“Hey Siri, you sound artificial to me” – variability and flexibility in the perception of synthetic voices

Research Project Proposal

DAAD Postdoc-Programm - Kurzstipendium

Submitted by: Dr. Christine Nussbaum
Date of Submission: 12.03.2024
Time of Project: 1. October 2025 – 31. March 2026 (6 months)
Hosting Institution: University College London, UK
Scientific Mentors: Prof. Dr. Carolyn McGettigan & Dr. Nadine Lavan

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 underresearched feature is the perceived **naturalness of a voice** (Nussbaum et al., 2025), i.e. whether a voice sounds monotonous, robotic or 'weird'. Listeners seem to be very sensitive to unnatural voice features, which can have tremendous 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 For example, individuals whose voices sound unnatural due to voice pathologies are often perceived as withdrawn or bored, which has a direct impact on their quality of life (Klopfenstein et al., 2020).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.

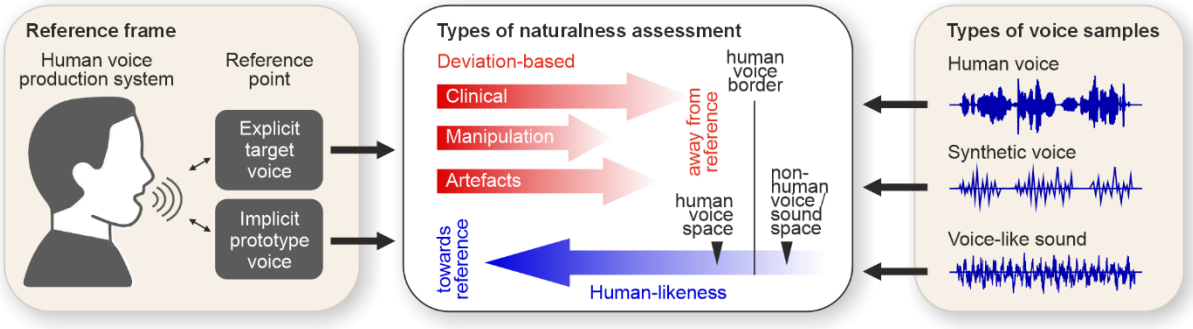
This may change in the future through increased exposure to synthetic voices. Our perceptual system is shaped based on experience (Belin et al., 2011). Individuals who were exposed only to human voices so far may find synthetic ones unnatural. For individuals who are highly accustomed to this technology, 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 February 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 may be shifted (towards synthetic voices) or broadened (i.e. a larger range of vocal features are accepted as natural) upon contact 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 rarely heard synthetic voices before would likely rate them both as 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 present research project, the focus lies on the variability of synthetic voice perception. I plan to address this topic from two 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. In what follows, I will describe the empirical design for both 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 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. It will assess the type and frequency of utilization, e.g. 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 human and various forms of synthesized voices. 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. 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 reach a diverse 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 rate synthetic voices as more natural (deviation-based measure) but not as more human-like (human-likeness-based measure). Further, they 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

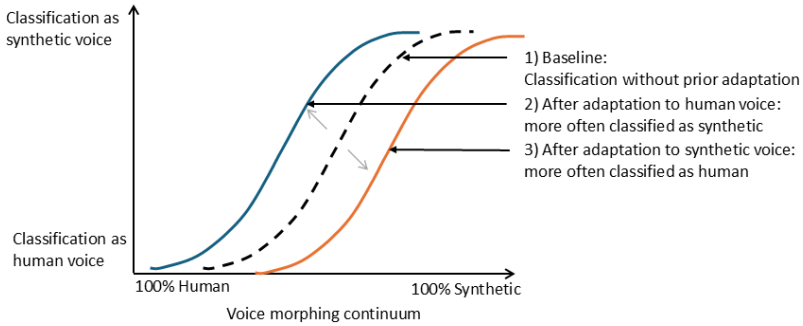
Research question: Can the perception of human-likeness be manipulated by short-term auditory exposure to human vs. synthetic voice features?

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 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**. 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 of 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



The study will be conducted online, with an approximate duration of 25 minutes and a target sample size of 40-50 (refined upon power calculations). Importantly, the creation of stimulus material requires me 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 be 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. Scientific value and quality assurance

Both studies will 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 therefore employs an experimental paradigm, with the potential to show that our inner reference for human-likeness-based naturalness can be manipulated via recent perceptual exposure.

Ethical approval is already in place at the host department. To ensure maximum transparency and reproducibility, both studies will be preregistered. Exact sample size calculations will be based on prior 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 (i.e. on the OSF platform: https://osf.io/). Further, I will aim for open access publication.

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**, leader of the Department for 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 on my specific empirical designs. Third, they give me access to an interdisciplinary network and the opportunity to discuss my research from many angles. I am a trained psychologist, but voice research also covers speech sciences, phonetics, linguistics, computer science and many more. I am therefore very excited to meet researchers with other specialties in Prof. McGettigan's lab and participate in their regular lab meetings. Finally, I am eager to acquire new skills by attending Master courses which are open to guest researchers (i.e. "Introduction to Deep Learning for Speech and Language Processing").

5. Impact, dissemination and outlook on future research projects

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 scientific understanding of the manifold consequences is lacking behind. This has recently been acknowledged by the award of an EU-MSCA doctoral network "Voice Communication Sciences" (https://www.vocs.eu.com/). The aim is to position Europe at the forefront of Voice Research. With the present research project, I contribute to this vision. I feel honored to be already part of this great network to share and discuss my research with. After my time in London, I will return to the Department of Prof. Dr. Stefan R. Schweinberger in Jena. There, I will continue co-supervising a PhD student on the topic "Neurocognitive processing of voice naturalness in human and synthetic voices". Further, I will share my insights with the Jena Voice Research Unit (VRU, https://www.voice.uni-jena.de/). As a special highlight in autumn 2026, we will host the third VoiceID conference in Jena and welcome voice researchers from all over the world. This provides me with ideal conditions and the necessary resources to pursue and share ground-breaking discoveries on the topic of voice naturalness.

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.* Advance online publication. 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