

# Voice Design Compass: Guidance for Developing and Reflecting on Voice Concepts for Embodied Characters

Johanna Magdalena Kuch

johanna.kuch@uni-a.de

University of Augsburg

Anna Lea Reinwarth

German Research Centre for Artificial Intelligence

Bernhard Hilpert

Leiden University

Laura Kanschä

University of Augsburg

Tanja Schneeberger

German Research Centre for Artificial Intelligence

Elisabeth André

University of Augsburg

---

## Research Article

**Keywords:** voice design, embodied characters, human-agent interaction, design framework, design reflection, formative evaluation

**Posted Date:** October 19th, 2026

**DOI:** <https://doi.org/10.21203/rs.3.rs-11221361/v1>

**License:**   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

**Additional Declarations:** Competing interest reported. E.A. is a member of the Editorial Board of the Journal on Multimodal User Interfaces. The authors declare no other competing financial or non-financial interests.

---

# Voice Design Compass: Guidance for Developing and Reflecting on Voice Concepts for Embodied Characters

## Abstract

An embodied character’s voice can shape how we perceive the whole character. Nevertheless, the voice is often overlooked during the planning process, or no explanation is given as to why certain approaches were chosen. We present the *Voice Design Compass*, a conceptual framework for developing, reviewing, and documenting voice concepts for embodied characters. A voice concept describes how a character’s voice should sound and be perceived. It provides a basis for voice selection and production. The Voice Design Compass offers questions to help designers consider early on what influences their voice choices and which unintended effects they might otherwise overlook. These questions address, for example, how a voice relates to the character’s appearance, whom it might remind users of, and whether it could reinforce stereotypes. We developed a literature-informed prototype. Ten domain experts from research and industry reviewed its written questions, examples and workflow in formative think-aloud interviews. Qualitative content analysis identified experts’ accounts of component relevance, process structure, potential support for design work, and reflection. Needs for change concerned missing content, wording, conceptual distinctions, ease of understanding, cognitive load, and overall organization. Experts suggested using the prototype to clarify requirements, communicate intentions, and revisit decisions across project phases, identifying researchers, practitioners, and non-experts as potential users. We used these findings to revise the prototype into the Voice Design Compass. Four work areas cover project context, voice-concept development with four modular Design Sets, review and practical testing, supporting designers in creating and maintaining their own Project Record.

**Keywords:** voice design, embodied characters, human-agent interaction, design framework, design reflection, formative evaluation

# 1 Introduction

The sound of an embodied character’s voice shapes how we perceive it, for example how credible, likable, and humanlike the character seems [10, 19]. Designing that voice involves relating intended impressions to the character’s appearance and role, its listeners, and its use context. However, convenience often drives voice selection [49], and research reports do often not even describe the voices used or explain why they were chosen [62, 75, 76].

There are already various approaches that address this problem, but none of them aims to provide specific guidance during the process of developing a voice concept. Cambre and Kulkarni organize voice-design research around user, device, and context [7]. Fritsch et al. support critical and generative exploration of synthetic voices through open questions, examples, and embodied listening [23]. For a particular character, designers need to connect these perspectives in a *voice concept*: an account of the voice’s intended identity, sound, and listener impressions. They must decide which qualities to pursue together, justify compromises, and identify assumptions to test. Explicit reasons provide a basis for comparing alternatives and revisiting decisions [45].

We developed the *Voice Design Compass*, a conceptual framework for developing, reviewing, and documenting voice concepts. Its modular sets of questions and literature-based examples connect voice intentions with ethical concerns, technical requirements, and reflection on possible tensions between choices. Designers retain decisions, reasons, open questions, and planned tests in their own Project Record. The resulting voice concept provides a basis for selecting or producing a voice and can be revised after trials or changes in the project. The Compass addresses general vocal characteristics across speech and other vocal utterances, whether recorded, live, or synthesized.

This paper makes three contributions:

1. A literature-informed conceptual prototype that organizes voice-design questions, examples, and review activities into a modular process (Section 3).

2. Formative findings from ten domain experts on the prototype’s guidance, difficulties of interpretation and use, and anticipated applications (Sections 5.2–5.4).
3. The revised Voice Design Compass, with an account linking expert feedback and further literature work to changes in its questions, organization, and workflow (Section 6).

## 2 Related Work

This section reviews research on character design, voice perception and its implications for design, existing voice-design approaches, and formats for supporting design reflection. These perspectives frame the design concerns addressed by the prototype and by its revision into the Voice Design Compass.

### 2.1 Character Design

Character design combines a character’s purpose, personality, appearance, and behavior to communicate who the character is and what role it serves [27, 80]. We use *character* to emphasize this designed identity, whereas *agent* emphasizes a system’s capacity to perceive and act in its environment [91]. An embodied agent can therefore also be designed as a character. Harder’s practice-oriented guide develops purpose and personality before visual choices such as body shape, clothing, color, and style [27]. Sloan’s account of virtual character design connects visual and auditory presentation with identity, emotional expression, narrative context, and role [80].

In this paper, *embodied characters* are designed characters with a physical or virtual body, such as social robots, virtual agents, and avatars. Both body and voice contribute to how a character is perceived. Rist and André describe virtual-character design in terms of visual and audible appearance, conversational setting, and behavior, including gestures, facial expressions, and utterances [66]. Listeners can associate particular sounds with bodily features. Fischer and Niebuhr showed that changes in fundamental frequency (related to pitch) and spectral center of gravity (related to vocal brightness) influenced which of two differently sized images of the same robot listeners associated with a synthetic voice [20].

Choices about a character’s vocal characteristics also relate to its persona and intended relationship with users. Wilson and Moore describe how vocal performance and signal processing shape fictional voices in film and television to suit a character’s persona and narrative context [90]. In a study with older adults, spoken content was held constant across serious and playful vocal styles. Perceptions of a robot’s social attractiveness and eeriness depended on how these styles were combined with assistant or companion roles [82].

How a character’s appearance and voice are combined can affect the resulting impression. In a video experiment, Mitchell et al. found that mismatches in face and voice realism were rated as eerier than matched combinations [51]. For a fictional character intended to feel unsettling, a designer might deliberately pair a lifelike face with an artificial-sounding voice. Deliberate departures from expectations may also be considered in relation to social roles. In voice-interface research, Cambre and Kulkarni propose exploring whether voices that challenge role stereotypes could change how users view a device and its associated tasks [7].

Voice can thus help shape who a character appears to be. Choosing vocal characteristics calls for considering both the intended portrayal and the meanings listeners may attach to what they hear.

## 2.2 Voice as a Design Material

We define *voice design* as the selection or shaping of vocal characteristics. We adopt the sociophonic perspective in Sutton et al.’s paper “Voice as a Design Material” because it connects choices about vocal characteristics with the social meanings listeners assign to voices through their experience and the context of use [83]. Such interpretations may draw on social categories, stereotypes, prejudices, and beliefs about which ways of speaking are considered appropriate or prestigious [83]. For voice design, this raises the question of which stereotypes a chosen voice may reinforce.

McAler et al. found that listeners formed similar personality impressions from recordings of “hello” lasting less than a second [47]. For voice design, this suggests that even a character’s brief

vocal output can contribute to its perceived personality. Listeners may agree in their impressions of a voice, but their judgments can also differ from person to person. Lavan and Sutherland found that the relative contributions of shared and listener-specific influences varied across the characteristics judged and the voices presented [39]. These findings support considering differences among the intended users and listeners when developing a voice concept and examining intended impressions with them.

Voice quality can convey information about a speaker beyond the verbal content, including biological, psychological, and social characteristics [40]. Listeners use vocal cues to judge speaker characteristics such as identity, biological sex and social gender, regional or socioeconomic background, and age [72]. The same vocal feature can carry different social meanings in different contexts. Podesva and Callier describe falsetto as a resource for constructing an expressive “diva” persona and as a way to show respect in Lachixío Zapotec [63].

Sutton et al. illustrate how voices can evoke personal memories. A voice quality may remind a listener of a difficult boss, while an accent may remind them of a place they regularly visited on holiday as a child [83]. These associations can arise from particular vocal features without reproducing a specific person’s voice. For a character, they motivate considering unintended associations even when its voice is newly created.

Lavan and McGettigan’s Person Perception from Voices (PPV) model offers a theoretical account of how listeners use prior experience to recognize people and form impressions from voices [38]. It proposes that features of a heard voice are compared with stored representations of familiar individuals, recognizable social types (*personae*), and characteristics such as age or personality. Listeners can therefore form impressions of an unfamiliar person without identifying who they are. In this model, impressions of a speaker’s characteristics can provide clues to who is speaking. The connection also works in the other direction: being familiar with a voice can help listeners judge the speaker’s characteristics. For example, listeners may judge a speaker’s sex more quickly when they already know the voice.

**Table 1** Functional comparison of related voice-design approaches and the Voice Design Compass.

| Approach                                | Primary purpose                                                    | Support provided                                                                                                                                 | Evidence base and scope                                                                                                                                                  |
|-----------------------------------------|--------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Voice Design Compass</b> (this work) | Support voice concept development for embodied characters          | Project context, four modular Design Sets, specific sound ideas, review and testing, with guidance for creating and maintaining a project record | Literature-informed development, qualitative content analysis of ten formative expert interviews on the prototype, and a documented revision to the Voice Design Compass |
| Cambre and Kulkarni (2019) [7]          | Organize smart-device voice design                                 | Three linked lenses: user, device, and context                                                                                                   | Literature-based research framework focused on smart-device voices                                                                                                       |
| Seaborn et al. (2021) [75]              | Organize experimental research on voice in human-agent interaction | Classification framework and voice-body continua                                                                                                 | Meta-synthesis of 89 experimental studies                                                                                                                                |
| Fritsch et al. (2025) [23]              | Expand the design space for synthetic VUI voices                   | Three dimensions and guiding questions for critical and generative exploration                                                                   | Theoretical foundations, three workshops with listening and sonic prototyping, and participant interviews in the third workshop                                          |
| Sutton et al. (2019) [83]               | Address social and cultural dimensions of VUI voices               | Strategies for individualization, context awareness, and diversification                                                                         | Sociophonetic literature and conceptual analysis                                                                                                                         |
| Kim et al. (2021) [34]                  | Describe how practitioners approach VUI naturalness                | Twelve naturalness characteristics and seven design challenges                                                                                   | Interviews with 20 VUI designers                                                                                                                                         |
| Li et al. (2025) [43]                   | Optimize humanoid robot appearance and voice jointly               | Parameter modeling, prediction, and optimization                                                                                                 | User evaluations, neural-network models, genetic-algorithm optimization, and validation                                                                                  |
| Noufi et al. (2023) [55]                | Support context-dependent expressive speech synthesis              | A continuous persona space for exploring tone of voice                                                                                           | Initial thematic analysis of ten expert interviews and a proposed synthesis model                                                                                        |
| Owens et al. (2022) [59]                | Characterize deceptive design in voice interfaces                  | Six interface properties and paired scenarios                                                                                                    | Author brainstorming exercise and a survey with 93 participants                                                                                                          |
| Zhang et al. (2025) [92]                | Test behavioral effects of vocal manipulations                     | Controlled variation of voice characteristics, intensities, and scenarios                                                                        | Preliminary study ( $N = 24$ ) and main experiment ( $N = 36$ ) using Mandarin female synthetic speech                                                                   |
| Wenzel and Kaufman (2024) [89]          | Identify harms from recognition failures                           | Six downstream harms and proposed communication-repair strategies                                                                                | Interviews and design activities with 16 multicultural participants                                                                                                      |

## 2.3 Related Voice-Design Approaches

We compare selected frameworks, strategies, and studies by how they organize voice-design considerations, support particular tasks, or identify risks. These functions can overlap.

Table 1 compares the purpose, support, and evidence base of these approaches.

Some approaches organize multiple considerations relevant to voice design. Cambre and Kulkarni structure smart-device voice design through the linked lenses of user, device, and context. They identify open research questions about personalization, device appearance and role, and changes across situations and time [7]. Seaborn et al. classify experimental research on voice in human-agent interaction across agent forms, contexts, user groups, and relations between voice and body [75].

Sutton et al. propose individualization and context awareness as initial sociophonetically informed strategies for voice output in voice user

interfaces (VUIs) [83]. They caution that adapting voices to personal preferences or particular activities can reproduce prejudices and caricatures. In response, they introduce diversification: considering a wider range of voices beyond socially privileged standards and examining how these choices may reinforce stereotypes or affect the people whose ways of speaking are represented [83].

Kim et al. identify twelve characteristics through which practitioners approach VUI naturalness and seven challenges they encounter [34]. Li et al. model and optimize combinations of humanoid robot appearance and voice within a defined audiovisual design space [43]. Noufi et al. propose a persona-informed synthesis model in which tone of voice occupies a continuous, context-dependent space [55].

Risk-focused work addresses deception, manipulation, and harm. Owens et al. identify six properties of voice interfaces that may enable deceptive design and examine users’ judgments

of corresponding scenarios [59]. Zhang et al. test how manipulated vocal characteristics affect behavior in Mandarin shopping and question-answering scenarios [92]. Wenzel and Kaufman identify downstream harms associated with recognition failures for multicultural users and propose communication-repair strategies [89].

Fritsch et al. provide a particularly close conceptual reference point. Their generative framework connects *Sonic Expressions*, *Situated Role*, and *Interaction*, drawing on sound studies to examine voices through embodied listening, affective experience, and sociocultural and political relations [23]. *Sonic Expressions* includes vocal qualities and their meanings. *Situated Role* addresses the circumstances, relationships, norms, and interests surrounding a voice. *Interaction* considers material form, change over time, and modalities. These dimensions overlap, keeping the sound of a voice connected to the situation in which it is heard. Three workshops combined listening, discussion and experiments with interactive prototypes, with participant interviews in the third. In the first workshop, participants used their own words to describe the voice samples when the suggested terms did not match what they heard. This informed the authors' shift from a fixed vocabulary to open questions and examples. The framework is intended for individual and collaborative work on new or existing VUIs, with guiding questions in the paper and written example answers and related literature references in its supplement [23].

Fritsch et al. propose extending their framework to further technological platforms and design processes involving other professional groups and everyday users. They also identify vocal experiences and relationships over longer periods as a topic for future research [23]. The Compass connects modular Design Sets, review, testing, and a Project Record to develop a revisable voice concept for a particular embodied character.

Pearl's practice-oriented guide discusses persona design for VUIs with or without a visual avatar and addresses voice selection [61].

## 2.4 Supporting Design Reflection

Design support can organize considerations and activities [22] or connect alternatives with reasons for choosing them [45]. Friedman and Hendry's

Envisioning Cards organize themes and activities around stakeholders, time, values, and pervasiveness. Their early field reports describe how selected cards drew attention to additional stakeholders, settings, and implications, followed by revisions to designs or analyses [22].

Questions are one way to direct this reflection. Fleck and Fitzpatrick distinguish describing a situation from explaining it, considering alternative perspectives, and revisiting assumptions. They identify questioning as a way to elicit reasons and explanations, while emphasizing that reflection also depends on the circumstances and support available [21]. This distinction is useful for voice design: specifying an intended quality and examining why it is desirable are related but different activities.

Design-rationale research provides a precedent for connecting questions with reasoned choices. MacLean et al.'s Questions, Options, and Criteria representation organizes design issues, possible answers, and grounds for comparing them [45]. This offers a rationale for asking designers to connect an answer with its reasons, while leaving the choice open to reconsideration.

Modular access makes it possible to select support for a current design concern. In their study of access to ethics-focused methods, Gray et al. examine how designers find methods relevant to their intentions and working conditions. They recommend making parts of larger toolkits usable as methods in their own right, while noting that numerous entry points or insufficient guidance can make selection difficult [26]. Alongside modular access, designers need guidance on which groups of questions are relevant and how they connect to the wider project.

The Voice Design Compass uses questions to specify voice intentions, review to examine possible consequences and alternatives, and a Project Record to retain reasons and planned tests. These activities turn voice-design considerations into a documented voice concept that can guide selection, testing, and revision.

## 3 Conceptual Prototype

We developed a conceptual prototype that organized voice-design questions, literature-based examples, and review activities into a modular process. Figure 1 summarizes the components and

workflow implemented in the study application. Its formative expert evaluation and the resulting Voice Design Compass follow in Sections 4 and 6.

The prototype grouped related design questions into six Question Sets.

### 3.1 Developing the Conceptual Prototype

Our literature work began with the frameworks and overviews of voice design discussed in Related Work. It included evidence from voice interfaces without a body, which we considered in relation to embodied characters and the expectations their appearance may evoke. We formulated questions as a starting point for developing a voice concept for a particular embodied character. The questions were intended to help designers specify desired voice qualities and effects, identify requirements, and explain their reasoning. We used Cambre and Kulkarni’s user, device, and context perspectives to structure the questions [7]. We adapted Fritsch et al.’s guiding questions for a particular embodied character [23], and drew on Sutton et al.’s sociophonetic strategies for voice identity and ethical reflection [83].

We organized the questions around considerations relevant before selecting or producing an embodied character’s voice. Context addressed

users and situations of use. Character Role concerned the character’s function and relationship with users. Embodiment addressed its visible form and the expectations this creates. Identity Relation concerned the identity represented or evoked by the voice and who could select or change it. Perception Effect addressed the impressions listeners should form when hearing the voice. Implementation Constraints and Resources concerned voice production, technical requirements, and project resources; we used literature on these topics to formulate its questions [1, 9, 83].

The prototype focused on helping designers specify a character’s general vocal characteristics and explain their choices. We excluded linguistic content and moment-to-moment variations in delivery, such as changes in emotional expression. These topics are addressed, for example, in research on natural language generation, including content planning and linguistic realization [25], and on vocal communication of emotion [70].

Table 2 maps the questions and strategies we adapted from Cambre and Kulkarni, Fritsch et al., and Sutton et al. to the prototype’s components. Cambre and Kulkarni’s distinction between *Appearance* and *Role* informed our separate Character Role and Embodiment Sets [7].

The prototype and Fritsch et al.’s supplementary framework share concrete examples. Fritsch

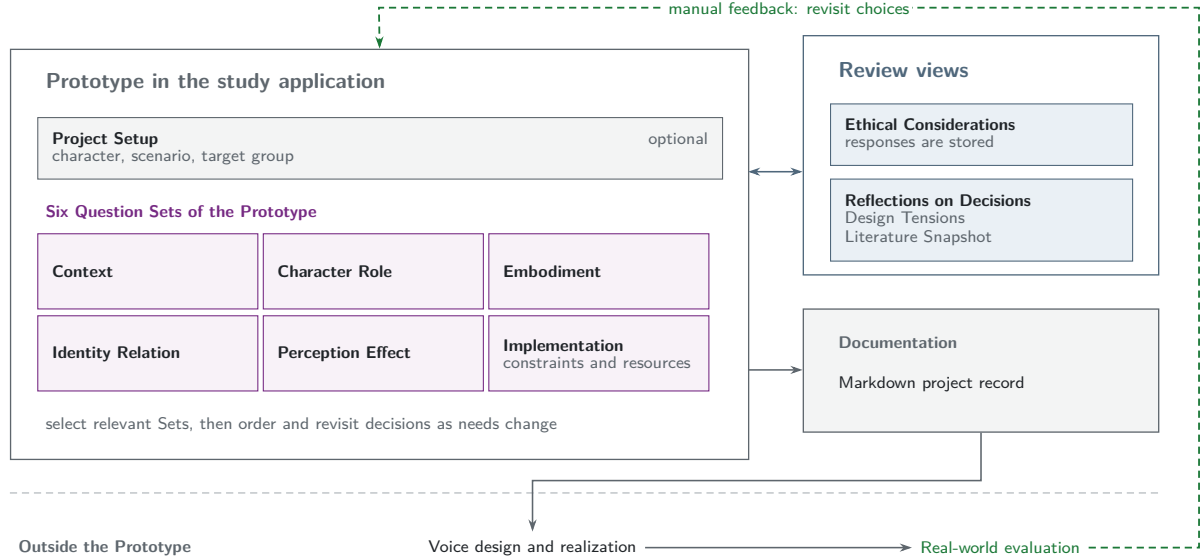


**Fig. 1** Components and workflow of the conceptual prototype as represented in the study application

**Table 2** Questions and strategies adapted from prior frameworks and their use in the prototype.

| Source question or strategy                                                                                                                                                         | Prototype component and adaptation                                                                                                                   |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|
| Cambre and Kulkarni, <i>Device</i> lens: appearance and role [7]                                                                                                                    | Separate Character Role and Embodiment Sets: what the voice should convey about the role, and how it should relate to the visible form               |
| Cambre and Kulkarni, <i>User</i> lens: tailoring a voice to its users, similar or complementary vocal characteristics [7]                                                           | Identity Relation: the relationship between the voice and the user                                                                                   |
| Cambre and Kulkarni: deliberate voice–body incongruence; overstated system capabilities, with Moore’s habitability gap [7, 52, 53]                                                  | Embodiment example; Ethical Considerations                                                                                                           |
| Fritsch et al., <i>Situated Role</i> , Place: “Where and when is the VUI situated?” [23]                                                                                            | Context: how, with whom, and under which acoustic conditions the interaction takes place                                                             |
| Fritsch et al., <i>Situated Role</i> , Social Relation: “What is the interlocutor’s relation to the VUI and its creators?” [23]                                                     | Character Role: the social role the character assumes; the relation to the creators was omitted                                                      |
| Fritsch et al., <i>Interaction</i> , Tangible Form: “What materials is the VUI shaped from, and how are they shaped?” [23]                                                          | Embodiment: which visible properties shape expectations of the voice, and how the voice should meet or deliberately contrast with those expectations |
| Fritsch et al., <i>Sonic Expressions</i> : vocal qualities, volume, affective expression, gender norms, intelligibility, the surrounding sonic space, and effects on listeners [23] | Perception Effect: how the voice should shape listeners’ perceptions                                                                                 |
| Fritsch et al.: persuasive voice design and alternatives to humanlike voices [23]                                                                                                   | Ethical Considerations source                                                                                                                        |
| Sutton et al., <i>Individualisation</i> : voice selection, modification, and system selection based on user information [83]                                                        | Identity Relation: “How is the character’s voice identity selected, and who may modify it after deployment?”                                         |
| Sutton et al., <i>Context Awareness</i> : geographical cues and particular activities [83]                                                                                          | Context: intended situations and their implications, without prescribing different voices for different activities                                   |
| Sutton et al., <i>Diversification</i> : caricatures, stereotyped social qualities, privilege, and representation [83]                                                               | Implicit across the framework; social-group identity in Identity Relation; exclusion and stereotyping in Ethical Considerations                      |

et al. list home and office under *Place*, large and small under *Tangible Form*, and assistant under *Imaginary*. The prototype includes Home / private living space and Workplace / office in Context, Large / Tall and Small / Short in Embodiment, and Assistant in Character Role. Fritsch et al.’s *Paralinguistic Features* examples include pitch, tempo, rhythm, and intonation range [23]. In Implementation Constraints and Resources, the prototype asks whether controls for speaking rate, rhythm, and intonation are required, connecting these sound considerations with production requirements.

Drawing on Porcheron et al.’s account of Amazon Echo use in shared domestic activities and conversation [64], we used the Context example

*Home / private living space* to encourage consideration of other people who might hear the character’s voice.

We took up the concern about capability overestimation through Cambre and Kulkarni, who link highly humanlike voices to inflated expectations of system intelligence [7]. Their account refers to Moore’s discussion of the habitability gap between users’ expectations and system capabilities [53]. Moore also explicitly warns of capability overestimation and argues for considering appearance, voice, and behavior together [52].

Descriptions of the settings, voices, and characters in the papers helped us develop examples of possible answers to the prototype’s questions. These descriptions and findings relevant

### a) Answering a Question Set

Where will the character be encountered?

• How to answer

Examples from the reviewed literature

Hover over or focus an option for details.

☒ Home / private living space

☐ Car / driving

☐ Hospital / healthcare

### b) Reviewing an ethical concern

#### Consent & Identity

Check how the project uses, stores, shares, or resembles a recognisable person's voice.

1. Does the voice refer to a real, known, familiar, or recognizable person?

No Yes Unclear N/A

⚠ This answer identifies an issue that should be reviewed before implementation.

2. Has informed consent been obtained for recording, storing, using, or sharing the voice?

No Yes Unclear N/A

**Fig. 2** Cropped views of the study application with the interface’s original wording: (a) a Context question with a selected literature-based example; (b) a response in Consent & Identity that calls for follow-up, while the next consent question remains unanswered. The excerpts show how choices and review needs were displayed

to voice choices also informed the questions. Questions arising during development guided further searches and reading. Boell and Cecez-Kecmanovic likewise describe how reading helps identify further relevant literature and refine subsequent searches [4].

More focused literature on voice–appearance relations, voice identity and cloning, and listener judgments informed the questions and literature-based examples within the Question Sets. For instance, the Embodiment material connects robot size with voice choices through Fischer and Niebuhr’s study [20]. The Identity Relation material includes cloning the user’s own voice as studied by Kuch et al. [36].

The six Question Sets contained 34 questions and 137 predefined answer options covering context, character role, embodiment, voice identity, intended perceptions, and implementation constraints and resources. These considerations were connected: ambient noise may alter intelligibility requirements, embodiment may create expectations about voice, and available data or expertise may constrain production. Two Review Views supported reflection across the answers: Ethical Considerations, with 28 questions about potential harms and responsibilities, and Reflections on Decisions, which related current intentions to literature findings and 25 rule-triggered Design Tensions.

In Sets 1–5, the literature-based examples were intended to inspire project-specific answers and explain why the questions were included. Set 6, Implementation Constraints and Resources, listed common implementation conditions.

## 3.2 Implementation as a Study Application

We implemented the conceptual prototype as a password-protected *study application*, a web application developed with Streamlit, so that experts could work through it using a voice-design scenario.

An opening page introduced voice concept development with an example use case and the rationale of the approach, and gave an overview of the prototype’s components. The application comprised the optional Project Setup, an editable route through the six Question Sets, the two Review Views, and a final page for exporting the project record and previewing a voice-generation prompt.

To establish the project context, the optional Project Setup captured a character name, a brief application scenario, and an intended user group. Users began with an empty Question Set route and could select, order, remove, and revisit Sets. Unselected Sets remained available, and both the route and recorded choices could change as the project developed. This reflects problem–solution co-evolution, in which problem framing and candidate solutions develop together [15].

Questions offered single or multiple selections, options labeled *not yet defined*, and optional free text for project-specific answers or unresolved questions. Set summaries kept current choices visible. Figure 2 shows example response states on a Question Set page and in Ethical Considerations.

*Ethical Considerations* served as the prototype’s Ethics Checkpoint and offered 28 questions about potential harms and responsibilities, organized into six topic groups. All questions

remained visible, and users chose which to answer. Responses used statuses such as Yes, No, Unclear and N/A, whose meaning depended on the question. The application highlighted critical and *Unclear* responses for follow-up.

*Reflections on Decisions* combined Design Tensions and the Literature Snapshot. The application displayed curated Design Tension entries inviting discussion of particular combinations of selected answers, with explanations, sources where available, and links back to the relevant questions. Proposed resolutions and alternatives were author-created reflection aids. The supporting studies did not necessarily examine the displayed combinations or resolutions.

The Literature Snapshot automatically assembled source-specific findings linked to selected predefined answers in Sets 1–5 and grouped them into five types of finding.

The final page compiled the recorded answers and reflections into a downloadable text summary in Markdown format. It also included an experimental preview of a voice-generation prompt (a text instruction for a possible generation tool), discussed as a software extension in Section 8.

The prototype’s six Question Sets, examples, Ethics questions, Design Tension entries, and review explanations are documented with selected accompanying sources in Online Resource 1.

## 4 Formative Expert Evaluation

To guide further development, we conducted a formative expert evaluation of the prototype’s preliminary questions, literature-based examples, and proposed process, identifying useful elements and needs for clarification, revision, or additional content.

The evaluation focused on participants’ understanding of the prototype, their judgments of its six Question Sets and Review Views, their use of the editable route and Markdown record, and their views on conceptual boundaries, possible revisions, and application contexts.

The following research questions retain their registered wording. In these questions, *Voice Design Compass* denotes the earlier conceptual prototype.

*RQ1:* Which elements of the Voice Design Compass do experts consider missing, redundant, unclear, or misplaced?

*RQ2:* Which elements do experts consider relevant, useful, and actionable for voice design?

*RQ3:* For which professional roles, expertise levels, design tasks, project stages, and goals do experts consider the Voice Design Compass suitable or unsuitable?

The study was registered on the Open Science Framework (OSF)<sup>1</sup> after the first of the ten interviews included in the main-study corpus and before manual qualitative coding. The prototype application overview, interview guide with its topic-coverage matrix, relevance-rating plots, revised Voice Design Compass Working Pack, and revision rationale and change mapping accompany this article as Online Resources 1–5.

### 4.1 Study Design and Recruitment

Before this evaluation, three developmental think-aloud walkthroughs of earlier prototype versions, presented through the study application, informed refinements to Implementation Constraints and Resources, the indication of optional elements and dependencies, the Review Views, and a Context question.

An online presurvey provided study information and supported informed consent. It collected participants’ professional or research context, prior experience, work region, a free-text description of expertise, and ratings of their own expertise in the topic areas covered by the prototype. We used these responses to assign interview topics according to expertise and to describe participants’ experience. Missing self-ratings were left blank.

We conducted individual semi-structured think-aloud walkthroughs using prepared questions, written examples, and the study application. The interview guide provided structure while allowing follow-up questions in response to each expert’s account [13]. Individual walkthroughs let the interviewer follow each expert’s interpretation of specific material and ask about its relevance, omissions, and possible revisions in relation to a scenario and the expert’s experience. We used the interview feedback to revise the prototype, similar to Kepper et al.’s use of think-aloud testing

---

<sup>1</sup>[https://osf.io/wzf87/overview?view\\_only=7aa519fa4158405c840079c253e242ec](https://osf.io/wzf87/overview?view_only=7aa519fa4158405c840079c253e242ec)

to refine the content and navigation of a digital planning tool [33].

The individual format kept the experts' accounts separate. Workshops can instead support shared exploration and negotiation [58].

The sessions combined verbalization during navigation with follow-up questions to elicit the reasons behind experts' interpretations and judgments and clarify ambiguous comments. Explanatory probing differs from nondirective concurrent think-aloud [5]. We therefore interpret the accounts as elicited expert feedback, including the influence of the interviewer's questions.

The written questions, examples and study application focused the review on the considerations and workflow involved in developing a voice concept. This scope is consistent with Lim et al.'s account of prototypes as selective representations of the aspects of a design being explored [44]. No recorded or synthesized voice samples were presented. Acoustic realization, the relation between written descriptions and sounds, and listener responses were outside the scope of this evaluation.

The study used criterion-based purposive sampling [60] to obtain complementary disciplinary and practical perspectives on the six Question Sets and two Review Views. The registered plan aimed at approximately ten experts with complementary backgrounds, a target informed by empirical work on sample sizes in qualitative interview studies [28]. As specified in the registered plan, thematic saturation was not used as a stopping criterion because participants reviewed different parts of the prototype according to their expertise. We report the components covered and discussion depth in Section 4.3.

Potential participants were identified through contacts from previous projects and conferences, publications on related research, and introductions through existing contacts. Some participants were already known to the researchers before recruitment.

We drew on the Standards for Reporting Qualitative Research (SRQR) when reporting this study [57].

Participant selection drew on known expertise and prioritized the relevance and breadth of experience and complementary perspectives across completed interviews.

## 4.2 Participants

Ten experts ( $N = 10$ ) each completed the online presurvey and one interview for the formative expert evaluation. These ten interviews constitute the qualitative analysis corpus. The three earlier developmental walkthroughs were excluded.

Relevant experience ranged from 3 to 25 years, with a median of 7 years, and all reported work regions were in Europe. Participants came from different fields of research and industry, contributing experience in HCI, XR, audio and speech, UX, media production, AI, and ethical or social questions concerning voice and identity.

Table 3 links these backgrounds to the expert IDs used in the results. Interview details and quotations are reported in deidentified form.

**Table 3** Expert backgrounds and self-reported years of relevant professional or academic experience.

| ID  | Background                                                                   | Years |
|-----|------------------------------------------------------------------------------|-------|
| E01 | Audio generation and XR research                                             | 7     |
| E02 | UX research/design in an agency; psychology and AI personas                  | 3.5   |
| E03 | HCI; development and use of AI systems in mobility                           | 7     |
| E04 | Audiovisual production; 3D audio and immersive experiences                   | 8     |
| E05 | HCI and XR research; avatars, embodiment and education                       | 7     |
| E06 | Sound recording, music production and sound design in higher education       | 25    |
| E07 | Text-to-speech and voice generation; technical implementation and evaluation | 5     |
| E08 | Conversational AI, speech interaction and HRI research                       | 4     |
| E09 | Computational linguistics, HCI and AI research                               | 11    |
| E10 | HRI research; gender-ambiguous voices, stereotypes and inclusion             | 3     |

Backgrounds summarize the presurvey responses. AI: artificial intelligence; HCI: human-computer interaction; HRI: human-robot interaction; UX: user experience; XR: extended reality.

### 4.3 Procedure

The first author conducted all ten interviews in the main study remotely as semi-structured think-aloud walkthroughs using screen sharing. Across all interviews, the guide retained a common core of questions about the framework’s purpose, workflow, scope, and comprehensibility.

At the start of each interview, participants were briefed as expert reviewers.

Participants received information about voluntary participation, withdrawal rights, pseudonymous analysis, and access-restricted institutional storage, and provided informed consent.

The interviewer asked experts to imagine advising someone on the initial voice-design considerations for an embodied character and to assess whether the prototype would help and where they would proceed differently. Participants could draw on their own research or practice and optionally enter scenario information in Project Setup. For example, E08 considered a robot that explains exhibits to museum visitors.

The interviewer assigned Question Sets and Review Views according to participants’ expertise and the components requiring additional coverage. Given the prototype’s scope, each interview covered selected components alongside the common questions. Participants explained their interpretations while navigating the assigned material and were encouraged to comment on content, workflow, and interface behavior throughout the session. Follow-up questions addressed clarity, relevance, completeness, conceptual boundaries, and underlying assumptions, asking for concrete examples, missing considerations, and proposed revisions.

Topic coverage sometimes shifted during the semi-structured interviews in response to participants’ interests or time constraints. For example, one participant agreed to review a second Set in full after the interviewer checked whether time remained. In some other interviews, later parts were shortened because time was limited.

Across the ten main interviews, each prototype Question Set was worked through in full twice, except Identity Relation and Perception Effect, which received three complete walkthroughs each. Ethical Considerations received one complete walkthrough and four partial reviews of selected questions within the remaining interview time,

and two interviews reviewed the framework as a whole (general framework reviews, Online Resource 2). Shorter scans within the main interviews provided additional coverage. A retrospective matrix in the supplementary interview guide describes discussion depth for E01–E10 separately from complete walkthrough counts (Online Resource 2).

In addition to the interview questions, participants rated the relevance of individual framework questions, Question Sets, and other components on a seven-point scale. We used the same rating question, with the item named as appropriate (English translation): *“How relevant do you find this [Set/question/component] on a scale from 1 to 7, where 1 means not at all relevant and 7 means very relevant?”*

In some sessions, participants asked how answers depended on the scenario or what the relevance ratings referred to. The interviewer explained how answers related to a specific scenario and clarified which component’s relevance to voice design was being rated.

Depending on the components covered in their interview, participants also reviewed Ethical Considerations, Design Tensions, the Literature Snapshot, and the Markdown export. We asked participants who saw the experimental preview whether they considered the idea of a voice-generation prompt relevant or useful.

The interviews concluded with questions about the framework’s potential value, risks, intended users, application contexts, and required revisions.

Interviews lasted between 32 and 90 minutes (median 56 minutes). Eight were conducted in [REDACTED] and two in English. The interviews were automatically transcribed using Microsoft Teams. The second and fourth authors manually revised the transcripts.

Quotations are reported in English and attributed using the participant identifiers E01–E10. The first author translated the non-English excerpts.

An edited English interview guide with the coverage matrix is provided as Online Resource 2.

Several aspects of the study differed from the registered plan. Most interviews took longer than the estimated 30 to 45 minutes. The upper anchor of the rating question was spoken as *very relevant*

rather than *essential*. The registered plan specified five coverage labels. After the semi-structured interviews, we chose four levels to describe the varying depth of participants’ actual contributions. Online Resource 2 lists both schemes and defines the labels used. The decision to represent verbal rating ranges by their midpoints was made during analysis.

## 4.4 Analysis

The analysis examined how experts assessed the prototype and what their feedback suggested for its development and use. We summarized the relevance ratings descriptively and analyzed the interview transcripts using qualitative content analysis (QCA) [46].

### 4.4.1 Descriptive analysis of ratings

We summarized relevance ratings of individual questions, Sets, Review Views and application features. Participants reviewed different material, so response counts varied by item. Four responses given as ranges were represented by their midpoints. All rating plots and the treatment of range responses are provided in Online Resource 3.

### 4.4.2 Qualitative analysis

We analyzed the interview transcripts using Mayring’s qualitative content analysis and the QCAmapp software [18]. For each of the three research questions, two researchers independently analyzed all ten transcripts in full and developed categories from the interview material (inductive category formation). They used the same coding settings and the following rules to select relevant passages, regardless of where these appeared in the interviews. For RQ1, they marked each passage indicating that an element of the prototype was missing, redundant, unclear, or assigned to an unsuitable component. For RQ2, they marked passages concerning the relevance or usefulness of elements for voice design, or their actionability, meaning whether they could be put into practice. For RQ3, any interview statement relevant to the prototype’s suitability and possible uses could be considered.

Category labels captured the central point of each RQ-relevant statement as concise, phrase-like entries. Coders interpreted each passage in

the context of the participant’s complete interview and the prototype as presented through the study application. Multiple relevant passages from the same transcript could be coded.

The first author coded all ten interviews for RQ1–RQ3. The second author coded all ten interviews for RQ1 and RQ3, and the fourth author coded all ten interviews for RQ2. To prepare for consolidation, the second and fourth authors also coded selected interviews within RQ1–RQ3 beyond their respective full coding assignments. The independently developed categories were then consolidated jointly by the first, second and fourth authors for RQ1 and RQ2, and by the first and second authors for RQ3.

The first author led the prototype’s development and also conducted the interviews. The second and fourth authors had contributed to the prototype’s development and were familiar with it, but were less deeply involved in its development than the first author. Contextual knowledge can help situate statements within the interview and design context, while requiring ongoing critical reflection on how prior understanding may influence interpretation [2].

Independent category development followed by joint consolidation provided a basis for comparing interpretations against the interview material. We did not compute an inter-coder agreement statistic. McDonald et al. discuss when agreement measures suit such processes [48]. Comparing analyses from researchers with different levels of involvement was intended to mitigate confirmation bias in interpreting feedback on the prototype. Shared assumptions arising from their involvement in its development could still influence the analysis.

## 4.5 Revision after the Expert Interviews

We used the consolidated interview findings to revise the conceptual prototype. Our design decisions combined interpretations of the feedback with subsequent literature work. Context became the shared starting point; Character Role, Embodiment, Identity Relation and Perception Effect were regrouped as Function and Relationship, Appearance and Embodiment, Voice Identity and Familiarity, and Intended Voice Impressions. Implementation Constraints and Resources

became Technical Reflection, and the Ethical Considerations review was retained. Eight questions were formulated after the interviews and were not reviewed during those interviews: two Appearance and Embodiment questions (further physical features and equipment; movement, posture and other visible signals), the optional Specific Sound Ideas question, and the five Shared Reflection questions (Section 6). Of the 178 sources cited in the Appendix, 37 were added during the subsequent literature work, including sources identified retrospectively for existing prototype questions. Online Resource 5 documents how expert findings, our interpretation, later literature work and design decisions relate to the changes, including alternatives we did not adopt. We also revised functions that experts valued (Section 5.2): we replaced the answer-triggered Design Tension notices in Reflections on Decisions with open Shared Reflection questions. These questions retain attention to conflicts and stereotypes while asking designers to identify relevant issues themselves. The automatic Literature Snapshot from that Review View remains a possible feature for a future software implementation. The revised Compass has not yet been evaluated in project use. Later informal feedback is reported separately in Section 7.4.

## 5 Results

The findings concern the prototype as explored through the study application, including feedback on software-specific features. The qualitative findings address perceived relevance, usefulness and actionability (RQ2), revision needs (RQ1), and anticipated uses and limits (RQ3). We first provide a brief descriptive summary of the relevance ratings as supplementary context. Question numbers in this section refer to the evaluated prototype. The revised questions are listed separately in the Appendix.

Descriptions of the consensus categories and their scope are provided in the supplementary material on the Open Science Framework (OSF). Our subsequent revision decisions are documented separately in Online Resource 5.

### 5.1 Descriptive Ratings

Higher ratings on the seven-point scale indicate greater perceived relevance of the item under discussion. The ten interviews yielded 129 relevance ratings across different items. Table 4 lists the overall ratings per component. Ratings of Implementation Constraints and Resources varied: the individual questions on recordings, budget and personnel received ratings ranging from 1 or 2 to 7. Character Role received middling overall ratings although each of its two questions received one rating of 5 and one of 6. Participants rated multiple items, so the values describe judgments by the same ten people on different items. Online Resource 3 provides all rating plots with distributions and response counts for individual questions and components.

**Table 4** Overall relevance ratings per prototype component (1 = *not at all relevant*, 7 = *very relevant*). Each value is one participant’s rating. Two participants gave ranges, represented here by their midpoints: 5–6 as 5.5 for Embodiment and 4–5 as 4.5 for Identity Relation. The same relevance question was also applied to individual questions within the prototype; those ratings are shown in Online Resource 3.

| Component                                | Ratings         |
|------------------------------------------|-----------------|
| Context                                  | 3, 6, 6, 6      |
| Character Role                           | 3, 3, 5, 5      |
| Embodiment                               | 4, 5, 5.5, 6    |
| Identity Relation                        | 4.5, 5, 5, 5, 7 |
| Perception Effect                        | 5, 5, 5, 6, 6   |
| Implementation Constraints and Resources | 1, 3, 6, 7      |
| Question Set selection                   | 6, 7            |
| Adjustable Set order                     | 4, 4            |
| Ethical Considerations                   | 5, 7            |
| Reflections on Decisions                 | 5, 5, 6, 7      |
| Design Tensions                          | 5, 6, 6, 7      |
| Literature Snapshot                      | 4, 4, 5, 6      |
| Final page                               | 4, 7, 7, 7      |
| Markdown export                          | 2, 5, 7, 7      |
| Voice-generation prompt idea             | 5, 7, 7         |

## 5.2 Perceived Relevance, Usefulness and Actionability (RQ2)

The researchers' consensus category system for RQ2 comprises four overarching groups: *Relevant Elements of Architecture*, *Structure*, *Enabling*, and *Reflection*. Within these groups, the hierarchy comprises 19 categories, 30 subcategories and 14 subsubcategories.

**Relevant Elements of Architecture.** This group concerns the framework elements that experts considered relevant or useful. Experts identified relevant or useful aspects of the introduction, all six Question Sets, the Literature Snapshot, Design Tensions, and Ethics Checkpoint. E02 particularly valued the notices about conflicts and stereotypes in a design and wanted them retained. The overall Question Set selection was considered useful. Experts also valued the final overview and its Markdown representation, as well as the modular organization with adjustable Set order and additional information. E09 described different orders for Character Role, Identity Relation, and Perception Effect depending on the starting perspective, explaining (translated): *"I think it makes sense that they are so modular."*

Experts explicitly identified five questions as relevant or useful in their qualitative feedback: the Identity Relation questions on voice reference (prototype Q1), intended identity relation (prototype Q2), and selection and control (prototype Q3); the Perception Effect question on social qualities (prototype Q2); and the Implementation question on technical control (prototype Q10). The corresponding ratings varied, for example 3, 4 and 5 for selection and control (Online Resource 3).

Experts described the study application's Markdown documentation download, experimental prompt idea, and automatically assembled literature as potentially useful. E02 considered some form of export important, and E09 found it valuable that exported decisions could be supported with literature, while noting that the final output logged the selected answers with their conclusions in quite a lot of text. Experts associated the prompt suggestion with potential time savings. These comments concern anticipated benefits of software support (Section 8).

**Structure.** Structure concerns experts' appraisals of the workflow and its role in organizing project work. Experts valued the process guidance and introduction to voice design. Experts described the prototype as a starting point for their work and a way to revisit decisions. They also appreciated being asked questions they might not otherwise have considered. Asked about the greatest benefit, E10 stated: *"I think the greatest benefit is that it helps me reflect on things that probably I didn't think about before."* Reflection was valued as an explicit activity with designated places in the process.

**Enabling.** Enabling concerns experts' accounts of support for accomplishing project tasks and for broader learning and professional activities. At the project level, experts described the Character Role questions as support for initial role framing and Implementation Constraints and Resources as support for anticipatory implementation planning. They also identified potential support for forming requirements, identifying evidence relevant to decisions and the final voice, and communicating results, ethical considerations, decisions, related work, and requirements. Exploring different paths was another form of support they described.

Experts also described potential support beyond the immediate project, including consultative support, making influences on voice design visible, acquiring scientific skills, and awareness of technology trends.

**Reflection.** Experts valued the prototype's support for reflection. They reconsidered or saw a need to reconsider ethical issues, decisions, priorities, possible revisions, and technical considerations. Ethical Considerations was explicitly associated with revisiting decisions. Prioritization also emerged as a topic of reflection, although it was not an explicit component of the prototype.

## 5.3 Revision Needs Identified in RQ1

The consolidated RQ1 findings cover six category groups: *Missing Content*, *Unclear Wording*, *Conceptual Boundaries*, *Conceptual Comprehension*, *Cognitive Load*, and *Architecture*.

**Missing Content.** Missing Content included both requests for additional considerations and uncertainty about their evidence base. Experts

described voice design as underexplored and noted that plausible considerations may lack the research needed to justify a recommendation. Suggestions concerned appearance, identity, personality, voice design in mental-health contexts, end-user safety, technical conditions, changing situations, and voice-signal manipulation. We treated these suggestions as candidates for further examination before deciding whether to add questions or examples (Section 4.5).

**Unclear Wording.** Unclear Wording concerns terms that experts found difficult to interpret and their suggestions for alternative labels.

**Conceptual Boundaries.** Conceptual Boundaries concerns unclear distinctions and overlaps between Sets, questions, and response options, together with misplaced or conflated concepts. Reported overlaps included Character Role and Perception Effect, Identity Relation and Perception Effect, and Embodiment and Perception Effect. Referring to Character Role (Set 2), Identity Relation (Set 4), and Perception Effect (Set 5), E09 explained (translated): *“I definitely see an overlap between 2, 4 and 5, but I can also see why they are separate.”* E09 expected to work through them together because considering a character also involves its identity and intended impression. Within Identity Relation, the distinction between the voice reference in prototype Q1 and the intended identity relation in prototype Q2 was also unclear.

**Conceptual Comprehension.** Conceptual Comprehension concerns difficulty interpreting questions, response options, and the framework’s functions and limits. Three Context concerns arose. A target-group concern asked whether the relevant people were those speaking to the character, those addressed by it, or others who could hear it. Another concern involved unclear distinctions between the available age-group options. A publicness concern asked how to distinguish the available listening situations. These concerns require clear explanations of the decisions being made.

At framework level, experts were not always clear about what an answer would lead to. During the discussion of Design Tensions, E06 asked (translated): *“Does the system then detect that some things do not quite fit together and point that out again? Then what you clicked would not be quite right after all, would it?”* This followed the

interviewer’s description of possible contradictions between answers. E06 accepted the subsequent clarification and considered the function useful. E08 likewise valued the notices but asked about their implications (translated): *“In general, I definitely think it makes sense to point out tensions like these. Do I have to resolve them?”* These exchanges indicate uncertainty about whether a tension signaled an incorrect choice or a required revision. The *Scope* findings within *Conceptual Comprehension* also concern what the prototype can deliver, how far literature findings can predict a response, and what Ethics, the Literature Snapshot, its five-part organization, and Design Tensions are meant to do. In discussing feedback on Ethical Considerations, E05 also requested positive feedback on choices that might be unproblematic, alongside notices about possible difficulties.

**Cognitive Load.** Experts criticized the cognitive load associated with complexity and abstractness, assumed expert knowledge, extensive text, the number of questions and reflection tasks, and the overall length. While reviewing Design Tensions, E08 wanted less visible text and suggested making details expandable when relevant.

**Architecture.** The Architecture findings concern instructions for use, the entry point and workflow, and candidates for removal. E01 argued that Context should be compulsory (translated): *“You really need to work through [Set 1], because otherwise you end up designing without taking the context into account.”* E07 wanted technical limitations considered after the character’s design choices, explaining: *“Once a user select[s] those things, the requirements and the constraints change a lot.”* When explicitly asked whether this should be before or after ethical considerations, E07 chose after. Asked where it would belong in the overview, E08 proposed a separate implementation item below Ethical Considerations.

Relevant ethical considerations were requested within the Sets while retaining an Ethics Checkpoint. E10 explained: *“I still think that these ethical considerations should be at the end. So what I was saying before is more like a separate ethical consideration related to a specific set.”* E10 also recognized that including Ethics in every section would increase complexity. E01 preferred initial creative freedom followed by review and possible

reconsideration. These comments express different preferences about when to reflect on ethical concerns.

Prioritization was requested for the broad social-qualities question in Perception Effect and for Implementation Constraints and Resources. E06 described the difficulty of choosing among many relevant qualities (translated): *“You would need either to construct a specific case or to set a priority. Otherwise, you could tick everything without any problem.”*

Further requests concerned documenting the project goal and making testing and iteration explicit. E02 asked for a clearer reminder that voice design needs testing, describing a design as initially an idea or hypothesis whose reception should be examined before retaining it for long-term use. Views on implementation also differed: the findings included both a request to make implementation considerations compulsory and suggestions to remove technical control (prototype Q10) or the operating-constraints question (prototype Q9). E07 considered technical control a downstream implementation concern. When asked about its possible removal, E08 argued for retaining it, for example to let listeners adjust speech rate or volume according to their needs.

## 5.4 Potential Uses and Limits Identified in RQ3

RQ3 concerns where experts thought the prototype could be used and where they questioned its usefulness. The findings cover five areas: *Users (Who)*, *Applications (What)*, *Point of use (When)*, *Motivation (Why)*, and *Contraindications*. They describe anticipated suitability during the interviews.

**Potential users.** This group concerns potential users of the design aid, distinct from the people who would interact with the resulting character. Experts named researchers, UX/UI practitioners, designers, audio professionals, and advertising professionals as possible users. Teams and non-experts were also identified. After the walk-through, E09 described a possible benefit for people who need to design a voice without working professionally in the area (translated): *“And for those who do not work professionally in this area but have a need for it, I think it is considerably more useful, because it brings this expertise*

*together and makes it available in a framework [...]”* E09 also considered it a useful possible addition for professionals.

**Applications.** Experts identified application domains and design purposes, including education, service, and video games, as well as persona development, accessibility, and verbal interaction.

**Point of use.** Experts described possible uses in early and conceptual project phases, including after the project scenario had been specified. Use during active voice design, repeated use through iterations, and retrospective consideration were also identified. On first viewing the opening overview, E09 understood the sequence as considering the target group, goals, and application before answering the questions, with use of the prototype preceding implementation.

**Motivation.** Reasons to consider using the prototype included a limited budget, a specific project context, ethically sensitive applications, and the use of AI voices. Anticipated activities included defining voice requirements and clarifying implementation constraints and resources, designing a suitable voice, and exploring different versions during prototyping. Experts also described possible uses for communicating design intentions, documenting and justifying decisions, grounding them in literature, and conducting ethical review. Becoming familiar with voice design, communicating research knowledge, and accessing relevant literature were further motivations. Literature access was considered in the setting of the study application, which included the automatically assembled Literature Snapshot.

**Contraindications.** Experts questioned the prototype’s usefulness for casual individual use, niche applications, projects with little design freedom, or work under time pressure. E03 questioned the value for narrow applications with many hard constraints because possible choices might not be implementable. E08 described the problem of time pressure (translated): *“So yes, if I am really under a lot of time pressure or something, I think it is not so helpful to have to think through so much first.”* High expertise and routine cases were also associated with limited added value. At the other end, low expertise and limited technical knowledge raised the possibility that support would be needed.

Other reservations concerned particular aesthetic or functional priorities, emotional expression, and linguistic content. E09 expressed reservations about using the prototype for an individual simply seeking a pleasant companion-bot voice, while still allowing that it might be worthwhile. E07 anticipated little value when the client’s requirement was simply a voice that spoke correctly and further voice-design choices were not a priority. E03 described a project with an already designed voice whose current work focused on linguistic content, but remained open to trying the prototype. E06, referring to that interview’s framing of emotion as outside the prototype’s scope, saw its use limited to cases such as translation or simple dubbing, where emotional expression is less central.

## 6 Voice Design Compass

Based on the interview findings (Section 5) and subsequent literature work (Section 4.5; detailed mapping in Online Resource 5), the Voice Design Compass supports developing, reviewing and documenting voice concepts for embodied characters.

The default route has four work areas (Figure 3): Define the Project and Use Context; Develop the Voice Concept; Review the Voice Concept; and Try Out and Revise. Concept development uses four modular Design Sets. Review brings together Ethics, Technical Reflection and Shared Reflection. Throughout the process, designers create and maintain their own Project Record: a working document of intentions, requirements, reasons, sources, assumptions, open questions, planned tests and observed results.

Designers can choose which Sets to address and reorder questions within the Sets and Technical Reflection. Known constraints can inform earlier work. Designers can return directly to relevant parts of any earlier work area when needed. The dashed return arrow shows one example. Existing review notes are carried forward when the concept is revised.

Designers can use the examples as inspiration for their own answers and consult the related literature. Questions may be skipped, left open, or answered with several responses or a mixture described in the designer’s own words.

The Voice Design Compass is intended to support individual reflection and interdisciplinary

teamwork (Figure 4). Its questions help designers discuss how character and sound design, technical capabilities and users’ needs shape the same voice concept. Designers can document their decisions and reasons in their own Project Record.

### 6.1 Questions and Supporting Materials

The Compass keeps questions about appearance, represented identity and intended listener judgments distinct but connected. Each component’s questions occupy one Appendix page, with explanations, examples and sources linked by question number (Appendix A), following the modular-access rationale in Section 2.4.

### 6.2 Define the Project and Use Context

The first step is to describe the character or system, the people and situation involved, and the project goal. This separates what the project is trying to achieve from the character’s task and the impression its voice should create. This context is required at entry. Existing project notes or a provisional description can provide it. Its seven questions offer optional help, and the description can develop as project conditions or knowledge change.

The seven Context questions cover setting, background sound, intended listeners and people who may overhear, patterns of use over time, listening demands, age, and access needs (Table A1).

Requests to household smart speakers, for example, become part of co-present talk and domestic activity [64], and robot-voice ratings can change with restaurant ambience [87]. In interviews and a survey, deaf and hard-of-hearing smart-assistant users described access barriers involving voice pitch and speaking rate, background noise, and insufficient visual feedback [3].

In Context, designers distinguish people who interact with the character, people the voice is intended to address, and others who may hear it. A robot answering one visitor while others overhear is different from a robot making an announcement to everyone nearby. The age of those people is also kept separate from the character’s apparent age and the age impression intended for its voice. The time question records

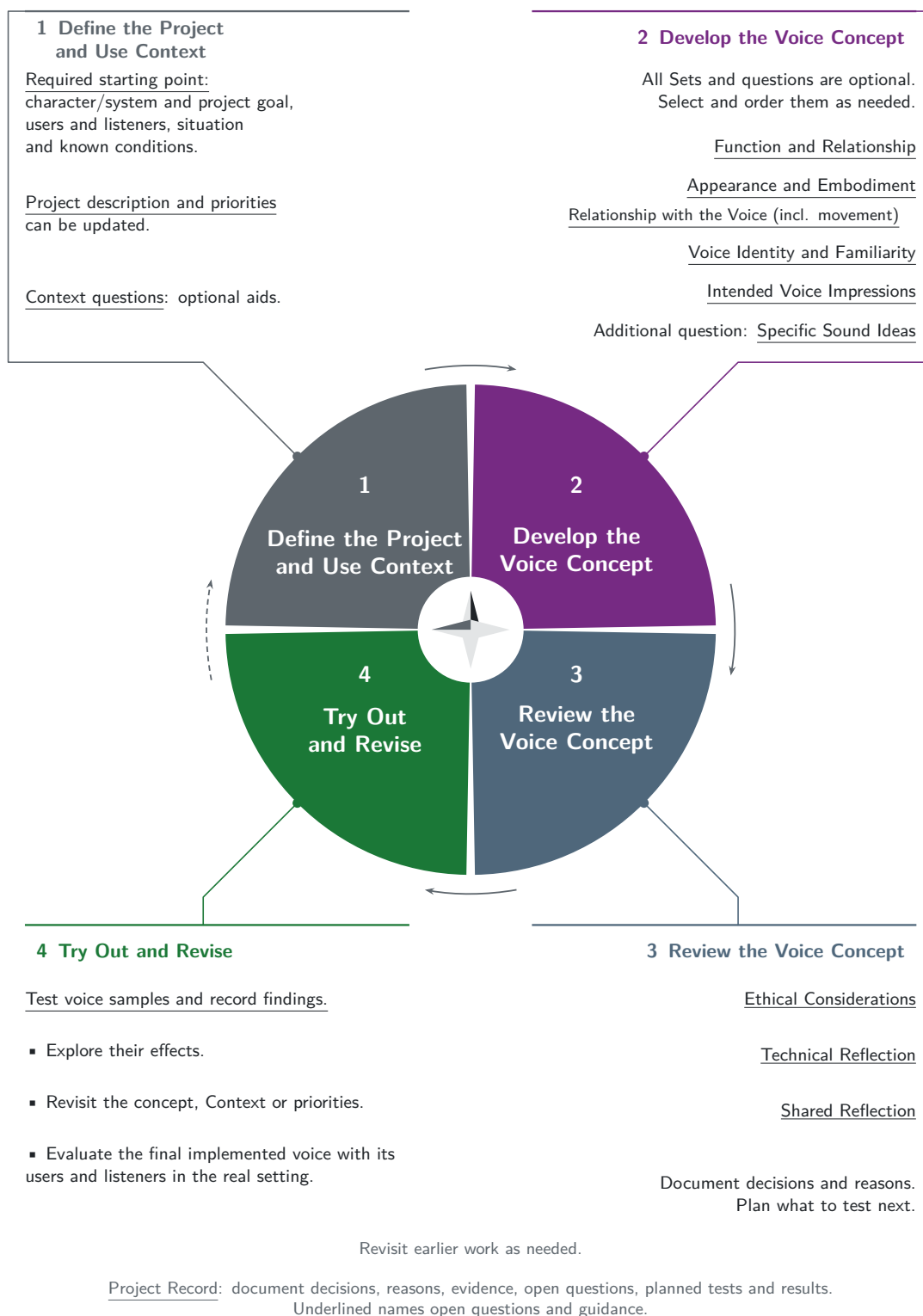


**Fig. 3** Voice Design Compass. Underlined names link to questions and guidance in the Appendix

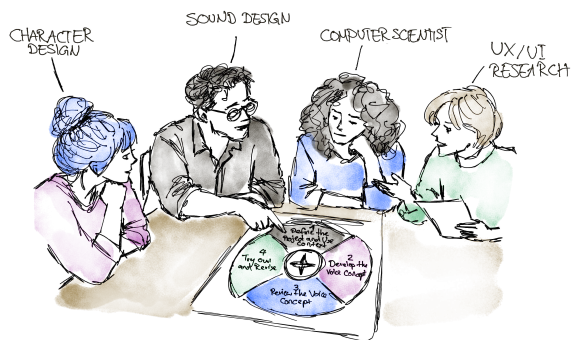


**Fig. 4** The Voice Design Compass as a shared basis for interdisciplinary discussion. Team members can use its questions to consider voice-related aspects together while developing the concept

the length of one interaction, how often it recurs, and the overall period of use separately. Listening demands describe the task, competing activities, time pressure and sustained attention.

Designers can also prioritize the questions before working through them. Prioritization concerns what needs attention now, what can wait, what is not needed for the project, and what remains unclear. Relevance, available information and dependencies between decisions can guide the order of work. An important unknown may need investigation now. Priorities can change as the project develops.

## 6.3 Develop the Voice Concept

Four Design Sets contain 18 questions about the character's function and relationship, appearance and embodiment, voice identity and familiarity, and intended listener impressions. Tables A3–A10 provide the questions, examples and related sources.

### 6.3.1 Function and Relationship

The two Function and Relationship questions distinguish what the character does from the relationship it should have with its users. The intended relationship also connects to the listener impressions considered in Intended Voice Impressions. Keeping function and intended impression distinct makes clear whether an entry describes what the character should do or how its voice should come across.

The relational examples draw on descriptions of hierarchy, exchange, and social bonds [69, 86]. The *relationship question* separates three description fields: *Authority*, who can decide what and in which situation; *Support*, the practical, social or emotional support planned; and *Formality*, how formal the relationship is. Each field can remain open independently.

### 6.3.2 Appearance and Embodiment

Appearance and Embodiment records aspects of the character's physical or virtual form that may shape expectations of its voice. Nine questions are organized under three internal headings. Body and Equipment covers embodiment configuration, shape, dimensions, and further physical features or equipment. Visible Expression covers visual style and human likeness, apparent age, and gender cues. Relationship with the Voice covers the intended voice–appearance relationship and coordination with movement. These headings are intended to help designers find questions about how the features might influence voice choice or its effects.

The questions distinguish physical–virtual configurations from combinations of human, animal-like, and mechanical form features. Multiple selections allow such combinations to be described. A further question covers physical features and equipment, including clothing, color, facial features, accessories, and disability-related representation.

Audiovisual studies motivate considering appearance and voice together, while their particular manipulations limit what can be attributed to a single feature [14, 19, 51].

The age examples use Child, Adult and Older adult as overlapping descriptions of appearance. Research on audiovisual age perception and perceived vocal age informs the consideration of age [17, 78].

A review documents inconsistent gender terminology in voice research, and an online survey of ten non-binary adults seeking to change their voices illustrates different desired voice impressions [11, 77]. For *appearance* and *voice*, we therefore distinguish *Design intention*, the planned choices and emphasis; *Intended perception*, the impression sought from viewers or listeners; and *Observed perception*, the impressions reported,

if examined. Giving gender little emphasis is a design intention. Seeking an ambiguous gender attribution is a perception target. Represented identity remains separate.

Two questions within *Relationship with the Voice* connect visible features and vocal intentions: one addresses movement, posture and other visible signals, and the other asks how the voice should fit with or deliberately contrast with appearance. They can be considered alongside intended impressions and identity intentions. Cambre and Kulkarni review research on voice fit and identify “deliberate misalignment in expectations” as a direction for future exploration [7]. Their discussion extends to users’ expectations about device functions and social roles. Our question focuses on voice and visible appearance, with deliberate contrast remaining an intention to justify and test.

### 6.3.3 Voice Identity and Familiarity

Voice Identity and Familiarity asks about the voice reference, intended identity relation, anticipated familiarity or similarity for listeners, and who can select or modify the stable voice identity. A reference may be familiar to one listener group and unknown to another. Ethical Considerations addresses consent and representation risks.

Studies of self- and familiar-voice cloning motivate recording the voice reference and the listener’s relationship to it [36, 68]. The perceptual account in Section 2.2 connects reference choice with anticipated listener impressions. Identity goals may include concealing social cues in unfamiliar interactions [6], signaling that the character is artificial, or distinguishing its voice from a named reference. Designers use Technical Reflection to examine suitable production methods and comparisons.

Familiarity means recognizing a known person’s voice. Similarity concerns features shared with listeners, such as accent or personality impression, without requiring recognition. Their reported effects are specific to the listening conditions and social relations examined [29, 54, 68]. The question combines Cambre and Kulkarni’s similarity perspective with familiar-voice considerations from the studies above, keeping the two answers separate. A familiar person’s voice need

not resemble the listener, and a newly created voice can share the listener’s speaking style.

Designers can consider whether either quality supports the character’s role or introduces unwanted expectations, or record that neither is an aim.

### 6.3.4 Intended Voice Impressions

Intended Voice Impressions records intended listener judgments and connects them with the role, represented identity, and listening demands documented elsewhere. Its three questions address vocal humanlikeness, social qualities such as warmth or competence, and perceived age and gender cues. Designers can identify several intended perceptions for prioritization, acoustic realization, and testing. Visible age or gender cues, an identity represented by the character, and an impression conveyed through its voice remain distinct considerations, even when considered together.

Seaborn and Urakami’s overview of quantitative voice UX measures distinguishes trust, likeability, personality, anthropomorphism, and perceived intelligence [74]. In the current Voice Design Compass, this overview supports the vocabulary for describing listener judgments, including the Likability example.

Evidence for these perceptual targets is context-specific [47, 71]. When comparing voice alternatives, designers consider the selected impressions together and record relevant changes beyond the primary target.

Optional OCEAN guidance makes psychological impressions more explicit. It covers openness, conscientiousness, extraversion, agreeableness, and neuroticism. Work on perceived vocal personality and synthetic voices supports examining such impressions, while also showing limits to deliberate control [24, 79]. Perceived Intelligence is considered alongside competence [50]. The guidance distinguishes seeming able to think or understand from seeming capable in a role, while allowing overlap. These remain listener attributions to examine, not claims about the character’s actual abilities.

The example headings separate judgments about the character, its information and the voice itself. Desired influence on decisions is recorded in

Ethics question 14 and revisited in Shared Reflection. The persuasiveness example is included in Intended Voice Impressions as a linked consideration.

For gender, the [question about vocal gender and age](#) uses the same distinction between design intentions, intended perceptions and observed perceptions as the [question about visible gender expression](#). Age remains a separate part of the question.

### 6.3.5 Specific Sound Ideas

The optional question *Specific Sound Ideas* gives concrete sound intentions a place of their own. It follows the general design considerations and asks, “Do you have a particular sound in mind for the voice?” Designers describe their idea in their own words, explain why they want this sound for the character and situation, and may refer to an audio example. Table A11 provides example sound ideas with related sources. The question can remain open, and designers may choose no additional stylization.

Sound intentions are distinguished from expected associations: “a crackling recording” describes a sound, while “a historical association” describes a response to examine with listeners. Reflection and practical trials address implementation feasibility and listener responses. When using reference voices, designers also consider intended changes and permission to use them.

## 6.4 Review the Voice Concept

The review brings ethical concerns and technical feasibility together to decide which choices to keep, revise or leave open, document the reasons, and plan what to test next.

### 6.4.1 Ethical Considerations

The review identifies possible harms, affected people, and follow-up needs. This includes people who hear the voice incidentally and people whose identities or work are represented, even if they are not present. Designers can record a concern and follow-up action, a reason for identifying no concern, or what remains unclear.

The 28 Ethics questions and their sources are grouped under six topic headings: Consent and Identity; Labor and Professional Practice;

Bias and Stereotypes; Influence and Manipulation; Information and Expectations; and Diversity and Inclusion (Table A12). The questions invite reflection, and designers choose how to respond. Legal requirements apply in addition, for example transparency obligations for AI systems that interact with people or generate synthetic audio under the EU Artificial Intelligence Act [16] and data-protection rules for voice recordings.

The labor questions address voice performance, the reuse of recordings, and work involved in developing synthesis models, including permission, control, attribution, and compensation. The bias questions address harmful stereotypes involving professions and gender identities. The Diversity questions also address exclusion, representation, access and participation, including which identities are treated as defaults (for example, treating one gender or accent as the unmarked normal voice) and whose choice, privacy and control are at stake. These perspectives overlap. Designers can refer to a shared explanation across related questions, adding any concerns or follow-up needs specific to each question. Information and Expectations begins with information about voice production and control of the system’s voice and actions. It then addresses possible mistaken impressions and relational concerns, alternatives, and plans for testing assumed social impressions.

Seaborn et al.’s survey reviews relationships between voice, gender, social roles, and listener judgments [75]. Our question about masculine associations with authority or expertise applies this perspective to the planned combination of voice and character role (Ethics question 12).

The reference to professions and gender identities in the bias questions is relevant in light of studies of British children’s occupational judgments after pitch manipulation and of robot roles in service settings [8, 14, 30]. These findings motivate examining voice–role associations in the intended project setting. Work on trans and non-binary users’ requirements and on voice labels informs the distinction between acoustic expression, attributed identity, and self-identification [12, 65].

Seaborn recommends pilot testing and manipulation checks to examine how people perceive an agent’s social identity, including possible differences across cultures [73]. Ethics questions 22 and 28 adapt this guidance to the planned voice:

what listeners associate with it, and how people from relevant cultural backgrounds interpret its social cues. These questions ask designers to plan how to examine their assumptions with the intended users and listeners.

### 6.4.2 Technical Reflection

Technical Reflection examines whether the intended design can be supported by the project’s technical, organizational, and material conditions. Four headings organize its eleven questions: Requirements, Resources, Operating Conditions, and Production Approach.

Before production or a trial, designers examine the feasibility and unresolved conditions relevant to that step. Operating details can remain open at entry. The extent of Technical Reflection that should be mandatory in every project remains open. When a later choice changes the voice, data use, or production method, designers return to the affected design and Ethics entries. Changes affecting many topics may require a broader review.

The recording question distinguishes material that exists, material that might be collected, permission to use it, and suitability for the intended method. Operating-condition guidance brings together service availability, dependence on a supplier, service-level commitments, capacity as usage grows, and AI governance. Related frameworks such as the NIST AI Risk Management Framework provide further guidance for organizing risk work [84].

Production methods entail different recording, licensing, and later revision costs [61, pp. 148–150], as well as recording and tool-integration work [9]. The resource questions address the project’s budget, available time and expertise. Their assessment depends on the project’s conditions. Prepared audio can include human recordings or audio synthesized and processed in advance. Preparation time and production method are separate considerations. Further technical trade-offs are discussed by Stan and Lőrincz [81].

Robinson et al. distinguish authored utterances, modular sound assets, and procedural generation in robot sound design [67]. We use this distinction to relate required control to possible voice-production methods.

The technical-control question distinguishes the desired sound from the control needed to select, adjust or compare voices.

### 6.4.3 Shared Reflection

After Ethical Considerations and Technical Reflection, five Shared Reflection questions revisit the concept’s goals, possible further effects, evidence and next actions (Table A15). Two explicitly address designers’ own stereotypes or prejudices and possible conflicting expectations or effects. The final question asks what to keep, change or leave open, why, and what to test next.

The questions accommodate freely described choices. Designers identify relevant conflicts and stereotypes themselves and justify what to keep or change, including deliberate contrasts.

Examples illustrate possible answers to the reflection questions and link to related research (Table A16).

## 6.5 Try Out and Revise

Testing makes the transition from written intentions to practical evidence explicit. After reflection, designers identify an assumption or requirement to examine, choose suitable voice material or a prototype, and state how they will judge the result. For an intended association or impression, they also identify other impressions that matter for the project and could change with the voice choice. The test record specifies which voice version, people, and situation were involved. Reported impressions and observed behavior are distinguished. Results can support keeping a decision, changing it, or leaving it open pending better evidence.

Focused early trials are possible once suitable voice material and the relevant ethical and technical conditions are available, while other Voice Design Compass questions remain open. Unexpected effects, unmet requirements, new risks or changed use conditions can lead to a return to the affected parts of the voice concept, the Context description or priorities. Further trials can then assess voice material against the revised intentions and requirements. Designers evaluate the final implemented voice with its actual users and listeners in the real setting.

## 6.6 Working with the Voice Design Compass

For individual review or group discussion, we recommend starting with existing or provisional project notes and using the overview in Figure 3 to select questions relevant to open decisions. The questions and examples in Appendix A support discussing and recording those decisions. A Project Record entry can link a proposed voice choice to its reason, sources, open questions and a next action, such as comparing candidate voices. Table A18 provides a template with an excerpt from the Rowan example (Section 6.6.1). The standalone *Voice Design Compass Working Pack* (Online Resource 4) brings the overview, questions, examples and full references together.

Two fictional examples illustrate how the Voice Design Compass could be used to develop and document a voice concept: a museum planning Rowan, an exhibition narrator, and a research team preparing a humanoid robot for a laboratory study of pointing gestures. Both examples follow selected questions from the project context through reflection to plans for testing and revision.

### 6.6.1 Rowan: An Exhibition Narrator

Suppose a museum plans *Rowan*, a fictional narrator with a stationary wooden radio body and a face (Figure 5). Visitors would use buttons to start, pause and replay short stories recorded by a human performer. The team wants an old-record effect to suit the historical exhibition.

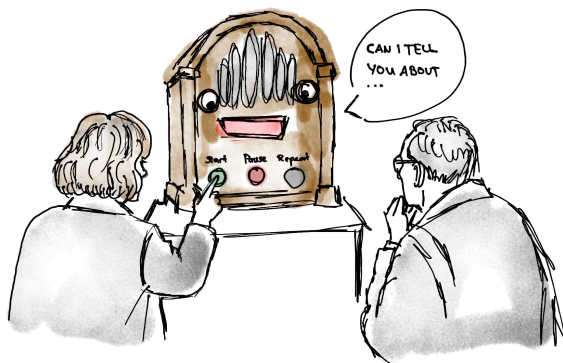


**Fig. 5** Sketch of the fictional Rowan scenario: the museum's radio character addresses exhibition visitors, who control its recorded speech with physical buttons

In **Project and Use Context**, the team records its goal: helping adult exhibition visitors, including older adults, understand and enjoy the stories. Companions and children may also overhear. Understanding the stories and using the controls take priority over the effect. Moving and talking visitors could create background noise, so the team plans to inspect the **listening conditions**.

In **Develop the Voice Concept**, the team defines Rowan as a professional narrator intended to convey social warmth and competence. They seek a new voice without imitating a historical or familiar person. To connect the radio body with old recordings, they propose subtle crackle and noise under *Specific Sound Ideas*. This association remains an untested aesthetic assumption. They leave **vocal gender and age** open until audition samples can be compared.

In **Ethical Considerations**, the team notes that no performer has been selected and no consent obtained. Before recording, they must agree on payment, recording, storage, public playback, effect processing and reuse, and document the performer's contribution and use rights. They mark the **question about AI's effects on paid work** as not applicable because no synthesis, cloning or AI processing is planned, and will revisit it if that changes.

In **Technical Reflection**, fixed stories suggest commissioned recordings with separately applied effects and local playback; budget and skills still need checking. In **Shared Reflection**, the team limits the effect to stories and keeps operating instructions clean. Visitors would also be able to choose complete stories without effects. Effect strength remains open.

For **Try Out and Revise**, they plan to compare the same recording with and without effects in the exhibition, once recordings and permissions are available. Intended visitors would recount the content and describe what the sound reminds them of. If only the version with effects is difficult to understand, the team would revise the effect; if both versions are difficult, they would revisit acoustic and playback conditions.

Their **Project Record** retains these choices, reasons, open questions and next actions, including the room inspection, performer agreements and audition samples. Table A18 shows the entry for the old-record effect.

### 6.6.2 Preparing a Robot Study

Suppose a research team plans a laboratory study with a small humanoid robot. They want to examine whether pointing helps adults select a named object more quickly and accurately. Both conditions use the same spoken instructions; one adds pointing (Figure 6). The team uses the Voice Design Compass to choose and justify the voice.

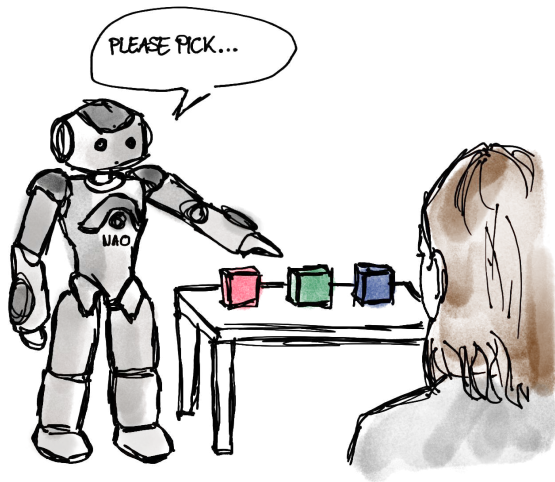


**Fig. 6** Sketch of the gesture condition in the fictional robot study: the robot gives a spoken instruction and points to an object for the participant to select

The **Project and Use Context** description specifies adult participants, repeated object-selection trials and a quiet laboratory. In **Develop the Voice Concept**, the team defines the robot as a matter-of-fact task instructor and prioritizes intelligibility and consistent delivery. They specify identical speech files and playback timing in both conditions, and when pointing begins in the gesture condition.

The **age and gender question** prompts two considerations: whether the small robot should sound childlike and whether gender ambiguity could avoid stereotypes. The team considers age and gender separately and consults the related Appendix sources. Fischer and Niebuhr found that acoustic cues affected which of two differently sized robot images listeners associated with a synthetic voice [20]. In Torre et al.'s robot-video study, a voice selected for gender ambiguity reduced feminine ratings but increased masculine

ratings compared with silent videos [85]. Gender ambiguity therefore does not guarantee freedom from gendered associations.

In **Ethical Considerations**, the team examines stereotypes about authority and considers alternatives that might challenge them. Unintended associations remain a question for listeners. **Technical Reflection** covers the required language, voice-use permissions, resources, and control over speech rate, pauses and playback timing. They plan to save synthesized instructions for repeated playback.

In **Shared Reflection**, the team provisionally chooses an available synthetic voice they hear as adult and feminine. It offers the required language and control over speech rate and pauses. They prioritize these requirements over gender ambiguity or a childlike impression and record clear articulation and restrained delivery as intentions to check.

For **Try Out and Revise**, a pilot with intended participants would check comprehension, voice impressions and speech timing with and without pointing, guiding revisions before the main study. Identical audio keeps the speech constant; its possible influence on the task remains to be examined.

The **Project Record** would retain the reasons for the choice, unresolved assumptions, audio files, production settings and timing relative to gestures. These entries support reporting the voice in the study's methods and considering its possible influence when interpreting results.

## 7 Discussion

Experts valued the prototype's guidance and reflection but also reported unclear consequences of answers, cognitive load and differing preferences about structure. We discuss what these findings imply for reflecting on voice choices, balancing guidance with flexibility and effort, and matching the Voice Design Compass to project needs, before reporting later feedback on the revised Compass and the scope and limitations of the study.

## 7.1 Reflecting on Voice Choices

A voice choice intended to convey one quality can also affect other impressions. In evaluations of isolated speech and humming, processing that adjusted pitch and formant frequencies reduced Gender Specificity and voice Anthropomorphism scores and increased Novelty scores, which combine ratings of perceived unfamiliarity and unusualness [35]. In an android interaction, the android received higher likability ratings with a clone of the user’s voice or a voice chosen to suit its appearance than with a dissimilar voice; the first two voice conditions did not differ significantly [36]. Familiar-voice cloning brought character ratings closer to ratings based on memories of the source person for selected outcomes, with different outcomes affected in the pilot and follow-up studies [37]. Torre et al.’s robot-video study (Section 6.6.2) likewise showed that a voice selected for gender ambiguity reduced feminine but increased masculine ratings [85]. These findings motivate checking the chosen voice across relevant impressions in its intended character presentation.

The reactions to Design Tensions suggest that reflection support should draw attention to possible consequences while leaving room for deliberate design contrasts (Section 5.3).

Ethical reflection also needs to distinguish recognizable role expectations from discriminatory generalizations. In discussing adaptive social roles for robots, Huber et al. present stereotypical roles as one possible way to align users’ expectations with a robot’s capabilities, while arguing for roles that can adapt to users’ changing needs [31]. In their account of gender-sensitive care robots, Weiss and Plöckig argue that categorization cannot simply be prevented and that researchers should avoid reinforcing stereotypes that lead to unequal treatment [88]. For the Voice Design Compass, we interpret these positions as a reason to examine what a voice–role association helps communicate and what social assumptions it may reproduce. A recognizable role does not by itself justify assigning a particular gendered voice to it. The Ethics questions and Shared Reflection provide places to document these assumptions, consider alternatives, and identify effects to examine with intended listeners.

## 7.2 Guidance, Flexibility, and Effort

Useful reflection depends on designers understanding the questions and being able to relate them to their project. Fritsch et al. report a related difficulty: participants used their own words when the suggested terms did not fit their impressions of the voice samples [23]. In the Voice Design Compass, literature-based examples accompany open questions so that designers can describe intentions beyond the examples provided. Experiments in engineering design found that participants reproduced features of supplied examples [32]. This suggests a risk that our examples could also steer the options designers consider.

Experts valued modularity and adjustable Set order (Section 5.2), while also requesting prioritization and a partly fixed route (Section 5.3). We interpret this combination as a distinction between being able to choose a path and knowing which questions matter for a project. Gray et al.’s study of selecting ethics-focused methods likewise highlights the need for guidance alongside flexibility [26]. Allowing questions to be reordered or skipped leaves the relevance judgment to the designer. Additional guidance may help with that judgment, but could also add reading and decisions to an already demanding process. In the Voice Design Compass, the four work areas, the shared Context description and optional prioritization notes are an organizational response to this tension.

Experts reviewed the prototype in a web application. The revised Compass is provided as documents in the Appendix and Working Pack. Experts saw potential uses for the study application’s Literature Snapshot and documentation export. In the Voice Design Compass, the questions are accompanied by explanations of possible answers and links to related literature (Appendix A). Designers choose the evidence relevant to their project and maintain the Project Record themselves. Literature assistance and automated output remain possibilities for a future software implementation.

## 7.3 Support for Different Project Needs

Support needs depend on who uses the Voice Design Compass and for what purpose. Our study

connects with Fritsch et al.’s interest in engaging practitioners beyond sound studies [23]. Our interviews with experts from research and industry (Table 3) examined how they interpreted and assessed the prototype’s prepared questions, examples, and workflow. This adds formative evidence about the guidance experts valued, the difficulties they encountered, and the revisions they requested.

The suitability findings suggest that the potential value of using the Voice Design Compass depends on project needs as well as professional role (Section 5.4). Experts saw potential for non-expert use but noted that applying the available knowledge may require support. Conversely, high expertise and routine cases may reduce the additional value of working through the questions. In our interpretation, unresolved voice intentions, requirements or reasons for decisions are therefore relevant when considering whether to use the Voice Design Compass.

Resources and design freedom further shape this judgment. Bringing voice-design knowledge together may make the approach attractive when budgets are limited, while reflection still demands time. Little freedom to change the voice can also limit which considerations lead to action. The early, iterative and retrospective uses that participants suggested support treating voice concept development as an activity that can be revisited when questions or project conditions change.

## 7.4 Feedback on the Revised Compass

We sent the revised Compass as a Working Pack, together with an initial draft of this article and accompanying materials, to the ten interview participants and additional researchers outside the author team, including authors whose work and concepts informed the Compass. Seven people provided substantive feedback: four interview participants and three additional researchers. Two other interview participants reported being unable to review the materials within the requested timeframe. The following account draws on the five responses for which permission covered the reported content. Respondents described possible future uses, including seeking guidance in a suitable voice-design project; one had no current project of this kind. Another saw potential value

for robot developers. One wrote, *“I believe it will be super useful to use when I need some guidance in my voice-related projects.”* Suggestions concerned the overview’s readability, clearer guidance for the examples-and-sources tables, multipart questions, and assumptions about bodies and social relationships. One respondent cautioned that *“the presence of certain questions and absence of others will also guide people’s thinking.”* In response, we clarified listener characteristics and body size in the materials, made explicit that voice includes vocal sounds other than speech, and discussed how questions and examples may steer design choices. These responses describe impressions and anticipated uses, with no reported application of the revised Compass to a concrete project.

## 7.5 Scope and Limitations

Our experience developing the Voice Design Compass suggests that general voice characteristics can overlap with linguistic content and emotional portrayal. In Table A11, emphasis on key words and relaxed or tense voice qualities connect sound choices with linguistic content and emotional portrayal.

Iterative literature work connected emerging design questions with relevant research and informed both the questions and their examples [4]. The resulting collection provides a documented rationale for the considerations included in the Voice Design Compass. It remains selective and may favor familiar perspectives or omit relevant alternatives and conflicting findings. Expert review identified omissions and unclear distinctions, but does not establish the completeness or balance of the literature base. The selection and wording of the questions can also direct designers’ attention toward some considerations and away from others.

The expert review informed the documented revision from prototype to Voice Design Compass (Online Resource 5). The revised Voice Design Compass, including the new questions and document format, has received informal feedback but has not yet been evaluated in project use. Its effects on the clarity of voice concepts, reflection effort, practical support needs and resulting voice designs remain empirical questions. We evaluated the prototype through formative expert

walkthroughs and illustrate the revised Compass through fictional worked examples, corresponding to usage and demonstration in Ledo et al.’s account of toolkit evaluation [41].

Participants gave feedback to the prototype’s developer, and some already knew the researchers, which may have encouraged favorable responses. Component coverage varied, with most Sets receiving two complete walkthroughs and Ethics one complete walkthrough plus partial reviews of selected questions. All participants worked in Europe. Transfer to other cultural settings therefore requires further examination, as does transfer to other professional groups. Written examples, the study application and interviewer probing may also have shaped their accounts. The individual interviews used written material without voice samples. Collective decision-making and the relation between written intentions and experienced sound therefore require study in use. Fritsch et al.’s workshops illustrate how listening to and manipulating voices can form part of that exploration [23].

## 8 Future Work

We invite researchers and practitioners to use the Voice Design Compass in their projects to develop, review and document voice concepts, and to relate their experiences and findings from voice-design research to its questions and examples. We particularly welcome contributions from researchers in linguistics, dialogue design and emotional speech that connect their findings and methods with the questions and literature-based examples in Appendix A. Sharing the supporting evidence and project conditions can help refine the guidance and extend it to further characters, audiences and settings.

### 8.1 Learning from Voice Design Compass Use

Comparisons with designers’ usual practice could examine the Voice Design Compass’s added value using comparable project briefs, time budgets, expertise and production tools. Studies should follow concept development and revision to assess how designers define and justify requirements, identify overlooked influences and possible unintended effects, and explain their decisions.

Project-relevant criteria should address clarity, documentation usefulness, design effort and support needs, including why designers answer, defer or omit questions. Studies should also examine whether the examples help designers express intentions while allowing alternatives. Whether resulting voices achieve intended effects with actual users and listeners should be tested separately in the project setting.

Workshops involving domain experts, prospective Voice Design Compass users with different levels of sound expertise, or intended users and listeners of the character could examine how participants discuss the Voice Design Compass’s questions, negotiate priorities, and revise a voice concept together. Individual sessions with voice samples could examine how designers connect written intentions to audible alternatives. These activities could combine the situated listening and sonic experimentation explored by Fritsch et al. with the Voice Design Compass’s process and the Project Records participants create [23]. They could trace how encounters with concrete voices generate new aims and inform requirements and revisions. Longer studies should distinguish revisions to the voice concept from changes in listeners’ experiences and relationships with the character through repeated use.

Further studies could identify which considerations are needed to meet a project’s voice-design goals and address its ethical and technical requirements. Comparisons of detailed and simplified Compass versions could examine whether shorter guidance still helps designers identify relevant requirements and risks and explain their choices. Comparisons with the prototype could assess differences in cognitive load.

Future software development could extend documentation export and literature assistance and develop a justified mapping from recorded intentions and requirements to voice-generation prompts. Evaluation should examine documentation effort, preservation and reuse of decisions and uncertainties, the relevance and traceability of suggested literature, and whether the generated voices express the intended qualities.

## 8.2 Strengthening the Evidence for Voice Design

The Voice Design Compass’s questions and literature-based examples provide a structure for connecting further research on voice–character relationships with existing knowledge. Focused systematic reviews could synthesize evidence for individual topics. The following directions illustrate how new studies and reviews could inform the questions and examples, clarify where findings apply, and extend the topics covered.

**Appearance and voice.** Research should clarify whether a visible feature changes the voice people expect, their interpretation of the same voice, or their response to alternative voices. Building on studies that combine color or clothing with other appearance changes [14, 43], focused comparisons could vary one feature while keeping others stable, compare controlled voice alternatives, and keep expected voice, perceived fit, preference, intelligibility and intended social impression as separate outcomes, including deliberate contrast. Authoring toolkits such as Anthropomorphic AI already combine character models, configurable synthetic voices, and nonverbal behavior [42]. Future work could examine whether incorporating the Compass’s questions into such tools helps designers consider voice and appearance decisions together.

**Identity cues and intended impressions.** Work on audiovisual age perception, gendered voices and perceived vocal personality already provides relevant distinctions [11, 17, 24, 77, 79]. Further studies could examine how voice and appearance cues interact across audiences and situations, compare intended profiles with listener judgments across several voices, texts and tasks, and keep perceived age, gender or intelligence separate from identity representation, personal preference and actual system capability [50]. Such results could improve examples for comparing candidate voices and recognizing trade-offs.

**Cultural meaning, specific sounds, and changing situations.** Cultural background and speech milieu have relevant conceptual and empirical literature [56, 83] that future reviews can connect more closely to the questions about how speech is interpreted in context. Studies could compare voice alternatives with intended listeners

and represented groups, separating comprehension, familiarity, belonging, preference and stereotypical attribution. An extension of the Voice Design Compass could address how a character’s voice should vary in particular situations, including moment-to-moment changes in emotional expression, for example a shift from entertainment to an urgent message. For concrete sound effects, comparisons with and without an effect could examine whether it supports the intended impression while preserving intelligibility in the planned setting.

**Application-specific guidance.** In the interviews, experts requested guidance on voice design in mental-health contexts and on end-user safety. These requests suggest examining how well the Voice Design Compass supports projects with specific requirements. Studies involving domain specialists and intended users could identify which questions need further explanation or examples in mental-health contexts or projects involving children or older people. Evaluation should distinguish support for this reflection from the effects of the resulting voice, including any claimed benefit or safety.

Contributions are most informative when they specify the voice alternatives, characters, listeners and settings studied, the observed effects and the conditions that may limit transfer.

## 9 Conclusion

We developed the Voice Design Compass through iterative literature work and formative think-aloud interviews in which ten domain experts reviewed the prototype. Experts valued guidance, reflection and documentation, while identifying missing content, unclear distinctions and substantial cognitive load. These findings informed the Voice Design Compass’s revised questions, required Context and the option to return directly to earlier work areas. The Voice Design Compass retains modular Design Sets and connects them with ethical and technical review and practical trials.

The Voice Design Compass provides structured questions and a documented rationale for developing, reviewing and documenting voice concepts for embodied characters. Experts’ anticipated uses and reservations help position this support around project needs, expertise, time and

design freedom. Further research and application can refine evidence about how appearance, identity cues and use situations affect voice expectations and responses.

**Supplementary information.** Online Resources 1–5<sup>2</sup> accompany this article: the prototype application overview, the interview guide with its topic-coverage matrix, the relevance-rating plots, the Voice Design Compass Working Pack, and *From the Prototype to the Voice Design Compass: Revision Rationale and Change Mapping*.

## References

- [1] Ada AA, Jørgensen SH, Fritsch J (2024) Cultures of the AI paralinguistic in voice cloning tools. In: Companion Publication of the 2024 ACM Designing Interactive Systems Conference. Association for Computing Machinery, pp 249–252, <https://doi.org/10.1145/3656156.3663708>
- [2] Bengtsson M (2016) How to plan and perform a qualitative study using content analysis. *NursingPlus Open* 2:8–14. <https://doi.org/10.1016/j.npls.2016.01.001>
- [3] Blair J, Abdullah S (2020) “it didn’t sound good with my cochlear implants”: Understanding the challenges of using smart assistants for deaf and hard of hearing users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4(4):118:1–118:27. <https://doi.org/10.1145/3432194>
- [4] Boell SK, Cecez-Kecmanovic D (2014) A hermeneutic approach for conducting literature reviews and literature searches. *Communications of the Association for Information Systems* 34:257–286. <https://doi.org/10.17705/1CAIS.03412>
- [5] Boren MT, Ramey J (2000) Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication* 43(3):261–278. <https://doi.org/10.1109/47.867942>
- [6] Byeon HJ, Lee C, Lee J, et al (2022) “a voice that suits the situation”: Understanding the needs and challenges for supporting end-user voice customization. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, <https://doi.org/10.1145/3491102.3501856>
- [7] Cambre J, Kulkarni C (2019) One voice fits all? social implications and research challenges of designing voices for smart devices. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW):1–19. <https://doi.org/10.1145/3359325>
- [8] Cartei V, Oakhill J, Garnham A, et al (2021) Voice Cues Influence Children’s Assessment of Adults’ Occupational Competence. *Journal of Nonverbal Behavior* 45:281–296. <https://doi.org/10.1007/s10919-020-00354-y>
- [9] Chan SWT, Gunasekaran TS, Pai YS, et al (2021) KinVoices: Using voices of friends and family in voice interfaces. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW2):1–25. <https://doi.org/10.1145/3479590>
- [10] Chiou EK, Schroeder NL, Craig SD (2020) How we trust, perceive, and learn from virtual humans: The influence of voice quality. *Computers & Education* 146:103756. <https://doi.org/10.1016/j.compedu.2019.103756>
- [11] De Cet M, Obaid M, Torre I (2025) Breaking the binary: A systematic review of gender-ambiguous voices in human-computer interaction. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, pp 1–17, <https://doi.org/10.1145/3706598.3713608>
- [12] De Cet M, Hope M, Torre I (2026) Queer, Nonbinary, or Ambiguous? Rethinking Voice Labels through Queer Theory in HCI. In: *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, pp 1–8, <https://doi.org/10.1145/3772318.3791246>

---

<sup>2</sup>[https://osf.io/q5x4h/overview?view\\_only=3c3faa70194b48a5b291e642dfc2eebf](https://osf.io/q5x4h/overview?view_only=3c3faa70194b48a5b291e642dfc2eebf)

- [13] DeJonckheere M, Vaughn LM (2019) Semistructured interviewing in primary care research: a balance of relationship and rigour. *Family Medicine and Community Health* 7(2):e000057. <https://doi.org/10.1136/fmch-2018-000057>
- [14] Dennler N, Kian M, Nikolaidis S, et al (2025) Designing robot identity: The role of voice, clothing, and task on robot gender perception. *International Journal of Social Robotics* 17(4):707–728. <https://doi.org/10.1007/s12369-025-01209-6>
- [15] Dorst K, Cross N (2001) Creativity in the design process: Co-evolution of problem-solution. *Design Studies* 22(5):425–437. [http://doi.org/10.1016/S0142-694X\(01\)00009-6](http://doi.org/10.1016/S0142-694X(01)00009-6)
- [16] European Parliament and Council of the European Union (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L, 2024/1689, 12 July 2024, URL <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- [17] Fagel S (2007) Auditory-visual integration in the perception of age in speech. In: *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, pp 725–728, URL <https://www.icphs2007.de/conference/Papers/1001/1001.pdf>
- [18] Fenzl T, Mayring P (2017) QCMap: eine interaktive Webapplikation für Qualitative Inhaltsanalyse. *Zeitschrift für Soziologie der Erziehung und Sozialisation* 37(3):333–340
- [19] Ferstl Y, Thomas S, Guiard C, et al (2021) Human or robot?: Investigating voice, appearance and gesture motion realism of conversational social agents. In: *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. ACM, New York, NY, USA, pp 76–83, <https://doi.org/10.1145/3472306.3478338>
- [20] Fischer K, Niebuhr O (2023) Which voice for which robot? designing robot voices that indicate robot size. *ACM Transactions on Human-Robot Interaction* 12(4):1–24. <https://doi.org/10.1145/3632124>
- [21] Fleck R, Fitzpatrick G (2010) Reflecting on reflection: Framing a design landscape. In: *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction*. Association for Computing Machinery, OZCHI '10, pp 216–223, <http://doi.org/10.1145/1952222.1952269>
- [22] Friedman B, Hendry DG (2012) The Envisioning Cards: A toolkit for catalyzing humanistic and technical imaginations. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, CHI '12, pp 1145–1148, <https://doi.org/10.1145/2207676.2208562>
- [23] Fritsch J, Karaturhan P, Jørgensen SMH, et al (2025) Towards a framework for exploring synthetic voices in VUI design. In: Nunes NJ, Nisi V, Oakley I, et al (eds) *Proceedings of the 2025 ACM Designing Interactive Systems Conference*. Association for Computing Machinery, ACM Digital Library, pp 724–739, <https://doi.org/10.1145/3715336.3735729>
- [24] Gao S, Aylett MP, Braude DA, et al (2023) Synthesising personality with neural speech synthesis. In: *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czechia, pp 393–399, <https://doi.org/10.18653/v1/2023.sigdial-1.36>
- [25] Gatt A, Krahmer E (2018) Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research* 61:65–170. <https://doi.org/10.1613/jair.5477>
- [26] Gray CM, Chivukula SS, Carlock T, et al (2023) Scaffolding ethics-focused methods for practice resonance. In: *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. Association for Computing Machinery, DIS '23, pp 2375–2391, <https://doi.org/10.1145/3588112.3588112>

[//doi.org/10.1145/3563657.3596111](https://doi.org/10.1145/3563657.3596111)

- [27] Harder J (2023) *Creative Character Design for Games and Animation*. CRC Press, Taylor & Francis Group, Boca Raton, FL, <https://doi.org/10.1201/9781003243533>
- [28] Hennink M, Kaiser BN (2022) Sample sizes for saturation in qualitative research: A systematic review of empirical tests. *Social Science & Medicine* 292:114523. <https://doi.org/10.1016/j.socscimed.2021.114523>
- [29] Holmes E, Johnsrude IS (2020) Speech spoken by familiar people is more resistant to interference by linguistically similar speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 46(8):1465–1476. <https://doi.org/10.1037/xlm0000823>
- [30] Hu Q, Pan X, Luo J, et al (2022) The effect of service robot occupational gender stereotypes on customers' willingness to use them. *Frontiers in Psychology* 13:985501. <https://doi.org/10.3389/fpsyg.2022.985501>
- [31] Huber A, Lammer L, Weiss A, et al (2014) Designing adaptive roles for socially assistive robots: A new method to reduce technological determinism and role stereotypes. *Journal of Human-Robot Interaction* 3(2):100–115. <https://doi.org/10.5898/JHRI.3.2.Huber>
- [32] Jansson DG, Smith SM (1991) Design fixation. *Design Studies* 12(1):3–11. [https://doi.org/10.1016/0142-694X\(91\)90003-F](https://doi.org/10.1016/0142-694X(91)90003-F)
- [33] Kepper MM, L'Hotta AJ, Shato T, et al (2024) Supporting teams with designing for dissemination and sustainability: the design, development, and usability of a digital interactive platform. *Implementation Science* 19(1):82. <https://doi.org/10.1186/s13012-024-01410-7>
- [34] Kim Y, Reza M, McGrenere J, et al (2021) Designers characterize naturalness in voice user interfaces: Their goals, practices, and challenges. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp 1–13, <https://doi.org/10.1145/3411764.3445579>
- [35] Kuch JM, Nasir J, Mertes S, et al (2024) Evaluating gender ambiguity, novelty and anthropomorphism in humming and talking voices for robots. In: *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp 2219–2225, <https://doi.org/10.1109/RO-MAN60168.2024.10731423>
- [36] Kuch JM, Heisler M, Klein S, et al (2025) Your robot, my voice: Enhancing android robot likability through personalization by cloning the user's voice. In: *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp 192–198, <https://doi.org/10.1109/RO-MAN63969.2025.11217611>
- [37] Kuch JM, Withanage Don DS, Heimerl N, et al (2026) What happens when a virtual character speaks with the cloned voice of a person you know? In: *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*. ACM, pp 1–5, <https://doi.org/10.1145/3772363.3799332>
- [38] Lavan N, McGettigan C (2023) A model for person perception from familiar and unfamiliar voices. *Communications Psychology* 1(1):1. <https://doi.org/10.1038/s44271-023-00001-4>
- [39] Lavan N, Sutherland CAM (2024) Idiosyncratic and shared contributions shape impressions from voices and faces. *Cognition* 251:105881. <https://doi.org/10.1016/j.cognition.2024.105881>
- [40] Laver JDM (1968) Voice quality and indexical information. *British Journal of Disorders of Communication* 3(1):43–54. <https://doi.org/10.3109/13682826809011440>
- [41] Ledo D, Houben S, Vermeulen J, et al (2018) Evaluation strategies for HCI toolkit research. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, CHI '18, pp 36:1–36:17, <https://doi.org/10.1145/3173574.3173610>

- [42] Li K, Mostajeran F, Rings S, et al (2026) Anthropomorphic AI: a toolkit for authoring and interacting with intelligent virtual agents for extended reality. *Frontiers in Virtual Reality* 7:1794720. <https://doi.org/10.3389/frvir.2026.1794720>
- [43] Li M, Zhang J, Guo F, et al (2025) Audiovisual affective design of humanoid robot appearance and voice based on Kansei Engineering. *International Journal of Social Robotics* 17(1):15–37. <https://doi.org/10.1007/s12369-024-01202-5>
- [44] Lim YK, Stolterman E, Tenenberg J (2008) The anatomy of prototypes: Prototypes as filters, prototypes as manifestations of design ideas. *ACM Transactions on Computer-Human Interaction* 15(2):7:1–7:27. <https://doi.org/10.1145/1375761.1375762>
- [45] MacLean A, Young RM, Bellotti VME, et al (1991) Questions, options, and criteria: Elements of design space analysis. *Human-Computer Interaction* 6(3–4):201–250. <https://doi.org/10.1080/07370024.1991.9667168>
- [46] Mayring P (2014) Qualitative content analysis: Theoretical foundation, basic procedures and software solution. Klagensfurt, open access monograph, URL <https://nbn-resolving.org/urn:nbn:de:0168-ssolar-395173>
- [47] McAleer P, Todorov A, Belin P (2014) How do you say “hello”? personality impressions from brief novel voices. *PLOS ONE* 9(3):e90779. <https://doi.org/10.1371/journal.pone.0090779>
- [48] McDonald N, Schoenebeck S, Forte A (2019) Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW):72:1–72:23. <https://doi.org/10.1145/3359174>
- [49] McGinn C, Torre I (2019) Can you tell the robot by the voice? an exploratory study on the role of voice in the perception of robots. In: 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, pp 211–221, <https://doi.org/10.1109/HRI.2019.8673305>
- [50] Miniota J, Wang S, Beskow J, et al (2023) Hi robot, it’s not what you say, it’s how you say it. In: 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp 307–314, <https://doi.org/10.1109/RO-MAN57019.2023.10309427>
- [51] Mitchell WJ, Szerszen KA, Lu AS, et al (2011) A mismatch in the human realism of face and voice produces an uncanny valley. *i-Perception* 2(1):10–12. <https://doi.org/10.1068/i0415>
- [52] Moore RK (2017) Appropriate voices for artefacts: Some key insights. In: 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2017), URL [https://vihar-2017.vihar.org/assets/papers/VIHAR-2017\\_paper\\_8.pdf](https://vihar-2017.vihar.org/assets/papers/VIHAR-2017_paper_8.pdf)
- [53] Moore RK (2017) Is spoken language all-or-nothing? implications for future speech-based human-machine interaction. In: Jokinen K, Wilcock G (eds) *Dialogues with Social Robots: Enablements, Analyses, and Evaluation*, Lecture Notes in Electrical Engineering, vol 427. Springer Singapore, p 281–291, [https://doi.org/10.1007/978-981-10-2585-3\\_22](https://doi.org/10.1007/978-981-10-2585-3_22)
- [54] Nass C, Lee KM (2000) Does computer-generated speech manifest personality? an experimental test of similarity-attraction. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, pp 329–336, <https://doi.org/10.1145/332040.332452>
- [55] Noufi C, May L, Berger J (2023) The role of vocal persona in natural and synthesized speech. In: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG). IEEE, pp 1–4, <https://doi.org/10.1109/FG57933.2023.10042714>
- [56] Noufi C, May L, Berger J (2025) A model of vocal persona: Context, perception, production. *Frontiers in Computer Science* 7:1575296. <https://doi.org/10.3389/fcomp.2025.1575296>

- [57] O'Brien BC, Harris IB, Beckman TJ, et al (2014) Standards for reporting qualitative research: A synthesis of recommendations. *Academic Medicine* 89(9):1245–1251. <https://doi.org/10.1097/ACM.0000000000000388>
- [58] Ørngreen R, Levinsen K (2017) Workshops as a research methodology. *Electronic Journal of e-Learning* 15(1):70–81. URL <https://files.eric.ed.gov/fulltext/EJ1140102.pdf>
- [59] Owens K, Gunawan J, Choffnes D, et al (2022) Exploring deceptive design patterns in voice interfaces. In: 2022 European Symposium on Usable Security (EuroUSEC 2022). Association for Computing Machinery, pp 64–78, <https://doi.org/10.1145/3549015.3554213>
- [60] Palinkas LA, Horwitz SM, Green CA, et al (2015) Purposeful sampling for qualitative data collection and analysis in mixed method implementation research. *Administration and Policy in Mental Health and Mental Health Services Research* 42(5):533–544. <https://doi.org/10.1007/s10488-013-0528-y>
- [61] Pearl C (2016) *Designing Voice User Interfaces: Principles of Conversational Experiences*. O'Reilly Media, URL <https://www.oreilly.com/library/view/designing-voice-user/9781491955406/>
- [62] Perugia G, Lisy D (2023) Robot's gendering trouble: A scoping review of gendering humanoid robots and its effects on HRI. *International Journal of Social Robotics* 15:1725–1753. <https://doi.org/10.1007/s12369-023-01061-6>
- [63] Podesva RJ, Callier P (2015) Voice quality and identity. *Annual Review of Applied Linguistics* 35:173–194. <https://doi.org/10.1017/S0267190514000270>
- [64] Porcheron M, Fischer JE, Reeves S, et al (2018) Voice interfaces in everyday life. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, pp 1–12, <https://doi.org/10.1145/3173574.3174214>
- [65] Rincón C, Keyes O, Cath C (2021) Speaking from experience: Trans/non-binary requirements for voice-activated AI. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW1):1–27. <https://doi.org/10.1145/3449206>
- [66] Rist T, André E (2003) Building smart embodied virtual characters. In: Butz A, Krüger A, Olivier P (eds) *Smart Graphics*, Lecture Notes in Computer Science, vol 2733. Springer, Berlin, Heidelberg, pp 123–130, [https://doi.org/10.1007/3-540-37620-8\\_12](https://doi.org/10.1007/3-540-37620-8_12)
- [67] Robinson FA, Bown O, Velonaki M (2022) Designing Sound for Social Robots: Candidate Design Principles. *International Journal of Social Robotics* 14:1507–1525. <https://doi.org/10.1007/s12369-022-00891-0>
- [68] Rosi V, Soopramanien E, McGettigan C (2025) Perception and social evaluation of cloned and recorded voices: Effects of familiarity and self-relevance. *Computers in Human Behavior: Artificial Humans* 4:100143. <https://doi.org/10.1016/j.chbah.2025.100143>
- [69] Sarigul B, Schneider FM, Utz S (2025) Believe it or not? investigating the credibility of voice assistants in the context of social roles and relationship types. *International Journal of Human-Computer Interaction* 41(10):6253–6265. <https://doi.org/10.1080/10447318.2024.2375797>
- [70] Scherer KR (2003) Vocal communication of emotion: A review of research paradigms. *Speech Communication* 40(1–2):227–256. [https://doi.org/10.1016/S0167-6393\(02\)00084-5](https://doi.org/10.1016/S0167-6393(02)00084-5)
- [71] Schreibelmayer S, Mara M (2022) Robot voices in daily life: Vocal human-likeness and application context as determinants of user acceptance. *Frontiers in Psychology* 13:787499. <https://doi.org/10.3389/fpsyg.2022.787499>
- [72] Schweinberger SR, Kawahara H, Simpson AP, et al (2014) Speaker perception. *WIREs*

- Cognitive Science 5(1):15–25. <https://doi.org/10.1002/wcs.1261>
- [73] Seaborn K (2025) Social identity in human-agent interaction: A primer. *ACM Transactions on Human-Robot Interaction* <https://doi.org/10.1145/3760500>
  - [74] Seaborn K, Urakami J (2021) Measuring Voice UX Quantitatively: A Rapid Review. In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, pp 1–8, <https://doi.org/10.1145/3411763.3451712>
  - [75] Seaborn K, Miyake NP, Pennefather P, et al (2021) Voice in human-agent interaction: A survey. *ACM Computing Surveys* 54(4):1–43. <https://doi.org/10.1145/3386867>
  - [76] Seaborn K, Urakami J, Pennefather P, et al (2024) Qualitative approaches to voice UX. *ACM Computing Surveys* 56(12):301:1–301:34. <https://doi.org/10.1145/3658666>
  - [77] Shefcik G, Tsai PT (2023) Voice-related experiences of nonbinary individuals (VENI) development and content validity. *Journal of Voice* 37(2):294.e5–294.e13. <https://doi.org/10.1016/j.jvoice.2020.12.037>
  - [78] Skoog Waller S, Eriksson M, Sörqvist P (2015) Can you hear my age? influences of speech rate and speech spontaneity on estimation of speaker age. *Frontiers in Psychology* 6:978. <https://doi.org/10.3389/fpsyg.2015.00978>
  - [79] Skuk VG, Jacob I, Wientzek R, et al (2025) Big Five personality perceptions from voices and faces: Impressions and kernels of truth. *Journal of Nonverbal Behavior* 49:125–154. <https://doi.org/10.1007/s10919-025-00480-5>
  - [80] Sloan RJS (2015) *Virtual Character Design for Games and Interactive Media*. CRC Press, Boca Raton, FL
  - [81] Stan A, Lőrincz B (2021) Generating the voice of the interactive virtual assistant. In: *Virtual Assistant*. IntechOpen, <https://doi.org/10.5772/intechopen.95510>
  - [82] Sundar SS, Jung EH, Waddell TF, et al (2017) Cheery companions or serious assistants? role and demeanor congruity as predictors of robot attraction and use intentions among senior citizens. *International Journal of Human-Computer Studies* 97:88–97. <https://doi.org/10.1016/j.ijhcs.2016.08.006>
  - [83] Sutton SJ, Foulkes P, Kirk D, et al (2019) Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction. In: Brewster S (ed) *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, ACM Digital Library, pp 1–14, <https://doi.org/10.1145/3290605.3300833>
  - [84] Tabassi E (2023) Artificial intelligence risk management framework (AI RMF 1.0). NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD, <https://doi.org/10.6028/NIST.AI.100-1>
  - [85] Torre I, Lagerstedt E, Dennler N, et al (2023) Can a gender-ambiguous voice reduce gender stereotypes in human-robot interactions? In: *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp 106–112, <https://doi.org/10.1109/ro-man57019.2023.10309500>
  - [86] Tschoop M, Gieselmann M, Sassenberg K (2023) Servant by default? how humans perceive their relationship with conversational AI. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 17(3):Article 9. <https://doi.org/10.5817/CP2023-3-9>
  - [87] Tuttösi P, Hughson E, Matsufuji A, et al (2023) Read the room: Adapting a robot’s voice to ambient and social contexts. In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp 3998–4005, <https://doi.org/10.1109/IROS55552.2023.10341925>
  - [88] Weiss A, Plöcknig M (2022) Gendersensitive Pflegeroboter: Unbedingt, aber wie? In: *Frauengesundheit und Digitalisierung: Sammelband des Wiener*

Programms für Frauengesundheit. No. 2 in Schriftenreihe Frauen\*Gesundheit\*Wien, Stadt Wien, Wien, p 127–130, URL <https://www.digital.wienbibliothek.at/wbrup/content/titleinfo/3836454>

- [89] Wenzel K, Kaufman G (2024) Designing for harm reduction: Communication repair for multicultural users’ voice interactions. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp 1–17, <https://doi.org/10.1145/3613904.3642900>
- [90] Wilson S, Moore RK (2017) Robot, alien and cartoon voices: Implications for speech-enabled systems. In: 1st International Workshop on Vocal Interactivity In-and-Between Humans, Animals and Robots (VIHAR-2017), pp 40–44
- [91] Wooldridge M, Jennings NR (1995) Intelligent agents: theory and practice. The Knowledge Engineering Review 10(2):115–152. <http://doi.org/10.1017/S0269888900008122>
- [92] Zhang S, Chen H, Wang Y, et al (2025) The manipulative power of voice characteristics: Investigating deceptive patterns in Mandarin Chinese female synthetic speech. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 9(3):1–32. <http://doi.org/10.1145/3749510>

## Appendix A Voice Design Compass: Questions and Literature-Based Examples

Voice material can include speech and other vocal sounds, recorded, performed live or synthesized. The examples illustrate possible answers to the questions. The explanations clarify what these answers can mean in a project, and the citations point to related research and guidance. Combinations, intermediate positions and other answers remain possible.

Where companion material is separate, question numbers link to the matching examples or sources and back. Ethics sources appear directly after each question. Table headings and outer rules follow the figures: gray for Context, purple for design questions, blue for reviews, and green for trying out and revising.

### Starting point and use

Start with the character, intended users and listeners, situation and project goal. The Context questions are optional aids. Questions may be prioritized, reordered, skipped or answered freely. Context and the voice concept remain revisable. The Compass helps you create and maintain your own Project Record, a working document of intentions, requirements, reasons, evidence, open questions and revisions.

**Table A1 Project and Use Context**

1. In what setting will people use or hear the character?
2. What will the surroundings sound like where people hear the voice?
3. Who is the voice intended for, and who else can hear it?
4. What pattern of use do you expect over time?
5. What will people need to understand or pay attention to while listening?
6. How old are the intended users and listeners?
7. What experience, abilities or support needs matter for these users and listeners?

Examples and sources: Table A2 on page 36.

**Table A2: Project and Use Context: Examples and sources**

Questions: Table A1 on page 36.

| Example answers                    | Explanation and sources                                                                            |
|------------------------------------|----------------------------------------------------------------------------------------------------|
| <b>Question 1</b>                  |                                                                                                    |
| <i>Home / private living space</i> | Domestic voice interaction can be part of shared activities and conversation. [136, 172, 219, 221] |
| <i>Car / driving</i>               | [108, 180, 211]                                                                                    |
| <i>Hospital / healthcare</i>       | [198, 232, 245, 262]                                                                               |
| <i>Workplace / office</i>          | [145, 172]                                                                                         |
| <i>School / education</i>          | [156, 172, 181, 187, 249]                                                                          |
| <i>Retail / service</i>            | [136, 160, 172, 198, 201, 245]                                                                     |

Continued on next page

**Table A2: Project and Use Context: Examples and sources (continued)**

Questions: Table A1 on page 36.

| Example answers                                                                                                                                                                                                                                                                                                                                                                                                                                                | Explanation and sources                                                                                                  |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|
| <i>Entertainment</i>                                                                                                                                                                                                                                                                                                                                                                                                                                           | [232, 263]                                                                                                               |
| Users operate or interact with the character. Intended listeners are addressed by it, while others may overhear [231].                                                                                                                                                                                                                                                                                                                                         |                                                                                                                          |
| <b>Question 2</b>                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                          |
| <i>Quiet</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                   | [172, 219, 257]                                                                                                          |
| <i>Loud / noisy</i>                                                                                                                                                                                                                                                                                                                                                                                                                                            | [143, 172, 257]                                                                                                          |
| <i>Changing background sound</i>                                                                                                                                                                                                                                                                                                                                                                                                                               | [172, 215, 227, 257, 258]                                                                                                |
| <b>Question 3</b>                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                          |
| <i>One person, with private listening</i>                                                                                                                                                                                                                                                                                                                                                                                                                      | No other listeners are expected, and the listening arrangement is intended to keep the exchange private. [160, 172, 192] |
| <i>One person, with others able to overhear</i>                                                                                                                                                                                                                                                                                                                                                                                                                | [172, 219, 222, 268]                                                                                                     |
| <i>A group sharing the interaction</i>                                                                                                                                                                                                                                                                                                                                                                                                                         | For example, household members, a class or a team. Not everyone has to speak. [172, 221, 234, 235]                       |
| <i>An open or passing audience</i>                                                                                                                                                                                                                                                                                                                                                                                                                             | A broad or open audience beyond a defined small group. [172]                                                             |
| <b>Question 4</b>                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                          |
| <i>Brief individual interaction</i>                                                                                                                                                                                                                                                                                                                                                                                                                            | [112, 117, 160, 257]                                                                                                     |
| <i>Repeated use</i>                                                                                                                                                                                                                                                                                                                                                                                                                                            | [115, 178, 268]                                                                                                          |
| <i>Use over a longer period by the same person</i>                                                                                                                                                                                                                                                                                                                                                                                                             | [181, 192, 198]                                                                                                          |
| Duration of one interaction, frequency of use and overall period of use are separate aspects.                                                                                                                                                                                                                                                                                                                                                                  |                                                                                                                          |
| <b>Question 5</b>                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                          |
| <i>Learning, understanding or remembering information</i>                                                                                                                                                                                                                                                                                                                                                                                                      | [134, 189, 240]                                                                                                          |
| <i>Listening while doing another task</i>                                                                                                                                                                                                                                                                                                                                                                                                                      | [270]                                                                                                                    |
| <i>Needing to respond quickly</i>                                                                                                                                                                                                                                                                                                                                                                                                                              | [166, 261]                                                                                                               |
| <i>Keeping attention on the voice for a long stretch</i>                                                                                                                                                                                                                                                                                                                                                                                                       | [153, 193]                                                                                                               |
| <b>Question 6</b>                                                                                                                                                                                                                                                                                                                                                                                                                                              |                                                                                                                          |
| <i>Adults</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                  | [136]                                                                                                                    |
| <i>Older adults</i>                                                                                                                                                                                                                                                                                                                                                                                                                                            | [109, 127, 222, 236, 262, 268]                                                                                           |
| <i>Children</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                | [141, 156, 218, 249]                                                                                                     |
| <i>Mixed-age group</i>                                                                                                                                                                                                                                                                                                                                                                                                                                         | [172]                                                                                                                    |
| Age descriptions can overlap. Listener age is separate from the character's apparent age and the voice's age impression. Consider age alongside hearing-related listening requirements: older respondents in a survey of deaf and hard-of-hearing users reported difficulty with the pitch or tone of the assistant's voice more often than younger respondents [107]. Language experience, hearing and support needs are considered separately in Question 7. |                                                                                                                          |

Continued on next page

**Table A2: Project and Use Context: Examples and sources (continued)**

Questions: Table A1 on page 36.

| Example answers                                       | Explanation and sources                                                                                                                                                 |
|-------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Question 7</b>                                     |                                                                                                                                                                         |
| <i>Deaf or hard-of-hearing users</i>                  | Listening requirements can differ with hearing characteristics and hearing devices. Relevant considerations include pitch, speaking rate and background noise. [107]    |
| <i>Little experience with this type of technology</i> | [109]                                                                                                                                                                   |
| <i>People listening in a non-native language</i>      | Consider listeners' language experience, support for understanding the character's speech, and the impressions that different voices evoke in this audience. [129, 262] |

### Table A3 Design Set: Function and Relationship

1. What should the character do for its users?
2. What kind of relationship should the character have with its users?

Examples and sources follow below (Table A4).

### Table A4: Function and Relationship: Examples and sources

Questions: Table A3 on page 39.

| Example answers                                                                                                                                                           | Explanation and sources                                                                                |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|
| <b>Question 1</b>                                                                                                                                                         |                                                                                                        |
| <i>Tool / Device</i>                                                                                                                                                      | Performs a specific task with little social interaction. [112, 128, 147, 160]                          |
| <i>Assistant</i>                                                                                                                                                          | Helps with tasks or responds to wishes and needs. [222, 243]                                           |
| <i>Service Agent</i>                                                                                                                                                      | Provides a particular service, such as reception, ordering or information. [123]                       |
| <i>Teacher / Tutor</i>                                                                                                                                                    | Supports teaching, practice, peer tutoring, or guided learning. [173, 235]                             |
| <i>Coach / Advisor</i>                                                                                                                                                    | Offers advice, helps people think through decisions, or supports a personal goal. [112, 165, 222, 268] |
| <i>Companion</i>                                                                                                                                                          | Offers ongoing company and social or emotional support. [195, 237]                                     |
| <i>Guardian / Monitor</i>                                                                                                                                                 | Guards people or property. [253]                                                                       |
| Roles can overlap and change with the situation.                                                                                                                          |                                                                                                        |
| <b>Question 2</b>                                                                                                                                                         |                                                                                                        |
| <b>Authority:</b> who can decide what, and in which situation?                                                                                                            |                                                                                                        |
| <b>Support:</b> what practical, social or emotional help is planned?                                                                                                      |                                                                                                        |
| <b>Formality:</b> how formal is the relationship?                                                                                                                         |                                                                                                        |
| <i>Authority: Subordinate</i>                                                                                                                                             | The user normally makes decisions. The character follows those decisions. [222, 230]                   |
| <i>Authority: Equal-status / non-hierarchical</i>                                                                                                                         | Neither character nor user has a stable position of authority over the other. [255]                    |
| <i>Support: Practical or emotional support</i>                                                                                                                            | Help with everyday tasks or emotional support. [130]                                                   |
| <i>Formality: Professional</i>                                                                                                                                            | The relationship is organized around a formal task or service. [110, 170, 173]                         |
| Other relationships are possible, for example a character that instructs or supervises its users, or an informal, personal relationship. Describe them in your own words. |                                                                                                        |

**Table A5 Design Set: Appearance and Embodiment**

**Body and Equipment**

1. Does the character have a physical form, a digital form, or both?
2. What overall shape does the character have?
3. How tall is the character, how much space does it take up, and how bulky or slender does it look?
4. What other aspects of the character's body, appearance or equipment might matter for its voice?

**Visible Expression**

5. What visual style does the character have, and which features look human?
6. How old does the character look?
7. How does the character's appearance express gender?

**Relationship with the Voice**

8. How should the character's movement, posture and other visible signals work together with the voice?
9. How should the voice fit with, or deliberately contrast with, the character's appearance?

Examples and sources: Table A6 on page 40.

**Table A6: Appearance and Embodiment: Examples and sources**

Questions: Table A5 on page 40.

| Example answers                 | Explanation and sources                                                                              |
|---------------------------------|------------------------------------------------------------------------------------------------------|
| <b>Body and Equipment</b>       |                                                                                                      |
| <b>Question 1</b>               |                                                                                                      |
| <i>Physical embodiment</i>      | A robot, kiosk or toy. [112, 155, 171, 237]                                                          |
| <i>Virtual</i>                  | An avatar, on-screen character or character in virtual reality. [111, 139, 142, 149, 154, 177, 269]  |
| <i>Hybrid</i>                   | Physical and virtual elements in one embodiment, such as a robot with a digital face. [95, 152, 155] |
| <b>Question 2</b>               |                                                                                                      |
| <i>Humanoid</i>                 | A human-like body with a head, torso, and limbs. [149, 236, 254]                                     |
| <i>Animal-like (zoomorphic)</i> | [112, 171]                                                                                           |

Continued on next page

**Table A6: Appearance and Embodiment: Examples and sources (continued)**

Questions: Table A5 on page 40.

| Example answers                                                                                                                                                                                                                                                                                                                                                                                     | Explanation and sources                                                                                                                                                                      |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Mechanical/Functional</i>                                                                                                                                                                                                                                                                                                                                                                        | A mechanical form or one primarily shaped by its practical function. [130, 155, 159, 164, 254, 260]                                                                                          |
| <i>Object-like: rounded or soft-looking</i>                                                                                                                                                                                                                                                                                                                                                         | Rounded contours, a soft-looking surface, or both, as in a toy or sphere. [112, 143, 198]                                                                                                    |
| <i>Object-like: angular or machine-like</i>                                                                                                                                                                                                                                                                                                                                                         | Angular contours, a machine-like appearance, or both. [198, 225]                                                                                                                             |
| <b>Question 3</b>                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                              |
| <i>Large or tall</i>                                                                                                                                                                                                                                                                                                                                                                                | Large overall size, substantial height, or both. [138, 143]                                                                                                                                  |
| <i>Small or short</i>                                                                                                                                                                                                                                                                                                                                                                               | Small overall size, low height, or both, as in a tabletop character. [143, 144]                                                                                                              |
| <i>Heavy-looking</i>                                                                                                                                                                                                                                                                                                                                                                                | A body that appears heavy. This adapts findings on weight impressions from voices to visible build. [138]                                                                                    |
| Overall size, height and visible build are separate aspects. Digital size refers to the displayed scale. Consider how the voice might fit or contrast with the body's size: a robot-image study found that pitch and vocal brightness influenced voice-size associations [143]. Record overall size, height and visible build where they are meaningful for the planned physical or displayed form. |                                                                                                                                                                                              |
| <b>Question 4</b>                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                              |
| <i>Represented disability and communication aids</i>                                                                                                                                                                                                                                                                                                                                                | A represented disability or communication aid that matters for the character's communication or identity. This adapts an account of an assistive voice as part of a person's identity. [224] |
| <i>Clothing</i>                                                                                                                                                                                                                                                                                                                                                                                     | Clothing that contributes to the character's intended identity or role, and how the voice should relate to that appearance. [133]                                                            |
| <i>Color</i>                                                                                                                                                                                                                                                                                                                                                                                        | The character's color scheme and the impression it is intended to create alongside the voice. [188]                                                                                          |
| <i>Facial features</i>                                                                                                                                                                                                                                                                                                                                                                              | The shape and detail of the character's face, including how humanlike it looks and what impression it should create together with the voice. [200]                                           |
| <i>Accessories and equipment</i>                                                                                                                                                                                                                                                                                                                                                                    | Items worn or carried by the character, such as a face covering, including any effects on how its voice is heard. [124]                                                                      |
| <b>Visible Expression</b>                                                                                                                                                                                                                                                                                                                                                                           |                                                                                                                                                                                              |
| <b>Question 5</b>                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                              |
| <i>Nearly human / photorealistic</i>                                                                                                                                                                                                                                                                                                                                                                | A nearly human android or photorealistic digital human. [142, 149, 269]                                                                                                                      |
| <i>Stylized</i>                                                                                                                                                                                                                                                                                                                                                                                     | A simplified, exaggerated, illustrated, anime, or cartoon aesthetic in 2D or 3D. [142, 149, 214, 232, 248, 269]                                                                              |
| <i>Humanoid / visibly mechanical</i>                                                                                                                                                                                                                                                                                                                                                                | Human anatomy or proportions with a clearly visible mechanical construction, such as Pepper or NAO. [142, 152, 162, 173, 229, 232]                                                           |
| <i>Clearly non-human</i>                                                                                                                                                                                                                                                                                                                                                                            | An abstract, function-oriented mechanical or animal-like form with little human anatomy. [135, 142, 162, 197, 201–203, 225, 252]                                                             |
| <b>Question 6</b>                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                              |
| <i>Child</i>                                                                                                                                                                                                                                                                                                                                                                                        | [159]                                                                                                                                                                                        |
| <i>Adult</i>                                                                                                                                                                                                                                                                                                                                                                                        | [140, 159]                                                                                                                                                                                   |
| <i>Older adult</i>                                                                                                                                                                                                                                                                                                                                                                                  | [140, 241]                                                                                                                                                                                   |
| Apparent character age is separate from listener age and voice age. The older-adult appearance example transfers findings from human audiovisual and vocal age perception.                                                                                                                                                                                                                          |                                                                                                                                                                                              |

Continued on next page

**Table A6: Appearance and Embodiment: Examples and sources (continued)**

Questions: Table A5 on page 40.

| Example answers                                                                                                                                               | Explanation and sources                                                                                                                                            |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Question 7</b>                                                                                                                                             |                                                                                                                                                                    |
| <i>Design intention</i>                                                                                                                                       |                                                                                                                                                                    |
| <i>Gender not emphasized</i>                                                                                                                                  | Give gender little emphasis in the appearance. [131]                                                                                                               |
| <i>Mixed visual cues</i>                                                                                                                                      | Combining features intended to suggest feminine and masculine expression. [131]                                                                                    |
| <i>Intended perception</i>                                                                                                                                    |                                                                                                                                                                    |
| <i>Feminine</i>                                                                                                                                               | [112, 198, 225]                                                                                                                                                    |
| <i>Masculine</i>                                                                                                                                              | [102, 112, 138, 143, 198, 225, 235]                                                                                                                                |
| <i>Gender-ambiguous impression</i>                                                                                                                            | Individual uncertainty, differing interpretations between viewers, or both. [131]                                                                                  |
| <b>Observed perception:</b> The gender impressions viewers report when asked about the character's appearance. These may differ from the intended impression. |                                                                                                                                                                    |
| <i>Represented identity (optional)</i>                                                                                                                        |                                                                                                                                                                    |
| <i>Non-binary</i>                                                                                                                                             | An identity not defined exclusively as a man or woman. [132]                                                                                                       |
| <i>Genderqueer</i>                                                                                                                                            | A deliberately chosen gender designation beyond conventional classifications. Its intended meaning may need explanation. [132]                                     |
| <i>Queer</i>                                                                                                                                                  | A chosen term that may refer to gender, sexuality or questioning these norms. [132]                                                                                |
| Identity may be described as woman, man or another chosen term, or left unspecified. A real reference person's self-description provides their identity.      |                                                                                                                                                                    |
| <b>Relationship with the Voice</b>                                                                                                                            |                                                                                                                                                                    |
| <b>Question 8</b>                                                                                                                                             |                                                                                                                                                                    |
| <i>Movement</i>                                                                                                                                               | How natural or mechanical the character's movements look alongside its voice. [142]                                                                                |
| <i>Posture</i>                                                                                                                                                | Body posture accompanying vocal expression, or head orientation in relation to how the voice sounds and where the sound travels. [216, 256]                        |
| <i>Facial expression, mouth movements and lights</i>                                                                                                          | Visible expressions, mouth movements or light signals that accompany the character's voice. They can be coordinated with its timing or expression. [151, 167, 207] |
| <b>Question 9</b>                                                                                                                                             |                                                                                                                                                                    |
| <i>Close match</i>                                                                                                                                            | Voice intended to match appearance and expressive style. [95, 135, 149, 152, 155, 196, 204, 229, 256]                                                              |
| <i>Strong contrast</i>                                                                                                                                        | Pronounced, deliberate voice-body contrast. [95, 112, 152, 200, 256]                                                                                               |
| The intended relationship can differ by feature.                                                                                                              |                                                                                                                                                                    |

### Table A7 Design Set: Voice Identity and Familiarity

1. Whose voice is the character's voice based on, or is it newly created?
2. What should the voice communicate or conceal about who or what it represents?
3. How familiar should the voice sound, and how similar should it be to its listeners?
4. Who chooses the voice, and who can change it once the character is in use?

Examples and sources follow below (Table A8).

### Table A8: Voice Identity and Familiarity: Examples and sources

Questions: Table A7 on page 43.

| Example answers                                                                                                                                                           | Explanation and sources                                                                                                                                        |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Question 1</b>                                                                                                                                                         |                                                                                                                                                                |
| <i>User's own voice</i>                                                                                                                                                   | The user's own voice, used in recordings or recreated through voice cloning. [174, 197, 228]                                                                   |
| <i>Familiar person's voice</i>                                                                                                                                            | A voice already known to intended listeners, used in recordings or as a model for the character's voice. [175, 197, 228]                                       |
| <i>Celebrity or public figure</i>                                                                                                                                         | A named celebrity's or public figure's voice, used in recordings or as a model for the character's voice. [175, 197]                                           |
| <i>New voice for the character</i>                                                                                                                                        | No particular real person serves as the voice's identity model. [103, 197, 228, 239]                                                                           |
| Familiarity depends on the listener (see question 3 in this Set).                                                                                                         |                                                                                                                                                                |
| <b>Question 2</b>                                                                                                                                                         |                                                                                                                                                                |
| <i>Signal that the character is artificial</i>                                                                                                                            | [112, 210, 233]                                                                                                                                                |
| <i>Connect with a chosen audience group</i>                                                                                                                               | Selected features, such as accent, are intended to connect the voice with a particular audience group. [108, 120, 137, 161, 163, 179, 210, 234, 235, 243, 245] |
| <i>Hide selected identity cues</i>                                                                                                                                        | Specified age, gender or personal-identity cues should remain hidden. Anonymity is not guaranteed. [110, 197, 226]                                             |
| Deliberate distinction from a reference voice is also possible. Speaker-similarity comparisons inform the technical follow-up (question 3 in Technical Reflection) [244]. |                                                                                                                                                                |
| <b>Question 3</b>                                                                                                                                                         |                                                                                                                                                                |
| <i>Familiarity</i>                                                                                                                                                        | The voice should be recognizable as that of a familiar person, such as a family member or friend. [158, 175, 228]                                              |
| <i>User similarity</i>                                                                                                                                                    | Intended resemblance to listeners in selected features: accent, speaking style, age or gender impression, or personality impression. [112, 209, 235, 267]      |
| <b>Question 4</b>                                                                                                                                                         |                                                                                                                                                                |
| <i>Fixed voice chosen by the organization</i>                                                                                                                             | One fixed voice for all users, without user adjustments. [97, 108, 110, 130, 243]                                                                              |
| <i>Voice selected by the system</i>                                                                                                                                       | Selection based on information about the user. [108, 243, 245]                                                                                                 |
| <i>User choice or adjustment</i>                                                                                                                                          | Including selection from prepared voices. [97, 108, 110, 137, 161, 197, 237, 243, 245, 268]                                                                    |

### Table A9 Design Set: Intended Voice Impressions

1. How human-like should the voice sound?
2. What impressions should listeners form of the character and its voice?
3. How should the voice express gender, and how old should it sound?

Examples and sources: Table A10 on page 44.

### Table A10: Intended Voice Impressions: Examples and sources

Questions: Table A9 on page 44.

| Example answers                                                                                                                                                    | Explanation and sources                                                                                            |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------|
| <b>Question 1</b>                                                                                                                                                  |                                                                                                                    |
| <i>Very human-like</i>                                                                                                                                             | The voice is intended to sound like a person's voice. [173, 176, 232, 251]                                         |
| <i>Low humanlikeness / clearly non-human</i>                                                                                                                       | The voice is intended to sound recognizably artificial or non-human. [93, 173]                                     |
| <b>Question 2</b>                                                                                                                                                  |                                                                                                                    |
| Intended impressions and their measures can overlap [104].                                                                                                         |                                                                                                                    |
| <b>Interpersonal impressions</b>                                                                                                                                   |                                                                                                                    |
| <i>Social warmth</i>                                                                                                                                               | The character should seem friendly, well intentioned, and approachable. [119, 146, 191, 228]                       |
| <i>Likability</i>                                                                                                                                                  | The character should be liked overall. [93, 111, 117, 136, 157, 174–176, 188, 191, 208, 213, 234]                  |
| <i>Trustworthiness</i>                                                                                                                                             | The character should seem reliable, honest, and well intentioned. [121, 122, 206, 228, 251]                        |
| <b>Credibility and capability</b>                                                                                                                                  |                                                                                                                    |
| <i>Credibility</i>                                                                                                                                                 | The character, its information, or both should seem believable. [122, 126, 137, 240]                               |
| <i>Competence</i>                                                                                                                                                  | The character should seem capable, knowledgeable, and effective in its role. [119, 123, 136, 206, 228]             |
| <i>Perceived Intelligence</i>                                                                                                                                      | The character should seem able to understand or reason in its task. [105, 199]                                     |
| <b>Perceived status</b>                                                                                                                                            |                                                                                                                    |
| <i>Authority</i>                                                                                                                                                   | The character should seem to have authority, a leadership role, or the right to give instructions. [136, 206, 268] |
| <b>Personality impressions</b>                                                                                                                                     |                                                                                                                    |
| <i>Introversion</i>                                                                                                                                                | Reserved and socially less expressive. [210]                                                                       |
| <i>Extraversion</i>                                                                                                                                                | Outgoing, energetic, and socially expressive. [210, 213, 235]                                                      |
| OCEAN covers Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism [242]. Their independent control through synthesis is not established [148]. |                                                                                                                    |
| <b>The voice itself</b>                                                                                                                                            |                                                                                                                    |
| <i>Vocal attractiveness</i>                                                                                                                                        | The voice itself should sound attractive. [176, 213, 228]                                                          |
| <b>Influence on decisions</b>                                                                                                                                      |                                                                                                                    |
| <i>Persuasiveness</i>                                                                                                                                              | A goal of influencing judgments or decisions (see Ethics question 14 and Shared Reflection). [210, 268]            |

Continued on next page

**Table A10: Intended Voice Impressions: Examples and sources (continued)**

Questions: Table A9 on page 44.

| Example answers                                                                                                                                                                              | Explanation and sources                                                                                              |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|
| <b>Question 3</b>                                                                                                                                                                            |                                                                                                                      |
| <b><i>Design intention</i></b>                                                                                                                                                               |                                                                                                                      |
| <i>Gender not emphasized</i>                                                                                                                                                                 | Give gender little emphasis in the voice design. [131]                                                               |
| <b><i>Intended gender perception</i></b>                                                                                                                                                     |                                                                                                                      |
| <i>Gender-ambiguous impression</i>                                                                                                                                                           | Intended uncertainty for individual listeners, differing interpretations between listeners, or both. [173, 246, 250] |
| <i>Masculine impression</i>                                                                                                                                                                  | Listeners should perceive the voice as masculine. [133, 205, 206]                                                    |
| <i>Feminine impression</i>                                                                                                                                                                   | Listeners should perceive the voice as feminine. [133, 194, 205]                                                     |
| <b>Observed perception:</b> The gender impressions listeners report when asked about the voice. These may differ from the intended impression.                                               |                                                                                                                      |
| <b><i>Age impression</i></b>                                                                                                                                                                 |                                                                                                                      |
| <i>Older adult voice impression</i>                                                                                                                                                          | The voice is intended to suggest an older adult. [117, 137, 194]                                                     |
| <i>Young voice impression</i>                                                                                                                                                                | A child or adolescent voice impression. [194]                                                                        |
| Intended and observed age impressions may differ. Represented identity is separate (question 7 in Appearance and Embodiment). Apparent character age is addressed in question 6 of that Set. |                                                                                                                      |

**Table A11 Specific Sound Ideas**

**1. Do you have a particular sound in mind for the voice?**

For example, a slightly rough voice, the sound of an old record, or a light echo. Describe your idea in your own words and explain why you want this sound for the character and situation. You can also refer to an audio example and indicate which aspect you have in mind. This question can remain open.

**Example sound ideas**

These examples link a desired sound to a possible reason from Context or the Design Sets, including Intended Voice Impressions. Technical Reflection addresses the controls and production methods needed to realize that sound.

| <b>Sound idea</b>                     | <b>Example reason and context</b>                                                                                                                                                                                    |
|---------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>Whispered delivery</i>             | Whispering for a character intended to be unobtrusive around nearby people or to create a sense of intimacy. [219]                                                                                                   |
| <i>Creaky or tight-sounding voice</i> | A lax, creaky quality for a character intended to sound relaxed, or a tense quality for one intended to sound stressed. [265]                                                                                        |
| <i>Wider pitch range</i>              | More varied pitch for a character intended to sound enthusiastic. [145]                                                                                                                                              |
| <i>Slower speaking rate</i>           | Slower speech for a robot guide, intended to help visitors follow its information while walking. [238]                                                                                                               |
| <i>Emphasis on key words</i>          | Slightly slower and louder key words intended to draw attention to important material in instructional narration. [134]                                                                                              |
| <i>Electronically altered timbre</i>  | An audibly processed sound for an artificial character: ring modulation as used for the Daleks [263], or vocoder and flanger effects explored in robot voice design. [225]                                           |
| <i>Echo or reverberation</i>          | Audible repetitions (echo) or a lingering sound (reverberation) for a stylized fictional voice, as documented for Ultron in <i>Ultimate Alliance</i> and Max in <i>Flight of the Navigator</i> , respectively. [263] |

**Table A12 Ethical Considerations: Questions and sources**

**Consent and Identity**

1. Does the voice refer to a real person, or could listeners recognize a particular person in it? [175, 197]
2. Have the people whose voices are involved agreed to the planned recording, storage, use and sharing after these uses were explained to them? [197]
3. Does that agreement cover this character, its role and the planned situations and uses? [197]
4. Could the voice give listeners a false idea of who is represented, their relationship to the listener or their authority? [175, 197]
5. Could someone misuse the voice to pretend to be another person, deceive others, damage a person's reputation or misuse their identity? [175, 266, 268]

**Labor and Professional Practice**

6. If the voice audio can be prepared in advance, have you considered hiring a voice performer as well as using synthesis? [96]
7. Could using AI to create the voice replace or reduce paid work by voice performers or other affected workers, or make their work less valued? [96, 106]
8. If recordings of people's voices are reused, can their sources be traced, and are permission, control for contributors, credit and fair payment secured? [96, 106, 185, 197]

**Bias and Stereotypes**

9. Could listeners associate the voice with a particular gender, age, accent, social class or status? [222]
10. Could the voice, the character's role and the social group associated with it reinforce stereotypes, for example about professions or gender identities? [112, 226]
11. Could this voice and role present care, service, assistance or emotional labor (managing emotions as part of a job) as feminine? [222]
12. Could this voice and role make authority, expertise, security or technical skill seem masculine? [235]
13. Have you considered a less stereotyped voice, one whose gender is not clearly perceived, or one that deliberately challenges a stereotype? [97, 253]

**Influence and Manipulation**

14. Is the voice intended to make people trust it more, comply with requests, feel greater urgency or closeness, accept its authority, or be more easily persuaded? [217, 268]
15. Where such influence is intended, will users be able to recognize how the voice is meant to influence them? [232, 268]
16. Will people hear the voice in situations where they are vulnerable, stressed, dependent on support or under time pressure? [175, 268]
17. Could the voice make one option seem more trustworthy, urgent, or desirable than another? [268]

**Information and Expectations**

18. Will people be clearly told how the voice is made and who or what controls the system's vocal output and actions? [232]
19. Could the voice and appearance make listeners overestimate how human the system is, what it can do or how independently it can decide and act? [112, 197, 201, 202]
20. Could the voice make people expect human understanding, care, or personal intentions beyond what the system can actually provide? [197]
21. Would a less human-sounding voice, a deliberately distinctive voice style or a clearly synthetic-sounding voice suit this project better? [157, 197]
22. Will you test your assumptions about what age, authority or other social characteristics the intended users and listeners associate with the voice? [233]

**Diversity and Inclusion**

23. Have you described what listeners may take for granted about the character from its voice? [245]
24. Could the voice leave some user groups out, reinforce stereotypes about them, portray them as strange or exotic, or represent them inaccurately? [222, 245]
25. For people whose voices or identities are represented, what choices will they have, how will their privacy be protected, and what will they be able to influence or change? [226]
26. Has the voice been tested for the relevant languages, accents, people's hearing and access needs, background sound and other acoustic conditions, and situations of use? [197, 245]
27. Will people from the groups affected or represented take part in designing, testing or evaluating the voice? [97, 222, 226]
28. Will you test how people from the relevant cultural backgrounds interpret the voice's social cues in the planned situations? [233]

**Table A13 Technical Reflection: Questions** Examples and sources: Table A14 on page 48.

**Requirements**

1. Which languages must the character be able to speak?
2. What is the longest acceptable delay before the character responds with its voice?
3. What do you need to be able to choose, adjust, or compare in the voice?

**Resources**

4. Which voice recordings do you have or could you obtain, and what are you allowed to use them for?
5. How much money is available to create, set up, and test the voice?
6. How much time is available to create, test, and revise the voice?
7. Who is available to work on the voice, and how much working time can they contribute?
8. What relevant skills can the project draw on?

**Operating conditions**

9. What computers or computing services are available for the voice system?
10. What practical limits and requirements apply when the voice system is used?

**Production approach**

11. How can you produce the character's voice with the resources available and the requirements you have set?

**Table A14: Technical Reflection: Examples and sources**

Questions: Table A13 on page 48.

| Example answers                                                                                                                          | Explanation and sources                                                                                                        |
|------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------|
| <b>Requirements</b>                                                                                                                      |                                                                                                                                |
| <b>Question 1</b>                                                                                                                        |                                                                                                                                |
| <i>One language</i>                                                                                                                      | Speech output in one specified language, for example French. [182]                                                             |
| <i>Two languages</i>                                                                                                                     | Speech output in two specified languages, for example English and Chinese. [100]                                               |
| <i>Multiple languages</i>                                                                                                                | Speech output in several required languages, using suitable recordings, voice generation, or a combination of both. [114, 223] |
| <b>Question 2</b>                                                                                                                        |                                                                                                                                |
| <i>Conversational turn-taking</i>                                                                                                        | Back-and-forth speech, where the wait for each reply matters to the flow of conversation. [168]                                |
| <i>Single queries or commands</i>                                                                                                        | One request followed by a spoken answer or confirmation, with a waiting time appropriate to that task. [168]                   |
| Response delay is the wait between the end of a user's utterance or another input event and the start of the character's voice response. |                                                                                                                                |
| <b>Question 3</b>                                                                                                                        |                                                                                                                                |
| <i>Selecting an existing voice</i>                                                                                                       | Choosing from voices already available in the intended system. [142, 266]                                                      |
| <i>Adjusting pitch, speaking rate or level</i>                                                                                           | Control over how high or low the voice sounds, how fast it speaks, or the audio level. [110, 116, 186]                         |
| <i>Selecting expressive speaking styles</i>                                                                                              | Choosing a delivery style, such as an emotional speaking style. [94, 258]                                                      |
| <i>Controlling emphasis, rhythm and pitch contours</i>                                                                                   | Control over which words stand out, word and phrase timing, and how pitch changes through an utterance. [99, 183, 184]         |

Continued on next page

**Table A14: Technical Reflection: Examples and sources (continued)**

Questions: Table A13 on page 48.

| Example answers                           | Explanation and sources                                                                                               |
|-------------------------------------------|-----------------------------------------------------------------------------------------------------------------------|
| <i>Comparing voice candidates</i>         | Listening to alternatives, for example in pairs, to decide which better fits the intended voice concept. [264]        |
| <i>Distinction from a reference voice</i> | Checking how similar a candidate sounds to a reference voice, through listening or speaker-similarity measures. [244] |

**Resources****Question 4**

|                                                                                                                      |                                                                                                                 |
|----------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------|
| <i>Short reference recordings</i>                                                                                    | Brief speech samples used as a reference when creating a voice with similar speaker characteristics. [114, 174] |
| <i>New recording sessions</i>                                                                                        | Prepared lines or other vocal sounds recorded for the project. [115, 182]                                       |
| <i>Larger speech corpora</i>                                                                                         | Collections of recorded speech and accompanying transcripts or annotations for model training. [183]            |
| Availability, technical suitability and permission for the intended use are separate considerations [113, 150, 197]. |                                                                                                                 |

**Question 5**

Possible costs include recording sessions, audio production and licenses for the materials or tools used [220].

**Question 8**

|                                                          |                                                                                                                        |
|----------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------|
| <i>Acoustics and sociophonetics</i>                      | Understanding acoustic voice properties and how speech variation relates to social identities and contexts. [245, 259] |
| <i>Voice performance, recording and audio production</i> | Skills to perform, record, edit and prepare the voice audio. [220]                                                     |
| <i>Speech synthesis and software integration</i>         | Skills to generate speech with software and connect it to the character's system. [258, 264]                           |
| <i>Permissions and data use</i>                          | Expertise to establish what recordings and voice data may be used for and under which conditions. [197]                |
| <i>User research</i>                                     | Skills to plan and conduct studies with intended users and listeners. [266]                                            |

**Operating conditions****Question 9**

|                                                                                                 |                                                                                                       |
|-------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|
| <i>Remote speech generation</i>                                                                 | Access to a server or hosted service that can generate the character's speech. [115]                  |
| <i>Local GPU-based synthesis</i>                                                                | A local computer with a graphics processor (GPU) available for running speech-synthesis models. [169] |
| <i>CPU-based audio generation</i>                                                               | A computer's main processor (CPU) available to convert a speech model's output into audio. [182]      |
| <i>Embedded speech synthesis</i>                                                                | A small computer built into the character or device and available to run speech synthesis. [125, 258] |
| For prepared recordings, consider the computer or device used to store and play back the audio. |                                                                                                       |

Continued on next page

**Table A14: Technical Reflection: Examples and sources (continued)**

Questions: Table A13 on page 48.

| Example answers                                                                                                                                                                                                                                                                                                                     | Explanation and sources                                                                                                                                                |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Question 10</b>                                                                                                                                                                                                                                                                                                                  |                                                                                                                                                                        |
| <i>Remote processing</i>                                                                                                                                                                                                                                                                                                            | Speech or dialogue processing runs on a separate server and needs a connection to that server during use. [115, 190]                                                   |
| <i>Offline speech synthesis</i>                                                                                                                                                                                                                                                                                                     | Speech is generated without requiring a network connection during use. [258]                                                                                           |
| <i>Local data processing</i>                                                                                                                                                                                                                                                                                                        | Sensor data that could identify a person are processed on the device. [125]                                                                                            |
| <i>Operation:</i> when the service needs to be available, the provider it depends on, agreed service commitments, and how many users it must serve at once [101]. <i>AI governance:</i> who is responsible for the system, how its use is monitored, and how risks from external services or components are handled [98, 212, 247]. |                                                                                                                                                                        |
| <b>Production approach</b>                                                                                                                                                                                                                                                                                                          |                                                                                                                                                                        |
| <b>Question 11</b>                                                                                                                                                                                                                                                                                                                  |                                                                                                                                                                        |
| <i>Prepared recordings or synthesized audio</i>                                                                                                                                                                                                                                                                                     | Human voice recordings or synthesized audio are saved before use and played back during the interaction. [174, 205, 250]                                               |
| <i>Live speech, optionally voice-converted</i>                                                                                                                                                                                                                                                                                      | A person speaks through a remotely controlled character during the interaction, optionally with the sound of their voice transformed before playback. [118]            |
| <i>Speech synthesized during use</i>                                                                                                                                                                                                                                                                                                | The system turns text into spoken audio when it is needed during the interaction. [258]                                                                                |
| <i>Prepared audio combined with speech synthesis</i>                                                                                                                                                                                                                                                                                | Prepared audio clips and speech synthesized during the interaction are used together, for example recorded fillers or laughter alongside synthesized utterances. [168] |

**Table A15 Shared Reflection**

1. What do you want these choices to achieve, for whom and in which situations?
2. Could your own stereotypes or prejudices have influenced these choices? If so, which ones, and how might they have shaped your reasoning?
3. Could these choices create conflicting expectations or effects? If so, what could conflict, and how does it relate to your goals?
4. What other effects could these choices have, and what is that expectation based on?
5. What will you keep, change, or leave open? Why, and what will you test next?

Examples and sources: Table A16 on page 51.

**Table A16: Shared Reflection: Examples and sources**

Questions: Table A15 on page 51.

These examples illustrate possible reflections and decisions. Use the related studies to examine the expectations relevant to your own project.

| Example answer                                                       | Explanation and sources                                                                                                                                                                              |
|----------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Question 1</b><br><i>Credibility and learning</i>                 | A credible voice for an educational character, with learning as a separate project goal. Specify whose learning matters, in what situation, and how each goal would be assessed. [122]               |
| <b>Question 2</b><br><i>Gender cues and authority</i>                | A preference for a masculine-sounding voice because it is assumed to convey authority. An answer can examine whether that assumption reflects a stereotype and how it influenced the choice. [179]   |
| <b>Question 3</b><br><i>Face and voice realism</i>                   | Different degrees of humanlikeness in the character's face and voice. The answer can explain whether this contrast is intended, which expectations it might create and what should be checked. [200] |
| <b>Question 4</b><br><i>Convincing delivery of false information</i> | A concern that an authoritative-sounding voice could make unreliable information seem credible. An answer can identify when this might occur and what evidence would help assess the concern. [268]  |
| <b>Question 5</b><br><i>Familiar and unfamiliar listeners</i>        | A decision to retain a familiar voice provisionally and compare impressions among listeners who recognize it and those who do not. [228]                                                             |

**Table A17 Try Out and Revise**

|                                            |                                                                                                                                                                                                                                                                                                                |
|--------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Before a test</b>                       | Record the relevant assumptions, the voice material to be used, the intended users and listeners, and the use situation.                                                                                                                                                                                       |
| <b>Comparing candidate voices</b>          | Present the same content in each candidate voice to intended listeners in the planned setting, vary the order, and record impressions and preferences separately. Listeners who know the person whose voice is used as a reference can judge a recognizable voice differently from listeners who do not [228]. |
| <b>Checking intended impressions</b>       | Ask listeners to describe the voice or rate the intended qualities, including age, gender and social qualities, and compare their answers with the intention. Pilot tests and manipulation checks are recommended for social identity cues [233].                                                              |
| <b>Checking intelligibility and access</b> | Test comprehension in the real acoustic conditions with the intended users and listeners, including people who overhear the voice and people with hearing-related access needs [107].                                                                                                                          |
| <b>Questionnaires</b>                      | Established scales support comparisons across voices, for example the Godspeed questionnaire series for anthropomorphism, likeability and perceived intelligence [104].                                                                                                                                        |
| <b>Interpreting small trials</b>           | Treat results from a few listeners as preliminary indications for revision or further testing. Record who took part and under which conditions.                                                                                                                                                                |
| <b>When to return</b>                      | Findings may lead to revisions to the voice concept, Context or priorities. Evaluate the final implemented voice with its intended users and listeners in the real setting.                                                                                                                                    |

**Table A18 Project Record: Template and excerpt**

One entry per decision or intention, revised as the project develops. The excerpt shows an entry from the Rowan example in the article.

| <b>Entry</b>                     | <b>What to record</b>                                                                 | <b>Excerpt: Rowan's old-record effect</b>                                                                                                                                       |
|----------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Decision or intention</b>     | The voice choice, quality or effect under consideration, and the question it answers. | Subtle crackle and noise on the recorded stories, not on operating instructions (voice–appearance relationship, question 9 in Appearance and Embodiment; Specific Sound Ideas). |
| <b>Reason</b>                    | Why this choice serves the character, listeners and project goal.                     | Connects the radio body with old recordings; comprehension and control take priority over the effect.                                                                           |
| <b>Source or evidence</b>        | Literature examples, project observations or earlier tests supporting the reason.     | The team's aesthetic assumption; no listener evidence yet.                                                                                                                      |
| <b>Assumption and confidence</b> | Assumptions about listeners' responses and your confidence in them.                   | Visitors will associate the effect with historical recordings and still understand the stories; untested.                                                                       |
| <b>Open question</b>             | What remains unclear or depends on other decisions.                                   | Effect strength; whether the effect makes stories harder to understand in the exhibition.                                                                                       |
| <b>Planned test</b>              | Voice material, people, situation and the criterion for judging the result.           | The same performance with and without effects, played in the exhibition to intended visitors, who recount the content and describe what the sound reminds them of.              |
| <b>Result</b>                    | What was observed, with the version, people and situation involved.                   | Pending recordings and permissions.                                                                                                                                             |
| <b>Status and date</b>           | Kept, changed or left open, and when the entry was last revised.                      | Open; revisit after audition samples are available.                                                                                                                             |

## References for the Appendix

- [93] Abdulrahman A, Richards D (2022) Is Natural Necessary? Human Voice versus Synthetic Voice for Intelligent Virtual Agents. *Multimodal Technologies and Interaction* 6(7):51. <https://doi.org/10.3390/mti6070051>
- [94] Ada AA, Jørgensen SH, Fritsch J (2024) Cultures of the AI Paralinguistic in Voice Cloning Tools. In: Companion Publication of the 2024 ACM Designing Interactive Systems Conference. Association for Computing Machinery, pp 249–252, <https://doi.org/10.1145/3656156.3663708>
- [95] Alimardani M, de Roode R, Vaitonyte J, et al (2024) Effect of a Virtual Agent’s Appearance and Voice on Uncanny Valley and Trust in Human-Agent Collaboration. In: Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents. Association for Computing Machinery, pp 1–7, <https://doi.org/10.1145/3652988.3673970>
- [96] Almeda S, Netzorg R, Li I, et al (2025) Labor, Power, and Belonging: The Work of Voice in the Age of AI Reproduction. In: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency. Association for Computing Machinery, pp 1219–1230, <https://doi.org/10.1145/3715275.3732082>
- [97] Amiri GA, Shahidi S, Darmel FA, et al (2024) Beyond Binary Gender: Navigating Gender Ambiguity in Voice User Interfaces. *Cognizance Journal of Multidisciplinary Studies* 4(6):79–86. <https://doi.org/10.47760/cognizance.2024.v04i06.008>
- [98] Autio C, Schwartz R, Dunietz J, et al (2024) Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1, National Institute of Standards and Technology, Gaithersburg, MD, <https://doi.org/10.6028/NIST.AI.600-1>
- [99] Aylett MP, Cowan BR, Clark L (2019) Siri, Echo and Performance: You Have to Suffer Darling. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, pp 1–10, <https://doi.org/10.1145/3290607.3310422>
- [100] Azzuni H, El Saddik A (2025) Voice Cloning: Comprehensive Survey. arXiv preprint arXiv:2505.00579, <https://doi.org/10.48550/arXiv.2505.00579>
- [101] Badger L, Grance T, Patt-Corner R, et al (2012) Cloud Computing Synopsis and Recommendations. NIST Special Publication 800-146, National Institute of Standards and Technology, Gaithersburg, MD, <https://doi.org/10.6028/NIST.SP.800-146>
- [102] Barfield JK (2024) The Role of Name, Origin, and Voice Accent in a Robot’s Ethnic Identity. *Sensors* 24(19):6421. <https://doi.org/10.3390/s24196421>
- [103] Barrington S, Cooper EA, Farid H (2025) People are poorly equipped to detect AI-powered voice clones. *Scientific Reports* 15(1):11004. <https://doi.org/10.1038/s41598-025-94170-3>
- [104] Bartneck C (2008) The Godspeed Questionnaire Series. URL <https://www.bartneck.de/2008/03/11/the-godspeed-questionnaire-series/>, updated 27 January 2023. Accessed 11 September 2026.
- [105] Becker D, Braach L, Clasmeier L, et al (2025) Influence of Robots’ Voice Naturalness on Trust and Compliance. *ACM Transactions on Human-Robot Interaction* 14(2):29:1–29:25. <https://doi.org/10.1145/3706006>
- [106] Berkowitz AE, Sweeney ME (2025) Look Who’s Talking: Voice Cloning as Tension Point Between Identity and Data. *Philosophy & Technology* 38(4):131. <https://doi.org/10.1007/s13347-025-00965-7>
- [107] Blair J, Abdullah S (2020) “It Didn’t Sound Good with My Cochlear Implants”: Understanding the Challenges of Using Smart Assistants for Deaf and Hard of Hearing Users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4(4):118:1–118:27. <https://doi.org/10.1145/3432194>
- [108] Braun M, Mainz A, Chadowitz R, et al (2019) At Your Service: Designing Voice Assistant Personalities to Improve Automotive User Interfaces. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp 1–11, <https://doi.org/10.1145/3290605.3300270>
- [109] Broadbent E, Stafford R, MacDonald B (2009) Acceptance of Healthcare Robots for the Older Population: Review and Future Directions. *International Journal of Social Robotics* 1(4):319–330. <https://doi.org/10.1007/s12369-009-0030-6>
- [110] Byeon HJ, Lee C, Lee J, et al (2022) “A Voice that Suits the Situation”: Understanding the Needs and Challenges for Supporting End-User Voice Customization. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, <https://doi.org/10.1145/3491102.3501856>

- [111] Cabral JP, Cowan BR, Zibrek K, et al (2017) The Influence of Synthetic Voice on the Evaluation of a Virtual Character. In: Interspeech 2017. ISCA, pp 229–233, <https://doi.org/10.21437/Interspeech.2017-325>
- [112] Cambre J, Kulkarni C (2019) One Voice Fits All? Social Implications and Research Challenges of Designing Voices for Smart Devices. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW):1–19. <https://doi.org/10.1145/3359325>
- [113] Casanova E, Weber J, Shulby CD, et al (2022) YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for Everyone. In: *Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research*, vol 162. PMLR, pp 2709–2720, URL <https://proceedings.mlr.press/v162/casanova22a.html>
- [114] Casanova E, Davis K, Gölge E, et al (2024) XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. In: *Interspeech 2024*, pp 4978–4982, <https://doi.org/10.21437/Interspeech.2024-2016>
- [115] Chan SWT, Gunasekaran TS, Pai YS, et al (2021) KinVoices: Using Voices of Friends and Family in Voice Interfaces. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW2):1–25. <https://doi.org/10.1145/3479590>
- [116] Chang CJ, Sohn SS, Zhang S, et al (2023) The Importance of Multimodal Emotion Conditioning and Affect Consistency for Embodied Conversational Agents. In: *Proceedings of the 28th International Conference on Intelligent User Interfaces*. Association for Computing Machinery, pp 790–801, <https://doi.org/10.1145/3581641.3584045>
- [117] Chang RCS, Lu HP, Yang P (2018) Stereotypes or golden rules? Exploring likable voice traits of social robots as active aging companions for tech-savvy baby boomers in Taiwan. *Computers in Human Behavior* 84:194–210. <https://doi.org/10.1016/j.chb.2018.02.025>
- [118] Chang YC, Rea DJ, Kanda T (2024) Investigating the Impact of Gender Stereotypes in Authority on Avatar Robots. In: *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, New York, NY, USA, HRI '24, pp 106–115, <https://doi.org/10.1145/3610977.3634985>
- [119] Chappuis C, Grandjean D (2022) Set the tone: Trustworthy and dominant novel voices classification using explicit judgement and machine learning techniques. *PLOS ONE* 17(6):e0267432. <https://doi.org/10.1371/journal.pone.0267432>
- [120] Chen J, Zhang Q (2025) Voice Design and Trust in Automated Vehicles: Findings and a Research Agenda. In: *2025 IEEE 7th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*. IEEE, pp 688–696, <https://doi.org/10.1109/TPS-ISA67132.2025.00085>
- [121] Chérif E, Lemoine JF (2019) Anthropomorphic virtual assistants and the reactions of Internet users: An experiment on the assistant’s voice. *Recherche et Applications en Marketing (English Edition)* 34(1):28–47. <https://doi.org/10.1177/2051570719829432>
- [122] Chiou EK, Schroeder NL, Craig SD (2020) How we trust, perceive, and learn from virtual humans: The influence of voice quality. *Computers & Education* 146:103756. <https://doi.org/10.1016/j.compedu.2019.103756>
- [123] Chong T, Yu T, Keeling DI, et al (2026) The Voice in the Machine: Designing Embodied Conversational Agents for Rapport in Complex Online Service Delivery. *European Journal of Marketing* 60(5):1142–1175. <https://doi.org/10.1108/ejm-04-2024-0340>
- [124] Corey RM, Jones U, Singer AC (2020) Acoustic effects of medical, cloth, and transparent face masks on speech signals. *The Journal of the Acoustical Society of America* 148(4):2371–2375. <https://doi.org/10.1121/10.0002279>
- [125] Cosache G, Salgado F, Jain R, et al (2023) Data Center Audio/Video Intelligence on Device (DAVID) - An Edge-AI Platform for Smart-Toys. In: *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, pp 66–71, <https://doi.org/10.1109/SpeD59241.2023.10314915>
- [126] Craig SD, Schroeder NL (2017) Reconsidering the voice effect when learning from a virtual human. *Computers & Education* 114:193–205. <https://doi.org/10.1016/j.compedu.2017.07.003>
- [127] Creer S, Cunningham S, Green P, et al (2013) Building personalised synthetic voices for individuals with severe speech impairment. *Computer Speech & Language* 27(6):1178–1193. <https://doi.org/10.1016/j.csl.2012.10.001>

- [128] Dautenhahn K (2007) Socially Intelligent Robots: Dimensions of Human–Robot Interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences* 362(1480):679–704. <https://doi.org/10.1098/rstb.2006.2004>
- [129] Davis RO, Vincent J, Park T (2019) Reconsidering the Voice Principle with Non-Native Language Speakers. *Computers & Education* 140:103605. <https://doi.org/10.1016/j.compedu.2019.103605>
- [130] De Cet M, Cvajner M, Torre I, et al (2024) Do Your Expectations Match? A Mixed-Methods Study on the Association Between a Robot’s Voice and Appearance. In: *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. Association for Computing Machinery, pp 1–11, <https://doi.org/10.1145/3640794.3665551>
- [131] De Cet M, Obaid M, Torre I (2025) Breaking the Binary: A Systematic Review of Gender-Ambiguous Voices in Human-Computer Interaction. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, pp 1–17, <https://doi.org/10.1145/3706598.3713608>
- [132] De Cet M, Hope M, Torre I (2026) Queer, Nonbinary, or Ambiguous? Rethinking Voice Labels through Queer Theory in HCI. In: *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, pp 1–8, <https://doi.org/10.1145/3772318.3791246>
- [133] Dennler N, Kian M, Nikolaidis S, et al (2025) Designing Robot Identity: The Role of Voice, Clothing, and Task on Robot Gender Perception. *International Journal of Social Robotics* 17(4):707–728. <https://doi.org/10.1007/s12369-025-01209-6>
- [134] Désiron JC, Schneider S (2024) Listen Closely: Prosodic Signals in Podcast Support Learning. *Computers & Education* 216:105051. <https://doi.org/10.1016/j.compedu.2024.105051>
- [135] Di Donato B, Romeo M, Aylett M (2025) Sound design for HRI: a case study exploring appropriateness, naturalness, and sociality of a zoomorphic robot. In: *Proceedings of the 20th International Audio Mostly Conference*. ACM, pp 141–153, <https://doi.org/10.1145/3771594.3771608>
- [136] Dou X, Wu CF, Lin KC, et al (2021) Effects of Different Types of Social Robot Voices on Affective Evaluations in Different Application Fields. *International Journal of Social Robotics* 13(4):615–628. <https://doi.org/10.1007/s12369-020-00654-9>
- [137] Edwards C, Edwards A, Stoll B, et al (2019) Evaluations of an artificial intelligence instructor’s voice: Social Identity Theory in human-robot interactions. *Computers in Human Behavior* 90:357–362. <https://doi.org/10.1016/j.chb.2018.08.027>
- [138] Efthymiou F, Hildebrand C, de Bellis E, et al (2024) The Power of AI-Generated Voices: How Digital Vocal Tract Length Shapes Product Congruency and Ad Performance. *Journal of Interactive Marketing* 59(2):117–134. <https://doi.org/10.1177/10949968231194905>
- [139] Ehret J, Bönsch A, Aspöck L, et al (2021) Do Prosody and Embodiment Influence the Perceived Naturalness of Conversational Agents’ Speech? *ACM Transactions on Applied Perception* 18(4):1–15. <https://doi.org/10.1145/3486580>
- [140] Fagel S (2007) Auditory-Visual Integration in the Perception of Age in Speech. In: *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, pp 725–728, URL <https://www.icphs2007.de/conference/Papers/1001/1001.pdf>
- [141] Fatima SM, Shehzad M, Murtuza SS, et al (2020) Neural Style Transfer Based Voice Mimicking for Personalized Audio Stories. In: *Proceedings of the 2nd International Workshop on AI for Smart TV Content Production, Access and Delivery*. ACM, pp 11–16, <https://doi.org/10.1145/3422839.3423063>
- [142] Ferstl Y, Thomas S, Guiard C, et al (2021) Human or Robot?: Investigating Voice, Appearance and Gesture Motion Realism of Conversational Social Agents. In: *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*. ACM, New York, NY, USA, pp 76–83, <https://doi.org/10.1145/3472306.3478338>
- [143] Fischer K, Niebuhr O (2023) Which Voice for which Robot? Designing Robot Voices that Indicate Robot Size. *ACM Transactions on Human-Robot Interaction* 12(4):1–24. <https://doi.org/10.1145/3632124>
- [144] Fischer K, Niebuhr O, Langedijk RM, et al (2019) I Shall Know You by Your Voice: Melodic and Physical Dominance in the Design of Robot Voices. In: *Proceedings 1st International Seminar on the Foundations of Speech: Breathing, Pausing, and the Voice*. Syddansk Universitet, pp 88–90, URL <https://portal.findresearcher.sdu.dk/en/publications/i-shall-know-you-by-your-voice-melodic-and-physical-dominance-in-/>

- [145] Fischer K, Niebuhr O, Asadi A (2022) The Voice of Creativity: Effects of Pitch Range in the Voice of a Robot Facilitator. In: *Studientexte zur Sprachkommunikation Band 103: Elektronische Sprachsignalverarbeitung 2022*, vol 103. TUDPress, pp 121–130, URL [https://www.essv.de/pdf/2022\\_121\\_130.pdf](https://www.essv.de/pdf/2022_121_130.pdf)
- [146] Fiske ST (2018) Stereotype Content: Warmth and Competence Endure. *Current Directions in Psychological Science* 27(2):67–73. <https://doi.org/10.1177/0963721417738825>
- [147] Fleisig E, Vargo J, Schwarz-Acosta NA, et al (2026) Identity and personality in the social perception of synthesized voices: perceptions of OpenAI’s text-to-speech technology. *Phonetica* <https://doi.org/10.1515/phon-2025-0065>
- [148] Gao S, Aylett MP, Braude DA, et al (2023) Synthesising Personality with Neural Speech Synthesis. In: *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czechia, pp 393–399, <https://doi.org/10.18653/v1/2023.sigdial-1.36>
- [149] Gao Y, Guo Z, Liang HN (2026) Is Seeing Believing and Hearing Trusting? Exploring the Role of Visual and Vocal Cues in Shaping User Trust in VR News Anchors. *IEEE Transactions on Visualization and Computer Graphics* 32(5):4021–4031. <https://doi.org/10.1109/tvcg.2026.3680672>
- [150] Gebbru T, Morgenstern J, Vecchione B, et al (2021) Datasheets for Datasets. *Communications of the ACM* 64(12):86–92. <https://doi.org/10.1145/3458723>
- [151] de Gelder B, Vroomen J (2000) The perception of emotions by ear and by eye. *Cognition and Emotion* 14(3):289–311. <https://doi.org/10.1080/026999300378824>
- [152] Gong L, Nass C (2007) When a Talking-Face Computer Agent is Half-Human and Half-Humanoid: Human Identity and Consistency Preference. *Human Communication Research* 33(2):163–193. <https://doi.org/10.1111/j.1468-2958.2007.00295.x>
- [153] Gülen SK, Çöğgeven NS, Büyükköse GT, et al (2026) Exploring the Experiences of Online and Distance Learners with Visual Impairments in Regard to AI-Assisted Narration. *Open Praxis* 18(2):365–380. <https://doi.org/10.55982/openpraxis.18.2.802>
- [154] Gurung N, Busby Grant J, Herath D (2024) The Uncanny Effect of Speech: The Impact of Appearance and Speaking on Impression Formation in Human–Robot Interactions. *International Journal of Social Robotics* 16(6):1265–1280. <https://doi.org/10.1007/s12369-023-00976-4>
- [155] Haresamudram K, Torre I, Behling M, et al (2024) Talking body: the effect of body and voice anthropomorphism on perception of social agents. *Frontiers in Robotics and AI* 11:1456613. <https://doi.org/10.3389/frobt.2024.1456613>
- [156] Hastie H, Lohan K, Deshmukh A, et al (2017) The Interaction Between Voice and Appearance in the Embodiment of a Robot Tutor. In: *Social Robotics, Lecture Notes in Computer Science*, vol 10652. Springer, Cham, pp 64–74, [https://doi.org/10.1007/978-3-319-70022-9\\_7](https://doi.org/10.1007/978-3-319-70022-9_7)
- [157] Higgins D, Zibrek K, Cabral J, et al (2022) Sympathy for the Digital: Influence of Synthetic Voice on Affinity, Social Presence and Empathy for Photorealistic Virtual Humans. *Computers & Graphics* 104:116–128. <https://doi.org/10.1016/j.cag.2022.03.009>
- [158] Holmes E, Johnsrude IS (2020) Speech Spoken by Familiar People Is More Resistant to Interference by Linguistically Similar Speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 46(8):1465–1476. <https://doi.org/10.1037/xlm0000823>
- [159] Huang G, Moore RK (2024) Freedom Comes at a Cost?: An Exploratory Study on Affordances’ Impact on Users’ Perception of a Social Robot. *Frontiers in Robotics and AI* 11:1288818. <https://doi.org/10.3389/frobt.2024.1288818>
- [160] Im H, Sung B, Lee G, et al (2023) Let voice assistants sound like a machine: Voice and task type effects on perceived fluency, competence, and consumer attitude. *Computers in Human Behavior* 145:107791. <https://doi.org/10.1016/j.chb.2023.107791>
- [161] Jaggy O, Schwan S, Meyerhoff HS (2025) AI-determined similarity increases likability and trustworthiness of human voices. *PLOS ONE* 20(3):e0318890. <https://doi.org/10.1371/journal.pone.0318890>
- [162] Jing B, Wu C, Pi Z, et al (2025) Examining the Voice-Image Matching for Pedagogical Agents Presented in Instructional Videos. *The Journal of Experimental Education* pp 1–21. <https://doi.org/10.1080/00220973.2025.2565179>
- [163] Jones A, Zellou G (2024) Voice accentedness, but not gender, affects social responses to a computer tutor. *Frontiers in Computer Science* 6:1436341. <https://doi.org/10.3389/fcomp.2024.1436341>

- [164] Kalegina A, Schroeder G, Allchin A, et al (2018) Characterizing the Design Space of Rendered Robot Faces. In: Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction. ACM, pp 96–104, <https://doi.org/10.1145/3171221.3171286>
- [165] Kannampallil T, Smyth JM, Jones S, et al (2020) Cognitive plausibility in voice-based AI health counselors. *npj Digital Medicine* 3(1):72. <https://doi.org/10.1038/s41746-020-0278-7>
- [166] Kannampallil T, Ronneberg CR, Wittels NE, et al (2022) Design and Formative Evaluation of a Virtual Voice-Based Coach for Problem-solving Treatment: Observational Study. *JMIR Formative Research* 6(8):e38092. <https://doi.org/10.2196/38092>
- [167] Kawahara M, Sawada Y, Tanaka A (2021) Multisensory Emotion Perception to Humanoid Robot. *Transactions of Japan Society of Kansei Engineering* 20(3):329–335. <https://doi.org/10.5057/jjske.TJSKE-D-21-00015>
- [168] Kawahara T, Saruwatari H, Higashinaka R, et al (2025) Spoken Dialogue Technology for Semi-Autonomous Cybernetic Avatars. In: Ishiguro H, Ueno F, Tachibana E (eds) *Cybernetic Avatar*. Springer, Singapore, p 71–105, [https://doi.org/10.1007/978-981-97-3752-9\\_3](https://doi.org/10.1007/978-981-97-3752-9_3)
- [169] Kim J, Kong J, Son J (2021) Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. In: Meila M, Zhang T (eds) *Proceedings of the 38th International Conference on Machine Learning, Proceedings of Machine Learning Research*, vol 139. PMLR, pp 5530–5540, URL <https://proceedings.mlr.press/v139/kim21f.html>
- [170] Kim Y, Reza M, McGrenere J, et al (2021) Designers Characterize Naturalness in Voice User Interfaces: Their Goals, Practices, and Challenges. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp 1–13, <https://doi.org/10.1145/3411764.3445579>
- [171] Ko S, Barnes J, Dong J, et al (2023) The Effects of Robot Voices and Appearances on Users’ Emotion Recognition and Subjective Perception. *International Journal of Humanoid Robotics* 20(01):2350001. <https://doi.org/10.1142/s0219843623500019>
- [172] Koike A, Okafuji Y, Song S (2026) Practical Insights into Designing Context-Aware Robot Voice Parameters in the Wild. In: *Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction*. ACM, pp 1140–1149, <https://doi.org/10.1145/3757279.3785546>
- [173] Kuch JM, Nasir J, Mertes S, et al (2024) Evaluating Gender Ambiguity, Novelty and Anthropomorphism in Humming and Talking Voices for Robots. In: *2024 33rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp 2219–2225, <https://doi.org/10.1109/ro-man60168.2024.10731423>
- [174] Kuch JM, Heisler M, Klein S, et al (2025) Your Robot, My Voice: Enhancing Android Robot Likability through Personalization by Cloning the User’s Voice. In: *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp 192–198, <https://doi.org/10.1109/ro-man63969.2025.11217611>
- [175] Kuch JM, Withanage Don DS, Heimerl N, et al (2026) What Happens When a Virtual Character Speaks With the Cloned Voice of a Person You Know? In: *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*. ACM, pp 1–5, <https://doi.org/10.1145/3772363.3799332>
- [176] Kühne K, Fischer MH, Zhou Y (2020) The Human Takes It All: Humanlike Synthesized Voices Are Perceived as Less Eerie and More Likable. Evidence From a Subjective Ratings Study. *Frontiers in Neurorobotics* 14:593732. <https://doi.org/10.3389/fnbot.2020.593732>
- [177] Kunimi Y, Kimura K, Matsumoto K, et al (2024) Character-Voice Embodiment Impacts on the Cognitive Task Performance with the Voice Ownership Illusion. In: Hasegawa S, Sakata N, Sundstedt V (eds) *ICAT-EGVE 2024 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*. The Eurographics Association, <https://doi.org/10.2312/egve.20241358>
- [178] Lavan N (2025) Time after time: Voice perception from first impressions to identity recognition. *Quarterly Journal of Experimental Psychology* 78(12):2583–2593. <https://doi.org/10.1177/17470218251379036>
- [179] Lee EJ, Nass C, Brave S (2000) Can computer-generated speech have gender? In: *CHI ’00 Extended Abstracts on Human Factors in Computing Systems*. ACM, pp 289–290, <https://doi.org/10.1145/633292.633461>
- [180] Lee SC, Sanghavi H, Ko S, et al (2019) Autonomous driving with an agent. In: *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct*

- Proceedings. ACM, pp 209–214, <https://doi.org/10.1145/3349263.3351515>
- [181] Leite I, Martinho C, Paiva A (2013) Social Robots for Long-Term Interaction: A Survey. *International Journal of Social Robotics* 5(2):291–308. <https://doi.org/10.1007/s12369-013-0178-y>
  - [182] Lenglet M (2023) Analysis of Latent Representations of Neural Text-To-Speech Models for Expressive Audio-Visual Synthesis. PhD thesis, Université Grenoble Alpes, URL <https://theses.hal.science/tel-04541736>
  - [183] Lenglet M, Perrotin O, Bailly G (2022) Speaking Rate Control of end-to-end TTS Models by Direct Manipulation of the Encoder’s Output Embeddings. In: *Interspeech 2022*, pp 11–15, <https://doi.org/10.21437/Interspeech.2022-759>
  - [184] Lenglet M, Perrotin O, Bailly G (2026) A Closer Look at Internal Representations of End-to-End Text-to-Speech Models: How Is Phonetic and Acoustic Information Encoded? *Computer Speech & Language* 100:101985. <https://doi.org/10.1016/j.csl.2026.101985>
  - [185] Leuenberger M (2025) The Role of the Voice for Identity and Implications for Voice Cloning Technology. *Philosophy & Technology* 38(4):175. <https://doi.org/10.1007/s13347-025-00969-3>
  - [186] Li J, Zhang L (2023) ZSE-VITS: A Zero-Shot Expressive Voice Cloning Method Based on VITS. *Electronics* 12(4):820. <https://doi.org/10.3390/electronics12040820>
  - [187] Li M, Guo F, Fang C, et al (2023) Multisensory integration effect of humanoid robot appearance and voice on users’ affective preference and visual attention. *Behaviour & Information Technology* 42(14):2387–2406. <https://doi.org/10.1080/0144929X.2022.2125830>
  - [188] Li M, Zhang J, Guo F, et al (2025) Audiovisual Affective Design of Humanoid Robot Appearance and Voice Based on Kansei Engineering. *International Journal of Social Robotics* 17(1):15–37. <https://doi.org/10.1007/s12369-024-01202-5>
  - [189] Li S, Liu Z, Qiu M, et al (2024) Examining the Effects of Communication Features of Educational Robots on Students’ Cognitive Load, Attitudes, and Learning Performance. *Journal of Educational Computing Research* 62(4):764–789. <https://doi.org/10.1177/07356331231226422>
  - [190] Lim S, Schmäzle R, Bente G (2025) Artificial social influence via human-embodied AI agent interaction in immersive virtual reality (VR): Effects of similarity-matching during health conversations. *Computers in Human Behavior: Artificial Humans* 5:100172. <https://doi.org/10.1016/j.chbah.2025.100172>
  - [191] Liu X, Dong J, Jeon M (2023) Robots’ “Woohoo” and “Argh” Can Enhance Users’ Emotional and Social Perceptions: An Exploratory Study on Non-lexical Vocalizations and Non-linguistic Sounds. *ACM Transactions on Human-Robot Interaction* 12(4):1–20. <https://doi.org/10.1145/3626185>
  - [192] Lubold N, Walker E, Pon-Barry H (2016) Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion. In: *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pp 255–262, <https://doi.org/10.1109/hri.2016.7451760>
  - [193] Luján Rubio M, Rojo López AM, Ramos Caro M, et al (2026) Listening to Dance: Gendered Voices and the Emotional Experience of Poetic Audio Description for Audiences with Visual Impairments. *Disabilities* 6(1):21. <https://doi.org/10.3390/disabilities6010021>
  - [194] Mandai Y, Seaborn K, Nakano T, et al (2025) Super Kawaii Vocalics: Amplifying the "Cute" Factor in Computer Voice. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, pp 1–19, <https://doi.org/10.1145/3706598.3713709>
  - [195] Manoli A, Pauketat JV, Anthis JR (2025) Characterizing Relationships with Companion and Assistant Large Language Models. In: *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing*. ACM, pp 312–319, <https://doi.org/10.1145/3715070.3749245>
  - [196] Mara M, Schreibelmayr S, Berger F (2020) Hearing a Nose: User Expectations of Robot Appearance Induced by Different Robot Voices. In: *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, pp 355–356, <https://doi.org/10.1145/3371382.3378285>
  - [197] McGettigan C, Bloch S, Bowles C, et al (2025) Voice Conversion and Cloning: Psychological and Ethical Implications of Intentionally Synthesising Familiar Voice Identities. *Journal of the British Academy* 13(3):a31. <https://doi.org/10.5871/jba/013.a31>
  - [198] McGinn C, Torre I (2019) Can you Tell the Robot by the Voice? An Exploratory Study on the Role of Voice in the Perception of Robots. In: *2019 14th ACM/IEEE International Conference on Human-Robot*

- Interaction (HRI). IEEE, pp 211–221, <https://doi.org/10.1109/HRI.2019.8673305>
- [199] Miniota J, Wang S, Beskow J, et al (2023) Hi Robot, It’s Not What You Say, It’s How You Say It. In: 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pp 307–314, <https://doi.org/10.1109/RO-MAN57019.2023.10309427>
  - [200] Mitchell WJ, Szerszen KA, Lu AS, et al (2011) A Mismatch in the Human Realism of Face and Voice Produces an Uncanny Valley. *i-Perception* 2(1):10–12. <https://doi.org/10.1068/i0415>
  - [201] Moore RK (2017) Appropriate Voices for Artefacts: Some Key Insights. In: 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2017), URL [https://vihar-2017.vihar.org/assets/papers/VIHAR-2017\\_paper\\_8.pdf](https://vihar-2017.vihar.org/assets/papers/VIHAR-2017_paper_8.pdf)
  - [202] Moore RK (2017) Is Spoken Language All-or-Nothing? Implications for Future Speech-Based Human-Machine Interaction. In: *Dialogues with Social Robots. Lecture Notes in Electrical Engineering*, Springer Singapore, p 281–291, [https://doi.org/10.1007/978-981-10-2585-3\\_22](https://doi.org/10.1007/978-981-10-2585-3_22)
  - [203] Moore RK, Mitchinson B (2017) A Biomimetic Vocalisation System for MiRo. In: *Biomimetic and Biohybrid Systems. Lecture Notes in Computer Science*, Springer International Publishing, p 363–374, [https://doi.org/10.1007/978-3-319-63537-8\\_30](https://doi.org/10.1007/978-3-319-63537-8_30)
  - [204] Moore RK, Marxer R, Thill S (2016) Vocal Interactivity in-and-between Humans, Animals, and Robots. *Frontiers in Robotics and AI* 3:61. <https://doi.org/10.3389/frobt.2016.00061>
  - [205] Mooshammer S, Etzrodt K (2022) Social Research with Gender-Neutral Voices in Chatbots – The Generation and Evaluation of Artificial Gender-Neutral Voices with Praat and Google WaveNet. In: *Chatbot Research and Design. Lecture Notes in Computer Science*, Springer International Publishing, p 176–191, [https://doi.org/10.1007/978-3-030-94890-0\\_11](https://doi.org/10.1007/978-3-030-94890-0_11)
  - [206] Mooshammer S, Etzrodt K, Weidmüller L (2025) Trust in gendered voice assistants—the special case of gender ambiguity. *Publizistik* 70(4):581–611. <https://doi.org/10.1007/s11616-025-00907-5>
  - [207] Nakata Y, Yagi S, Yu S, et al (2022) Development of ‘ibuki’ an electrically actuated childlike android with mobility and its potential in the future society. *Robotica* 40:933–950. <https://doi.org/10.1017/S0263574721000898>
  - [208] Nass C, Brave S (2005) *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. The MIT Press, URL <https://mitpress.mit.edu/9780262140928/wired-for-speech/>
  - [209] Nass C, Lee KM (2000) Does Computer-Generated Speech Manifest Personality? An Experimental Test of Similarity-Attraction. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Association for Computing Machinery*, pp 329–336, <https://doi.org/10.1145/332040.332452>
  - [210] Nass C, Lee KM (2001) Does Computer-Synthesized Speech Manifest Personality? Experimental Tests of Recognition, Similarity-Attraction, and Consistency-Attraction. *Journal of Experimental Psychology: Applied* 7(3):171–181. <https://doi.org/10.1037/1076-898x.7.3.171>
  - [211] Nathwani K, Richard G, David B, et al (2017) Speech intelligibility improvement in car noise environment by voice transformation. *Speech Communication* 91:17–27. <https://doi.org/10.1016/j.specom.2017.04.007>
  - [212] National Institute of Standards and Technology (n.d.) NIST AI RMF Playbook. AI Resource Center, URL <https://airc.nist.gov/airmf-resources/playbook/>, accessed 11 September 2026.
  - [213] Niculescu A, van Dijk B, Nijholt A, et al (2013) Making Social Robots More Attractive: The Effects of Voice Pitch, Humor and Empathy. *International Journal of Social Robotics* 5(2):171–191. <https://doi.org/10.1007/s12369-012-0171-x>
  - [214] Niu J, She W, Cheng W, et al (2026) Character style and relational judgments in human–AI romance: trust, commitment, intimacy, and passion. *Frontiers in Psychology* 17:1819889. <https://doi.org/10.3389/fpsyg.2026.1819889>
  - [215] Noufi C, May L, Berger J (2023) The Role of Vocal Persona in Natural and Synthesized Speech. In: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG). IEEE, pp 1–4, <https://doi.org/10.1109/FG57933.2023.10042714>
  - [216] Otsuka T, Nakadai K, Takahashi T, et al (2010) Voice-awareness control for a humanoid robot consistent with its body posture and movements. *Paladyn, Journal of Behavioral Robotics* 1(1):80–88. <https://doi.org/10.2478/s13230-010-0009-x>
  - [217] Owens K, Gunawan J, Choffnes D, et al (2022) Exploring Deceptive Design Patterns in Voice Interfaces. In: 2022 European Symposium on Usable Security (EuroUSEC 2022). Association for Computing Machinery,

- pp 64–78, <https://doi.org/10.1145/3549015.3554213>
- [218] Pantoja LS, Diederich K, Crawford L, et al (2019) Voice Agents Supporting High-Quality Social Play. In: *Proceedings of the 18th ACM International Conference on Interaction Design and Children*. ACM, pp 314–325, <https://doi.org/10.1145/3311927.3323151>
  - [219] Parviainen E, Søndergaard MLJ (2020) Experiential Qualities of Whispering with Voice Assistants. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, pp 1–13, <https://doi.org/10.1145/3313831.3376187>
  - [220] Pearl C (2016) *Designing Voice User Interfaces: Principles of Conversational Experiences*. O’Reilly Media, URL <https://www.oreilly.com/library/view/designing-voice-user/9781491955406/>
  - [221] Porcheron M, Fischer JE, Reeves S, et al (2018) Voice Interfaces in Everyday Life. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, pp 1–12, <https://doi.org/10.1145/3173574.3174214>
  - [222] Pradhan A, Erete S, Chopra S, et al (2025) “No, not that voice again!”: Engaging Older Adults in Design of Anthropomorphic Voice Assistants. *Proceedings of the ACM on Human-Computer Interaction* 9(2):1–30. <https://doi.org/10.1145/3711039>
  - [223] Pratap V, Tjandra A, Shi B, et al (2024) Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research* 25(97):1–52. URL <https://jmlr.org/papers/v25/23-1318.html>
  - [224] Preece J, Sullivan E, Tams-Gray F, et al (2024) Making my voice and owning its future. *Medical Humanities* 50(4):624–634. <https://doi.org/10.1136/medhum-2024-013021>
  - [225] van Rijn P, Mertes S, Janowski K, et al (2024) Giving Robots a Voice: Human-in-the-Loop Voice Creation and open-ended Labeling. In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, pp 1–34, <https://doi.org/10.1145/3613904.3642038>
  - [226] Rincón C, Keyes O, Cath C (2021) Speaking from Experience: Trans/Non-Binary Requirements for Voice-Activated AI. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW1):1–27. <https://doi.org/10.1145/3449206>
  - [227] Robinson F (2023) *Designing Sound for Social Robots: Advancing Professional Practice through Design Principles*. PhD thesis, UNSW Sydney, <https://doi.org/10.26190/unsworks/25231>
  - [228] Rosi V, Soopramanien E, McGettigan C (2025) Perception and Social Evaluation of Cloned and Recorded Voices: Effects of Familiarity and Self-Relevance. *Computers in Human Behavior: Artificial Humans* 4:100143. <https://doi.org/10.1016/j.chbah.2025.100143>
  - [229] Sarigul B, Saltik I, Hokelek B, et al (2020) Does the Appearance of an Agent Affect How We Perceive his/her Voice? In: *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, pp 430–432, <https://doi.org/10.1145/3371382.3378302>
  - [230] Sarigul B, Schneider FM, Utz S (2025) Believe It or Not? Investigating the Credibility of Voice Assistants in the Context of Social Roles and Relationship Types. *International Journal of Human-Computer Interaction* 41(10):6253–6265. <https://doi.org/10.1080/10447318.2024.2375797>
  - [231] Schober MF, Clark HH (1989) Understanding by Addressees and Overhearers. *Cognitive Psychology* 21(2):211–232. [https://doi.org/10.1016/0010-0285\(89\)90008-X](https://doi.org/10.1016/0010-0285(89)90008-X)
  - [232] Schreibelmayer S, Mara M (2022) Robot Voices in Daily Life: Vocal Human-Likeness and Application Context as Determinants of User Acceptance. *Frontiers in Psychology* 13:787499. <https://doi.org/10.3389/fpsyg.2022.787499>
  - [233] Seaborn K (2025) *Social Identity in Human-Agent Interaction: A Primer*. *ACM Transactions on Human-Robot Interaction* <https://doi.org/10.1145/3760500>
  - [234] Seaborn K, Urakami J (2021) Measuring Voice UX Quantitatively: A Rapid Review. In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, pp 1–8, <https://doi.org/10.1145/3411763.3451712>
  - [235] Seaborn K, Miyake NP, Pennefather P, et al (2021) Voice in Human-Agent Interaction: A Survey. *ACM Computing Surveys* 54(4):1–43. <https://doi.org/10.1145/3386867>
  - [236] Seaborn K, Sekiguchi T, Tokunaga S, et al (2023) Voice Over Body? Older Adults’ Reactions to Robot and Voice Assistant Facilitators of Group Conversation. *International Journal of Social Robotics* 15(2):143–163. <https://doi.org/10.1007/s12369-022-00925-7>

- [237] Shi Z, Chen H, Velentza AM, et al (2023) Evaluating and Personalizing User-Perceived Quality of Text-to-Speech Voices for Delivering Mindfulness Meditation with Different Physical Embodiments. In: Castellano G, Riek L, Cakmak M, et al (eds) *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, New York, NY, USA, pp 516–524, <https://doi.org/10.1145/3568162.3576987>
- [238] Shimada M, Kanda T (2012) What is the appropriate speech rate for a communication robot? *Interaction Studies* 13(3):408–435. <https://doi.org/10.1075/is.13.3.05shi>
- [239] Shiramizu VKM, Lee AJ, Altenburg D, et al (2022) The role of valence, dominance, and pitch in perceptions of artificial intelligence (AI) conversational agents’ voices. *Scientific Reports* 12(1):22479. <https://doi.org/10.1038/s41598-022-27124-8>
- [240] Siegle RF, Craig SD (2024) The voice quality of pedagogical agent impacts learning and agent perceptions. *Journal of Computer Assisted Learning* 40(5):2278–2291. <https://doi.org/10.1111/jcal.13027>
- [241] Skoog Waller S, Eriksson M, Sörqvist P (2015) Can you hear my age? Influences of speech rate and speech spontaneity on estimation of speaker age. *Frontiers in Psychology* 6:978. <https://doi.org/10.3389/fpsyg.2015.00978>
- [242] Skuk VG, Jacob I, Wientzek R, et al (2025) Big Five Personality Perceptions from Voices and Faces: Impressions and Kernels of Truth. *Journal of Nonverbal Behavior* 49:125–154. <https://doi.org/10.1007/s10919-025-00480-5>
- [243] Snyder EC, Mendu S, Sundar SS, et al (2023) Busting the one-voice-fits-all myth: Effects of similarity and customization of voice-assistant personality. *International Journal of Human-Computer Studies* 180:103126. <https://doi.org/10.1016/j.ijhcs.2023.103126>
- [244] Stanton D, Shannon M, Mariooryad S, et al (2022) Speaker Generation. In: *ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, pp 7897–7901, <https://doi.org/10.1109/ICASSP43922.2022.9747345>
- [245] Sutton SJ, Foulkes P, Kirk D, et al (2019) Voice as a Design Material: Sociophonetic Inspired Design Strategies in Human-Computer Interaction. In: Brewster S (ed) *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, ACM Digital Library, pp 1–14, <https://doi.org/10.1145/3290605.3300833>
- [246] Székely É, Gustafson J, Torre I (2023) Prosody-controllable Gender-ambiguous Speech Synthesis: A Tool for Investigating Implicit Bias in Speech Perception. In: *Proceedings of Interspeech 2023*. International Speech Communication Association, pp 1234–1238, <https://doi.org/10.21437/Interspeech.2023-2086>
- [247] Tabassi E (2023) Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD, <https://doi.org/10.6028/NIST.AI.100-1>
- [248] Takahashi H, Maeshiro T (2023) Effects of Visual and Personality Impressions on the Voices Matched to Animated Characters. In: *Human Interface and the Management of Information. Lecture Notes in Computer Science*, Springer Nature Switzerland, p 431–444, [https://doi.org/10.1007/978-3-031-35132-7\\_33](https://doi.org/10.1007/978-3-031-35132-7_33)
- [249] Tielman M, Neerinx M, Meyer JJ, et al (2014) Adaptive emotional expression in robot-child interaction. In: *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, pp 407–414, <https://doi.org/10.1145/2559636.2559663>
- [250] Tolmeijer S, Zierau N, Janson A, et al (2021) Female by Default? – Exploring the Effect of Voice Assistant Gender and Pitch on Trait and Trust Attribution. In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, pp 1–7, <https://doi.org/10.1145/3411763.3451623>
- [251] Torre I, Goslin J, White L, et al (2018) Trust in Artificial Voices: A “Congruency Effect” of First Impressions and Behavioural Experience. In: *Proceedings of the Technology, Mind, and Society*. ACM, New York, NY, USA, pp 1–6, <https://doi.org/10.1145/3183654.3183691>
- [252] Torre I, Holk S, Carrigan E, et al (2021) Dimensional perception of a ‘smiling McGurk effect’. In: *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, pp 1–8, <https://doi.org/10.1109/acii52823.2021.9597423>
- [253] Torre I, Lagerstedt E, Dennler N, et al (2023) Can a gender-ambiguous voice reduce gender stereotypes in human-robot interactions? In: *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp 106–112, <https://doi.org/10.1109/ro-man57019.2023.10309500>

- [254] Trovato G, Ramos JG, Azevedo H, et al (2017) A receptionist robot for Brazilian people: study on interaction involving illiterates. *Paladyn, Journal of Behavioral Robotics* 8(1):1–17. <https://doi.org/10.1515/pjbr-2017-0001>
- [255] Tschopp M, Gieselmann M, Sassenberg K (2023) Servant by Default? How Humans Perceive Their Relationship with Conversational AI. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 17(3):Article 9. <https://doi.org/10.5817/CP2023-3-9>
- [256] Tsiourti C, Weiss A, Wac K, et al (2019) Multimodal Integration of Emotional Signals from Voice, Body, and Context: Effects of (In)Congruence on Emotion Recognition and Attitudes Towards Robots. *International Journal of Social Robotics* 11(4):555–573. <https://doi.org/10.1007/s12369-019-00524-z>
- [257] Tuttösi P (2025) I Know You’re Listening: Adaptive Voice for HRI. In: 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, pp 1897–1899, <https://doi.org/10.1109/HRI61500.2025.10974224>
- [258] Tuttösi P, Mehta S, Syvenky Z, et al (2025) EmojiVoice: Towards long-term controllable expressivity in robot speech. In: 2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). IEEE, pp 657–664, <https://doi.org/10.1109/RO-MAN63969.2025.11217645>
- [259] Tymburiba Elian M, Masuko S, Yamanaka T (2022) Who are you talking to? Considerations on Designing Gender Ambiguous Voice User Interfaces. In: 9th International Conference on Kansei Engineering and Emotion Research. KEER2022. Proceedings, pp 711–720, <https://doi.org/10.5821/conference-9788419184849.74>
- [260] Walters ML, Syrdal DS, Koay KL, et al (2008) Human approach distances to a mechanical-looking robot with different robot voice styles. In: RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication. IEEE, pp 707–712, <https://doi.org/10.1109/ROMAN.2008.4600750>
- [261] Wen S, Ping S, Wang J, et al (2024) AdaptiveVoice: Cognitively Adaptive Voice Interface for Driving Assistance. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. ACM, pp 1–18, <https://doi.org/10.1145/3613904.3642876>
- [262] Wenzel K, Kaufman G (2024) Designing for Harm Reduction: Communication Repair for Multicultural Users’ Voice Interactions. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp 1–17, <https://doi.org/10.1145/3613904.3642900>
- [263] Wilson S, Moore RK (2017) Robot, Alien and Cartoon Voices: Implications for Speech-Enabled Systems. In: 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2017), URL [https://vihar-2017.vihar.org/assets/papers/VIHAR-2017\\_paper\\_1.pdf](https://vihar-2017.vihar.org/assets/papers/VIHAR-2017_paper_1.pdf)
- [264] Withanage D, Lingenfelser F, Kuch J, et al (2025) VoiceX as a Design Tool for Virtual Agents’ Voices. In: Proceedings of the 25th ACM International Conference on Intelligent Virtual Agents. Association for Computing Machinery, pp 1–4, <https://doi.org/10.1145/3717511.3749296>
- [265] Yanushevskaya I, Gobl C, Ní Chasaide A (2013) Voice quality in affect cueing: does loudness matter? *Frontiers in Psychology* 4:335. <https://doi.org/10.3389/fpsyg.2013.00335>
- [266] Zhang L, Jiang L, Washington N, et al (2021) Social Media through Voice: Synthesized Voice Qualities and Self-presentation. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW1):1–21. <https://doi.org/10.1145/3449235>
- [267] Zhang Q, Yang XJ, Robert LP (2025) Voice Similarity and its Impact on Cognitive and Affective Trust in Automated Vehicles. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 69(1):1040–1045. <https://doi.org/10.1177/10711813251364804>
- [268] Zhang S, Chen H, Wang Y, et al (2025) The Manipulative Power of Voice Characteristics: Investigating Deceptive Patterns in Mandarin Chinese Female Synthetic Speech. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9(3):1–32. <https://doi.org/10.1145/3749510>
- [269] Zibrek K, Cabral J, McDonnell R (2021) Does Synthetic Voice alter Social Response to a Photorealistic Character in Virtual Reality? In: Motion, Interaction and Games. ACM, pp 1–6, <https://doi.org/10.1145/3487983.3488296>
- [270] Zou Z, Alnajjar F, Lwin M, et al (2025) Driving with Voices: How Virtual Agent Affects Driver Stress and Performance. In: Proceedings of the 13th International Conference on Human-Agent Interaction. Association for Computing Machinery, pp 479–481, <https://doi.org/10.1145/3765766.3765878>