to more realistic settings. More research on contextual framing effects is needed so that we can design more realistic listening tasks without introducing confounds.

6. Bibliographical References

Samer Al Moubayed, Jens Edlund, and Joakim Gustafson. 2013. Analysis of gaze and speech patterns in three-party quiz game interaction. pages 1126–1130. The International Speech Communication Association (ISCA).

Adam L. Alter and Daniel M. Oppenheimer. 2009. Uniting the Tribes of Fluency to Form a Metacognitive Nation. *Personality and Social Psychology Review*, 13(3):219–235. Publisher: SAGE Publications Inc.

Robert F. Bornstein and Paul R. D'Agostino. 1994. The Attribution and Discounting of Perceptual Fluency: Preliminary Tests of a Perceptual Fluency/Attributional Model of the Mere Exposure Effect. *Social Cognition*, 12(2):103–128. Publisher: Guilford Publications Inc.

Glenn H Bracht and Gene V Glass. 1968. The external validity of experiments. *American educational research journal*, 5(4):437–474.

S. E. Brennan and M. Williams. 1995. The Feeling of Anothers Knowing: Prosody and Filled Pauses as Cues to Listeners about the Metacognitive States of Speakers. *Journal of Memory and Language*, 34(3):383–398.

Herbert H. Clark and Jean E. Fox Tree. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition*, 84(1):73–111.

Herbert H. Clark and Thomas Wasow. 1998. Repeating Words in Spontaneous Speech. *Cognitive Psychology*, 37(3):201–242.

Bram De Keersmaecker, Robert J Hartsuiker, and Aurelie Pistono. 2024. (don't) believe me, i'm telling the truth! speech disfluency and eye contact as cues to veracity, intention, and truth judgement. *Language, Cognition and Neuroscience*, 39(10):1263–1277.

Marko Dragojevic and Howard Giles. 2016. I Don't like You Because You're Hard to Understand: The Role of Processing Fluency in the Language Attitudes Process. *Human Communication Research*, 42(3):396–420.

Jens Edlund, Christina Tånnander, Sébastien Le Maguer, and Petra Wagner. 2024. Assessing the impact of contextual framing on subjective tts quality. In *Interspeech 2024*, pages 1205–1209. : International Speech Communication Association.

Hillel J. Einhorn and Robin M. Hogarth. 1986. Decision Making Under Ambiguity. *The Journal of Business*, 59(4):S225–S250. Publisher: University of Chicago Press.

Joakim Gustafson, Jonas Beskow, and Eva Szekely. 2021. Personality in the mix-investigating the contribution of fillers and speaking style to the perception of spontaneous speech synthesis. In *Proc. SSW 2021*, pages 48–53.

Jinni A. Harrigan, Ivette Suarez, and Joyce S. Hartman. 1994. Effect of Speech Errors on Observers Judgments of Anxious and Defensive Individuals. *Journal of Research in Personality*, 28(4):505–529.

Bing'er Jiang, Erik Ekstedt, and Gabriel Skantze. 2023. What makes a good pause? Investigating the turn-holding effects of fillers. In *20th International Congress of Phonetic Sciences (ICPhS). August 7-11 2023, Prague, Czech Republic*, pages 3512–3516. International Phonetic Association.

Ambika Kirkland and Jens Edlund. 2025. Who knows best? Effects of speech disfluencies on incentivized decision-making. pages 4508–4512.

Ambika Kirkland, Joakim Gustafson, and Éva Székely. 2023a. Pardon my disfluency: The impact of disfluency effects on the perception of speaker competence and confidence. pages 5217–5221.

Ambika Kirkland, Harm Lameris, Eva Szekely, and Joakim Gustafson. 2022. Where's the uh, hesitation? The interplay between filled pause location, speech rate and fundamental frequency in perception of confidence. pages 4990–4994.

Ambika Kirkland, Shivam Mehta, Harm Lameris, Gustav Eje Henter, Éva Székely, and Joakim Gustafson. 2023b. Stuck in the mos pit: A critical analysis of mos test methodology in tts evaluation. In *SSW*.

Ambika Kirkland, Marcin Wlodarczak, Joakim Gustafson, and Éva Székely. 2023c. Evaluating the impact of disfluencies on the perception of speaker competence using neural speech synthesis. In *International Congress of Phonetic Sciences (ICPhS)*, pages 550–554.

Harm Lameris, Ambika Kirkland, Joakim Gustafson, and Eva Szekely. 2023. Situating Speech Synthesis: Investigating Contextual Factors in the Evaluation of Conversational TTS. pages 69–74.

Clarry H. Lay and Bryan F. Burron. 1968. Perception of the Personality of the Hesitant Speaker. *Perceptual and Motor Skills*, 26(3):951–956. Publisher: SAGE Publications Inc.

Lucy J. MacGregor, Martin Corley, and David I. Donaldson. 2009. Not all disfluencies are are equal: The effects of disfluent repetitions on language comprehension. *Brain and Language*, 111(1):36–45.

Johannah O'Mahony, Pilar Oplustil Gallegos, Catherine Lai, and Simon King. 2021. Factors affecting the evaluation of synthetic speech in context. In *SSW*.

Daniel M. Oppenheimer. 2006. Consequences of erudite vernacular utilized irrespective of necessity: problems with using long words needlessly. *Applied Cognitive Psychology*, 20(2):139–156.

Daniel M. Oppenheimer. 2008. The secret life of fluency. *Trends in Cognitive Sciences*, 12(6):237–241.

Norbert Schwarz and Gerald Clore. 1996. Feelings and Phenomenal Experiences. In *Social Psychology: Handbook of Basic Principles 2nd Edn*, volume 2, pages 433–465. Journal Abbreviation: Social Psychology: Handbook of Basic Principles 2nd Edn.

Norbert Schwarz, Madeline Jalbert, Tom Noah, and Lynn Zhang. 2021. Metacognitive experiences as information: Processing fluency in consumer judgment and decision making. *Consumer Psychology Review*, 4(1):4–25.

Hyunjin Song and Norbert Schwarz. 2008. If it's hard to read, it's hard to do: Processing fluency affects effort prediction and motivation. *Psychological science*, 19(10):986–988.

Debora V. Thompson and Elise Chandon Ince. 2013. When Disfluency Signals Competence: The Effect of Processing Difficulty on Perceptions of Service Agents. *Journal of Marketing Research*, 50(2):228–240. Publisher: SAGE Publications Inc.

Charles R Tittle and Richard J Hill. 1967. Attitude measurement and prediction of behavior: An evaluation of conditions and measurement techniques. *Sociometry*, pages 199–213.

Christian Unkelbach and Rainer Greifeneder. 2013. A general model of fluency effects in judgment and decision making. *The Experience of Thinking*, pages 11–32.

Gil Verbeke. 2024. On the role of ecological validity in language and speech research. In *Taalkunde nu*, pages 69–95. Skribis.