Guiding Oral Conversations: How to Nudge Users Towards Asking Questions?

Marcel Gohsen
marcel.gohsen@uni-weimar.de
Bauhaus-Universität Weimar
Weimar, Germany

Johannes Kiesel
johannes.kiesel@uni-weimar.de
Bauhaus-Universität Weimar
Weimar, Germany

Mariam Korashi
mariam.korashi@uni-weimar.de
Bauhaus-Universität Weimar
Weimar, Germany

Jan Ehlers
jan.ehlers@uni-weimar.de
Bauhaus-Universität Weimar
Weimar, Germany

Benno Stein
benno.stein@uni-weimar.de
Bauhaus-Universität Weimar
Weimar, Germany

ABSTRACT

How could an envisioned voice-based conversational information system assist the information seeker when the seeker does not know how to continue the conversation? The system could explicitly suggest a question to ask after each of its responses, but this approach quickly feels restrictive, repetitive, and interrupts immersion in the conversation. In this paper, we explore, for the first time, unobtrusive syntactic and auditive modifications of oral system responses to nudge information seekers towards asking about specific topics. We report the results of a crowdsourcing study with 965 participations that investigated the effectiveness and drawbacks of different modifications in three information scenarios.

CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces**; • **Information systems** → *Crowdsourcing*; • **Applied computing** → Psychology.

KEYWORDS

conversational user interfaces, information-seeking, nudging, search, voice

ACM Reference Format:
Marcel Gohsen, Johannes Kiesel, Mariam Korashi, Jan Ehlers, and Benno Stein. 2023. Guiding Oral Conversations: How to Nudge Users Towards Asking Questions?. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '23), March 19–23, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3576840.3578291

1 INTRODUCTION

Paradoxical at first glance, information seekers occasionally do not know what to ask. Especially when they have little knowledge on the topic under investigation or in case of a spontaneous information need, seekers are frequently unable to carry out explicit requests [3, 12, 21]. Conversational systems are supposed to support their users in such a situation, however, to the present day, it is still unclear how to provide this support efficiently. Moreover, conversational systems usually present some hurdle to asking questions, as information seekers do not know which types of requests the system can handle [14]. This may lead users to oversimplify their question, overwhelm the system, or simply hesitate and abort. In voice-based conversational search, seekers might be put under additional pressure when the system allows input for only a short time, creating a kind of decision paralysis that makes it difficult to issue a command before the system stops listening.

Complex search tasks require the acquisition of complex information; for example, when we try to comprehend consequences of a political decision. Here, the need for information changes dynamically during the search and moments of uncertainty are thus common. At best, system responses provoke new requests and the conversation continues to a satisfactory end. If no new questions arise in time, a conversational system may offer a certain preselection; however, explicit suggestions from auditory-only systems involve particularly high costs, as making explicit suggestions is time-consuming and imposes significant cognitive demands on the information seeker to comprehend each option.

In human-to-human oral conversations, humans tend to apply subtle techniques to direct their listener's attention to specific content, for example emphasizing selected content terms, repeating the terms, or placing the terms at the beginning or end of their utterance rather than in the middle. Each of these techniques indicates the importance of the content without incurring much additional processing time like explicit suggestions. In naive users, the latter may create a sense of safety when operating the system; however, as a whole, implicit suggestions can provide for a smoother conversation by mimicking human-to-human dialogue.

The paper at hand evaluates the potential of implicit techniques for guiding a seeker's attention in auditory-only conversational systems. We therefore adopt a term from the decision-making literature: *nudges*, using the term as an intervention "that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid" [22]. This definition aligns with the concept outlined above: to suggest certain queries without imposing them, and to support the creation of an internal locus of control.

CHIIR '23, March 19–23, 2023, Austin, TX, USA

ACM ISBN 979-8-4007-0035-4/23/03.
https://doi.org/10.1145/3576840.3578291

This paper investigates the following research questions:
RQ1 How effective are different techniques in nudging people towards a specific target?
RQ2 How does nudging effect the seeker's mental load when listening to information?
RQ3 How does nudging effect the perceived naturalness of the synthesized speech?

After a brief review of related work, Section 3 introduces and classifies different techniques to nudge seekers towards certain topics. Section 4 then describes our experimental crowdsourcing setup for investigating the research questions, with results and insights from the experiments being detailed and discussed in Section 5.

2 RELATED WORK

Our work is connected to existing literature in several areas, specifically to nudging and biases, voice interfaces, and conversational search.

2.1 Nudging and Biases

A few studies also analyzed the effect of nudges in interactions with computer systems. Son Nguyen et al. [20] introduces biases and nudging in their study on consumer attitudes and behavior intentions. Specifically, the study investigated whether and how anthropomorphism and associated biases in voice assistants affects how consumers assess things and decide whether to make purchases while using voice assistants. Myers [13] analyze how different ways to provide feedback in a voice interface can assist the user. Depending on the detected user proficiency with the system, they nudge users towards the use of more advanced verbal commands and actions to complete their desired tasks. Kankane et al. [9] analyzed the effect of nudges on self-reported measures in password selection. Participants underwent a variety of nudges (incentives, norms, default, salience, and ego) before stating their level of comfort with an auto-generated password and intentions to establish a new password. These preliminary results suggest that various psychological effects-based nudges may be useful design cues that facilitate the performance of desired activities. To the best of our knowledge, the paper at hand presents the first study of the effect of different nudging techniques in a conversational setting, and if they have any effects on the users' mental load.

2.2 Voice Interfaces

Only few works in literature demonstrate the further continuity of conversations with voice interfaces. The recent study of Fischer et al. [6] looked at sequences of voice interaction captured in people's homes to see how a strong desire for progressivity in conversation actually plays out in practice. Their data demonstrates how non-answer responses hinder progress, how explanations for non-answer responses can promote recovery, how participants seek out answers, and how, in the end, moving an interaction forward does not always require a fitted answer but also other types of responses. Zheng et al. [28] suggested that a common issue in conversational interfaces is lack of engagement. This is especially true for scenarios involving group projects and online communities, where more contributions and increased community involvement are required. As outlined in their work, it is difficult to promote productive conversation among students in peer learning settings, such as by having students explain to peers and restate others' statements, and to offer others effective learning support, such as by having a positive attitude and causing tension release. Related to the information scenarios of this work are studies of auditory websites [25–27]. Zhang et al. [27] explored how voice in a website reader could be augmented with "audio styling" to convey aesthetics, user engagement, or interactivity. For example, they utilized techniques such as changing tones of voice, accent, personality to convey the aesthetics and information displayed on websites. These techniques were developed by 14 professional sound designers. Our study borrows some of these auditory modifications to nudge users towards specific topics.

Studies using crowdsourcing techniques to evaluate voice assistants remain limited. The previously mentioned study by Kankane et al. [9] adopted a crowdsourcing methodology, so as Randhawa et al. [15] who introduced a voice-based crowdsourcing platform to 725 workers in low-resource locations using low-end phones. Translations, the creation of data-sets, and surveys on demographics are examples of tasks that were performed in this study. While this study addresses a voice-based system, it lacks the presence of conversational systems. Since that remains an open question, we try to explore it in our crowdsourcing study.

Our study looks into how conversations could be actively sustained and continued with voice assistants, in three example scenarios: in a museum, in a product-comparison system, and argumentation.

2.3 Conversational Search

Several works exist to study conversational search. A good overview is provided by the report on the corresponding Dagstuhl seminar [1]. We here briefly point out a few studies that focus on dialogue analysis. In the largest dialogue analysis we are aware of, Vakulenko et al. [24] use 150K transcripts from 16 freely accessible dialogue data-sets, establishing connections between conversational search and other conversational AI tasks. Different to the dialogues analyzed in this work, our setup does not enable coherent dialogues, but uses the method of simulating parts of a conversation with gaps in between [10]. Trippas et al. [23] analyze different conversational techniques that people use in a search-related audio-only communication channel. Specifically targeting follow-up inquiries, Rosset et al. [19] analyze the effects of conversational question suggestions. Similar to the nudging techniques introduced in our work, but more direct and explicit, these suggestions direct seekers to more engaging experiences by offering amusing, educational, and practical follow-up inquiries. However, we see a clear lack of studies on how to create engaging conversational search dialogues.

The potential for conversations between users and voice assistants to go beyond the question-and-answer format has not been largely present in research on voice assistants. This study, motivated by this prospect, intends to investigate the possibilities of deeper conversation design for voice assistants beyond what is currently available.

3 NUDGING METHODOLOGY

The core concept of this publication is nudging in oral information-seeking conversations, which is, for this purpose, to suggest to the user certain questions without imposing them and to support the creation of an internal locus of control. Users of a conversational information system often have difficulty selecting a question to ask because the number of options is overwhelming. Moreover, it usually takes too much time for oral systems to fully reveal to the user which questions they could answer. To overcome the decision paralysis of a user as they try to formulate a question, and thus to continue the conversation, the system could highlight in each answer it gives the bits for which it would be beneficial for the user to ask questions on—for example, because the system has important or more in-depth information about these bits. The goal of highlighting bits of the answer is to keep them in the user's mind, and thus provide for an easy target for a follow-up question—hence, to nudge the user towards asking about the targeted information. As Thaler and Sunstein [22] put it: "if you want to encourage some action or activity, make it easy."

Each channel of communication comes with its own possibilities for nudging. Visual interfaces can employ typographical features like color and underline, and build on established appearances, e.g., for hyperlinks, to highlight interaction possibilities. For example, when an information seeker reads a Wikipedia article, they can always click on highlighted subtopics to learn more. For an oral conversational system, on which this study focuses, it is however a challenge to highlight such possibilities without compromising the intelligibility or increasing the length of the spoken content.

For this study, we analyze six different nudging techniques, suitable for oral conversations and described in Section 3.1. Some of these techniques are very subtle and work subconsciously. We discuss ethical implications in Section 3.2.

3.1 Employed Nudging Techniques

In addition to a baseline of explicitly suggesting the target concept to the user, the analyzed techniques fall into the following three categories: linguistic emphasis, natural voice emphasis, and artificial voice emphasis. Linguistic emphasis uses linguistic devices such as word choice or sentence structure to draw attention to the target. Natural speech emphasis leaves the spoken content unchanged and attempts to mimic how people emphasize things in a dialogue (e.g., speaking louder). Unlike natural speech emphasis, artificial speech emphasis can not be reproduced by humans. It means to modify the audio using unconventional methods to make the target stand out. We analyze the following six techniques:

Explicit. This baseline technique suggests to the user to ask a question about the target (e.g., "What else do you want to know about [target]"). Thereby, it considerably extends the spoken content's length, feels quickly repetitive, and should thus be used irregularly. However, it is very direct and thus likely very effective.

Repetition. This linguistic emphasis technique repeats the target within the response, for example by resolving a reference (she/he/it) or adding a short phrase. Exact and conceptual repetition of information significantly improves human recall from memory [18]. Repetition is also a common device in argumentation and increases the persuasiveness of arguments [4].

Last. This linguistic emphasis technique changes the information order in the system response to have the target occur at the response's end. People are more likely to remember spoken content from short-term memory that they heard first or last [2]. These phenomena are referred to as primacy and recency effects, respectively. We selected "last" since the recency effect is stronger [2].

Ensemble. This natural voice emphasis technique slows down the target's speaking rate by approximately 10% and increases the perceived volume by approximately 2 dB. It tries to imitate human emphasis and values are chosen accordingly. It is inspired by the work of Chuklin et al. [5], who used such prosody modifications to highlight search terms in spoken search engine result snippets.

Breaks. This natural voice emphasis technique places pauses of 500 ms before and after the target in the spoken response. This technique is inspired by the work of Chuklin et al. [5], too.

Reverb. This artificial voice emphasis technique adds a short reverb to the target in the response. In the context of auditory websites, professional sound designers suggest adding reverb to emphasize important content [27]. However, the effects of adding the reverb were not analyzed by the authors.

3.2 Ethical Considerations of Nudging

In our research, we analyze methods so that they can be employed with the intent to help users who are unsure or even paralyzed with continuing the conversation. As the oral channel is narrow when it comes to transferring words, techniques that use few words to suggest how to continue are clearly preferable. Hence, suggestion techniques that do not add more words but use other effects should be preferred. Additionally, we would like to use techniques that do not impair or distract users with a clear mindset. Thus, subtle or even subliminal techniques are attractive.

However, the less obvious a nudge, the more the question arises whether such manipulations can be ethically justified. Our experiments show the effectiveness of some proposed nudging techniques. Thus, to some extent, the techniques could be misused to maliciously change a user's behavior. Some examples of misuse include manipulating users to buy certain products, influencing a user's political opinion (e.g., by nudging only towards arguments of one stance), or nudging users to commit unethical or criminal acts.

We take the view of Thaler and Sunstein [22], which discuss this issue in detail. Here, we provide a brief summary of the main points. First, it is simply not possible to not use nudging at all. Consider the technique "last," which is one of the most effective in our evaluation. Obviously, there is always some information presented last in a sentence. With this consideration in mind, we argue that analyzing the effects of nudging is much needed to better understand its dangers and possibilities. Second, analyzing nudging and its effect allows pinpointing these effects in actual systems, and thus to evaluate whether these techniques are misused in specific situations. As a clear guideline for authors and designers of conversational systems, we follow Thaler and Sunstein [22] and endorse the usage of John Rawl's publicity principle [16]. The principle loosely reads as: only nudge when you would be both able and willing to defend this nudging publicly. We think this is—at least in the sense considered in this paper—the case for nudging information seekers towards specific targets of a response.

4 EVALUATING NUDGING TECHNIQUES

Since the effectiveness of some proposed nudging techniques is linked to the physiological abilities of the study participants, it is necessary to collect data from a diverse and representative set of people in different situations. We therefore decided to use a crowdsourcing setup with different scenarios. In the following sections, we describe our setup in detail and present the measures we applied to the collected data to answer the research questions.

4.1 Crowdsourcing Responses to Nudges

To gather data on the effects of the nudging techniques described in Section 3.1 we conducted a large-scale crowdsourcing study with 8574 questions asked in response to 30 different information snippets, each in one of 18 different nudging variants. Each participant listened to 10 informative audio snippets in sequence. For each snippet, the participant was asked to name the first question that would come to their mind by typing it in as soon as they have one. They were specifically told to avoid replaying the snippet unless they got distracted or similar, as we assume people would also do when interacting with voice apps in real life. Moreover, after naming the question, they were asked to fill out a "raw TLX" form [7] to measure mental task load,[1] to rate their curiosity in the answer to their own question as a self-reported measure of interest level, and to rate the snippet clarity and naturalness as a measure of possible side effects a nudging technique might have. Figure 1 shows the interface.

Scenarios. To make it easier for the participants to focus on coming up with questions over getting into topics, we avoided topic switches and selected the information snippets to be topically coherent for each study participation. However, to avoid restricting our results to a single topic, we prepared three sequences of 10 snippets each, which we refer to as scenarios. Each scenario has its own introductory text, that is displayed to the participant throughout the study, and which the participant is asked to read before listening to the first snippet (cf. Figure 1). The three scenarios are: (1) argumentation, in which the participant is asked to imagine their government is calling for a public vote on whether to reduce a universal basic income, and is talking to a trustworthy audio app that exists to inform them on the matter; (2) museum, in which the participant is asked to imagine using an audio exhibition app on the German Bauhaus (style in both architecture and product design) to inform themselves on its history; and (3) product comparison, in which the participant is asked to imagine the need to buy several items and asking a voice-based comparison app to provide information to assist them in the buying decision. The argumentation and museum scenarios are inspired by recent work in conversational information seeking [10, 11] and are designed to be as cohesive as possible: each snippet but the first one is the answer to a question that one could ask on what was said in the previous snippet. Whereas the information conveyed in the argumentation scenario are mostly reasons in favor and against (conjectures, opinions, and statistics), the information conveyed in the museum scenario are mostly historical facts. We chose product comparison as the third scenario to include a mundane everyday topic. To make it even more different from the others, we chose to focus on a different product category with every snippet. To get the participants into the product comparison topics quickly, we told them before each snippet that they would now hear the response to them asking about buying an item of the respective category.

Snippets and Targets. We selected the 30 informative snippets (10 for each scenario) to cover a wide variety of information types within the respective scenarios and used Amazon's Polly for text-to-speech synthesis, automating the use of the nudging techniques as much as possible. The presented information was collected from the respective Kialo discussion page[2] for the argumentation scenario, the information used in the study by Kiesel et al. [10] and the Bauhaus Wikipedia page[3] for the museum scenario, and various review websites for the product comparison scenario. We ensured that each information snippet contained at least three points to ask follow-up questions on. From these points, we selected three for each snippet to be the targets we would then nudge towards. We created for each snippet three variants, each with a different target at its end, corresponding to the technique "last." One of these variants is used as the base for other techniques. For the repetition technique we manually repeated the word(s) of the respective target with as few modifications as possible, if possible only by resolving co-references. For the ensemble and reverb techniques we used audio processing software, whereas we introduced the breaks of the breaks' technique in the SSML used for text-to-speech synthesis. For the explicit technique, we employ ten different patterns of asking or hinting at the target that we randomly select for each snippet and append to it (e.g., "What else do you want to know about [target]?" or "Feel free to ask about [target]."). We provide the scenario descriptions, SSML files, and audio files as supplementary material along with this paper.[4] Each participation used one scenario only, with the ten snippets being always in the same order, the "last" technique applied on two randomly chosen snippets to a randomly chosen target, and the same other technique applied on the other eight snippets to a randomly chosen target. We ensure that every participant hears snippets with the "last" technique to use these two for within-subject normalization purposes, as we expect that technique to be the least obtrusive.

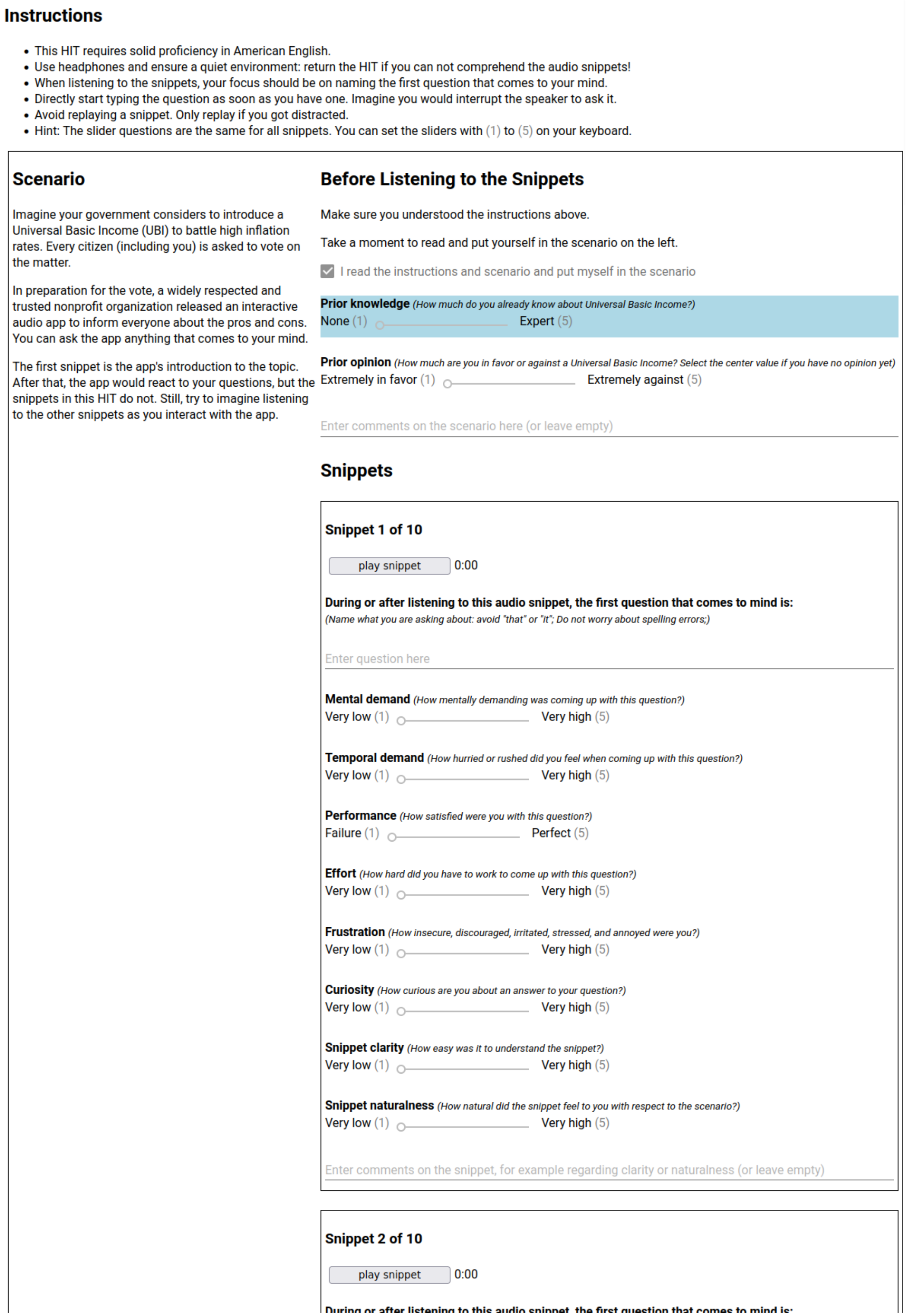
Crowdsourcing. We employed Amazon's Mechanical Turk to reach a diverse group of English speakers for participation. As Huff and Tingley [8] have shown, the demographics of workers on Amazon's Mechanical Turk is comparable to other established survey platforms. Participants were allowed to participate multiple times, but only once for each scenario. We enforced this restriction using Mechanical Turk's qualification system. Furthermore, to increase data quality and general language proficiency, we required that participants had a track record of at least 100 approved HITs and be located in Australia, Canada, the UK, or the US. By steadily increasing the number of participants, we ensured that for every combination of snippet, target, and nudging technique there are at least 10 participants who named a valid question for it. In total, our task attracted 2237 participations, of which we rejected 600 in

[1] We omitted the question on physical demand as coming up with a question has no physical component. To avoid confusing participants, we decided to drop the question completely.

[2] https://www.kialo.com/should-there-be-a-universal-basic-income-ubi-1634

[3] https://en.wikipedia.org/wiki/Bauhaus

[4] Supplementary material: https://doi.org/10.5281/zenodo.7226308



Instructions

- This HIT requires solid proficiency in American English.
- Use headphones and ensure a quiet environment: return the HIT if you can not comprehend the audio snippets!
- When listening to the snippets, your focus should be on naming the first question that comes to your mind.
- Directly start typing the question as soon as you have one. Imagine you would interrupt the speaker to ask it.
- Avoid replaying a snippet. Only replay if you got distracted.
- Hint: The slider questions are the same for all snippets. You can set the sliders with (1) to (5) on your keyboard.

Scenario

Imagine your government considers to introduce a Universal Basic Income (UBI) to battle high inflation rates. Every citizen (including you) is asked to vote on the matter.

In preparation for the vote, a widely respected and trusted nonprofit organization released an interactive audio app to inform everyone about the pros and cons. You can ask the app anything that comes to your mind.

The first snippet is the app's introduction to the topic. After that, the app would react to your questions, but the snippets in this HIT do not. Still, try to imagine listening to the other snippets as you interact with the app.

Before Listening to the Snippets

Make sure you understood the instructions above.

Take a moment to read and put yourself in the scenario on the left.

☑ I read the instructions and scenario and put myself in the scenario

Prior knowledge *(How much do you already know about Universal Basic Income?)*
None (1) ○ Expert (5)

Prior opinion *(How much are you in favor or against a Universal Basic Income? Select the center value if you have no opinion yet)*
Extremely in favor (1) ○ Extremely against (5)

Enter comments on the scenario here (or leave empty)

Snippets

Snippet 1 of 10

play snippet 0:00

During or after listening to this audio snippet, the first question that comes to mind is:
(Name what you are asking about: avoid "that" or "it"; Do not worry about spelling errors;)

Enter question here

Mental demand *(How mentally demanding was coming up with this question?)*
Very low (1) ○ Very high (5)

Temporal demand *(How hurried or rushed did you feel when coming up with this question?)*
Very low (1) ○ Very high (5)

Performance *(How satisfied were you with this question?)*
Failure (1) ○ Perfect (5)

Effort *(How hard did you have to work to come up with this question?)*
Very low (1) ○ Very high (5)

Frustration *(How insecure, discouraged, irritated, stressed, and annoyed were you?)*
Very low (1) ○ Very high (5)

Curiosity *(How curious are you about an answer to your question?)*
Very low (1) ○ Very high (5)

Snippet clarity *(How easy was it to understand the snippet?)*
Very low (1) ○ Very high (5)

Snippet naturalness *(How natural did the snippet feel to you with respect to the scenario?)*
Very low (1) ○ Very high (5)

Enter comments on the snippet, for example regarding clarity or naturalness (or leave empty)

Snippet 2 of 10

play snippet 0:00

During or after listening to this audio snippet, the first question that comes to mind is:

Figure 1: The study interface as displayed on Mechanical Turk, after checking the box that instructions have been read.

which the instructions were not followed and further excluded 672 of doubtful quality (especially bad English skills), resulting in 965 valid participations with 10 questions each. Participants required on average 21 minutes to complete the HIT, for which they were paid USD 3.00, resulting in an average hourly wage of USD 8.57.

Curation. Having collected 9650 questions, we manually annotated each question for whether it should be discarded due to quality concerns (e.g., being irrelevant to the topic, repeated several times by the participant, or simply unclear: 1274 total), and which of the three targets it is about, if any. Since we specifically use the explicit technique as baseline, we also used it to define aboutness for this annotation: we say a question is about the target if the question would fit to "What else do you want to know about [target]?" For an unbiased annotation, we ensured we were aware of neither the applied technique nor the target. Of the 8376 questions we did not discard and use in our analysis below, 4340 (52%) were about none of the three targets, 3657 (44%) on one, 301 (4%) on two, and the remaining 78 questions (1%) were so generic that they were on all three targets.

4.2 Effectiveness Metrics

The effectiveness of a nudging technique calculates how often a user is persuaded to ask a question about a particular target. We measure the "raw" effectiveness e of a technique τ as the ratio between the number of questions n_τ about the nudging target of τ, and the number of all questions N_τ we collected for τ.

$$e_\tau = \frac{n_\tau}{N_\tau} \tag{1}$$

In pilot studies, we found that there are hierarchical biases in nudging effectiveness based on the difficulty of the scenario, task, and targets. Regardless of the nudging technique, we found a large variance in nudging effectiveness across targets. To eliminate effects caused by an imbalance of the number of questions per nudging target, we estimate the normalized effectiveness $\hat{e}$ as a relative metric to the nudging effectiveness per target of the baseline (i.e., the "explicit" technique). Let a target be denoted by t with $t \in T$ and $|T| = 90$ (i.e., three scenarios times ten tasks times three targets). Let $e_{\tau,t}$ be the effectiveness of a nudging technique for an arbitrary target. The relative effectiveness of a technique τ is calculated as follows.

$$\hat{e}_\tau = \frac{1}{|T|}\sum_{t\in T}\frac{e_{\tau,t}}{e_{\text{explicit},t}} = \frac{1}{|T|}\sum_{t\in T}\frac{n_{\tau,t}\cdot N_{\text{explicit},t}}{N_{\tau,t}\cdot n_{\text{explicit},t}} \tag{2}$$

4.3 Speech Quality Evaluation Metrics

To obtain an unbiased view of how the application of nudging techniques affects the perceived speech quality and intelligibility of the audio material, we include automatic metrics for quality estimation. The "Perceptual Evaluation of Speech Quality" (PESQ) [17] is a model that predicts speech quality based on perceptual features that is susceptible to faulty codecs, packet loss, or other network-related distortions. We use an unprocessed audio snippet to compare it with snippets that have been processed using non-linguistic nudging techniques.

Table 1: Raw (e) and relative ($\hat{e}$) effectiveness estimations of the evaluated nudging techniques grouped by scenarios of the crowdsourcing study.

Technique	Scenario							
	Argument.		Museum		Prod. comp.		Total	
	e	$\hat{e}$	e	$\hat{e}$	e	$\hat{e}$	e	$\hat{e}$
Explicit	0.503	1.000	0.481	1.000	0.517	1.000	0.500	1.000
Repetition	0.328	0.639	0.200	0.430	0.155	0.310	0.228	0.459
Last	0.278	0.539	0.227	0.442	0.127	0.283	0.211	0.422
Ensemble	0.247	0.509	0.194	0.389	0.113	0.221	0.183	0.373
Breaks	0.222	0.465	0.189	0.393	0.125	0.242	0.178	0.366
Reverb	0.218	0.403	0.151	0.268	0.088	0.185	0.151	0.287
Total	0.294	0.586	0.235	0.480	0.179	0.356		

5 RESULTS

To answer our research questions, we analyze crowdsourced and annotated data in terms of effectiveness, efficiency, and side effects. To get a clear picture of the influence of nudging techniques, we discard questions written by crowdworkers who did not listen to the audio snippets far enough to be exposed to the nudging techniques. This means that crowdworkers exposed to the "last" technique must listen to the end of the snippet. The "repetition" technique is considered exposed when the participant has listened to the nudging target at least twice. Of the 8376 questions that were not discarded, 797 (~10%) were formulated without being exposed to a nudging technique. These are not included in the following analyses.

5.1 RQ1. How effective are different techniques in nudging people towards a specific target?

To evaluate nudging capabilities of the different techniques in order to answer RQ1, we compare the raw and relative effectiveness scores. Table 1 presents the effectiveness scores of the techniques with respect to the scenarios.

As expected, explicitly suggesting questions about a nudging target is the most effective technique, which results in 50% of the questions about the target. In comparison, repeating the same nudging target is less than half as effective overall. The least effective nudging technique is adding reverb to highlight targets with an effectiveness of about 15% overall.

The results show that there is a strict effectiveness ranking between the technique categories. Linguistic emphasis is more effective than natural voice emphasis. Natural voice emphasis is again more effective than artificial voice emphasis.

How salient a technique is does not seem to affect its effectiveness. Two of the most noticeable techniques, "explicit" and "reverb", are the most and least effective techniques, respectively. However, the interpretation of the salience might differ. For the "explicit" technique, there is no interpretation needed to understand that the participant is suggested to ask about a specific target. With other techniques, it might be unclear to a participant what to do with the information that certain targets are being emphasized. However, the improved recall potential also improves the nudging effectiveness,

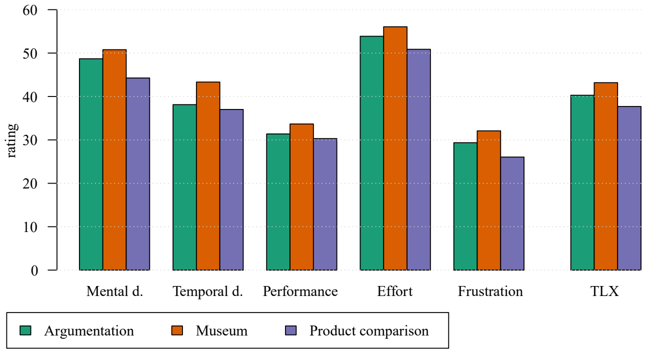


Figure 2: Average participant rating (100 is worst) for each TLX question and TLX score by scenario.

which we see in Table 1. Presumably, being transparent about the intention why certain targets are being emphasized could improve the effectiveness of more subtle techniques even further. This is a pathway for a potential follow-up study.

The effectiveness values vary greatly between the different scenarios. This suggests that the context in which a nudging technique is used plays an important role in its effectiveness. For example, "repetition" is a more effective nudging technique in the argumentation scenario, but mentioning nudging targets last is more effective in a museum context. In our personal estimation, the content of the argumentation scenario and the museum scenario is much more challenging than the content of the product comparison scenario. The overall better effectiveness in these more challenging scenarios indicates that our hypothesis is true, that nudging helps to find questions when it is difficult to find any. Figure 2, which shows the average participant rating of TLX values per scenario, proves that crowdworkers agree that argumentation and museum is actually more demanding than product comparison.

5.2 RQ2. How does nudging effect the seeker's mental load when listening to information?

Figure 3 shows the average participant ratings of TLX values by the studied nudging techniques. It can be seen that the influence of a nudging technique on the demands of a task is rather subtle. The most demanding technique is repetition with the largest delta in frustration. Hearing the same target over and over again seemed to be frustrating for crowdworkers. On average, frustration was three points higher for the repetition technique than for the second most frustrating technique.

Apart from this outlier, there are no non-negligible differences between the techniques in terms of their impact on the demands of a task. Upon further investigation, we found that these TLX values are approximately uniformly distributed for most study configurations. This raises the question whether TLX is an appropriate metric to quantify the influence of nudging on mental demand in a crowdsourcing setting. Further research is needed to obtain more reliable results and definitively answer RQ2.

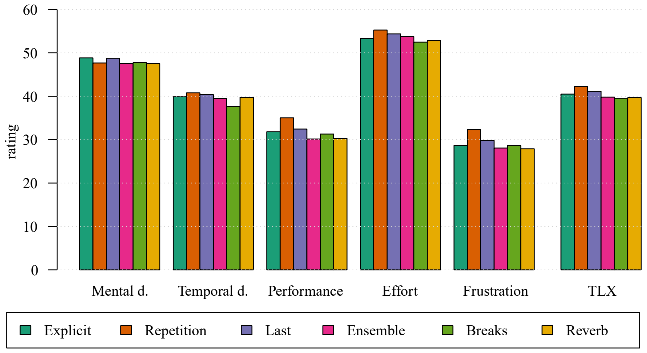


Figure 3: Average participant rating (100 is worst) for each TLX question and TLX score by nudging technique.

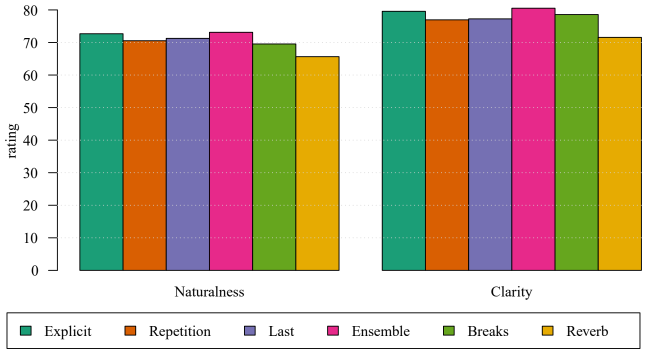


Figure 4: Average participant rating (100 is best) for snippet naturalness and clarity by nudging technique.

5.3 RQ3. How does nudging effect the perceived naturalness of the synthesized speech?

Figure 4 shows how users perceived the clarity and naturalness of the audio snippets after applying the nudging techniques according to their own ratings. Overall, both naturalness and clarity are rated quite high, which is a testament to the quality of the synthesized speech. The deltas are comparably small, similar to the differences with regard to mental load (RQ2).

For clarity, average ratings are roughly between 70 and 80 (where 100 is best), with the average rating for the reverb technique being the only one below 75. Some participants in the comments described the reverb effect as "weird," or conjectured that the reverb (often called "echo") was due to an error. It thus seems this nudging technique needs to be explained to the seekers before it is applied, in order to not confuse them.[5] However, participants also took note of the "breaks" technique similarly, which was still rated as one of the most clear. We see this as evidence that participants were indeed able to distinguish the concepts of clarity and naturalness.

For naturalness, average ratings are roughly between 65 and 75, and thus a bit lower than clarity. The reverb technique received the

[5] Note that some studies show that the effect of some nudging techniques does not vanish when they are explained beforehand [22]. Whether this is also the case for reverb, or whether its effect might even be increased through such explanations, still needs to be investigated.

Table 2: Objective evaluation for perceived speech quality of the different nudging techniqes computed as average PESQ at a sample rate of 16kHz between the unprocessed and the audio with the applied nudging technique.

Technique	PESQ	
	Avg.	Min.
Reverb	3.70	2.52
Ensemble	3.15	1.55
Breaks	2.28	1.17

lowest rating also for naturalness, which seems unsurprising as it is the least natural effect, especially if only applied on a few words of a sentence. We assume the rating is still comparably high, as it is only applied to such a small part of the entire audio snippet. At least in the setting compared in this study, reverb thus seems to be the least suited for nudging seekers.

In addition to the self-reported perceived naturalness of the audio, we also investigated the use of automatic measures of sound quality for this purpose, as these can be applied objectively and at scale. As such measures compare the modified audio to a "perfect" audio, these measures are only applicable for techniques that work without lexical changes (i.e., reverb, ensemble, and breaks). However, as Table 2 shows, the PESQ measure (cf. Section 4) actually rates the audio of the reverb technique to be of higher quality than for the other two techniques. We assume this is because the ensemble and breaks technique introduce temporal modifications, which seem to have a higher impact on PESQ than the additional sound effect introduced with reverb. We thus conclude that PESQ, a standard automated measure of sound quality, is not suited to evaluate the naturalness of nudging techniques, and human ratings are still needed to investigate the question of naturalness in the future.

6 CONCLUSION

How can we offer guidance to users in oral conversations in an unobtrusive way (i.e., without annoying them)? The problem of engaging information seekers in conversation and then sustaining that conversation has hardly been addressed in the literature so far. This paper proposes and analyzes for the first time the use of precise syntactic and auditive modifications of oral system responses to nudge information seekers. Using a large-scale online crowdsourcing study with 965 participations, we employ a contrastive setup that compares the effectiveness of six different nudging techniques in different information scenarios and using different targets for nudging. We find that explicitly suggesting asking about a certain topic causes the participants to do just that in about half the cases for our setup. However, also techniques of linguistic emphasis (word order and word repetition) are quite effective in nudging seekers, causing seekers to ask about the nudged topic in up to 33% of cases depending on the scenario, while being much less obtrusive. Also, natural voice emphasis clearly has a guiding effect. On the other hand, the effect of nudging techniques on the seeker's task load and perceived clarity and naturalness of the audio snippets is rather small. With this study, we hope to have contributed a first effort towards developing conversational methods to start and continue information-seeking conversations, showing the applicability of nudging for offering guidance to seekers.

7 LIMITATIONS

Our study is limited in some regards, which might call for further investigations in the future. First, we did not analyze participants' demographic factors. However, such factors could influence the effectiveness of nudging. For example, some of the analyzed nudging techniques are linked to physiological factors such as the ability to recall information, which is likely correlated with participant age. Second, due to the associated time and economic costs, we limited the number of investigated information scenarios and nudging techniques to three and six, respectively. As we have seen, the design of an information scenario has a major impact on the effectiveness of nudging, which deserves to be explored in more detail. In addition, other interesting nudging techniques should be tested in the future. We aimed to select a diverse set of techniques, but of course our list is not exhaustive. For example, other techniques could be to add an exaggerative adjective to emphasize the target, another technique could work with variations in the pitch of the voice. Third, we did not analyze how the effectiveness of nudging would change if the participants were informed. For example, one could tell the participants that they would hear a reverb effect on topics that other people frequently asked about. Thaler and Sunstein [22] note that the effectiveness of nudging does not decrease when being open about it, but it is unclear whether that holds true also in the context of oral conversations. Moreover, different ways of motivating the nudging techniques could lead to different effects. For example, people might behave differently if told that the highlighted information is frequently visited or if told that it is important in some regard. Also here, more investigations seem necessary.

REFERENCES

[1] Avishek Anand, Lawrence Cavedon, Matthias Hagen, Hideo Joho, Mark Sanderson, and Benno Stein. 2020. Dagstuhl Seminar 19461 on Conversational Search. *SIGIR Forum* 54, 1 (June 2020), 11 pages. https://doi.org/10.1145/3451964.3451967

[2] Alan Baddeley, Michael W. Eysenck, and Anderson Michael C. 2020. *Memory, Third Edition.* Routledge, 2 Park Square, Milton Park, Abingdon, Oxon.

[3] Nicholas J. Belkin, Robert N. Oddy, and Helen M. Brooks. 1982. Ask for Information Retrieval: Part I. Background and Theory. *Journal of Documentation* 38, 2 (1982), 61–71. https://doi.org/10.1108/eb026722

[4] John T. Cacioppo and Richard E. Petty. 1989. Effects of message Repetition on Argument Processing, Recall, and Persuasion. *Basic and Applied Social Psychology* 10, 1 (1989), 3–12. https://doi.org/10.1207/s15324834basp1001_2

[5] Aleksandr Chuklin, Aliaksei Severyn, Johanne R. Trippas, Enrique Alfonseca, Hanna Silén, and Damiano Spina. 2018. Prosody Modifications for Question-Answering in Voice-Only Settings. *CoRR* abs/1806.03957 (2018). arXiv:1806.03957 http://arxiv.org/abs/1806.03957

[6] Joel E. Fischer, Stuart Reeves, Martin Porcheron, and Rein Ove Sikveland. 2019. Progressivity for Voice Interface Design. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19).* Association for Computing Machinery, New York, NY, USA, Article 26, 8 pages. https://doi.org/10.1145/3342775.3342788

[7] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908. https://doi.org/10.1177/154193120605000909

[8] Connor Huff and Dustin Tingley. 2015. "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics* 2, 3 (2015), 1–12. https://doi.org/10.1177/2053168015604648

[9] Shipi Kankane, Carlina DiRusso, and Christen Buckley. 2018. Can We Nudge Users Toward Better Password Management? An Initial Study. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI EA '18).* Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3170427.3188689

[10] Johannes Kiesel, Volker Bernhard, Marcel Gohsen, Josef Roth, and Benno Stein. 2022. What is That? Crowdsourcing Questions to a Virtual Exhibition. In *2022 Conference on Human Information Interaction & Retrieval (CHIIR 2022)*, David Elsweiler (Ed.). ACM, 358–362. https://doi.org/10.1145/3498366.3505836

[11] Johannes Kiesel, Damiano Spina, Henning Wachsmuth, and Benno Stein. 2021. The Meant, the Said, and the Understood: Conversational Argument Search and Cognitive Biases. In *3rd Conference on Conversational User Interfaces (CUI 2021)*, Stephan Schlögl, Martin Porcheron, and Leigh Clark (Eds.). ACM, New York, 5 pages. https://doi.org/10.1145/3469595.3469615

[12] Carol Collier Kuhlthau. 1993. A Principle of Uncertainty for Information seeking. *Journal of Documentation* 49, 4 (1993), 339–355. https://doi.org/10.1108/eb026918

[13] Chelsea M. Myers. 2019. Adaptive Suggestions to Increase Learnability for Voice User Interfaces. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 159–160. https://doi.org/10.1145/3308557.3308727

[14] Filip Radlinski and Nick Craswell. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Human Information Interaction and Retrieval*. ACM, New York, NY, USA, 117–126. https://doi.org/10.1145/3020165.3020183

[15] Shan M Randhawa, Tallal Ahmad, Jay Chen, and Agha Ali Raza. 2021. Karamad: A Voice-Based Crowdsourcing Platform for Underserved Populations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 569, 15 pages. https://doi.org/10.1145/3411764.3445417

[16] John Rawls. 1971. *A Theory of Justice*. The Belknap press of Harvard University Press, Cambridge, Mass. Eleventh printing, 1981.

[17] Antony W. Rix, John G. Beerends, Mike Hollier, and Andries P. Hekstra. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)* 2 (2001), 749–752 vol.2.

[18] Henry Roediger and Bradford Challis. 1992. Effects of Exact Repetition and Conceptual Repetition on Free Recall and Primed Word-Fragment Completion. *Journal of experimental psychology. Learning, memory, and cognition* 18 (02 1992), 3–14. https://doi.org/10.1037/0278-7393.18.1.3

[19] Corbin Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul Bennett. 2020. Leading Conversational Search by Suggesting Useful Questions. In *Proceedings of The Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 1160–1170. https://doi.org/10.1145/3366423.3380193

[20] Hai Son Nguyen, Andreas Mladenow, Christine Strauss, and Katharina Auer-Srnka. 2021. Voice Commerce: Anthropomorphism Using Voice Assistants. In *The 23rd International Conference on Information Integration and Web Intelligence* (Linz, Austria) *(iiWAS2021)*. Association for Computing Machinery, New York, NY, USA, 434–442. https://doi.org/10.1145/3487664.3487724

[21] Robert S. Taylor. 1962. The process of asking questions. *American Documentation* 13, 4 (Oct. 1962), 391–396. https://doi.org/10.1002/asi.5090130405

[22] Richard H. Thaler and Cass R. Sunstein. 2021. *Nudge: The Final Edition*. Yale University Press, New Haven, CT.

[23] Johanne R. Trippas, Damiano Spina, Paul Thomas, Mark Sanderson, Hideo Joho, and Lawrence Cavedon. 2020. Towards a Model for Spoken Conversational Search. *Inf. Process. Manage.* 57, 2 (March 2020), 19 pages. https://doi.org/10.1016/j.ipm.2019.102162

[24] Svitlana Vakulenko, Evangelos Kanoulas, and Maarten De Rijke. 2021. A Large-Scale Analysis of Mixed Initiative in Information-Seeking Dialogues for Conversational Search. *ACM Trans. Inf. Syst.* 39, 4, Article 49 (Aug. 2021), 32 pages. https://doi.org/10.1145/3466796

[25] Fredrik Winberg and Sten Olof Hellstrom. 2002. Designing Accessible Auditory Drag and Drop. *SIGCAPH Comput. Phys. Handicap.* 73–74 (June 2002), 152–153. https://doi.org/10.1145/960201.957235

[26] Pavani Yalla and Bruce N. Walker. 2008. Advanced Auditory Menus: Design and Evaluation of Auditory Scroll Bars. In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility* (Halifax, Nova Scotia, Canada) *(Assets '08)*. Association for Computing Machinery, New York, NY, USA, 105–112. https://doi.org/10.1145/1414471.1414492

[27] Lotus Zhang, Jingyao Shao, Augustina Ao Liu, Lucy Jiang, Abigale Stangl, Adam Fourney, Meredith Ringel Morris, and Leah Findlater. 2022. Exploring Interactive Sound Design for Auditory Websites. In *Conference on Human Factors in Computing Systems (CHI'22)* (New Orleans, LA, USA). ACM, New York, NY, USA, Article 222, 16 pages. https://doi.org/10.1145/3491102.3517695

[28] Qingxiao Zheng, Yiliu Tang, Yiren Liu, Weizi Liu, and Yun Huang. 2022. UX Research on Conversational Human-AI Interaction: A Literature Review of the ACM Digital Library. In *Conference on Human Factors in Computing Systems (CHI'22)* (New Orleans, LA, USA). ACM, New York, NY, USA, Article 570, 24 pages. https://doi.org/10.1145/3491102.3501855