**Naturalness in voices**

Christine Nussbaum, XXX, and Stefan R. Schweinberger

**Abstract** (120 words)

Perceived naturalness of a voice is a prominent property emerging from vocal sounds, and one which affects our interaction with both human and artificial agents. Despite its importance, a systematic understanding of voice naturalness is currently lacking, due to (a) conceptual underspecification, (b) inconsistent operationalization, (c) a lack of exchange between research on human and synthetic voices and (d) insufficient anchoring in voice perception theory. In this work, we reflect on 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 remaining gaps in our current understanding of voice naturalness and discuss different approaches for future research.

Word Limit: 3500 words

Inhalt

[1. Introduction – voice naturalness (450) 3](#_Toc160791725)

[2. Current Problems (800) 3](#_Toc160791726)

[2.1. Conceptual Underspecification (300) 3](#_Toc160791727)

[2.2. Inconsistent Operationalization (200) 4](#_Toc160791728)

[2.3. Lack of exchange between different research domains (150) 5](#_Toc160791729)

[2.4. Insufficient anchoring in voice perception theory (150) 5](#_Toc160791730)

[*3.* Proposition of a concise framework for voice naturalness (900*)* 6](#_Toc160791731)

[3.1. Definitions of naturalness (500) 6](#_Toc160791732)

[3.2. Differentiation from other concepts (400) 7](#_Toc160791733)

[4. Progressing in conjunction (400) 7](#_Toc160791734)

[5. Naturalness research rooted in voice perception theory (400) 8](#_Toc160791735)

[6. Open questions and future/outlook (400) 9](#_Toc160791736)

# Introduction – voice naturalness (450)

When listening to a voice, we form an instant impression about it (Quelle). In a world that is both dominated by voice-to-voice social interaction and its more recent variants that are implemented by digital technology in its various forms, this impression 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, naturalness plays a crucial role in **human voices**. 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). Similarly, deliberate acoustic manipulations and distortions of human voices can reduce naturalness and disrupt effective communication (Quelle). On the other hand, naturalness plays a crucial role in **synthetic voices**. In the era of artificial intelligence, one can hardly keep up with the rapid developments in voice synthesis, which make indefatigable efforts to resemble human vocal expression (Quelle). 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 impacts user satisfaction across different domains of application (Schreibekmayer 2022, andere). Since artificial voices now form an integral part of our daily lives (Quelle), there is a strong need to close these knowledge gaps.

From both the perspective of human and artificial voices, perceived naturalness carries relevant implications for our everyday life and is therefore of scientific interest. But although many recent works provide useful empirical insights, close inspection of these studies reveals that we are currently looking at a rug rag rather than a research field, and hence are far from a systematic understanding of voice naturalness. For the present work, we therefore took a step back and reflected on how this research is currently conducted. To anticipate, we identify four problems: (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. We argue that these problems have so far precluded a systematic understanding of vocal naturalness, impeded the visibility of this research to a wider readership, made us overlook some crucial research questions, and has led to a divergence between theory and practice. In the next section, we will first elaborate on each of these problems, before proposing concrete measures to address them, and then proposing an initial concise conceptual framework for voice naturalness. The aim of this paper is to provide a useful basis for systematic, theory-driven, and profitable research on voice naturalness in the future.

# Current Problems (800)

## 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/empirical design. If definitions are provided, they vary tremendously across research contexts. In speech-language pathology, several 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 or anthropomorphism 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. For example, defining naturalness as human-likeness in a study which encompasses human voices only would make little sense. Unfortunately, despite covering relevant aspects, the current conceptualizations make it impossible to compare and integrate different insights into 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):



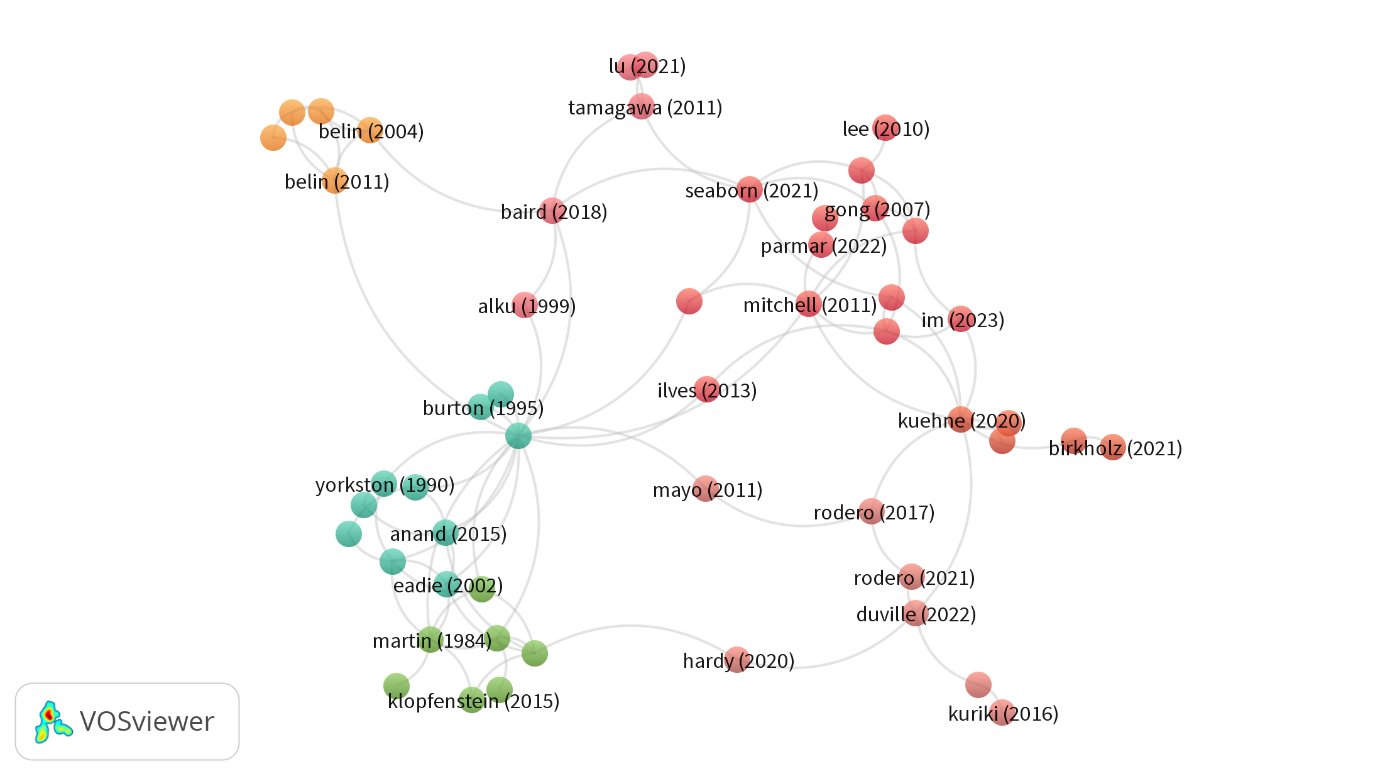
## Inconsistent Operationalization (200)

A common byproduct of inconsistent conceptualization is inconsistent operationalization. Primarily, this concerns 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), and gender identities (Quelle). In addition, it concerns the measurements and experimental designs. For example, although many studies use rating scales to assess 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 impedes a meaningful integration of findings. Specifically, it is usually not stated how naturalness and the related experimental task was explained to the listeners. Further, the precise acoustic 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 provided by few studies only (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 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. One may argue that this is not problematic, 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). In fact, 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 and 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 publications to the VOSViewer analysis (Fig X, ToDo), showing that they are rarely 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 voice naturalness (900*)*

In the previous section, we identified key problems that impede a systematic understanding of naturalness in voices. In the following sections, we propose concrete measures to address these issues. We start with a conceptual framework for the explicit definition of naturalness in voices.

## Definitions of naturalness (500)

We propose a taxonomy with two/three (to be discussed) distinct types: Deviation-based naturalness and human-likeness-based naturalness **(**Fig X**)**. 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 (see Quelle). However, in many studies, raters are instructed to use an inner implicit reference which is based on their experience and expectations (c.f. Yorkstons definition, section 2.1). The type of deviation is specified through the vocal material. It can virtually cover all acoustic features, ranging from specific manipulations (e.g. fundamental frequency variation, Quelle) to complex multivariate vocal patterns (e.g. in distorted or pathological voices Quelle).

**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.

With this taxonomy, we provide a flexible and intuitive reference for the explicit definition of naturalness alongside with its underlying assumptions. With future research anchored/grounded/united in one conceptual framework, systematic integration and comparison of findings could be greatly facilitated.

In fact, both conceptualizations are already prevalent, but often implicitly through certain design choices. For example, comparing human to synthetic voices implies human-likeness based naturalness. Of XX studies on naturalness (see Box 1), an independent raterfound that XX used a deviation-based, XX human-likeness-based and XX a combination of conceptualizations. One study deserves particular mention: Diel (2023, preprint) studied the uncanny valley effect in different types of unnatural voices. They found that impressions of uncanniness resulted from “deviation from familiar categories” rather “categorical ambiguity”. We interpret this as the empirical realization of the abovementioned conceptual differentiation, providing initial evidence for distinguishable perceptual outcomes.

* Hier wären jetzt noch paar Wörter Platz für evtl. authenticity, oder halt im nächsten Abschnitt 😉
* Evtl nochmal unterschied zwischen „Is this a natural voice“ vs. „is this a natural emotional voice” spezifizieren – to be discussed.

## Differentiation from other concepts (400)

* Distinctiveness/typicality
  + Deviation-based conceptualization is very similar
  + And both depend on the experience and learning history of the listener
  + Is probably correlated
  + Assumption: unnatural voices are always distinct; but natural voices can be both distinct and typical
  + For human-likeness, concepts can actually be separated: e.g. people with a smart-speaker device at home are used to that synthetic voice, therefore its less distinctive, but still clearly non-human
* Voice Pathology
  + part of voice naturalness
  + give some examples
  + almost all pathologies result in a reduction of naturalness, but not all reduction of naturalness are the result of voice pathologies
* Authenticity?
  + Can be argued to be another special case of voice naturalness
  + Familiarity-based authenticity
  + „does this voice sound like the real voices from speaker X?”
  + Again, a special case of deviation-based naturalness: reference is a speaker with a certain identity, deviation is the difference auditory input und stored representation.
  + But needs familiarity with the speaker (speaker characteristics and within-person variability of these characteristics)
  + This is very important with the now very prevalent danger of spoofing
* Any other? (evtl. "inadequate” prosody)

(this is one of the sections where I hope for substantial input from collaborators)

# Progressing in conjunction (400)

We believe that a systematic understanding of voice naturalness is only possible by pooling evidence from all available angles. Because as different as perspectives on voice naturalness may appear, they are united by several overarching questions: How is an impression on voice naturalness formed? Which acoustic features affect it? How does voice naturalness impact perception, interaction, and communication? Are there differences across individuals and listening contexts?

Fortunately, voice naturalness research is already a highly interdisciplinary field, with great potential to provide us with the answers. Now, it has to start progressing in conjunction. This can be achieved by two means: (a) converting the empirical heterogeneity (presented in section 2.2) from an impediment into an advantage and (b) fostering an active and profitable exchange between disciplines, especially on human and synthetic voices. For the implementation of both, a sensible awareness for the interdisciplinary nature of the field is crucial. First, publications need to be findable and accessible for others, i.e. through the establishment of common keywords. Second, findings need to be communicated inclusively enough for readerships from very diverse backgrounds. This may entail to provide some explicit definitions, avoid technical jargon, incorporate scientific standards from other fields where deemed fit, and discuss one’s findings against the backdrop of a wider interdisciplinary literature (oder – tie active knots to findings from other fields). Finally, all conceptual and empirical aspects need to be reported with a sufficient level of detail to allow comparability. In Box 2, we have converted these suggestions into a number of practical recommendations.

If conducted appropriately, this exchange can lead to mutual inspiration between research on human and synthetic voices. A successful example concerns current insights into naturalness of androgynous and transgender voices. (ToDo: elaborate in 2-3 sentences) […]

# Naturalness research rooted in voice perception theory (500)

Several authors have pointed out that research on naturalness is not sufficiently rooted in theory (Quellen). As discussed in section 2.4, the strongly applied orientation of the field comes at the expense of basic research, although several influential models on voice perception offer good staring points: The voice-space model proposed by Quelle represents voices in terms of their acoustic deviation from one another or a potential reference. The functional model by Belin 2011 assumes that an initial structural analysis of voices is followed by dissociable pathways processing vocal speech analysis, vocal affect analysis and voice recognition. Recently, Lavan (2023) integrated these previous models in a unifying framework, explaining how listeners form multiple impressions about both familiar and unfamiliar voices. Commonly studied person characteristics include identity, gender, age, emotion and personality of speakers.

Although voice naturalness is in principle covered by these models, it is never explicitly mentioned. This is particularly intriguing against the backdrop of a questions that has prompted extensive debate and empirical efforts in basic voice research: Are voices special (Belin 2011)? In other words, do voices recruit network and resources in the brain that are not recruited by other types of acoustic stimuli? Voices with varying degrees of naturalness provide a powerful tool to shed new light on this debate. What makes human voices special? What makes natural voices special? In a nutshell, trying to understand the impact of naturalness on voice perception means trying to answer these questions.

This is not all. Rooting naturalness research in voice perception theory prompts further crucial questions that are not fully answered yet. First, to which degree is naturalness a threat to ecological validity (Nussbaum 2023)? Many voice researchers use acoustic manipulation such as voice morphing which could have unintended side effects on perceived naturalness. If this cannot be avoided, it should be at least quantified. Second, how does naturalness interact with the processing of other voice characteristics? For example, first insights into the interplay of naturalness and emotionality suggest that […] (Quellen, shall I go into detail?). Third, [ToDo, brain data?, or “is naturalness always better than unnaturalness”, role of experience and learning history] Note that all of these questions are of relevance beyond the vocal modality. For faces, many of these aspects are covered in recent meta-analysis (Miller 2023)

(grade noch ein ziemlicher Flickenteppich…,this is one of the sections where I hope for substantial refinement from collaborators)

# Open questions and outlook (400)

- putting the conceptualizations to the test and compare whether they lead to different outcomes  
- systematic comparison of human-pathological, human-distorted, and synthetic voices  
- categorical perception between human- and non-human voices?  
- naturalness implications for ecological validity?  
- in naturalness always better?  
- individual differences  
- neurocognitive insights (aaaaaall kinds of brain data)

Box 1 (400 words): A field in numbers

* Give information about literature review
* XX studies
* On which fields, topics
* How many deliver an actual definition of naturalness
* Year range
* Common methods

Box 2 (400 words): - recommendations

* Offer a concise definition to both readers as participants of studies
* USE PROPER KEYWORDS to make research findable (Recommendations: Naturalness OR Human-likeness)
* 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:

* Synthetic/artificial voices
* Uncanny valley