“Hey Siri, let’s talk for real” – variability and flexibility in the perception of synthetic voices

Research Project Proposal

Postdoctoral Researchers International Mobility Experience

PRIME 2025/26

Submitted by:	Dr. Christine Nussbaum
Date of Submission:	12.08.2025
Time of Project:	1. October 2026 – 31. March 2028 (18 months)
Hosting Institutions:	University College London, UK
	Friedrich Schiller University Jena, Germany
Scientific Mentors:	Prof. Dr. Carolyn McGettigan (London)
	Dr. Nadine Lavan (London)
	Prof. Dr. Stefan R. Schweinberger (Jena)

1. Introduction: The importance of perceived naturalness in voices

When we hear a voice, we form an instant impression about it (Lavan & McGettigan, 2023). The characteristics that we infer are manifold, including age, sex, health, origin, attractiveness, and even personality traits like trustworthiness and dominance (Lavan, 2023). An important, but under-researched feature is the perceived **(un)naturalness of a voice**, i.e. whether a voice sounds monotonous, robotic or 'weird' (Nussbaum et al., 2025). Listeners seem to be very sensitive to unnatural voice features, which can have widespread implications for communicative quality. For example, individuals whose voices sound unnatural due to voice pathologies are often perceived as withdrawn or bored, which has a direct negative impact on their quality of life (Klopfenstein et al., 2020). Therefore, it is of practical importance to understand how such impressions are formed.

Now, in the digital era, questions of voice naturalness gain significance from a new angle. **Voice synthesis technology** quickly invades everyday life, e.g. in smart-home-devices, customer calls, gaming environments or support platforms (Rodero & Lucas, 2023). Fueled by rapidly evolving artificial intelligence technology, synthetic voices now approach an almost human-like auditory quality (Lavan et al., 2024). Nevertheless, most types of synthetic voices are still perceived as less natural than human voices, making them appear less pleasant, likeable and trustworthy (Kühne et al., 2020). Thus, as of today, listeners clearly prefer human over synthetic voices across many areas of application.

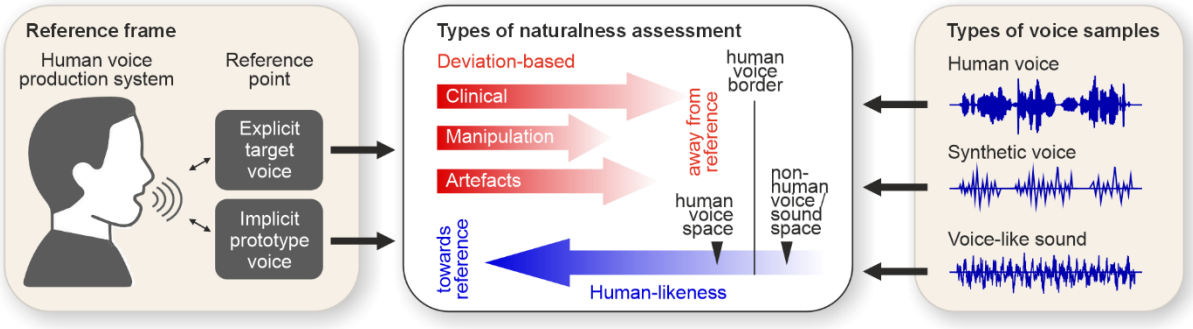
This may change in the future through increased exposure to synthetic voices. Our perceptual system is shaped based on experience (Belin et al., 2011). Individuals who were exposed only to human voices so far may find synthetic ones unnatural. For individuals who are highly accustomed to this technology, however, interacting with synthetic voices could become a fully natural experience. In Western, industrialized countries, most children of the next generation will likely grow up in a household with a smart-speaker device and will either interact with synthetic voices themselves or frequently observe others (i.e. their parents) in doing so. So far, it is not understood how these developments will affect the processing of synthetic (and human) voices. Therefore, the key objective of the present research proposal is to understand how **experience and exposure to synthetic voices shape the perception and evaluation of unnatural voice features**.

2. Theoretical background: The conceptual framework for voice naturalness

Despite its importance, a systematic understanding of voice naturalness is elusive, mostly due to conceptual underspecification. Until recently, there was no consistent framework for the definition of voice naturalness and no substantial efforts to link this to voice perception theory. We addressed this gap in a Review in *Trends in Cognitive Sciences,* published in May 2025 (Nussbaum et al., 2025). It offers the first conceptual framework for voice naturalness and the necessary starting point for systematic and theory-based empirical efforts. Specifically, we proposed a taxonomy with two types: deviation-based naturalness and human-likeness-based naturalness (**Figure 1**).

Figure 1

A conceptual framework for the definition of voice naturalness



Reprinted from Nussbaum et al. (2025).

In **deviation-based naturalness**, naturalness is defined as the deviation from a reference that represents maximum naturalness. The reference frame is commonly represented by the human voice production system. This can either be provided through an explicit target voice (i.e. comparison or baseline stimulus), or listeners are instructed to rely on an implicit prototype based on their experience and expectations. **Human-likeness-based naturalness** defines naturalness by its resemblance to a real human voice and is particularly well-suited for research on synthetic voices. Critically, this definition requires the assumption of a non-human voice space, which is not necessary for the deviation-based measure. However, both have in common that voice samples are assessed against a **reference for voice naturalness**. For a more detailed elaboration, please refer to Nussbaum et al. (2025).

The present project targets the **reference frame**. The individual inner reference for 'what sounds natural' is presumably shaped through our learning history and therefore may be shifted (towards synthetic voices) or broadened (i.e. a larger range of vocal features are accepted as natural) via greater experience with synthetic voices. Consequently, synthetic voice features may be perceived as less deviating and hence more natural. Additionally, individual differences in the amount of experience with synthetic voices could reveal an empirical distinction between the two types of naturalness: a person who has rarely heard synthetic voices before would likely rate them as both deviating from their natural norm as well as very non-human-like. Conversely, someone who is used to synthetic voices would rate them as less deviating/rare but may still perceive them as clearly non-human.

3. Planned empirical project

In the proposed research project, my focus lies on the variability and flexibility of synthetic voice perception. I plan to address this topic from three angles: **Study 1** will focus on long-term effects by exploring individual differences in experience with synthetic voices. **Study 2** will test whether synthetic voice perception is amenable to short-term perceptual manipulation. Finally, **Study 3** will combine both approaches in an intervention study, testing whether perception of synthetic voice features can be altered via three weeks of regular

exposure to synthetic vs. human voices (by listening to audiobooks). In what follows, I will describe the empirical design for all studies in more detail.

3.1. Study 1 – impact of long-term exposure

Research question: Are individual differences in the use of synthetic speaker devices linked to the perception of naturalness in voices?

Design: This is an exploratory perceptual rating study. It will consist of two parts. In part one, participants fill out a questionnaire on their experience/contact with synthetic voices in daily life (i.e. whether participants own a smart-speaker device at home and how often they talk to it). As additional control variables, I will assess the exposure to other types of voice deviations (e.g. pathological voices or digitally manipulated voices) and several personality traits, including openness towards technical innovations. Currently, there is no standardized test to assess familiarity with synthetic voices. Therefore, the development of this questionnaire forms one major task of this project.

In part two, participants will listen to a set of voices and provide ratings of naturalness (both deviation-based and human-likeness-based), pleasantness, eeriness and trustworthiness. The vocal material will be comprised of sentences spoken by human voices and various forms of synthesized ones. Human voices will be taken from pre-existing databases. Synthetic voices will be created with openly available synthesis tools, covering a broad range of human-likeness, such that some synthetic voices will sound not human at all, while others will be almost indistinguishable from human voices. Here, I will greatly benefit from the expertise in my host lab concerning the ethical, legal and practical issues around research with synthetic voices. The overall duration of the study will not exceed 45 minutes. To efficiently reach a suitably large and diverse participant sample, it will be conducted online. To ensure sufficient statistical sensitivity for individual differences, the target sample size will be between 150-200 participants (specific numbers refined upon power calculations).

Hypothesis: Individuals with more exposure to modern voice technology in their daily life will rate synthetic voices as more natural (deviation-based measure) but not as more human-like (human-likeness-based measure). Further, they will rate synthetic voices as more pleasant, more trustworthy and less eerie compared to individuals less experienced with synthetic voices.

3.2. Study 2 – impact of short-term manipulation via perceptual adaptation

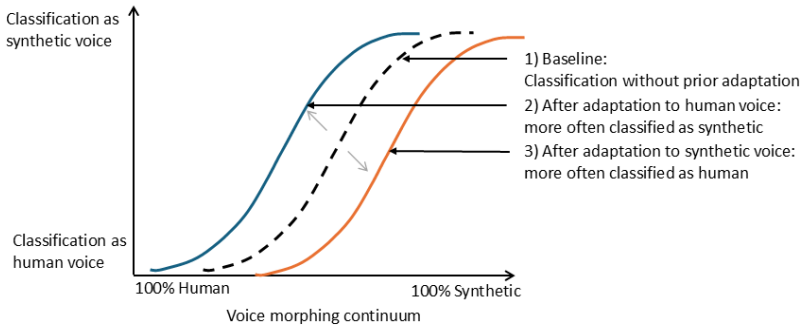
Research question: Can the perception of human-likeness be manipulated by short-term auditory exposure to human vs. synthetic voice features?

Design: This study employs a perceptual adaptation paradigm. Adaptation refers to a perceptual shift towards opposite stimulus features after prolonged exposure: For example, after adaptation to an angry voice, a subsequently presented ambiguous voice (i.e. lying in the middle of a continuum between angry and fearful voices) is more often classified as

fearful. Conversely, after exposure to fearful voices, the very same ambiguous voice is more often perceived as angry (Nussbaum et al., 2022). This perceptual shift is called a contrastive aftereffect. The present study will test whether it exists for perceived human-likeness. To this end, I will use voice morphing to create stimuli from a continuum between human and synthetic voices. When participants are asked to classify these voices as either human or synthetic (Baseline task), the response pattern usually resembles an S-shaped curve depicted in **Figure 2** (dashed-line curve). This is followed by two adaptation blocks: In one block, participants are repeatedly exposed to synthetic voices before they perform the classification task again. It is expected that this will lead to an increase in classifications as human, resulting in a shift of the curve (orange curve). In the second block, conversely, participants are exposed to human voices, resulting in a shift of classification as more likely synthetic (blue curve).

Figure 2

Visualization of a perceptual adaptation paradigm



I have piloted this paradigm in the context of a student project in 2025, which revealed several challenges that need consideration: First, the technical equipment seems to play a major role for this subtle auditory manipulation, such that having comparable equipment dramatically reduces noise in the data. Therefore, this study will be conducted in a lab under controlled conditions, using the same equipment for all participants. The study will last approximately 25 minutes and the target sample size is 40-50 (based on power calculations from the pilot study). The second major challenge concerns the creation of the stimulus material itself, via voice morphing. I have recently created the first set of morphed voices, comprised of 7 human-to-synthetic continua per speaker sex (male and female), uttering the two pseudowords /aba/ and /igi/. While I successfully created this stimulus set, I also gained a realistic idea of the complexity that comes with synthetic voice processing. The far-reaching phonetic experience of my host lab in London will be invaluable to further fine-tune the creation of stimulus materials. Thus, for the creation of valid and high-quality stimulus material, it is crucial to combine my extensive experience with voice morphing with the hosts' practical expertise on research using synthetic voices.

Hypothesis: I predict a contrastive aftereffect for human-likeness of voices: After adaptation to synthetic voices, ambiguous voices lying on a human-synthetic-morphing

continuum will be more likely classified as human. After adaptation to human voices, the same ambiguous voices will be more likely classified as synthetic. This shows that our inner reference for human-likeness-based naturalness is amenable to perceptual manipulation.

3.3. Study 3 – effects of an audiobook intervention on synthetic voice perception.

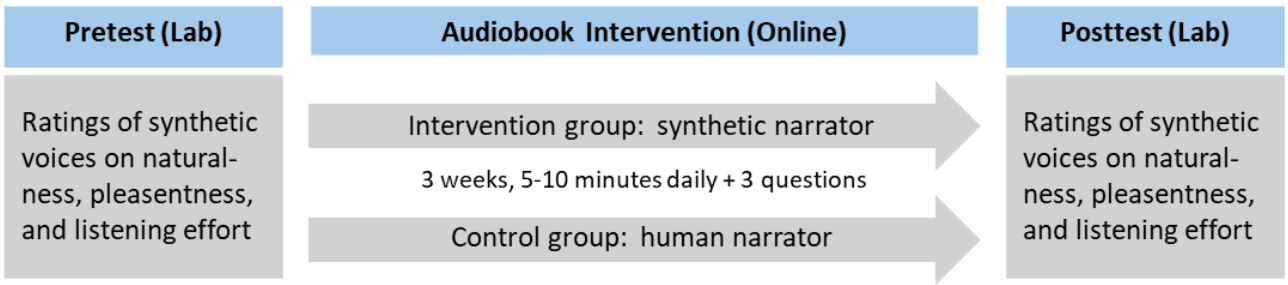
The first two studies provide unique and complementary insights. The strength of Study 1 lies in its ecological validity because it links daily-life experience of participants to synthetic voice perception. However, it is limited by its correlational design. Study 2 employs an experimental paradigm, with the potential to establish causal evidence that our inner reference for synthetic voice features can be manipulated via recent perceptual exposure. However, it is limited to lab-based, short-term effects. In the third and most ambitious study, I plan to combine the best of both of these approaches, capitalizing on ecologically valid familiarization alongside experimentally controlled data collection.

Research question: Does regular exposure to synthetic voices over a course of three weeks affect the perception and evaluation of synthetic voice features?

Design: Study 3 is a three-week audiobook intervention study, where one group of participants will regularly listen to an audiobook read by a synthetic narrator (**Figure 3**), using voices created with state-of-the-art tools, but which still sound clearly synthetic. The control group will listen to the same audiobook but read by a human narrator. Before and after intervention, participants will rate a set of synthetic voices on naturalness, pleasantness and listening effort. I will select a subset of the voice material from Study 1. Pre- and post-testing of participants will take place in the lab to ensure a controlled testing environment and to enhance commitment to the intervention via personal contact. The intervention itself is online. Over a course of three weeks, participants will be listening to an audiobook for 5-10 minutes each day. They will receive a link every day in the morning and another reminder in the afternoon. After listening to the track, they will be prompted with 3 multiple-choice questions about the content of the day's audiobook excerpt as an attention check. Only participants with >90% completed days and >90% correct answers will be kept in the final sample and will receive additional compensation as a motivator. The voices of narrators from the audiobook will not be included in the stimulus set for pre- and post-testing to avoid familiarity effects with specific voices.

Figure 3

Depiction of the intervention design for Study 3



I aim for 40-50 participants per group (specific numbers refined upon power calculations). Pre- and post-testing will take about 30-60 minutes per session in the lab, the daily intervention should not exceed 10 minutes, including the control questions. Longitudinal designs require thorough planning, careful monitoring and adequate time scheduling. I am aware of the ambitious nature of this project, and I am looking forward to taking on this challenge, building on the experience of my German host Prof. Dr. Stefan Schweinberger, who currently runs several online intervention studies on auditory perception. Realistically, data collection for this study could take at least 6 months; therefore, I will start with designing and planning it from the start of the project (more details in document *06-timeschedule*).

Hypothesis: Compared to the pretest, individuals who completed the audiobook intervention with a synthetic narrator will rate synthetic voices as sounding more natural, more pleasant and report less listening effort in the posttest. This effect will not be found in individuals who listened to the audiobooks with a human narrator.

3.4. Quality assurance

This project involves human participants and the processing of personal data. Further, it entails AI-generated voices, which may come with novel challenges regarding copyright or the scope of application. Thus, a conscious reflection on ethical and legal aspects of the project and careful research data management is a priority throughout.

Ethical approval is already in place at both host departments. Regarding the legal aspects around AI-generated voices, I am grateful to build on the prior experience of my host labs in London. Following the EU guidelines, I will ensure informed and voluntary consent of participants at all times and treat all data in accordance with the General Data Protection Regulation (GDPR). To protect the privacy of participants, data will be collected in a pseudonymous form and published in fully anonymized form only.

To ensure maximum transparency and reproducibility, all studies will be preregistered. Exact sample size calculations will be based on a priori power analyses. As I am deeply committed to the principles of Open Science, all research materials, including raw data, analysis scripts, and stimuli, will be made available on a public repository (e.g. on OSF). Further, I aim for open access publication.

3.5. Budget estimation

I estimate the following minimum budget requirements: participant reimbursement in Experiment 1, Experiment 2, and Experiment 3 (150 x 16 + 40 x 11.50 + 80 x 25 = 5660 EUR). The minimum research budget will be provided by the department of Prof. Stefan R. Schweinberger, while UCL will waive their standard bench fee for academic visitors (£6035 per annum in 2024/25). In addition, I will apply for research funding from local and national providers (e.g., the ProChance career program of the University Jena, or Research Grants by DFG).

4. Research environment and suitability of the hosts

The foreign host institution will be the University College London (UCL), where I will work under the mentoring of **Prof. Carolyn McGettigan**, who is Chair in Speech and Hearing Sciences at the Department of Speech, Hearing and Phonetic Sciences. The project will further be carried out in close collaboration with **Dr. Nadine Lavan**, Senior Lecturer at Queen Mary University of London. Both are world-leading voice researchers, with outstanding expertise in variability and individual differences in impression formation of voices.

For this specific project, they offer several competencies and resources which are crucial for its success: First, while I have focused on theoretical work so far, Prof. McGettigan and Dr. Lavan have extensively worked with synthetic voices already and can offer valuable practical support in my endeavors to translate my conceptual framework of voice naturalness into empirical insights. For example, they recently provided initial evidence that the perception of synthetic voice features can be highly context-specific (Lavan et al., 2024), contributing to my motivation for the present research project. Second, they provide the technical infrastructure and longstanding experience with large-scale online data collection (Eerola et al., 2021). While I have some expertise with online research myself (i.e. Nussbaum et al., 2023), the current empirical project presents a new level of complexity, and I will therefore profit very much from their insights. Third, I plan to break new ground by using voice morphing on synthetic voices. But in order to succeed, I need to discuss my work from many angles with researchers from speech science, phonetics, linguistics, computer science and many more. As mentioned in Section 3.2, the technical equipment has a large impact on quality assessment. Therefore, it is absolutely crucial to sit together in person in front of a standardized setup and explore different approaches. Coming to Prof. McGettigan's lab offers me precisely this opportunity, in state-of-the art lab facilities. Finally, I am eager to acquire new skills by attending Master courses which are open to guest researchers.

The German sending institution is the Friedrich Schiller University Jena, specifically the Department for General Psychology, led by **Prof. Stefan R. Schweinberger**, where I currently pursue my habilitation after completing my PhD in 2023. There, I have access to all laboratory facilities for high-quality data collection, as well as the fully-equipped audio-video-lab. Further, the department will cover the funding for the compensation of participants. Notably, our department is already in regular and productive exchange with the host institution in London, because we all collaborate in the Voice Communication Sciences (VoCS) network. Prof. Schweinberger supported me in the development of my conceptual framework for voice naturalness (Nussbaum et al., 2025). For the present project, I am also counting on his expertise for online interventions targeting voice perception and his ongoing support in shaping my individual academic profile. While I therefore plan to stay in the group where I completed my PhD, this does not impede my academic independence (for more details, please refer to document 17 – supplementary information). In fact, I would argue that Jena is the ideal place for me to continue my work, and is advantageous, primarily because of my role in the VoCS network and the Jena Voice Research Unit.

5. Impact, dissemination and outlook on future research projects

5.1. Impact and benefit

Right now, research on voice perception and synthetic voices in particular is more relevant than ever. These technologies already form a part of our daily life, and the scientific understanding of the manifold consequences is lagging behind. This has recently been acknowledged by the field, culminating in the formation of a big research network (VoCS) with the aim of positioning Europe at the forefront of Voice Research. With the present research project, I contribute to this vision.

All studies will provide unique but complementary insights, targeting long-term, mid-term and short-term aspects of interindividual variability and flexibility in the perception of synthetic voices. Studies thus conceptually build on one another but can be run independently, where no study is essentially contingent on the outcome of another. Together, they will provide a nuanced picture of how different forms of perceptual experiences could shape our perception and evaluation of synthetic voices and thus equip us with valuable knowledge on how daily life in the future is affected by the increasing number of smart-speaker devices surrounding us. At the same time, this series of studies represents a structured and systematic investigation of the recently proposed theoretical framework for voice naturalness. Therefore, the current research project is of high value both to applied as well as basic research.

5.2. Dissemination

The main scientific output will be open-access publications of all three studies, as well as national and international conference contributions. Further, I will share my insights with the Jena Voice Research Unit (VRU) and the VoCS network. Since all my research materials will be openly available on online repositories, they are also open to other colleagues for further scientific exploitation. Additionally, I am very enthusiastic about research communication to non-scientific audiences, and I am looking forward to sharing my insights in formats like the Lange Nacht der Wissenschaften or via my science communication platform *PhDScicom e.V*.

5.3. Outlook on future research projects

With the DAAD PRIME project focusing on the variability of listeners, a next logical step could be to focus on the variability of the context: Are natural voices always preferred, or is naturalness preference dependent on the specific application of synthetic voices? Can natural-sounding voices disrupt rather than promote successful communication under some circumstances? Are there potential interactions between the listener and the context in which they are exposed to synthetic voices? Ultimately, I am striving for a systematic understanding of voice naturalness and how it affects communicative quality. With this work, I hope to contribute valuable knowledge that capacitates us to navigate through a rapidly changing world full of technical advances in a human-friendly and sustainable manner.

References

Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding voice perception. *Br J Psychol*, *102*(4), 711–725. https://doi.org/10.1111/j.2044-8295.2011.02041.x

Eerola, T., Armitage, J., Lavan, N., & Knight, S. (2021). Online Data Collection in Auditory Perception and Cognition Research: Recruitment, Testing, Data Quality and Ethical Considerations. *Auditory Perception & Cognition*, *4*(3-4), 251–280. https://doi.org/10.1080/25742442.2021.2007718

Klopfenstein, M., Bernard, K., & Heyman, C. (2020). The study of speech naturalness in communication disorders: A systematic review of the literature. *Clinical Linguistics & Phonetics*, *34*(4), 327–338. https://doi.org/10.1080/02699206.2019.1652692

Kühne, K., Fischer, M. H., & 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*, 1–16. https://doi.org/10.3389/fnbot.2020.593732

Lavan, N. (2023). How do we describe other people from voices and faces? *Cognition*, *230*, 105253. https://doi.org/10.1016/j.cognition.2022.105253

Lavan, N., Irvine, M., Rosi, V., & McGettigan, C. (2024). *Voice deep fakes sound realistic but not (yet) hyperrealistic.* https://doi.org/10.31234/osf.io/jqg6e

Lavan, N., & McGettigan, C. (2023). A model for person perception from familiar and unfamiliar voices. *Communications Psychology*, *1*(1), 1–11. https://doi.org/10.1038/s44271-023-00001-4

Nussbaum, C., Frühholz, S., & Schweinberger, S. R. (2025). Understanding voice naturalness. *Trends in Cognitive Sciences*, *29*(5), 467–480. https://doi.org/10.1016/j.tics.2025.01.010

Nussbaum, C., Pöhlmann, M., Kreysa, H., & Schweinberger, S. R. (2023). Perceived naturalness of emotional voice morphs. *Cognition & Emotion*, 1–17. https://doi.org/10.1080/02699931.2023.2200920

Nussbaum, C., von Eiff, C. I., Skuk, V. G., & Schweinberger, S. R. (2022). Vocal emotion adaptation aftereffects within and across speaker genders: Roles of timbre and fundamental frequency. *Cognition*, *219*, 104967. https://doi.org/10.1016/j.cognition.2021.104967

Rodero, E., & Lucas, I. (2023). Synthetic versus human voices in audiobooks: The human emotional intimacy effect. *New Media & Society*, *25*(7), 1746–1764. https://doi.org/10.1177/14614448211024142