Naturalness of voices - how humans and artificial agents could learn from one another

Christine Nussbaum
Voice Research Unit, 16.02.2024

Motivation

“Impairments in speech naturalness can lead to communication partners perceiving the affected individuals as unhappy, cold, withdrawn, introverted, or bored. These false perceptions can interrupt participation in regular life roles, leading to loss of employment and independence. Thus, impaired speech naturalness can result in social isolation, reduced quality of life, and depression.” (Stepp & Voijtech, 2019)

“The growing popularity of speech interfaces goes hand in hand with the creation of synthetic voices that sound ever more human. Previous research has been inconclusive about whether anthropomorphic design features of machines are more likely to be associated with positive user responses or, conversely, with uncanny experiences. To avoid detrimental effects of synthetic voice design, it is therefore crucial to explore what level of human realism human interactors prefer and whether their evaluations may vary across different domains of application.” (Schreibelmayer & Mara, 2022)

“It is like my toaster is speaking to me.” (Kühne et al. 2020)

Abstract

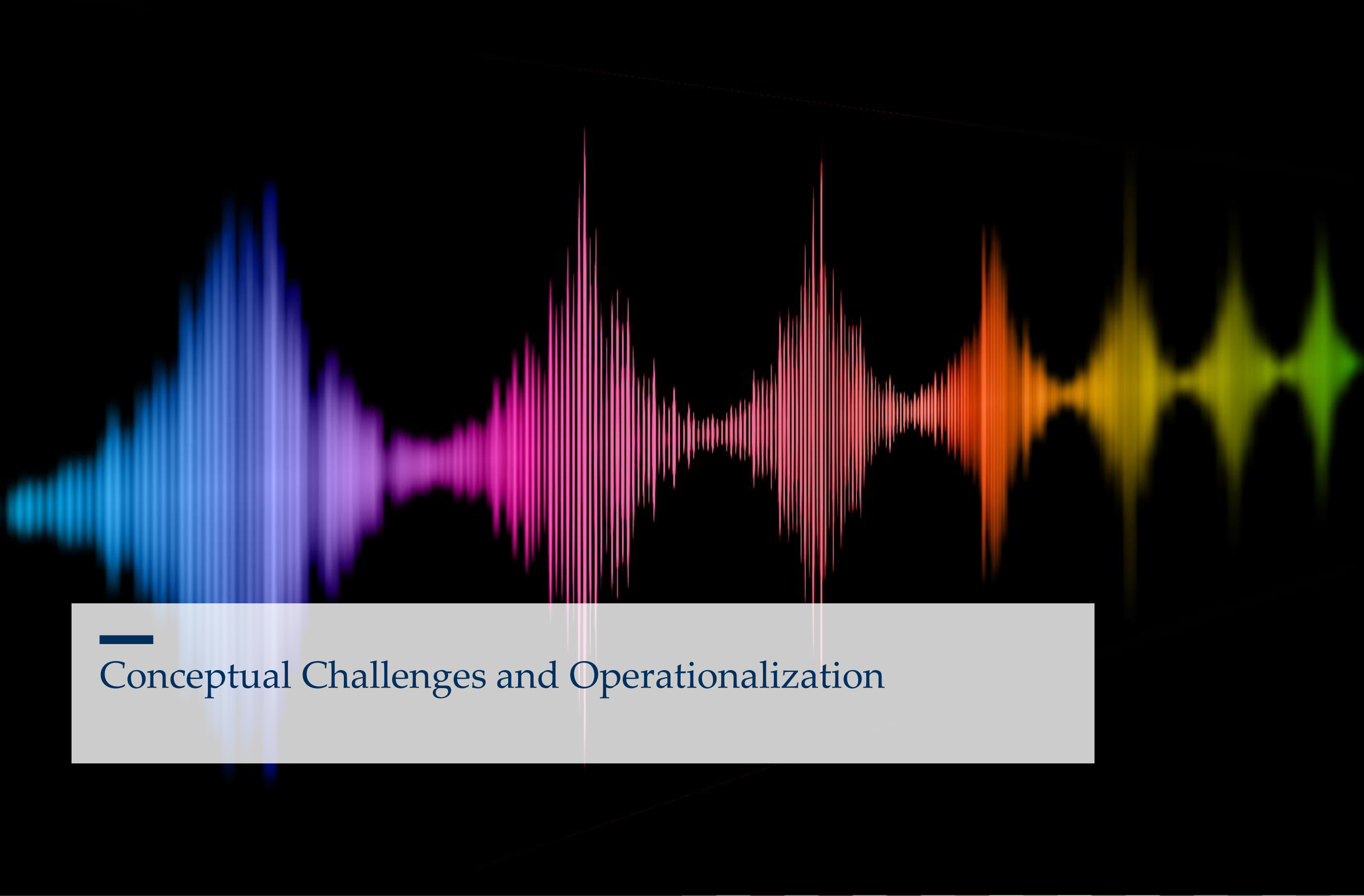
Perceived naturalness of a voice is a prominent feature which affects our interaction with both human and artificial agents. Despite its importance, (a) conceptual underspecification, (b) inconsistent operationalization, (c) a lack of exchange between research on human and synthetic voices and (d) insufficient anchoring in voice perception theory has precluded a systematic understanding of voice naturalness. In this work, we reflect on 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 future research.

Current problems:

1. (1) Conceptual underspecification
2. (2) Inconsistent operationalization
3. (3) Lack of exchange between different research domains
4. (4) Insufficient anchoring in voice perception theory

- ➢ Precluded a systematic understanding of vocal naturalness
- ➢ Impeded the visibility of this research to a wider readership
- ➢ Has kept us from asking some crucial research questions
- ➢ Has led to a divergence between theory and practise

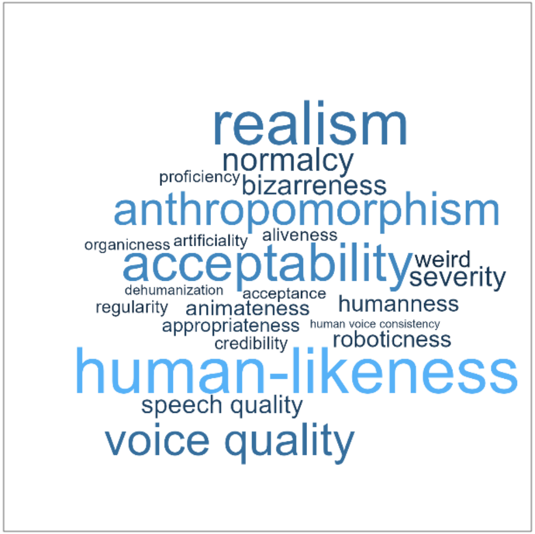
1. Introduction – voice naturalness (450)
2. Current Problems (800)
    - i. Conceptual Underspecification (300)
    - ii. Inconsistent Operationalization (200)
    - iii. Lack of exchange between different research domains (150)
    - iv. Insufficient anchoring in voice perception theory (150)
3. Proposition of a concise framework for voice naturalness (900)
    - i. Definitions of naturalness (500)
    - ii. Differentiation from other concepts (400)
4. Progressing in conjunction (400)
5. Naturalness research rooted in voice perception theory (400)
6. Open questions and future/outlook (400)



Conceptual Challenges and Operationalization

Voice Naturalness

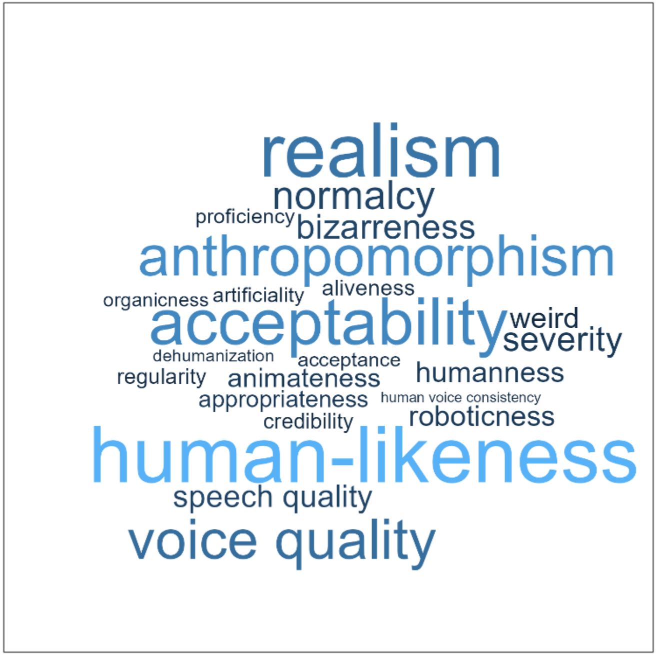
- pathological human voices
- manipulated human voices
- synthesized/artificial voices



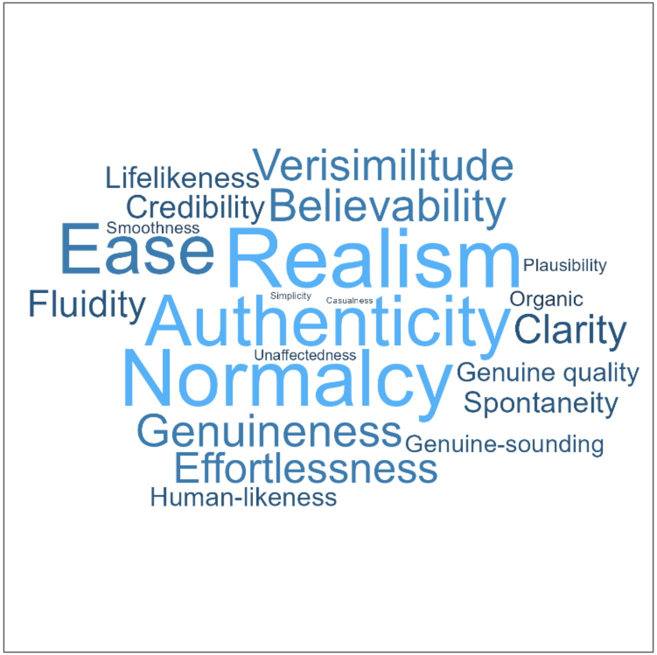
“Naturalness was defined as conforming to the listener’s standards of rate, rhythm, intonation, and stress patterning [...] ” [e.g. Yorkston et al. 1990]

“Natural speech is the speech most closely perceived as a human voice“ [e.g. Mawalim et al. 2022]

*"By naturalness, we understand the voice stimulus to be perceived as a **plausible outcome of the human speech production system**.“* [Nussbaum et al. 2023]



Naturalness Papers



ChatGPT

Challenges with operationalization

How naturalness is explained to the listener (what they should attend to)

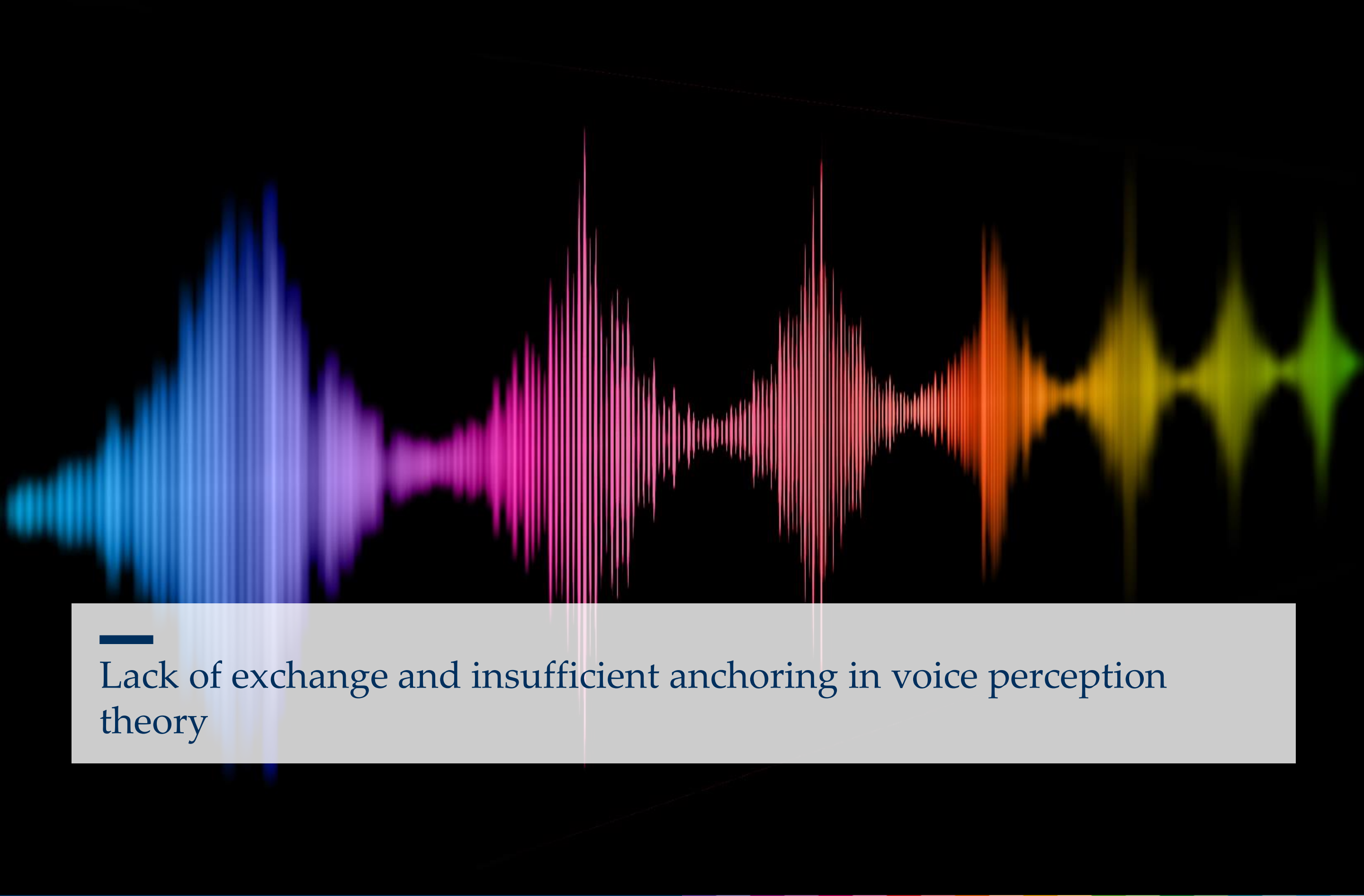
Reliability of measurements

The appropriate scale

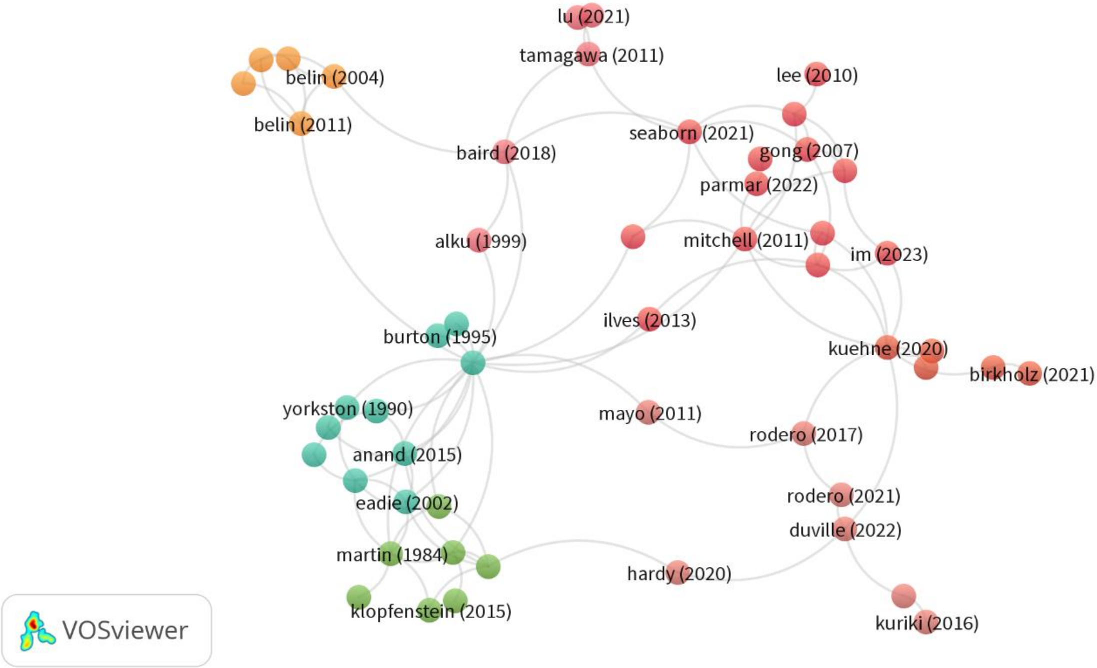
The precise properties of the voice material

Potential confounds

Insufficient report of empirical details



Lack of exchange and insufficient anchoring in voice perception theory


lu (2021)
tamagawa (2011)
lee (2010)
belin (2004)
belin (2011)
seaborn (2021)
gong (2007)
baird (2018)
parmar (2022)
alku (1999)
mitchell (2011)
im (2023)
ilves (2013)
burton (1995)
kuehne (2020)
birkholz (2021)
yorkston (1990)
mayo (2011)
rodero (2017)
anand (2015)
eadie (2002)
rodero (2021)
duville (2022)
martin (1984)
hardy (2020)
klopfenstein (2015)
kuriki (2016)
VOSviewer

Box 1: A field in numbers (mini literature review)

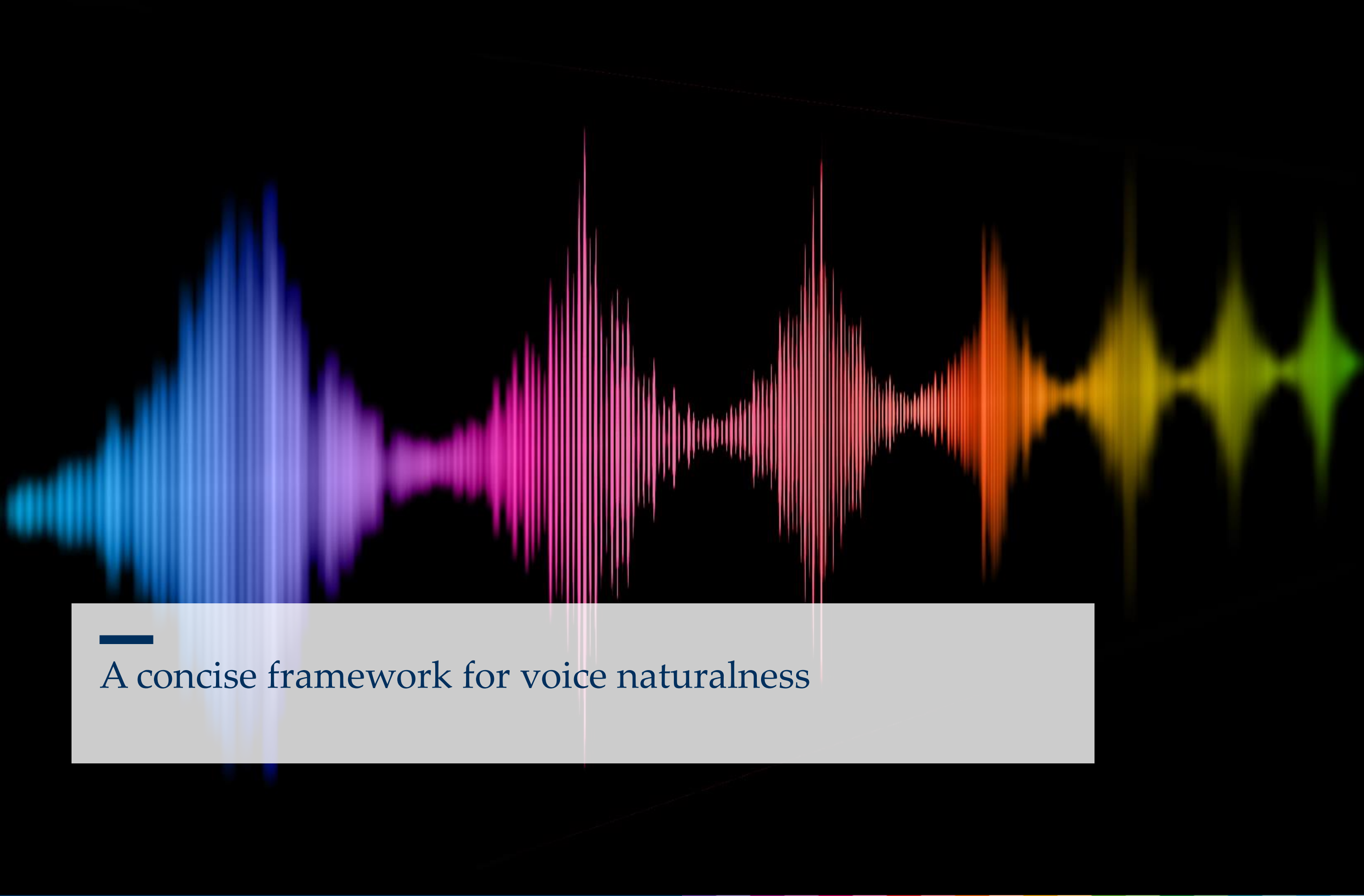
- Literature search (26.04.2024): naturalness AND voice + human-likeness AND voice + cited references

- Inclusion criteria:
  - Naturalness/Human-likeness was either measured or manipulated
  - Language: English
  - Published in journal or conference contribution
  - Human performance/perception data
  - Quantitative empirical analysis
  - OR integrative works of such works
  - Spoken voice (no singing) -> to be discussed
  - Preprint? -> to be discussed
  - Research paper (not presentation of a dataset)

-> 66 paper

Box 1: A field in numbers (mini literature review)

- Year range (objective)
- Voice type: synthetic, human-pathological, human-manipulated, mixture (semi-objective)
- Use Naturalness or synonyms in keywords (semi-objective)
- Rating data, performance measures, neuronal measures (objective)
- Provide an explicit definition of naturalness (objective)
- Use which kind of conceptualization from our framework (subjective)
- Citations (objective)
- Synonyms for naturalness (subjective)



A concise framework for voice naturalness

(1) Human-likeness-based naturalness		(2) Deviation-based naturalness
Human-likeness i.e. resemblance to real human voice	Conceptualization	Deviation from an exemplar/reference/ expectation/model that represents maximum naturalness
„Does this voice sound like a real human speaker?“	Example definitions for participants or readers	“Does this voice sound distorted?”/ “Does this voice sound untypical/rare/ unexpected?”

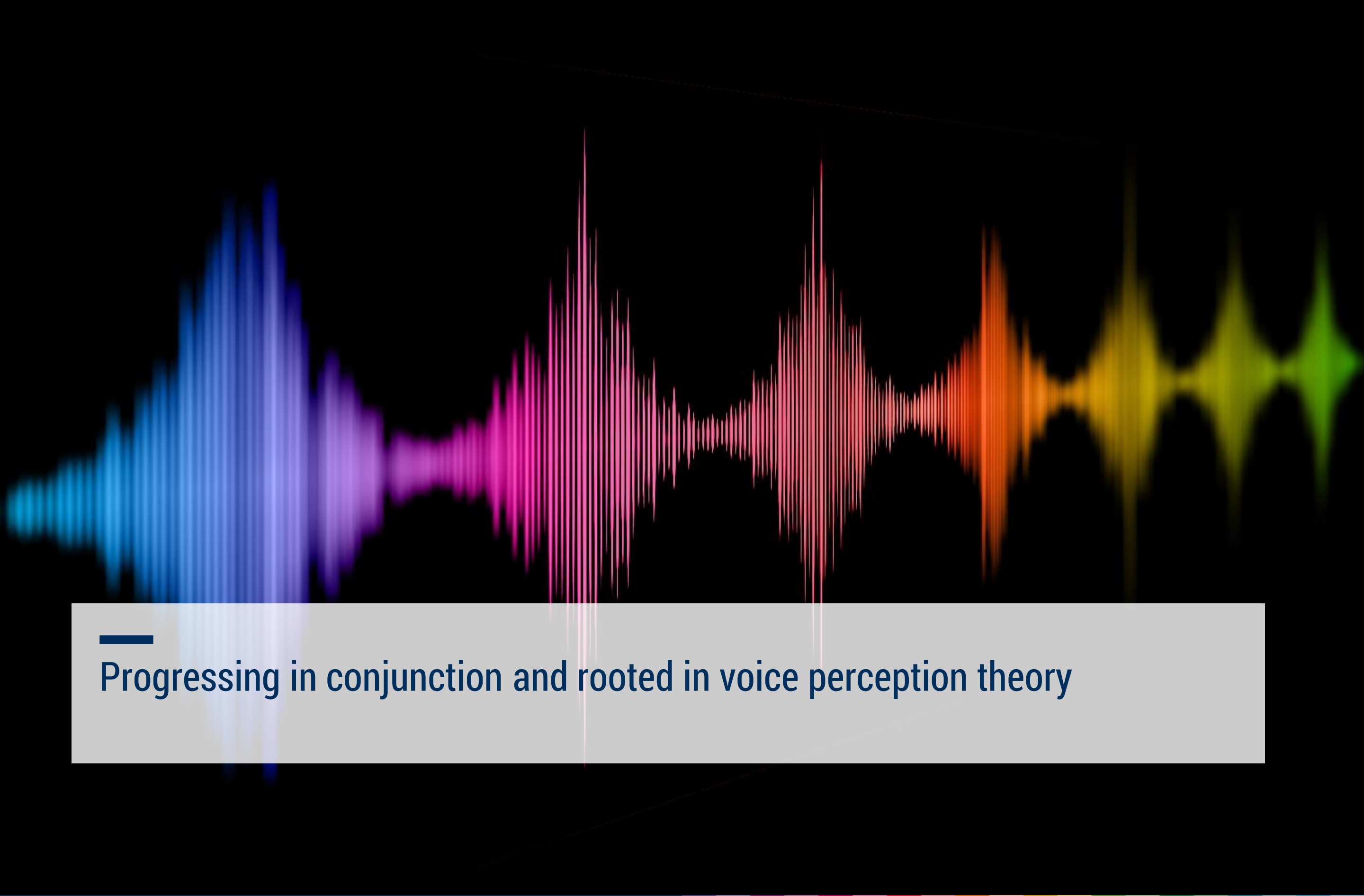
Differentiation from other concepts

Distinctiveness / Typicality

Voice Pathology

Authenticity/Genuineness

...to be discussed!



Progressing in conjunction and rooted in voice perception theory

Box 2: 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

Understanding voice perception

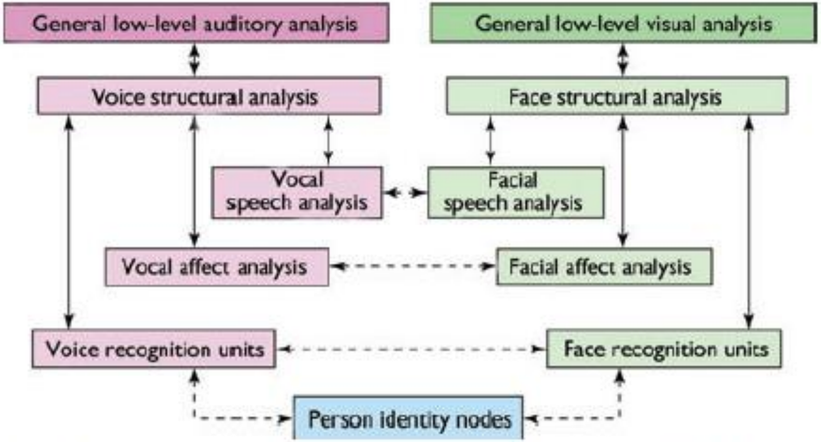
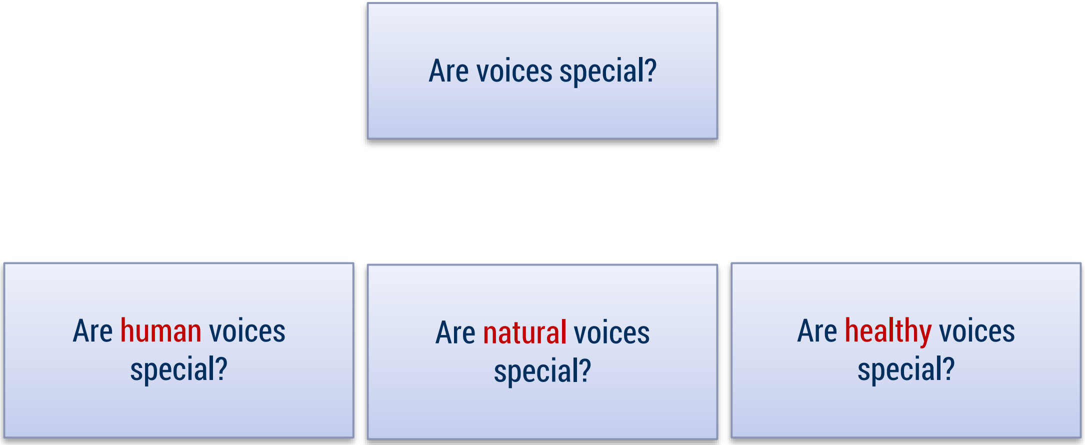


Figure 1. A model of voice perception. Reproduced from Belin *et al.* (2004). After a stage of voice structural encoding restricted to vocal sounds, three partially dissociable functional pathways are proposed to process the three main types of vocal information: speech, identity, and affect. These pathways are analogous to and interacting with equivalent functional pathways involved in facial processing.

Belin et al (2011)

Understanding voice perception



Are voices special?

Are **human** voices special?

Are **natural** voices special?

Are **healthy** voices special?

(1) Human-likeness-based naturalness

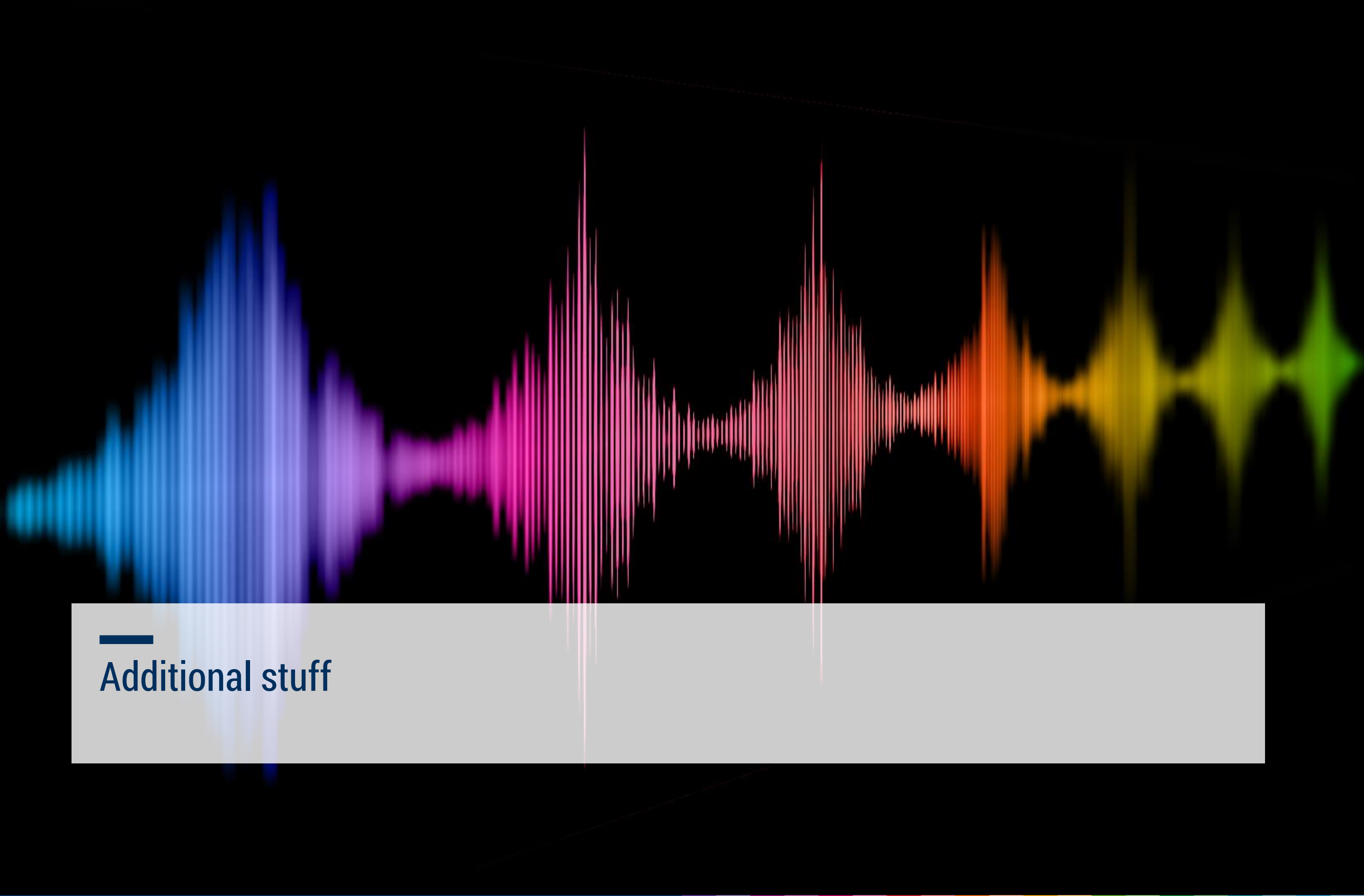
(2) Deviation-based-naturalness

Human-likeness i.e. resemblance to real human voice

Conceptualization

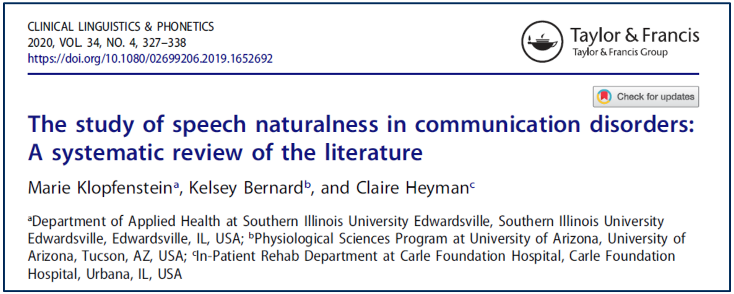
Deviation from an exemplar/reference/ expectation/model that represents maximum naturalness

- Is the perception between human and non-human voices categorical?
- Do similar rules/patterns apply to naturalness variation within human voices compared to human/non-human voices?
- How does naturalness affect the processing in the brain?
- Which role does experience play? / adaptability
- Does reduced naturalness due to stimulus manipulation have implications for ecological validity?



Additional stuff

How about any integrative works?



CLINICAL LINGUISTICS & PHONETICS
2020, VOL. 34, NO. 4, 327–338
https://doi.org/10.1080/02699206.2019.1652692

Taylor & Francis
Taylor & Francis Group

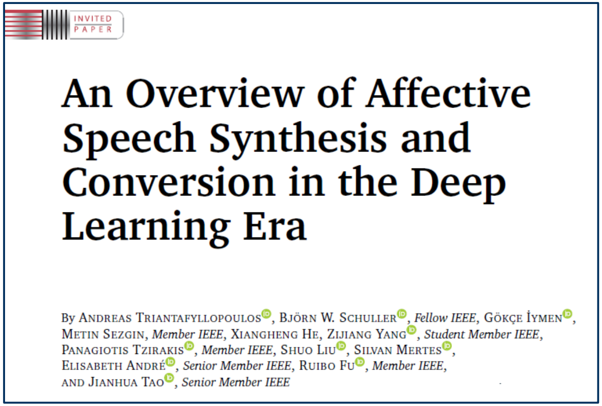
Check for updates

The study of speech naturalness in communication disorders: A systematic review of the literature

Marie Klopfenstein[a], Kelsey Bernard[b], and Claire Heyman[c]

[a]Department of Applied Health at Southern Illinois University Edwardsville, Southern Illinois University Edwardsville, Edwardsville, IL, USA; [b]Physiological Sciences Program at University of Arizona, University of Arizona, Tucson, AZ, USA; [c]In-Patient Rehab Department at Carle Foundation Hospital, Carle Foundation Hospital, Urbana, IL, USA

2020

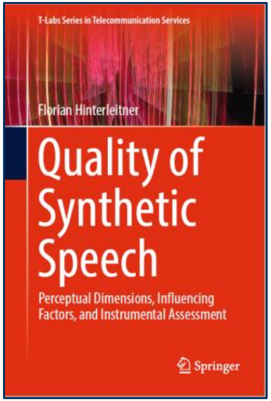


INVITED PAPER

An Overview of Affective Speech Synthesis and Conversion in the Deep Learning Era

By Andreas Triantafyllopoulos, Björn W. Schuller, *Fellow IEEE*, Gökçe İymen, Metin Sezgin, *Member IEEE*, Xiangheng He, Zijiang Yang, *Student Member IEEE*, Panagiotis Tzirakis, *Member IEEE*, Shuo Liu, Silvan Mertes, Elisabeth André, *Senior Member IEEE*, Ruibo Fu, *Member IEEE*, and Jianhua Tao, *Senior Member IEEE*

2023



T-Labs Series in Telecommunication Services

Florian Hinterleitner

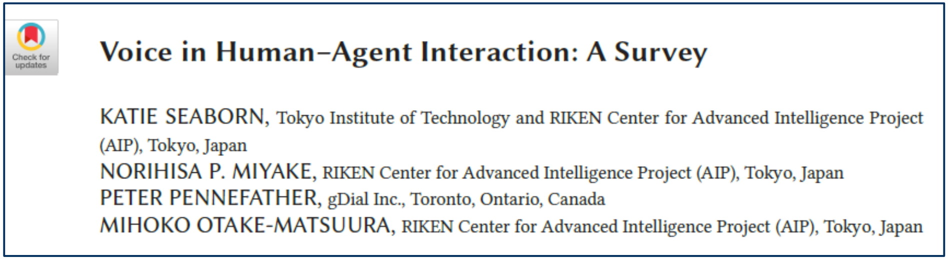
Quality of Synthetic Speech

Perceptual Dimensions, Influencing Factors, and Instrumental Assessment

Springer

2017

2021

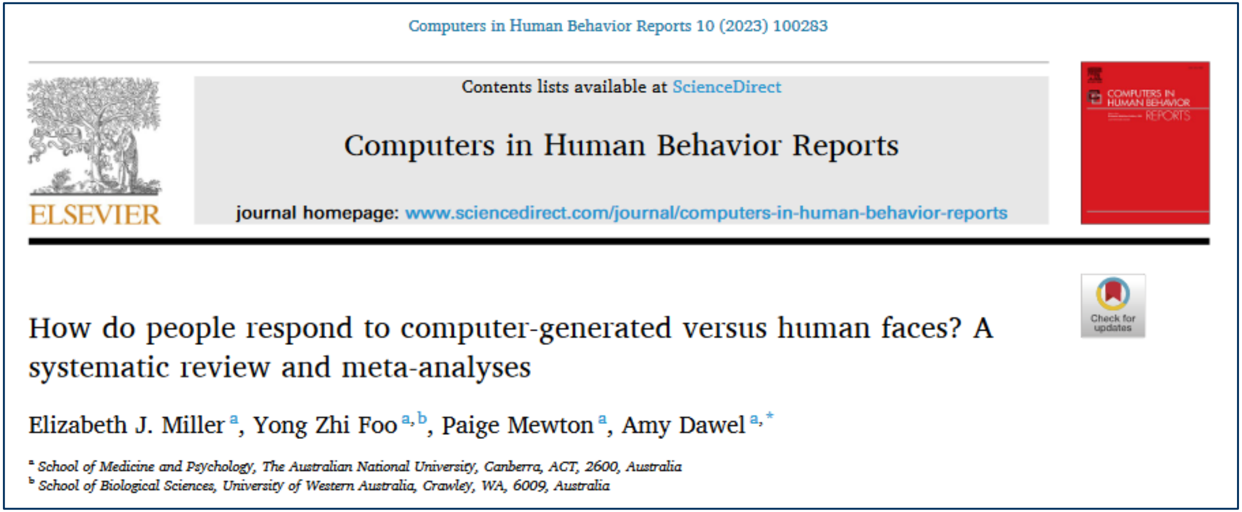


Check for updates

Voice in Human–Agent Interaction: A Survey

KATIE SEABORN, Tokyo Institute of Technology and RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan
NORIHISA P. MIYAKE, RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan
PETER PENNEFATHER, gDial Inc., Toronto, Ontario, Canada
MIHOKO OTAKE-MATSUURA, RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan

How about any integrative works?



Computers in Human Behavior Reports 10 (2023) 100283

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

How do people respond to computer-generated versus human faces? A systematic review and meta-analyses

Elizabeth J. Miller [a], Yong Zhi Foo [a,b], Paige Mewton [a], Amy Dawel [a,*]

[a] School of Medicine and Psychology, The Australian National University, Canberra, ACT, 2600, Australia
[b] School of Biological Sciences, University of Western Australia, Crawley, WA, 6009, Australia

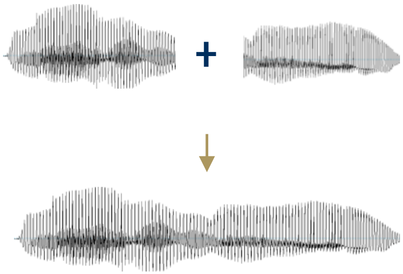
2023

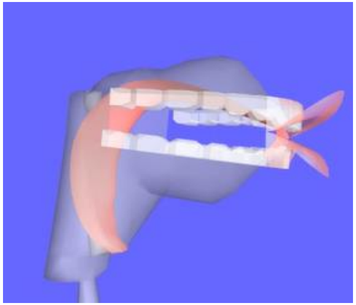
Overview over voice synthesis methods

Concatenative synthesis

Articulatory synthesis

Statistical parametric speech synthesis





https://www.vocaltractlab.de/

Hidden Markov Models

Deep Learning Methods

Text-to-Speech (TTS)

https://www.ibm.com/products/text-to-speech