**Naturalness in voices**

Christine Nussbaum, XXX, and Stefan R. Schweinberger

**Abstract** (120 words)

Perceived naturalness of a voice is a prominent feature which affects our interaction with both human and artificial agents. Despite its importance, (a) conceptual underspecification, (b) inconsistent operationalization, (c) lack of exchange between research on human and synthetic voices and (d) insufficient anchoring in voice perception theory has precluded a systematic understanding of voice naturalness. In this work, we review the current insights into voice naturalness by pooling evidence from a wider interdisciplinary literature. Against that backdrop, we develop a concise definition of naturalness and propose a conceptual framework rooted both in empirical findings and theoretical models. Subsequently, we identify core gaps in our current understanding of voice naturalness and discuss different approaches for further research.

Word Limit: 3500 words

Inhalt

[1. Introduction – voice naturalness (500 words) 2](#_Toc160438454)

[2. Current Problems (800 words) 2](#_Toc160438455)

[2.1. Conceptual Underspecification (200) 2](#_Toc160438456)

[2.2. Inconsistent Operationalization (200) 2](#_Toc160438457)

[2.3. Lack of exchange between different research domains (200) 2](#_Toc160438458)

[2.4. Insufficient anchoring in voice perception theory (200) 2](#_Toc160438459)

[3. Proposition of a concise framework for naturalness (1000 words) 2](#_Toc160438460)

[3.1. Definitions of naturalness (500) 2](#_Toc160438461)

[3.2. Differentiation from other concepts (500) 2](#_Toc160438462)

[4. Pooling evidence from different sources (“progressing in conjunction”) (400 words) 2](#_Toc160438463)

[5. Naturalness research rooted in voice perception theory (400 words) 3](#_Toc160438464)

[6. Open questions and future/outlook (400 words) 3](#_Toc160438465)

# Introduction – voice naturalness (500 words)

* why, what, how, by whom
* Clear identification of the importance and motivation for studying naturalness
* Identification of the key fields that are interested in voice naturalness
* Motivation for the current paper

Different sound characteristics affect how we perceive and interact with other agents through the vocal modality. One of these characteristics is the perceived naturalness of a voice.

* Naturalness is relevant and of scientific interest across several research domains, with tremendous importance for our everyday life
* Human: speech-language pathology; transgender voices -> people are basically worse off
* Synthetic voices: everywhere now; cite a few key findings
* Naturalness as threat to ecological validity

In humans, naturalness can be affected by voice distortion or transformation.

In artificial agents, its tremendous importance has been acknowledged through multiple efforts to create and constantly improve synthetic voices to resemble human vocal expression.

Far from a systematic understanding of naturalness.

Artificial voices now form an integral part of our daily lives, as we encounter them in navigation systems, telecommunication services, advertisement, entertainment media, educational resources, or smart-home devices (ToDo, Order 2017, 2023, Quellen).

When listening to a voice, we form an instant impression about it (Quelle). This includes perceived naturalness. Listeners seem to be very sensitive to (un-)natural voice features, which in turn can have a tremendous effect on the evaluation of and interaction with voices.

On the one hand, consistent evidence from different **speech-language pathologies** shows that impairments in speech naturalness affect everyday interaction to a degree that can result in social isolation, reduced quality of life, and even depression (Stepp & Voitech 2019, Quellen). In fact, even subtle acoustic manipulations can disrupt (maybe add 1-2 other examples from fully human voices, voice distortions).

One the other hand, in the era of artificial intelligence, one can hardly keep up with the rapid developments in **artificial voice synthesis,** which make indefatigable efforts to resemble human vocal expression. However, as of today, synthetic voices are consistently rated as less natural than human voices, which simultaneously affects perceived likeability, trustworthiness, and pleasantness (Quellen). Yet, it is not fully understood, how this affects user satisfaction across different domains of application (Schreibekmayer 2022, andere). Artificial voices now form an integral part of our daily lives, there is a strong need to close these knowledge gaps.

At the same time, several researchers put into questions, whether

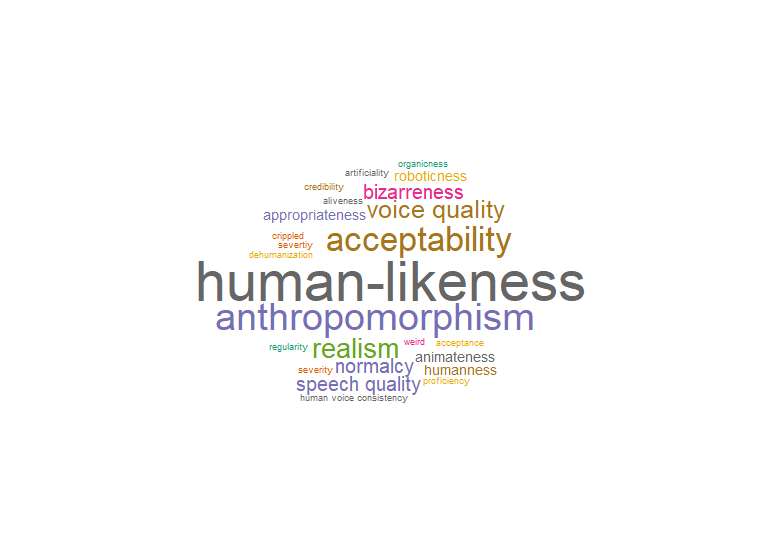
# Current Problems (800 words)

## Conceptual Underspecification (300)

Voice naturalness lacks a consistent definition in the literature. In fact, after reviewing a number of XX studies, the majority does not provide an explicit definition of naturalness at all (ToDo, see Box 1). In these studies, the conceptualization of naturalness can only be drawn implicitly from the operationalization. If definitions are provided, they vary tremendously across research contexts. In speech-language pathology, many researchers refer to the definition provided by Yorkston (1990): “*Naturalness is defined as conforming to the listener’s standards of rate, rhythm, intonation, and stress patterning and to the syntactic structure of the utterance being produced*”. In contrast, research on synthetic and non-human voices usually defines naturalness as “*speech most closely perceived as a human voice*“ (Mawalim 2022) or “*the degree to which a user feels a certain technology or system is human-like*” (Hu 2021). Accordingly, many studies using synthetic voices do not refer to naturalness but to human-likeness of voices. Finally, in an effort to bridge these different conceptualizations, we previously stated that “*by naturalness, we understand a voice stimulus to be perceived as a plausible outcome of the human speech production system.*“ (Nussbaum, et. al. 2023)

Interestingly, these definitions seem to share two important assumptions: First, voice naturalness is a perceptual and therefore subjective measure. Second, listener’s naturalness perception is the result of a complex multifactorial impression formation, most certainly based on the integration and weighting of many acoustic cues (Quelle). Beyond that, however, the conceptualizations are very heterogeneous because they are tailored to the respective empirical focus. Undoubtedly/For example, defining naturalness as human-likeness in a study which encompasses fully human voices only would make little sense. Unfortunately, despite covering relevant aspects, the current conceptualizations make it impossible to compare and integrate insights for a systematic understanding of voice naturalness. We therefore see a strong need to unite them under a concise conceptual framework, which we provide in Section 3.

Questions: also mention the verbal space we are dealing with here? See Figure (nur Platzhalter):



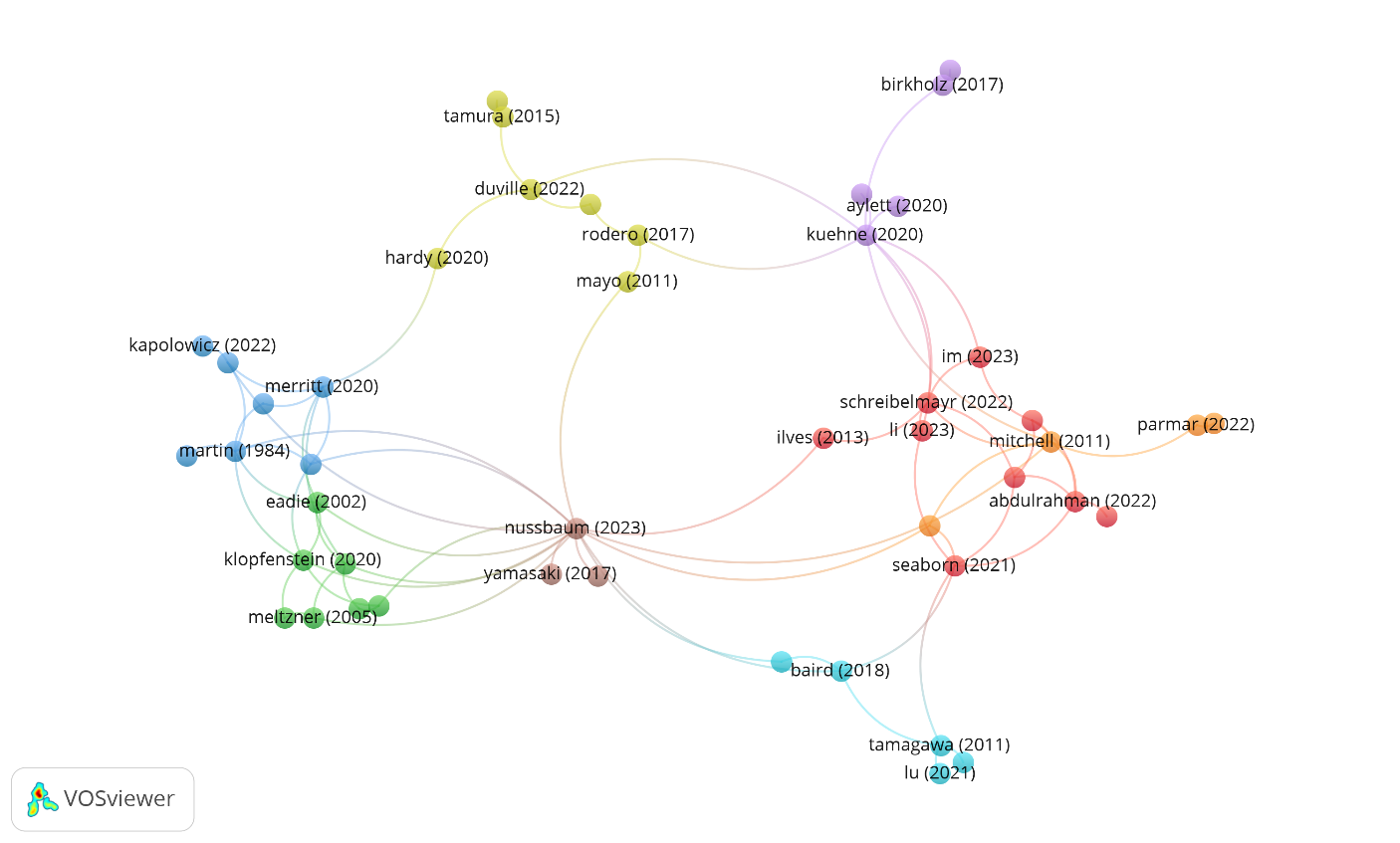
## Inconsistent Operationalization (200)

A common byproduct of inconsistent conceptualization is inconsistent operationalization. Primarily, this concerns to the studied vocal categories and features, which include human vs. synthetic voices (Quelle); pathological voices such as … (Quellen); acoustically manipulated human voices (Quellen), acted speech (Quelle), vocal fry (Quelle), as well as different accents (Quelle), age groups (Quelle), or gender identities (Quelle). In addition, it concerns the measurements and experimental designs. For example, although many studies use rating scales to asses perceived naturalness, they differ in the number of levels and denominations of endpoints (e.g. Quellen). In principle, such empirical heterogeneity can be a powerful source of insight. However, an insufficient report of empirical details in the majority of studies impede a meaningful integration of findings. Specifically, it is usually not explained how naturalness and the related experimental task was explained to the listeners. Further, the precise properties of the voice material often remain elusive, bearing a risk for potential undetected confounds. Finally, because perceived naturalness is based on subjective impression, a report on measurement reliability would be essential, but is only provided by few studies (e.g. Quelle). To address these issues, we collected some practical recommendations as a guidance for future research in Box 2.

## Lack of exchange between different research domains (150)

Research on voice naturalness is inherently interdisciplinary (evtl. Grafik dazu, ToDo), with two main domains: speech-language pathology and synthetic/artificial voices. However, while the scientific output is well-received within their own disciplines, they are poorly interconnected. This is impressively illustrated by a cross-citation analysis conducted using VOSViewer (Quelle, Fig X), showing several distinct clusters of studies, which rarely cite each other. Currently, research on voice naturalness is a rag rug rather than a research field. One may argue that this is not a problem, because the different disciplines simply have different interests and readerships. However, some intriguing commonalities and systematic patterns only emerge when pooling evidence from all available angles. For example, across synthetic, pathological and acoustically manipulated voices there is consistent evidence for a strong effect of fundamental frequency/prosody/pitch variation on perceived naturalness (Quellen). Further, while several studies failed to find an uncanny valley effect for synthetic voices, a recent study suggest it might exist for pathological ones (Quelle). Thus, we argue that the lacking exchange between research fields has not only precluded relevant insights, but has impeded the visibility and impact of voice naturalness research as a whole.

Platzhalter:



## Insufficient anchoring in voice perception theory (150)

The majority of naturalness research comes from applied fields, aiming to optimize artificial agents or improving the quality of life in patients with voice disorders. These findings equip us with valuable practical knowledge, but they are insufficiently anchored in voice perception theory. As an illustration, we added ten influential, theory-building voice perception publication to the VOSViewer analysis (Fig X, ToDo, marked in green, not yet there), showing that they are hardly cited by the field. Further, there is very little work on naturalness from basic voice research (some exceptions include ToDo). This leaves us with an intriguing divergence between rapidly increasing applied knowledge in rapidly developing branches (especially synthetic voices) on the one hand, but a simultaneous lack of understanding on basic voice mechanisms on the other. In order to fully understand how naturalness affects our perception and response to voices, this void needs to be filled.

# Proposition of a concise framework for naturalness (1000 words)

## Definitions of naturalness (500)

* provides equal flexibility and offers a useful reference for future

## Differentiation from other concepts (500)

* Distinctiveness
* Typicality
* Pathology
* Authenticity?

# Pooling evidence from different sources (“progressing in conjunction”) (500 words)

* Although the sometimes target similar questions and get to similar conclusions.
* give some recommendations for report/design standards
* Recommend some keywords (and a consistent use of proper definitions)
* Foster exchange between domains
* Examples, where pooling makes sense: androgynous voices
* Uncanny valley (Diel)

# Naturalness research rooted in voice perception theory (500 words)

* Relate naturalness to voice perception /person perception models
* Discuss where the topic is underexplored from the perspective of basic research
* Mention here that there is also cool stuff from face perception literature (Miller meta analysis)

# Open questions and future/outlook (500 words)

Box 1 (400 words):

* Give information about literature review

Box 2 (400 words): - recommendations

* Offer a concise definition to both readers as participants of studies
* USE PROPER KEYWORDS to make research findable (maybe give some recommendations)
* Full report of everything, especially reliability, instructions to listeners and acoustic manipulation/measurements
* Wherever possible provide stimulus examples (auditory impression simply tells you more than just acoustic measurements and descriptions) (bridging different publication culture, different scientific standards etc).
* Keep the wide readership in mind (very interdisciplinary field), avoid very technical jargon

Glossary:

* Uncanny valley