# DAAD PRIME Research Proposal

**Working Title:** “Hey Siri, you sound fine to me” – flexibility/variability and adaptability in the perception of unnatural/synthetic voice features

## 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). Another important, but underresearched feature is the perceived **naturalness of a voice** (Nussbaum et al., 2023). Listeners seem to be very sensitive to unnatural voice features, which can have tremendous implications for communicative quality, for example in individuals with speech-language pathologies (Klopfenstein et al., 2020). 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). Despite indefatigable efforts to resemble human vocal expression, synthetic voices are consistently perceived as less natural than human voices, making them appear less pleasant and likeable (Kühne et al., 2020). Thus, as of today, listeners clearly prefer human over artificial voices for most areas of application.

This may change in a future where we all are exposed to artificial voices on a daily basis. Our perceptual system is shaped based on experiences and past exposure. For a person who encountered only human voices during their whole lifespan, synthetic voices undoubtedly will have a rare and unnatural voice quality. But for someone highly accustomed to the use of modern voice technology, listening and 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. The proposed research project makes a first step towards closing this gap. My key objective it to understand how **experience and exposure to synthetic voices shape the perception and evaluation of unnatural voice features**.

This proposal is structured as follows: First, I will elaborate on the theoretical basis of my planned research (section 2), followed by a detailed description of my empirical project, comprised of three experiments (section 3). […]

1. A concise framework for the definition of voice naturalness

Despite its intuitive appeal, voice naturalness is an underspecified concept, which lacks consistent definition and terminology in the literature. Different research communities (i.e. on voice pathologies vs. synthetic/manipulated voices) unfortunately use the same term to refer to very different concepts and measurements. To address this issue, I recently developed the first conceptual framework that is rooted both in voice perception models and current empirical findings from a broad interdisciplinary literature (add citation and a bit of background, when submitted).

I proposed a taxonomy with two distinct types: Deviation-based naturalness and human-likeness-based naturalness (Fig 1). In **deviation-based naturalness**, naturalness is defined as the deviation from a reference that represents maximum naturalness. Example instructions for raters could be “Does this voice sound distorted?”, “Does this voice sound rare?”, or just “Does this voice sound natural?”. This conceptualization needs two important specifications: the **reference** representing maximum naturalness, and the **type of deviation**. In some cases, the reference is explicitly provided e.g. through a comparison or baseline stimulus. However, in many studies, raters are instructed to use an inner implicit reference which is based on their experience and expectations, e.g. whether “*the voice stimulus is perceived as a plausible outcome of the human speech production system*“ (Nussbaum et al., 2023). The type of deviation is specified through the vocal material (i.e. pathological, synthetic or manipulated voices).

**Human-likeness-based naturalness** defines naturalness by its resemblance to a real human voice. An instruction for raters could be “Does this voice sound like a real human speaker?” or “How human-like does the voice sound to you?” Compared to the deviation-based definition, it comes with an important additional assumption: the existence of a non-human voice category, and hence a categorical boundary to human voices (although the transition between categories can be continuous). In other words, a definition of human-likeness is only meaningful if we assume that voices can be non-human in principle. Apart from this important distinction, human-likeness-based naturalness can be seen as a special case of deviation-based naturalness: the reference is a human voice (or listeners representation of a human voice), and the deviation lies on the human/non-human spectrum.

In the present research project, the focus lies on the role of experience with and exposure to synthetic voices. In the context of this framework, this targets the **inner reference** that represents maximum naturalness. This inner reference 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 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.

1. Empirical research project

In the context of the DAAD Prime Fellowship, I will pursue an empirical project comprised of three voice perception experiments that target […]

(I will add these details – but for now they are in the Power Point because I felt that makes it easier to discuss them).

1. Research environment and suitability of the hosts

The foreign host institution will be the University College London (UCL), where I will work under mentoring of **Prof. Carolyn McGettigan**, who is the leader of the Department for Speech, Hearing and Phonetic Sciences. My project will further be carried out in close collaboration with **Dr. Nadine Lavan**, lecturer in psychology at Queen Mary University London. They are our partners in the Voice Communication Sciences (VoCS) project, a HORIZON-MSCA Doctoral Network funded by the EU from 2025-2028. Both are world-leading voice researchers, with particular expertise in variability and individual differences in impression formation of voices. Noteworthy, they have advanced the field of voice research not only through high-quality and innovative empirical research, but through rigorous development of theory, which has recently culminated in the proposal of the “person perception from voices” model (Lavan & McGettigan, 2023). This model provided an excellent starting point for my own conceptual work on voice naturalness. In order to gain a systematic understanding of voice naturalness, I strongly believe that it is crucial to root empirical efforts in voice perception theory. With their expertise, Prof. McGettigan and Dr. Lavan will provide outstanding support in my endeavors to translate my conceptual framework of voice naturalness into empirical designs.

Further, voice research is always inherently interdisciplinary and covers not only psychology and neuroscience, but speech sciences, phonetics, linguistics, computer science and many more. I am a trained psychologist with only little experience in phonetics and linguistics. I am therefore very excited that the department of Prof. McGettigan is specialized on speech sciences with links to computer linguistics. In London, I want to profit as much as possible from the interdisciplinary network and take the opportunity to discuss my research from many angles.

Finally, I plan to conduct my empirical research fully online. Online research has several advantages over lab research, e.g. by decreasing participation barriers, but it is challenging when it targets perceptual phenomena and especially when audiovisual stimulus material is involved. While I do have substantial experience with online research myself (two experiments during my PhD were carried out online), the current empirical project presents with a new level of complexity. Here, the technical support by Dr. Lavan will prove invaluable, as she has extensive experience with online testing in the context of auditory research (Eerola et al., 2021).

The German sending institution is the Friedrich Schiller University Jena. After my time in London, I will return to the Department for General Psychology, led by **Prof. Stefan R. Schweinberger**. I completed my PhD in 2023 under supervision of Prof. Schweinberger and have worked there as a post-doc since, where I started to pursue my habilitation. Note that staying in the group where I completed my PhD does not impede my academic independence. Instead, I deliberately stayed in Prof. Steinberger’s lab because it still provides the perfect environment for the development of my academic career. I have been given great independence and freedom in the development of my habilitation topic. When my choice fell on voice naturalness, Prof. Schweinberger supported me thoroughly and in collaboration with Prof. Sascha Frühholz from the University Oslo we have been working on the conceptualization I outlined in section 2. Recently, I was given the unique opportunity to become part of the EU-funded Voice Communication Sciences (VoCS). From 2025 to 2028, I will be co-supervising a PhD student on the topic “Neurocognitive processing of voice naturalness in human and synthetic voices” in collaboration with the University of Oslo, the Universidad Pompeu Fabra and Cochlear Inc. Interdisciplinary exchange is furhter actively practiced in the Jena Voice Research Unit (VRU, <https://www.voice.uni-jena.de/>), of which I am currently the principal coordinator. Thus, while my work is primarily rooted in Jena, I am well-connected with several world-leading institutions in the voice field. This provides me ideal conditions and the necessary resources to pursue ground-breaking discoveries on the topic of voice naturalness. To enhance my academic independence, Prof. Schweinberger explicitly encouraged me in the persuasion of individual research collaborations, of which one already culminated in a publication and two are in preparation.

1. Budget estimation

I estimate the following minimum budget requirements: participant reimbursement in Experiment 1, Experiment 2, and Experiment 3 (50xX + 200xX + 100xX = XXXX EUR). The minimum research budget will be provided by Prof. Stefan R. Schweinberger and the additional bench fees required by University College London (~500 GDP) will be covered by XXX.

1. Impact, benefits, dissemination, and future research and application
2. Training opportunities and alignment with long-term career goals

References

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*, 593732. 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., & McGettigan, C. (2023). A model for person perception from familiar and unfamiliar voices. *Communications Psychology*, *1*(1). https://doi.org/10.1038/s44271-023-00001-4

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

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