The Influence of Voice Similarity on Cognitive Processes

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Oliver Jaggy geb. Pawelko

aus Mühlacker

Tübingen

2025

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	03.11.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Stephan Schwan
2. Berichterstatter:	PD Dr. Jürgen Buder

Table of Contents

Summary

Text-to-speech (TTS) systems are ever more embedded in our daily interactions: In social media and other content platforms, TTS technologies are used to create voiceovers or to narrate videos, and they enable the creation of sophisticated voice assistants integrated into (digital) devices. The proliferation of TTS systems is due to their ever-improving ability to generate audio that mimics human speech. Moreover, over the past few years, they also gained the ability to clone the voice of a target individual. This specific capability raises the question of how voices similar to the listeners influence cognitive processes.

Existing research has consistently demonstrated that subjects that share attributes with ourselves are generally perceived as more favorable. The *Similarity-Attraction* theory has not only been validated across various domains – including physical appearance, attitudes, ethnicity, and origin – but research has also demonstrated that similarity alters our perception and influences attitudes and behavior. This is especially the case when we process information peripherally or heuristically – for example when we lack motivation or relevant skills. In such circumstances, similarity can serve as an influential cue. In line with the *Computers Are Social Actors* hypothesis, which suggests that humans extend social rules and expectations to interactions with computers and similar devices, it is plausible to infer that artificial voices perceived as similar to our own might evoke comparable reactions. Despite this theoretical foundation, research on the effects of voice similarity remains relatively scarce. In 12 experiments, this thesis investigates the effects of voice similarity on trait evaluation, information reception, and decision-making.

In the first manuscript, I tested whether a speaker recognition system trained with deep learning methods can predict human similarity judgments and whether the derived similarity of voices affects trait evaluation (Chapter 2). In all five experiments in this series, I employed an open-source speaker recognition system that generates *d*-vectors based on brief audio samples. D-vectors are numerical representations of the unique auditory features of a speaker and can be used to assess the similarity of two different speakers, for example, by calculating the cosine similarity value between their feature vectors. To get a German model of the SV system, I combined several datasets from within the web and trained the SV with audio samples from approximately 10.000 speakers. The first three experiments revealed a modest yet significant correlation between the system's cosine similarity values and human similarity judgments. Therefore, I demonstrated that the cosine values could be utilized as a proxy for human-perceived similarity. In the fourth experiment, I did not investigate the consequences of similarity but used the cosine values to address the question of whether average voices

elicit higher likeability and trustworthiness ratings. Existing research has revealed such a *Beauty-in-Averageness* effect for faces and several other stimuli, but there is little research on average voices. However, my findings did not support its presence in voice perception, showing no link between vocal averageness and trait evaluation. Conversely, the final experiment revealed a positive relationship between trustworthiness and likeability judgments with the degree of voice similarity to the listener. Consequently, the results indicate that subjects with a similar voice are evaluated more favorably.

In the second series of experiments, I investigated whether voice similarity affects decision-making (Chapter 3). For the first experiment, I adopted a standard probabilistic inference paradigm that has been used in various studies on decision-making. Participants are incentivized to find treasures hidden behind three houses and receive guidance from three advisors, each making predictions about the treasure's location. Past research has demonstrated the inability of participants to make optimal decisions, especially when two advisors with low predictive ability make coherent predictions that differ from the predictions of the advisor with high predictive ability. This phenomenon highlights the influence of heuristic or peripheral cues on decision-making – here, the majority rule. Before conducting the experiment, I gathered audio recordings from approximately 600 individuals articulating phrases such as 'There is a treasure.' and 'There is a spider.' In the initial experiment, I manipulated whether the advisor with a high prediction capability possessed a voice similar to the participant's or one of average similarity. Contrary to expectations, voice similarity did not significantly increase the likelihood of following the predictions of the high-capability advisor. In the two subsequent experiments, I simplified the setup, incorporated only two advisors with equal prediction capabilities, and varied whether the advisor with a similar voice was on the top or the bottom of the display. In both experiments, I found an interaction effect between voice similarity and the advisor's position on how often the participants followed the advice of the similar advisor.

My final experimental series (Chapter 4) aimed to test the impact of voice similarity on information processing, especially on truth perception. Previous studies have shown that peripheral cues, such as familiarity with a statement, can enhance the likelihood of perceiving that statement as accurate. This phenomenon is called *the Illusory-Truth* effect and can be caused, for instance, by presenting a statement repeatedly. I investigated whether voice similarity modulates the Illusory-Truth effect. The first experiments were designed to validate the experimental materials and assess whether the Illusory-Truth effect emerges with auditory statements generated by a TTS system. In the final experiment in this series, we presented

trivia statements (‘Paris is the capital of Germany.’) either once or twice and either with a similar or an average voice. The results revealed a marginally significant interaction between the repetition of a statement and the voice similarity on truth evaluation. Moreover, false trivia statements were more often judged as true when presented twice and when a voice similar to the participant’s voice delivered the statement. Therefore, the results revealed a significant effect of voice similarity on truth perception.

Overall, my thesis introduced the use of d-vectors as a new approach to investigate the effects of voice similarity on various cognitive processes. My results consistently demonstrated a (subtle) effect of voice similarity on trait evaluation, decision-making, and truth perception and raised critical questions about the possible integration of user-tailored voices into TTS systems.

Zusammenfassung

Text-to-Speech-Systeme (TTS) sind zunehmend in unsere alltägliche Kommunikation integriert. Sie werden auf sozialen Medien und anderen digitalen Plattformen genutzt, um etwa Videos zu vertonen oder Inhalte auditiv aufzubereiten. Darüber hinaus ermöglichen sie die Entwicklung leistungsstarker Sprachassistenten, die in einer Vielzahl von Geräten eingebettet sein können. Der rasante Anstieg ihrer Verbreitung ist auch auf die stetig verbesserte Fähigkeit zurückzuführen, synthetische Sprache zu erzeugen, die der menschlichen Stimme immer näherkommen. In den letzten Jahren wurde TTS-Systemen darüber hinaus die Fähigkeit verliehen, Stimmen gezielt zu „klonen“ – also eine synthetische Stimme zu erzeugen, die einer bestimmten Person stark ähnelt. Diese Entwicklung wirft die grundlegende Frage auf, wie sich Stimmen, die der eigenen ähnlich sind, auf kognitive Prozesse auswirken.

Zahlreiche Studien haben gezeigt, dass Objekte oder Personen, die dem Selbst ähnlich sind, im Allgemeinen positiver wahrgenommen werden. Die Similarity-Attraction-Theorie ist in unterschiedlichen Kontexten – darunter physische Erscheinung, Einstellungen, ethnische Zugehörigkeit und Herkunft – empirisch bestätigt worden. Ähnlichkeit kann die Wahrnehmung, Einstellung und das Verhalten gegenüber einer Person insbesondere dann beeinflussen, wenn Informationen heuristisch oder peripher verarbeitet werden, etwa bei eingeschränkter Motivation oder begrenzten kognitiven Ressourcen. In solchen Fällen fungiert Ähnlichkeit als wirksamer Hinweisreiz. Die „Computers-are-Social-Actors“-Hypothese postuliert, dass Menschen soziale Regeln und Erwartungen auch auf Interaktionen mit Computern oder digitale Akteure übertragen. Daraus ergibt sich die Annahme, dass auch künstliche Stimmen, die als der eigenen ähnlich empfunden werden, vergleichbare Reaktionen auslösen könnten. Trotz dieser theoretischen Grundlage existiert bislang nur eine begrenzte Anzahl empirischer Studien zur Auswirkung von Stimmähnlichkeit. Die vorliegende Dissertation widmet sich daher in insgesamt zwölf Experimenten der Frage, welchen Einfluss wahrgenommene Stimmähnlichkeit auf Urteile, Informationsverarbeitung und Entscheidungsverhalten hat.

In der ersten Studie (Kapitel 2) habe ich untersucht, ob ein mit Deep-Learning-Methoden trainiertes Sprechererkennungssystem menschliche Ähnlichkeitsurteile vorhersagen kann und ob Personen, die eine, der eignen Stimme ähnliche Stimme besitzen, positiver bewertet werden. In allen fünf Experimenten dieser Reihe kam ein Open-Source-System zur Sprecheridentifikation zum Einsatz, das sogenannte *d*-Vektoren generiert, die als numerische Repräsentationen der individuellen akustischen Merkmale eines Sprechers angesehen werden

können. Die Ähnlichkeit zweier Stimmen kann auf Basis dieser Vektoren beispielsweise durch Berechnung deren Kosinusähnlichkeitswerts, ermittelt werden. Um ein deutschsprachiges Modell zu trainieren, kombinierte ich mehrere, öffentlich zugängliche Datensätze und trainierte das System mit ca. 10.000 Sprecher und Sprecherinnen. Die ersten drei Experimente zeigten eine signifikante, wenngleich moderate Korrelation, zwischen den Kosinunsähnlichkeitswerten und den menschlichen Ähnlichkeitsurteilen – was die Validität von *d*-Vektoren als Maß für die menschliche Wahrnehmung von Stimmähnlichkeit unterstreicht. In Anlehnung an den vor allem aus Untersuchungen zur ästhetischen Wirkung von Gesichtern bekannten *Beauty-in-Averageness*-Effekt, bei denen sich zeigte, dass durchschnittliche Gesichter positiver bewertet werden, wurde im vierten Experiment untersucht, ob auch durchschnittliche Stimmen positiver wahrgenommen werden. Dies konnte jedoch nicht gezeigt werden. Hingegen zeigte sich im abschließenden Experiment dieser Studie, dass SprecherInnen umso vertrauenswürdiger und sympathischer bewertet wurden, je ähnlicher deren Stimme der Stimme der Versuchspersonen war.

Die zweite Studie (Kapitel 3) widmete sich der Frage, ob Stimmähnlichkeit Entscheidungsprozesse beeinflussen kann. Im ersten Experiment wurde auf ein etabliertes probabilistisches Inferenzparadigma zurückgegriffen, das bereits vielfach zur Untersuchung menschlichen Entscheidungsverhaltens eingesetzt wurde. Die Teilnehmenden sollten entscheiden, hinter welchem von drei Häusern sich ein Schatz befindet, wobei sie auf Vorhersagen dreier RatgeberInnen zurückgreifen konnten. Frühere Untersuchungen haben gezeigt, dass Personen suboptimale Entscheidungen treffen, insbesondere wenn zwei weniger kompetente RatgeberInnen miteinander übereinstimmende Vorhersagen abgeben, die im Widerspruch zur Vorhersage eines einzelnen, kompetente Ratgebers stehen. Dies stellt ein klassisches Beispiel für den Einfluss heuristischer Hinweisreize – in diesem Fall des *majority bias* – dar. Vor der Durchführung des Experiments wurden Sprachaufnahmen von ca. 600 SprecherInnen gesammelt, die Sätze wie „Da ist ein Schatz“ oder „Da ist eine Spinne“ äußerten. In der ersten Studie dieser Reihe wurde variiert, ob der kompetente Ratgeber eine Stimme mit hoher oder durchschnittlicher Ähnlichkeit zur Stimme der jeweiligen Versuchsperon hatte. Entgegen der Erwartung erhöhte die Stimmähnlichkeit nicht die Wahrscheinlichkeit, den Empfehlungen des kompetenten Ratgebers zu folgen.

In zwei nachfolgenden Experimenten wurde das Paradigma vereinfacht: Es wurden jeweils zwei gleich kompetente RatgeberInnen eingesetzt, deren Stimmen entweder ähnlich oder durchschnittlich ähnlich waren. In beiden Studien zeigte sich zunächst keine Auswirkung der Stimmähnlichkeit auf die Entscheidungen der Teilnehmenden. In explorativen Analysen

zeigte sich jedoch, dass sich, wenn die horizontale Position des Ratgebers auf dem Bildschirm (oben vs. unten) berücksichtigt wird, sich in den beiden letzten Experimenten ein Interaktionseffekt zwischen Stimmähnlichkeit und räumlicher Position zeigt: Versuchspersonen folgten hierbei häufiger den Vorhersagen von ähnlichen Ratgebern, wenn sich diese über dem Ratgeber mit einer durchschnittlich ähnlichen Stimme befand.

Die letzte Versuchsreihe (Kapitel 4) zielte darauf ab, die Auswirkung von Stimmähnlichkeit auf die Informationsverarbeitung, insbesondere auf die Beurteilung der Korrektheit von Aussagen, zu testen. Frühere Studien haben gezeigt, dass periphere Hinweisreize, wie z. B. die Vertrautheit mit einer Aussage, die Wahrscheinlichkeit erhöhen können, dass diese Aussage als wahr wahrgenommen wird. Dieses Phänomen wird als *Illusory-Truth-Effekt* bezeichnet und kann z. B. durch wiederholte Präsentation einer Aussage hervorgerufen werden. Ich habe untersucht, ob die Ähnlichkeit der Stimme den Illusory-Truth-Effekt verstärkt. Die ersten Experimente dienten der Validierung des Versuchsmaterials und der Feststellung, ob der Illusory-Truth-Effekt bei auditiver Präsentation von Aussagen auftritt, wenn diese von einem TTS-System generiert werden. Im letzten Experiment dieser Reihe wurden deklarative Aussagen wie „Paris ist die Hauptstadt von Deutschland.“ entweder einmal oder zweimal in einer ähnlichen oder einer durchschnittlichen Stimme präsentiert. Die Ergebnisse zeigten eine marginal signifikante Interaktion zwischen der wiederholten Präsentation einer Aussage und der Ähnlichkeit der Stimme auf die Wahrheitszuschreibung. Außerdem zeigte sich in beiden untersuchten Haupteffekten, dass falsche Aussagen häufiger als wahr bewertet wurde, wenn sie zweimal präsentiert wurden und wenn eine Stimme, die der Stimme des Teilnehmers ähnlich war, die Aussage vortrug. Die Ergebnisse zeigten also einen signifikanten Effekt der Stimmähnlichkeit auf die Wahrheitszuschreibung

Insgesamt führt die Dissertation *d*-Vektoren als neues methodisches Werkzeug zur Untersuchung von Stimmähnlichkeit ein und zeigt ihr Potenzial zur Erklärung subtiler Effekte auf zentrale kognitive Prozesse. Die Ergebnisse belegen konsistent, dass wahrgenommene Stimmähnlichkeit Bewertungen, Entscheidungsverhalten und Wahrheitsurteile beeinflussen kann. Damit wirft die Arbeit auch grundlegende Fragen zur zukünftigen Gestaltung personalisierter TTS-Systeme auf – insbesondere im Hinblick auf die Auswirkungen einer möglichen Integration von auf BenutzerInnen zugeschnittene Stimmen.

Chapter 1: General Introduction

Over the past decade, the development of artificial intelligence (AI) has seen remarkable advancements, and one can argue that the consequences can only be compared to major inventions like electricity (Lynch, 2017). Even though the full scope of AI's transformative potential on the global economy and societies has yet to be clear, they have begun to shape societal interactions significantly. For instance, ChatGPT, a generative AI based on a large language model, was only introduced to the public in November 2022 and has already achieved a milestone by being the first AI recognized in Nature's 2023 list of the most influential scientists (Nature, 2023). This achievement not only underscores the growing significance of AI and its capabilities but also the increasing acceptance of incorporating AI into our daily lives.

Such an integration can also be observed with text-to-speech (TTS) systems that are increasingly able to generate human-like speech. They are, for instance, essential accessibility tools for individuals with visual impairments or reading difficulties. Integrated into navigation systems, they provide spoken directions to drivers, which allows them to focus their attention on the road. And as a tool for content creators, they can help produce voiceovers without needing professional voice actors, which is often used on social media platforms. Among the most impactful applications, however, is the combination of TTS with generative AI, for instance in voice assistants.

Since Apple introduced Siri, the first voice assistant available to the consumer market in 2011, many companies have developed devices integrating voice assistants. Today, almost every smartphone, computer, and smart device is equipped with a built-in, proprietary voice assistant. The continuous advancements in generative AI and TTS technologies, coupled with their growing accessibility, point to an inevitable increase in user interactions. However, the expanding use of voice assistants raises important considerations. Voice assistants are not standalone products designed to monetize themselves. Instead, they are tools to promote a broader collaborative strategy. Alphabet, Amazon, and, to a lesser extent, Apple are so-called platform businesses (Y. Zhao et al., 2020) that use user data to generate revenue: Alphabet, for example, to "deliver relevant ads at just the right time" (Alphabet, 2020), and Amazon to achieve its vision "to be Earth's most customer-centric company" (Amazon, 2020). Voice assistants are primarily used to play music, search the web, and operate IoT devices, such as smart light bulbs, thermostats, and cameras (Ammari et al., 2019). Logging those interactions generates various information about the user. Although some of this information was already accessible to some companies, they also get information about the customers' voices. Like

fingerprints, the human voice can be used to distinguish individuals from one another with a high degree of accuracy. However, they can also be used as additional information in TTS systems to modify the generated voice output. This opens up the possibility of adapting the voice used by voice assistants to the user's voice, for instance, to make it more similar and potentially more likeable and trustworthy.

Even if the use of customized voices in voice assistants may be one of the most obvious applications, customized voices hold potential in various other domains. For instance, they could be employed in election campaigns or propaganda. Similarly, in educational settings, voice adaptation could enhance engagement and facilitate learning. To identify possible consequences of such adaptations, this thesis investigates if voice similarity affects cognitive processes, particularly trait evaluation, truth-perception, and decision-making.

In the following sections of the introduction, I will elaborate on the motivation of my research, the technical methods used, and the phenomena we selected to study the similarity effects. I then present three sets of experiments, each described in separate manuscripts: In Chapter 2, "AI-determined similarity increases likability and trustworthiness of human voices", I introduce cosine similarity values as a measure of human-perceived voice similarity and examine whether prototypical or similar voices alter trait evaluations. Chapter 3, "The Impact of Voice Similarity on Decision-Making: Do We Follow Advisors with Similar Voices?", addresses whether voice similarity influences decision-making processes. Whether information is more likely to be perceived as true when presented by a similar voice is examined in Chapter 4, "Can I Believe My Voice?". In Chapter 5, the concluding section, I summarize the findings, discuss theoretical and practical implications, and reflect on the strengths and limitations of my dissertation project and possible directions for future research.

Evidence of a Similarity Attraction Effect

The concept of people having an inherent preference for those who resemble them is a notion that has been present since the early days of contemporary social and psychological research (e.g., Richardson, 1940). However, it truly captured the attention of the academic community with the seminal work of Donn Byrne on the *Similarity-attraction hypothesis* (Byrne, 1961; Byrne et al., 1967; Byrne & Griffitt, 1973). Byrne and colleagues discovered that individuals are more attracted toward those who share their attitudes, and they explained this phenomenon through reinforcement theory. When we encounter someone who mirrors

our views or attitudes, it validates our self-concept, leading to positive emotions. This validation reinforces our liking for the person who provided it (Byrne & Griffitt, 1973; Clore & Byrne, 1974).

While his research within the similarity-attraction paradigm, how he initially labeled it, concentrated on attitudinal similarities (for reviews, see Montoya et al., 2008; Montoya & Horton, 2013), subsequent studies broadened the scope to include various other dimensions of similarity. This includes incidental similarities such as shared birthdays (Jiang et al., 2010), first names, or even fingerprints (Burger et al., 2004) but also extends to visually (Anderson & McMillion, 1995) and auditorily perceived ethnic similarities (Ivanič et al., 2014; Tsalikis et al., 1991), political beliefs and affiliations (Nelson & Garst, 2005; Roth et al., 2020), shared values (Lewis & Walsh, 1980; Silvia, 2005), personality traits (Carli et al., 1991), and social backgrounds (Feng & MacGeorge, 2010). Indeed, even similarity in music preferences has demonstrated to effect social attraction where music preferences act as a cue of a similar value orientation, which leads to social attraction (Boer et al., 2011). In the context of voice similarity, research has found that voices similar to one's own are perceived as more attractive (Peng et al., 2019).

The persuasive power of similarity is well-documented, with similar individuals often regarded more convincing (Faraji-Rad et al., 2015). This effect extends across various contexts, from the influence of similarity in online dating (Finkel et al., 2012) to its role in workplace dynamics, where perceived similarities can predict employer attractiveness (Devendorf & Highhouse, 2008). In personal relationships, the presence of similarity is linked to the formation of friendships (Urberg et al., 1998) and satisfaction within same-sex friendships (Morry, 2005). It shapes our opinions (Berscheid, 1966), judgments (Gino et al., 2009), attitudes (Jiang et al., 2010), compliance (Burger et al., 2004), shopping behavior (Fu et al., 2018), health-related self-efficacy (Phua, 2016), and applicant selection (Orpen, 1984; Roth et al., 2020).

Similarity not only affects interactions between humans, but also interactions between humans and machines. For example, greater similarity can increase the willingness of humans to cooperate with robots (You & Robert Jr., 2018) and increase their persuasiveness (Winkle et al., 2019). Along similar lines, Narayanan et al. (2023) found that an AI that exhibits comparable ethical principles is more likely to persuade humans.

However, why can seemingly irrelevant information, like the similarity of a person, affect us in such diverse ways?

Explanations for Similarity Effects

Byrne's explanation, which attributes attraction to the reinforcement we feel when others affirm our beliefs, is just one perspective among many theories from various fields that could give insights into the mechanisms driving similarity effects. I will provide an overview of these underlying mechanisms in the following sections.

Emotions

As mentioned by Byrne (1961) the relationship between similarity and attraction can be explained by the influence of emotions. However, the association between similarity and positive emotions could not only be grounded in a need for self-affirmation. Indeed, irrespective of similarity, emotions can play a significant role in guiding many important life choices (for review, see Loewenstein et al., 2003) and can influence perception, attention, memory, information processing, and evaluative judgements (Bless, 2003; Brosch et al., 2013; Vuilleumier, 2005). Winkielman and colleagues (2003) suggested that familiarity typically correlates with positive feelings because unfamiliar scenarios or objects are often perceived as potentially dangerous. Since similar others can elicit a sense of familiarity, the positive feelings associated with familiarity, could render the opinions and advice of similar others more influential (Lerner et al. 2015).

Implicit Egotism

Another explanation for similarity effects can be found in the concept of *Implicit Egotism* (Pelham et al., 2002; Pelham et al., 2005). It suggests an unconscious preference for subjects and objects that resemble ourselves or are associated with us. For example, such a preference has been demonstrated by the *name-letter effect*: a preference for letters that belong to one's name, independent of other letter characteristics (Nuttin, 1985). This letter preference can also affect important life decisions such as career choice and place of residence: A person named Dennis is disproportionately likely a dentist in Denver (Pelham et al., 2002). It is assumed that the positive associations and biased behaviors are unconscious because they are linked to the *Implicit Self-Esteem*, which is defined as “an automatic, overlearned, and nonconscious evaluation of the self that guides spontaneous reactions to self-relevant stimuli” (Bosson et al., 2000, p. 631). This implicit evaluation of the self is part of a more generalized theory of *Implicit Social Cognition* that posits that social behavior is often unconsciously motivated or influenced (Greenwald & Banaji, 1995). The reason that self-

associated entities lead to positive attitudes and behaviors towards them – which is not explained by the theory of Implicit Social Cognition itself – could be due to the self's attempt to maintain a positive self-image (Dunning, 1999; Gosling et al., 1998; Kruger & Dunning, 1999). Thus, the positive feelings and actions toward entities related to ourselves could come from a natural desire to keep and boost our self-esteem.

Even though the concept of Implicit Egotism has received some criticism – notably regarding the cause of the name-letter effect (Pelham & Carvallo, 2011; Simonsohn, 2011) – the empirical findings underpins the general tendency to pursue positive self-evaluations (for a review of the underlying mechanisms, see Pyszczynski et al., 2004). This causes several biases regarding self-enhancement and self-verification, for instance, the well-known *self-serving bias*, which describes the tendency of people to attribute desired outcomes internally and undesired outcomes externally (Miller & Ross, 1975; Shepperd et al., 2008). Another example is the *mere ownership effect*: simply owning an object leads individuals to evaluate it more positively (Beggan, 1992). Despite being often discussed in the context of the more economic-centric *endowment effect*, which emphasizes loss aversion's role in the attribution process, the mere ownership effect highlights the disposition towards self-favorable assessments that can spill over to objects associated with oneself (Morewedge & Giblin, 2015; Shu & Peck, 2011). Taken together, the concept of Implicit Egotism indicates that we might instinctively favor those who are similar to us, because a subconscious association with ourselves could activate self-enhancing biases.

An explanation related to implicit egotism, assumes that the similarity effect occurs because the valence and weight of inferred information about a person influences our evaluation of the other person (Ajzen, 1974; Kaplan & Anderson, 1973; Montoya & Horton, 2013). Thereby, attraction is influenced by the positive or negative information derived from similarities or differences in attitudes, personality traits or other attributes. Kaplan and Anderson (1973) state that similar attitudes lead to attraction not as a direct response, but through the expectation of other positive personality aspects of the similar other. This is due to individuals own positive self-assessment, which makes them assume that the similar other must have additional positive traits (Ajzen, 1974; Stalling, 1970). This is underpinned by the finding that the significance of an attribute in shaping attraction is determined by the amount and salience of the information it provides about the other person, with more informative or salient attributes having a greater influence on attraction or repulsion (E. E. Jones & Davis, 1965).

Social Influences

Beyond being the result of implicit egotism, similarity effects may also stem from humans' inherent social nature. An underlying mechanism could be found, for instance, in ingroup favoritism. *Social Identity Theory* (Tajfel & Turner, 1986) posits that we prefer and exhibit more positive behavior toward people who we perceive as similar or belonging to the same group. These ingroup biases (Mullen et al., 1992) stems from the belief and the desire that the own group is superior compared to other groups. Since group identity can easily emerge – as demonstrated by the minimal group paradigm (Brown, 2000; Tajfel et al., 1971) – the similarity of another person could serve as a clue for a common group belonging. Following the *Self Categorization Theory* (Turner et al., 1987), which is an elaboration and refinement of the Social Identity Theory (Hornsey, 2008), one can argue that the similarity of another person guides the focus on the own social identity and leads to socially desirable behavior.

Another explanation could be grounded in peoples need to be connected to others. The *Belongingness Hypothesis* (Baumeister & Leary, 1995) articulates a fundamental and universal human drive to establish and sustain positive, and meaningful interpersonal connections. Even though Baumeister and Leary do not mention the psychoanalytic contributions of Heinz Kohut, it should be mentioned that he had in fact already identified the need to belong as one of the three primary self-needs (Kohut, 1984; see also Lee & Robbins, 1995). Jiang et al. (2010) argue that the effects of (incidental) similarity can be linked to this need for belongingness. They suggest that similarities can act as subtle cues promoting social connectedness, for instance, in the initial stages of a relationship. To preserve the perceived connection between oneself and a similar other, and therefore, to satisfy the need for connectedness, people behave socially desirable.

Expanding on this, the tendency to extend perceived superficial similarities to deeper shared characteristics can lead to an exaggerated sense of congruence, such as assuming that an advisor with similar superficial characteristics will also have similar preferences, making their advice more relevant or diagnostic (Hovland et al., 1953). Filieri and colleagues (2018) link this phenomenon to the *Theory of Social Comparison* (Festinger, 1954). Festinger assumes that individuals have an intrinsic need to evaluate their opinions and abilities, and often rely on social comparisons when objective measures are absence. While the theory recognizes that there can be both upward and downward comparisons, it suggests that such evaluations are most likely to occur when the differences in individuals' opinions and abilities are minimal. Building on this premise, Filieri and colleagues (2018) argue that perceived

similarities between oneself and another not only facilitate these comparisons but also lead to the implicit assumption of shared needs and preferences between the individuals involved.

Somewhat on the intersection between emotions and socially driven behavior is an explanation provided by Condon and Crano (1988). They found that similarity attraction was mediated by people's belief of the partners evaluation of them. In other words, people are attracted towards similar others because they expect that they are liked more by them. This reciprocal sympathy suggests that the underlying mechanisms behind the dynamics of similarity attraction could go beyond mere social constructs and may have their roots in the more fundamental aspects of our biological constitution.

Biological Drivers

Another explanation related to ingroup favoritism focuses on the underlying biological mechanisms of similarity effects. In this context, the mechanism of attraction towards similar others is usually termed homophily and is often investigated in regard of its role in shaping social networks (McPherson et al., 2001). One explanation why similarity between social actors can positively affect their relation is the assumption that, in evolutionary terms, altruistic actions are often disadvantageous, except when the individuals involved are genetically related (Burnstein et al., 1994; Hamilton, 1964). Since a genetic relationship between individuals is often correlated with phenotypic similarity, the evolutionary pressure could have promoted a socially desirable behavior towards similar others.

Additionally, the *Niche Construction Theory* (Lewontin & Levins, 1997; Scott-Phillips et al., 2014) suggests that the environments created by organisms subsequently shape the evolutionary pressures for their descendants. Because cohesive social groups that share common traits or values are more effective at securing diverse resources, people who prefer to interact with similar others had a higher fit, which leads to a natural selection favoring those with a tendency to seek out similarity (Bahns et al., 2017).

Cognitive Explanations

Dual process models of reasoning have become popular beyond the scientific community since Daniel Kahneman published his book "Thinking, fast and slow" (Kahneman, 2011), where he suggested the existence of two distinct (cognitive) systems: a fast, automatic, and emotional system 1, and a slower, more deliberate, and logical system 2. However, the roots of dual process models trace back further and span multiple domains (for

reviews, see Evans, 2008; Gawronski & Creighton, 2013; for an overview, see Smith & DeCoster, 2000).

In the 1980s, Richard E. Petty and John Cacioppo introduced the *Elaboration Likelihood Model* (ELM), a pioneering paradigm explaining attitude formation and persuasion via two pathways: a central route, requiring analytical processing of information, and a peripheral route, influenced by superficial factors such as the attractiveness or credibility of the source of the information (Petty et al., 1986). Petty and Cacioppo believe that although people are motivated to have correct attitudes, "the amount and nature of issue-relevant elaboration in which people are willing or able to engage to evaluate a message vary with individual and situational factors" (Petty et al., 1986, p. 128). The ELM posits three intertangled variables affecting attitude change: argument quality, peripheral cues, and elaboration quality. However, personal relevance, motivation, and the ability to process arguments are additional crucial variables: If they are low, peripheral cues like source credibility or message likability become more significant, and the argument processing is less objective and more biased. This is particularly the case when the arguments are presented in an audio or video format where processing is less self-controlled than in written format (Chaiken & Eagly, 1983), so that simple cues "should be more powerful determinants of persuasion" (Petty et al., 1986, p. 162). Therefore, arguments presented in a similar voice could be processed more peripherally, and the similarity might serve as a peripheral cue that enhances the speaker's credibility or attractiveness, leading to more persuasive communication.

Similarly to the ELM, the *Heuristic-Systematic Model* (HSM) suggests two routes of how people process persuasive communication and form judgments: systematic and heuristic processing (Chaiken, 1980). Systematic processing entails thoroughly evaluating relevant information, which requires motivation and cognitive effort. In contrast, heuristic processing utilizes cues or simple rules, which demand less mental effort but can lead to less accurate or biased judgments. Heuristic processing is often engaged when we have low motivation, for instance, because the decision or judgment is perceived as trivial or when we lack the ability to process necessary information. However, the two processing modes can co-occur and can interact with one another: both systems could come to the same conclusions (additivity hypothesis), bolstering resulting judgments; heuristic cues could modulate systematic processing, especially in cases where information is ambiguous (bias hypothesis); conclusions drawn from systematic processing could contradict implications from heuristic processing, leading to a suppression of heuristically gained judgments (attenuation hypothesis),

particularly when motivation is high (Chen & Chaiken, 1999). When listeners hear a voice similar to their own, familiarity with it could serve as a heuristic cue to judge the speaker's credibility or likability and, therefore, influence decision outcomes.

Even though the *Continuum Model of Impression Formation* (Fiske & Neuberg, 1990; Neuberg & Fiske, 1987) does not suggest the existence of two distinct systems or processes, it is usually described as a dual process model. It posits that the information we have available about a person forms initial impressions based on the automatic categorization of individuals (into social groups). Therefore, the opinion arises based on salient cues, such as gender, age, and clothing. Without additional information about a person or the motivation to engage with them, the final impression will be built on this categorization process. If we believe the individual is relevant to us, or our goals, we shift from category-based processing to more individuated processing. In this case, a person's unique characteristics and behaviors are increasingly considered relevant. However, the extent to which we process person-specific information depends not just on the motivation and availability of information: Since the top-down, category-based processing requires less effort, we also need cognitive resources. When we hear a voice similar to our own, we might rely on category-based impressions, where the voice-similarity triggers a favorable automatic response based on in-group favoritism or assumed shared characteristics.

The *Motivation and Opportunity as Determinants* (MODE) model (R. H. Fazio, 1990; R. H. Fazio & Olson, 2014) describes two general categories of processes by which attitudes can guide behavior: a spontaneous and a deliberative process. It posits that if an attitude object is encountered and it relates to an existing cognitive representation of an attitude object, it comes to an automatic activation of the relevant attitude and the spontaneous process is engaged. This top-down process can either lead to immediate reactions or can serve as a filter shaping the perception of the object. If an attitude does not activate automatically, or if individuals are inclined to develop a nuanced opinion, they may engage in gathering evaluative information to form a more thorough judgment. However, this deliberative process is more effortful compared to the spontaneous process and requires the opportunity to access relevant information. Encountering a similar voice could lead to an automatic, positive response to that voice.

A dual process model with a broader scope compared to the previously mentioned is the *Reflective-Impulsive Model* (RIM) introduced by Strack and Deutsch (2004). It posits that social behavior, in general, is driven by two information processing systems: the reflective

system and the impulsive system. Even though these systems can operate in parallel, there is an asymmetry between those two: While the impulsive system is always active, the reflective system is only engaged if the processed information highly activates the impulsive system and receives significant attention. The impulsive system is characterized by its independence from cognitive resources. It initiates behaviors automatically via spreading activation, which can be triggered, for instance, by similarity or contiguity. In the reflective system, behavior results from a decision-making process where the value and likelihood of outcomes are evaluated. Once a decision is reached, the reflective system starts the chosen behavioral schemata. Since this process is cognitively demanding, it is only activated when enough resources are available. Since the Impulsive System of the RIM operates on associative links, a similar voice could activate positive associations or memories, leading to an unconscious preference for or trust in the speaker.

While findings from different experimental paradigms demonstrated that logical thinking could be performed relatively effortlessly and therefore challenging some of the foundational assumptions of dual-process models (De Neys, 2018; De Neys & Pennycook, 2019; Handley et al., 2011), they have high instrumental utility in empirical research. For our purpose, it is neither necessary to engage in the ongoing debate on a new conceptualization of dual process models nor to gather evidence in favor of a specific model. In accordance with Capraro (2024) , we use a pragmatic approach by assuming the existence of a rather generic dual process model (J. Evans, 2008). Additionally, it is somewhat irrelevant to us whether the two processes correspond to two distinct cognitive systems or evolutionary-developed modules (J. Evans, 2003). More important is the basic assumption of two different ways to process information: A more automatic, heuristic, and associative processing (system 1) and a more deliberate, systematic, and reflective processing (system 2). Only because people often process information with system 1, we are prone to biases and rather unrelated influences such as the similarity of another individual.

Research Areas

While in the first manuscript (Chapter 2), the applicability of speaker recognition systems as an objective measure for the perceived similarity of voices was established, I also showed that similar voices are perceived more favorably. Since this is a relatively basic finding and directly demonstrates similarity attraction, which is already described in detail, I will not discuss it further in this section. In contrast, I will briefly present research on

decision-making processes as well as on the Illusory Truth Effect since they represent two important domains where voice similarity could impact humans' life.

Decision-making

In cognitive psychology, decision-making is a central area that examines how people evaluate options and make decisions (for a review, see Fischhoff & Broomell, 2020). It has been significantly shaped by the work of Tversky and Kahneman, which examined how people make decisions, especially under conditions of risk or uncertainty (Tversky & Kahneman, 1973; Tversky & Kahneman, 1974). One of their fundamental findings was that people often disregard accurate probabilistic methods in favor of intuitive judgments and heuristics. Heuristics are mental shortcuts or rules of thumb that simplify decision-making processes and can lead to biases or errors. Such heuristics are, for instance, the *availability heuristic*, the *affect heuristic*, or the *anchoring heuristic*. The availability heuristic is applied if people rely on the number of examples that come to mind when evaluating a specific decision (Tversky & Kahneman, 1973). For instance, after hearing about a series of airplane accidents, one might overestimate the danger of air travel because those incidents are readily available in one's memory. The affect heuristic expresses the influence of positive and negative emotions on decisions (Slovic et al., 2007a). For instance, positive feelings associated with specific options can lead to decisions that disregard more objective attributes. The anchoring heuristic is applied if people rely too heavily on the first piece of information they encounter (Tversky & Kahneman, 1974).

The prevalence of heuristics and the susceptibility to systematic errors often lead to decisions that deviate from the utility-maximizing ideal as proposed by the *Utility Theory* (Edwards, 1954; Fishburn, 1979). Yet, these deviations do not imply irrationality per se. The decision-making process is inherently constrained by the concept of bounded rationality (Simon, 1955). Therefore, individuals' limited cognitive resources and the limited information available to them significantly influence their decision-making capabilities. The focus on achieving satisfactory results, for example, by using heuristics or other shortcuts (Payne et al., 1993), can be seen as reasonable in light of these constraints (Gigerenzer & Goldstein, 1996; Gigerenzer & Todd, 1999). While the rationality of using heuristics in decision-making is an ongoing subject of debate (Bröder & Newell, 2008; Costello & Watts, 2014), they do offer a realistic description of how decisions are made (Bhatia, 2017).

However, decisions are not solely shaped by knowledge and cognitive abilities but also by subjective evaluations of success (Frisch & Baron, 1988). The recognition of these

subjective dimensions has paved the way for the development of *Subjective Expected Utility Theory*, which integrates these personal factors and can be regarded as the leading framework in decision-making research (for reviews, see Fishburn, 1981; Hastie, 2001).

Since the evaluation of decision alternatives is rarely conducted in isolation but often influenced by other individuals, research on decision-making must acknowledge social influences (Bonaccio & Dalal, 2006; Kämmer et al., 2023). The connection between decision-making and social influence is significant as it situates decision-making within a broader social context, highlighting the interplay between personal and social factors. As described above, there are multiple mechanisms by which similarity can bias the perception and evaluation of others. This bias towards similar others, combined with the susceptibility to heuristics, especially the affect heuristic, opens up the possibility for an impact of similarity on decision-making.

Even though the concept of a free will has been questioned in the history of philosophy, psychology, and neuroscience, our societies essentially believe that humans are capable of making autonomous decisions (for a historic overview, see Dilman, 1999). If subtle, barely perceptible manipulations, such as making a voice similar to ones own, could affect people in a capacity that is considered inherently human, it is necessary to investigate and educate about the consequences of such manipulations – which is a central goal of my thesis and the reason, why I choose decision-making as one of my research areas.

Illusory Truth

Misinformation campaigns and conspiracy narratives spread by public figures and users on (social) media during the Covid-19 pandemic have probably not only cost countless people their lives by refusing to be vaccinated but have also permanently shattered trust in vaccinations in general, in the healthcare system and public institutions, the consequences of which are unforeseeable (Ferreira Caceres et al., 2022). However, not only since Covid-19 has it become apparent that we live in a post-truth era (Lewandowsky et al., 2017), in which evidence often appears less conclusive and credible than repetitive claims.

However, how can the repetition of a claim foster our belief in this claim? How do we judge truth? Brashier and Marsh (2020) posit that we infer from base rates, feelings, and consistency with memories when we make truth judgments. Although they all help us to arrive at a correct assessment, they are also connected to specific biases and pitfalls. For instance, base rates can inform us about the likelihood of an event or statement. However, when not the base rate of the statement's content is assessed but the probability of

encountering accurate information in everyday life – which is higher than the base rate of false information – the truth assessment of newly emerged information gets skewed towards a *true* judgment (Unkelbach, 2007). Moreover, while people should draw from their knowledge to judge truth, they often disregard their knowledge in favor of an increased ease of cognitive processing (L. K. Fazio et al., 2015). This preference for avoiding the costs of cognitive computation increases the influence of feelings on truth assessments, especially the influence of emotions connected to processing fluency.

Ecker and colleagues (2022) shifted the focus to how false beliefs arise. They identified different cognitive and socio-affective drivers of misinformation. Some of the main socio-affective drivers are source-related cues (e.g., group membership, attractiveness), emotions (e.g., emotional state of the perceiver, emotions connected to the message), and values and beliefs. Cognitive drivers are, among others, intuitive thinking, cognitive failures (e.g., neglect of one's knowledge), and illusory truth. The illusory truth effect occurs when information is encountered multiple times, which increases the probability of judging this information as accurate, regardless of its actual veracity (Hasher et al., 1977; Nadarevic et al., 2020; Pillai & Fazio, 2021). It is a robust effect often replicated (Dechêne et al., 2010; Henderson et al., 2022; Unkelbach et al., 2019). Research demonstrates, for instance, that repeated exposure bolsters the perceived truthfulness of information regardless of cognitive ability (De keersmaecker et al., 2020), across development (L. K. Fazio & Sherry, 2020), and persists over substantial timeframes (Boehm, 1994; Brown & Nix, 1996). Moreover, the illusion of truth occurs even for questionable statements (Lacassagne et al., 2022) when individuals have contradicting knowledge (Fazio et al., 2015; Fazio et al., 2019) or are offered financial incentives to discern truth from falsehood (Speckmann & Unkelbach, 2022). It can influence the perception of arguments (Moons et al., 2009), trivia (Bacon et al., 1979), rumors (DiFonzo et al., 2016), product and marketing claims (Hawkins et al., 2001; Venkataramani Johar & Roggeveen, 2007), news stories (Polage, 2012), and opinion statements (Arkes et al., 1989).

One proposed explanation for the illusory truth effect involves information processing and, as such, the dual process models. As discussed above, people often process information more peripherally or heuristically so that simplicity and cognitive ease, rather than analytical rigor, guide judgments (Gigerenzer & Gaissmaier, 2011; Pennycook & Rand, 2019). This tendency can lead to the application of the familiarity heuristic when encountering familiar information, where recognition becomes a marker for truth (Begg et al., 1992), even in the absence of an actual repetition of the statement (Hawkins et al., 2001; Law et al., 1998).

Recognizing information requires that it matches content that we have previously encountered and that is stored in our memory, regardless of whether this match is a true positive (accurate recall) or a false positive (false familiarity). While mere recognition caused by false familiarity can prompt the illusory truth effect, accurate recall is more likely to result in the effect and should exert a stronger response. This is because information stored in memory is more likely to be perceived as true than information not retrieved from memory (Ozubko & Fugelsang, 2011). Furthermore, although a single reference in memory can lead to a false attribution of truth, a larger amount of coherent information in memory increases its probability, as stated by the referential theory (Unkelbach & Rom, 2017).

On the other hand, familiarity with the information could increase processing fluency, which significantly influences evaluations across various domains, including truth judgments (for reviews, see Alter & Oppenheimer, 2009; Winkielman et al., 2003). While processing fluency is widely recognized as a key factor in the illusory truth effect (Ecker et al., 2022; Unkelbach, 2007), the debate continues on whether perceptual or conceptual fluency is critical for the emergence of this effect. Manipulating the readability of presented text influences truth perception even in the absence of repetition and, therefore, underlines the significance of perceptual fluency (Reber & Schwarz, 1999). However, research has demonstrated that the ease of processing the concept or semantic content of presented information is also a relevant aspect (Parks & Toth, 2006; Silva et al., 2017; Thapar & Westerman, 2009; Whittlesea, 1993). Since processing the underlying concept of information is often influenced by its presentation and vice versa, both types of processing fluencies are intertwined and difficult to decouple (Winkielman et al., 2012).

The importance of the influence of processing fluency on the illusory truth effect is not doubted by the referential theory described above. Instead, the underlying memory processes are emphasized as an essential factor influencing perceived cognitive fluency. This is apparent regarding conceptual fluency, which is increased if presented information fits well with existing knowledge or expectations but is also relevant regarding the ease with which content can be retrieved from memory. This ease of retrieval is also crucial in the aforementioned availability heuristic (Tversky & Kahneman, 1973) and could partially contribute to the illusory truth effect.

Despite the established impact of the illusory truth effect across various domains and the large body of research conducted, the specific role of audio and speech characteristics in reinforcing or mitigating this effect remains underexplored. Only two studies investigated the consequences of speaker characteristics on truth perception or credibility. Both of them

manipulated a rather superficial aspect of speech: the accents of a speaker. Frances and colleagues (2018) reported that regional accents did not significantly alter memory or perceived credibility. Conversely, Lev-Ari and Keysar (2010) found that speakers with non-native accents were generally considered less credible.

However, building upon the current research as discussed, voice similarity could enhance the illusory truth effect due to its impact on emotions, processing fluency, and the familiarity heuristic. Since feelings significantly influence truth judgments (Brashier and Marsh, 2020) and similar voices may be perceived as familiar and potentially more attractive, the positive feelings associated with their familiarity and attractiveness could influence people's truth judgments. Furthermore, in relation to the research of Ecker and colleagues (2022), voice similarity could also influence other socio-affective factors that contribute to misinformation, such as the perceived attractiveness of the source and their group affiliation. The familiarity with the physical attributes of the voice could additionally increase processing fluency and thus increase the belief in misinformation.

In summary, investigating the interplay between voice similarity and the illusory truth effect could improve our comprehension of how speech characteristics impact information processing and truth judgments. This research direction addresses a gap in the current understanding of the illusory truth effect and offers potential pathways for countering misinformation in an increasingly digital and audio-rich communication environment.

Speech Processing Technologies

Various speech processing technologies were required to address the research questions posed in the manuscripts in Chapters 2 to 4. First, a technique was needed to quantify the similarity between two different speakers, for which a speaker recognition system was used. Next, a system was needed to clone voices or at least to produce speech that is similar to target voices; this was achieved by a one-shot multi-speaker, multi-lingual text-to-speech (TTS) system. The following sections provide an introduction to these technologies and an overview of their development.

Determine the Similarity of Voices with Speaker Recognition Systems

The advent of the telephone as a widely accessible tool sparked interest in developing a voice-based identification system. Notably, one of the first significant methodologies in this area was pioneered by Lawrence Kersta during his tenure at the Bell Telephone Laboratories (Kersta, 1962). Kersta's motivation was driven by the escalating use of the telephone in

criminal activities, with the goal of aiding law enforcement in identifying criminals (Kersta, 1973). His approach involved the use of contour spectrograms, which are graphical representations of sound signals, plotting amplitude against frequency and time. Drawing a parallel with fingerprints, Kersta coined the term 'voiceprints' for these contour spectrograms. Visually comparing these voiceprints, much like fingerprints, was the method used to determine a speaker's identity.

Shortly after Kersta's pioneering work, the focus shifted towards developing automated systems for speaker identification (for reviews, see Atal, 1976; Doddington, 1985; Rosenberg, 1976). These automatic systems concentrated on fundamental acoustic attributes of the speech signal, such as "spectral amplitudes, voice pitch frequency, formant frequencies and bandwidths, and characteristic voicing aperiodicities" (Doddington, 1985, p. 1653). Even then, the idea came up of not only analyzing their characteristics over a certain period of time, but also evaluating them as a long-term average.

Speaker recognition systems have two primary forms: speaker identification, where the speaker's identity is unknown, and speaker verification, where the speaker's identity is compared to a known speaker (Hansen & Hasan, 2015; Mohd Hanifa et al., 2021; Sztahó et al., 2021). While speaker recognition systems are predominantly used in forensics to identify a speaker, speaker verification systems can be used to verify a speaker's identity for access to sensitive areas or information as part of a security process.

Another key distinction in speaker recognition systems is the method of audio presentation: text-dependent systems require users to say a specific word or phrase, while text-independent systems are indifferent to the content of the speech signal. While text-dependent systems are more straightforward to implement and more reliable, text-independent systems offer greater flexibility and robustness.

The process of speaker recognition can be structured into three main stages: development, enrollment, and evaluation (Hansen & Hasan, 2015; Salehghaffari, 2018; Variani et al., 2014). In the development phase large data sets are used to train background models that capture speaker characteristics in a numerical form, the voiceprints. During enrollment, speaker-specific models are created by extracting the unique features from new speaker utterances using the background models. In the evaluation phase, test utterances are matched against these speaker models using a scoring system, with the highest-scoring model indicating the speaker's identity. The system's accuracy and reliability are then assessed by the Equal Error Rate, a measure where the rate of false acceptances matches the rate of false rejections.

In the early days of this technology, text-independent systems could not keep up with text-dependent systems. However, with the advancements in computer technology and the implementation of Gaussian mixture models, hidden Markov models, and universal background models (Bimbot et al., 2004; Ohi et al., 2021; Reynolds et al., 2000), speaker recognition systems significantly improved the ability to solve more complex pattern-matching problems. Therefore, more and more systems used text-independent recognition methods, especially when Joint Factor Analysis based models (Kenny et al., 2007) and *i*-vector based models (Dehak et al., 2011) were introduced.

Despite delivering convincing results, text-independent speaker recognition systems using traditional computing methods involved a complex pipeline of generative models for several distinct tasks: a universal background model to collect speaker-independent feature characteristics based on Gaussian mixture models, a projection matrix to extract *i*-vectors, and a probabilistic linear discriminant analysis to calculate the similarity between *i*-vectors (Garcia-Romero & Espy-Wilson, 2011; Reynolds et al., 2000; D. Snyder et al., 2017). This complexity was reduced by introducing end-to-end deep neural network embeddings that directly process the audio input to output a similarity score (D. Snyder et al., 2017; Sztahó et al., 2021; Variani et al., 2014). End-to-end modeling requires minimal or no manual feature engineering and, therefore, contrasts with traditional methods, where distinct components or steps handle specific tasks, such as feature extraction, feature processing, and classification. However, there were also non end-to-end attempts to combine deep learning methods with an *i*-vector based approach – either by using deep learning to extract *i*-vectors or to classify them after being extracted (Tirumala & Shahamiri, 2016).

Contemporary systems commonly employ *d*-vectors or analogous speaker embedding types (Ohi et al., 2021; Sztahó et al., 2021). They are computed from the activations of layers within a deep neural network that is trained on speaker recognition task. Similar to *i*-vectors, *d*-vectors represent a set of features capable of differentiating between speakers (for a comparison of their performance, see Doddipatla et al., 2017). These voiceprints or feature vectors are numerical representations of voices, and their similarity can be determined, for instance, by calculating their cosine similarity (Ohi et al., 2021; Sztahó et al., 2021).

In addition to *d*-vectors, various deep speaker embeddings were developed, such as x-vectors, r-vectors, and ECAPA-TDNN. While each embedding type offers unique advantages, performance disparities among these embeddings are likely minimal (Z. Zhao et al., 2022). However, *d*-vectors benefit from their generation through a lightweight speaker encoder

model, which is popular in the open-source community for its straightforward implementation.

While many deep learning approaches could not outperform the more traditional method based on *i*-vectors regarding the Equal Error Rate, they are, nevertheless, now the state-of-the-art method in speaker recognition tasks (Sztahó et al., 2021) and were also used in the relevant experiments described in Chapter 2 to 4.

Text-to-Speech Systems: Components and Developments

Generating speech from text is a challenging undertaking. *Text-to-Speech* (TTS) systems have to solve several tasks to come from text to comprehensible speech: text analysis, acoustic modeling, and vocoding (Tan et al., 2021).

In the first step of the text analysis, text is *normalized* to get a unified form, for example, by transforming digits ("1" to "one") or abbreviations ("e.g." to "for example") into words. In languages like Chinese or Japanese, where white space does not necessarily define word boundaries, a word segmentation task is required to identify individual words. Afterward, the text can be further broken down into smaller units called tokens, which can be words, characters, or symbols – a process called *tokenization* (Webster & Kit, 1992). While tokens representing symbols, such as question marks, are often not vocalized on their own, they can alter the pronunciation of other tokens or entire sequences of tokens.

However, tokens representing words are often ambiguous: the word "record", for instance, refers as a noun to an item that keeps information, and as a verb, it means to capture information or performance. Since the pronunciation of words or tokens is influenced by their grammatical role, *part-of-speech tagging*, where the tokens or words are assigned to grammatical categories, can improve the quality of TTS systems (Schlünz, 2010).

Even though alphabetic writing systems such as English utilize orthography to symbolize the sounds of spoken words, there is often no exact correspondence between a word's written and spoken form. How one should pronounce a grapheme, the smallest meaningful written unit of a language, often depends on contextual influences. For a TSS to determine a word's pronunciation based on its written format, a *grapheme-to-phoneme conversion* (G2P) is necessary (Bisani & Ney, 2008), where phonemes are defined as the smallest perceptually distinct units of sounds that distinguish words from another. This G2P could be accomplished by a knowledge-based or data-driven approach. In a knowledge-based approach, a look-up dictionary is used to generate the phonemes – either by handcrafting a comprehensive dictionary or by defining rules to map graphemes to phonemes. A more data-

driven approach assumes that "given enough examples, it should be possible to predict the pronunciation of unseen words purely by analogy" (Bisani & Ney, 2008, p. 435).

After written text is converted into phonetic representations, *prosody prediction* is needed to predict their rhythm, stress, and intonation. Different languages utilize specific prosody tagging systems, such as ToBI for English (Silverman et al., 1992), to predict the prosody of syllables, words, and phrases.

Before introducing more sophisticated methods to synthesize speech, such as *statistical parametric speech synthesis* (SPSS), *concatenation-based speech synthesizers* were used to generate the speech waveform based on the linguistic features generated in the text analysis stage (Hunt & Black, 1996). Such a speech synthesis is achieved by piecing together sub-word segments from a database of annotated speech recordings. By forming a comprehensive inventory of speech units, the method groups similar phonetic units based on their phonetic features and prosodic contexts and selects the most appropriate unit to generate the sound signal (Black & Taylor, 1997; Black et al., 2007). Even though concatenative speech synthesizers could produce natural-sounding speech, their reliance on prerecorded sound samples renders them resource-intensive and inflexible in changing voice characteristics or even speaker (Zen, 2015).

With the introduction of SPSS, the linguistic features derived through pre-processing are used to compute the acoustic features of the text using an *acoustic model* (Zen et al., 2009). Acoustic modeling requires mathematical models that represent how linguistic elements are converted into acoustic properties of speech, such as pitch, duration, and timbre. Therefore, in acoustic modeling, the sequence of discrete symbols is transformed into a real-valued time series (Zen, 2015). One critical element of the sequence-to-sequence mapping problem is the differences in sequence length between linguistic and acoustic features. The most popular technique in SPSS to solve this mapping is based on hidden Markov models (HMM; Rabiner, 1989; Yoshimura et al., 1999). HMM are "trained to model the conditional distribution of an acoustic feature sequence given a linguistic feature sequence," which can be utilized at the synthesis stage to find "the most probable acoustic feature sequence for the linguistic feature sequence" (Zen, 2015, p. 2). Speech synthesis utilizing HMMs offers enhanced flexibility in modifying speaker characteristics, emotions, and speech styles compared to concatenation-based speech synthesis (Tokuda et al., 2000).

With the advancements in deep learning techniques, SPSS systems were outperformed by neural speech synthesis methods, such as WaveNet, the first architecture that produced more natural-sounding speech compared to concatenation and parametric-based models (Oord

et al., 2016). Since the introduction of WaveNet – which is in fact not only an acoustic model, but also a vocoder – a vast number of different deep-learning acoustic models have been developed, for example, Tacotron (Wang et al., 2017), Tacotron 2 (Shen et al., 2018), Deep Voice (Arik, Chrzanowski, et al., 2017), Glow-TTS (Kim et al., 2020), and FastSpeech (Ren et al., 2019).

Neural-based end-to-end acoustic modeling requires less pre-processing by eliminating the need for pre-aligning linguistic and acoustic features, with the possibility of just using characters or phonemes as input. Additionally, they have more computational power and, therefore, can generate more detailed, high-dimensional spectrograms compared to conventional SPSS systems (Tan et al., 2021). These spectrograms are often mel-spectrograms, which apply a non-linear frequency scaling that accentuates the frequencies to which human hearing is most adapted, making differences in these bands more pronounced. Therefore, it is designed to mimic the human ear's response to different frequencies, and their utilization can enhance the perceived naturalness of synthesized speech.

In the last part of the speech synthesis process, the *vocoding* step, the mathematical representations of audio or acoustic features are synthesized into a comprehensible speech waveform. The lack of high-quality vocoders was a significant factor preventing SPSS from synthesizing speech at a quality comparable to concatenation-based TTS (Black et al., 2007). Even though WaveNet was designed to generate waveform based on linguistic features, and therefore, delivering an integrated end-to-end system, WaveNet was soon used as a neural-based vocoder by conditioning it *not* on the linguistic features but instead on the mel-spectrograms derived from an acoustic model (Govalkar et al., 2019; Shen et al., 2018). Since then, several different model architectures have been developed to improve the vocoding of acoustic features, such as Glow (Kingma & Dhariwal, 2018), HiFi-GAN (Kong et al., 2020), and WaveGrad (N. Chen et al., 2020).

How to Clone a Voice

In the history of voice cloning, the question was not whether this was possible but at what cost. Technically speaking, concatenation-based speech synthesizers were already tools that made it possible to clone a specific voice, as they were nothing more than an exhaustive collection of speech segments from a single person. Although not designed for cloning per se, cloning a voice was initially the most effective strategy for achieving realistic speech synthesis.

As mentioned above, TTS systems based on SPSS offered greater flexibility to customize the generated speech, mainly through modeling spectrum, pitch and state duration simultaneously with HMM (Yoshimura et al., 1999). This made it possible to generate speech with specific speech characteristics or to transfer speech characteristics (Masuko et al., 1997; Shichiri et al., 2002; Yamagishi et al., 2009). However, due to the comparatively low speech quality of SPSS-based systems, such cloned voices could not be confused with real recordings of the replicated speaker.

With the development of deep learning-based systems and their ability to produce natural-sounding, human-like speech, the possibility of cloning a voice became more significant. The most basic method of cloning a voice using deep learning approaches is to train a system exclusively with annotated voice recordings of a specific person, so-called speaker-dependent TTS systems. Thereby, not only the mapping between linguistic features and waveform is learned but also the person's voice characteristics. However, the lack of sufficient training data for most individuals and the resource-intensive nature of this process made it an impractical solution for widespread use (for a single speaker model approx. 20 hours of speech data is required, see Wang et al., 2017). A more efficient approach requires training a TTS system to a competent level of speech generation before fine-tuning it with a target speaker's voice data – a method called speaker adaption (Fan et al., 2015; Taigman et al., 2017).

Even though this method is more efficient and requires less speech material, ideas already used in SPSS were soon adopted to improve the efficiency further. Instead of using only linguistic features or graphemes as the basis for training, one can also provide information about the speaker's identity. This method enabled the development of speaker-independent multi-speaker systems where the speaker used to generate the speech could easily be changed. Such systems used, for instance, predetermined *i*-vectors to identify speakers (Yang et al., 2016). Additionally, speaker information can also be trained together with the rest of the model (Arik, Diamos, et al., 2017; Ping et al., 2018). Using this approach, a single neural TTS system can learn hundreds of unique voices from less than half an hour of data per speaker while maintaining high audio quality and speaker identities. While multi-speaker systems represented a milestone in the development of TTS systems, individual speaker characteristics could only be generated for speakers used to train the model. While those multi-speaker systems mainly used speaker information data as a type of identification number, to achieve speech generation for unseen speakers, speaker information had to be captured more profoundly. Therefore, Arik and colleagues (2018) introduced a method called

speaker encoding, where speaker embeddings are directly estimated from audio. These embeddings should capture voice characteristics comparable to voiceprints derived from speaker recognition systems. While training, the TTS should learn to generate audio typical for speakers with a specific embedding. Afterward, the TTS can produce speech for unseen speakers based on short, un-transcribed utterances of audio.

Jia and colleagues (2018) used a similar method to achieve this zero-shot capability but trained the speaker encoder used to learn the speaker embeddings in isolation of the TTS system. Combining a neural TTS with a *d*-vector based speaker recognition system was a milestone in the history of voice cloning. This allowed the generation of convincing speech by a speaker unseen in training with only a few seconds of audio data. Moreover, the model architecture used was lightweight, resulting in the fast adoption by the open-source community (Jemine, 2019). Building on these ideas, several models were developed to improve the speech's quality and reduce the time to generate it (Casanova, Shulby, et al., 2021; Cooper et al., 2020; Rebryk & Beliaev, 2020). Moreover, based on the development of multi-lingual TTS systems (Nekvinda & Dušek, 2020; Y. Zhang et al., 2019), there are now also multi-lingual systems capable of generating speech of unseen speakers (Casanova, Weber, et al., 2021). The idea of using additional information while training also allowed the development of TTS systems where several other aspects of speech can be controlled, such as speed, speaking style, prosody, or emotions (Skerry-Ryan et al., 2018; Um et al., 2019; Wang et al., 2018; Y.-J. Zhang et al., 2018).

I used the open-source tool coqui-ai to generate speech in the experiments described below (Eren & The Coqui TTS Team, 2023). For simple speech generation, I utilized the VITS model (Kim et al., 2021). To clone the voices of our participants, I used the YourTTS model (Casanova, Weber, et al., 2021), which enhances the VITS model.

As described in this section, voice cloning was challenging in the past due to insufficient audio samples, limited computing power, or insufficient expertise. Today, however, numerous companies offer voice cloning services, making it available for a broader audience. While users usually are required to obtain consent from the people whose voices they intend to clone, the mechanisms for verifying this consent are not always clear, raising questions of transparency and ethical compliance, and thus the need for research into the impact of voice similarity.

Dissertation Overview

As outlined so far, manipulating voices to match the recipient's voice could have a significant effect. Based on the extensive research on similarity attraction, voices similar to one's own should have various effects on the recipient. However, as the ability to generate a specific voice has only recently become available, research in this area is still sparse. My dissertation, therefore, represents a first attempt to use these new TTS technologies to investigate what impact their implementation could have on our lives.

The research is structured into three main empirical chapters, each designed to test a series of hypotheses about the effects of voice similarity on trait evaluation, decision-making, and truth perception.

In the first manuscript (Chapter 2), my research focused on the capabilities of a speaker recognition system to approximate human judgments of voice similarity and its impact on trait evaluations. Specifically, I utilized over 1,000 hours of spoken audio from four extensive German speech datasets, featuring readings from audiobooks and Wikipedia articles by approximately 10,000 speakers. This data was employed to train a deep learning-based speaker recognition system, which generates *d*-vectors from short audio samples. These *d*-vectors are instrumental in identifying speakers and evaluating the similarity between them through the computation of cosine similarity values.

In the initial three experiments, participants assessed the similarity between systematically selected speaker pairs from my dataset. This included comparisons between the participant's own voice and that of another speaker. The results confirmed that cosine similarity values serve as a reliable objective measure of perceived voice similarity. Subsequently, this measure was used to explore the similarity-attraction hypothesis in voices, by having participants rate the likability and trustworthiness of voices that varied in their cosine similarity to the participants' own voices.

Additionally, I employed the *d*-vectors from the speakers in the dataset to compute the average similarity of each speaker to other same-gender speakers, using the mean cosine similarity value as a measure of each voice's prototypicality. I then examined whether speakers with varying levels of average cosine similarity scores were perceived differently in terms of likability and trustworthiness. This analysis aimed to explore the presence of a beauty-in-averageness effect in voices, specifically whether voices that more closely align with the average prototypicality are perceived as more likable and trustworthy.

The second manuscript (Chapter 3) expands the research into decision-making contexts. I applied a probabilistic inference paradigm to explore whether voice similarity

influences decision-making. In detail, participants must decide behind which of three houses a treasure is located and get predictions from three advisors about where the treasure is to be found. This initial experiment was set up to investigate how voice similarity interacts with heuristic cues like majority bias, which are known to influence decision outcomes. Specifically, the study aimed to determine whether participants, when incentivized monetarily, are more likely to follow the advice of an expert who sounds similar to themselves, especially when two other non-experts provide conflicting advice.

In two follow-up experiments, I employed a simplified design to examine whether voice similarity could influence decisions in the absence of conflicting information, thereby eliminating the confounding effect of majority bias. The findings indicated that participants were not generally more inclined to follow predictions from advisors with a similar voice compared to those with a generic voice. While this suggests that the similarity-attraction effect observed in Chapter 2 may not straightforwardly translate into decision-making behavior with real-life consequences, such as bonus payments for accurate choices, exploratory analyses hinted at a potential interaction between voice similarity and spatial position when visual cues were minimized. Thus, the overall impact of voice similarity appears to be subtle and context dependent.

Finally, the third manuscript (Chapter 4) addresses the influence of repetition of statements on truth perception. The first experiments were mainly deployed to test whether the statements I used in our study and the auditory information I generated through a TTS system could elicit the illusory truth effect. Most importantly, in the conducting experiment, I generated half of the employed statements in a voice that mimics the participant's voice, and I found that those statements are more often believed to be accurate compared to statements presented in a generic voice.

Overall, my thesis integrates psychology and computer science methodologies, builds upon a well-established effect in cognitive and social psychology, the similarity attraction effect, and uses recently developed technologies to investigate their possible impacts on our lives. Furthermore, my research goes beyond merely assessing the fundamental effects of voice similarity; it systematically investigates how voice similarity influences critical aspects of human autonomy: our decision-making processes and our perceptions of truth.

Chapter	Manuscript Title	Experiment	Research Question
2	AI-determined similarity increases likability and trustworthiness of human voices	1	Can cosine similarity values derived from a speaker recognition system be used to predict human (dis-) similarity judgments of voice pairs?
		2	How reliable are human (dis-) similarity judgments of voice pairs?
		3	Can cosine similarity values be used to predict human (dis-) similarity judgments when one of the voices is one's own voice?
		4	Are speakers with average voice characteristics perceived as more likable and trustworthy?
		5	Are voices similar to one's own voice perceived more favorably?
3	The Impact of Voice Similarity on Decision-Making: Do We Follow Advisors with Similar Voices?	1	Can decision quality be improved when advice is received from a similar-sounding advisor amidst conflicting information from less reliable sources?
		2	Are participants more inclined to follow the advice of a similar-sounding advisor?
		3	Are participants more inclined to follow the advice of a similar-sounding advisor when no other information is present?
4	Can I Believe My Voice?	1	Are repeated false statements more likely to be perceived as true compared to novel false statements?
		2	Are auditorily presented false statements more likely to be perceived as true when tested in written form?
		3	Are auditorily presented and tested false statements more likely to be perceived as true?
		4	Are false statements more likely to be perceived as true when presented in a similar voice?

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Declaration according to § 5 Abs. 2 No. 8 of the PhD regulations of the Faculty of Science -Collaborative Publications

The following chapter (Chapter 2) consists of a manuscript that is published and was co-authored by Stephan Schwan and Hauke S. Meyerhoff. The proportional contributions to this manuscript are presented in the subsequent table.

Author	Author position	Scientific ideas	Data generation	Analysis & interpretation	Paper writing
Oliver Jaggy	First author	80 %	100 %	80 %	70 %
Stephan Schwan	Second author	10 %	0 %	10 %	15 %
Hauke S. Meyerhoff	Third author	10 %	0 %	10 %	15 %

Title of paper: AI-determined similarity increases likability and trustworthiness of human voices

Status in publication process: Published. Jaggy O, Schwan S, Meyerhoff HS (2025). AI-determined similarity increases likability and trustworthiness of human voices. *PLoS ONE 20*(3): e0318890. https://doi.org/10.1371/journal.pone.0318890

Chapter 2: AI-Determined Similarity Increases Likability and Trustworthiness of Human Voices

Abstract

Modern artificial intelligence (AI) technology is capable of generating human sounding voices that could be used to deceive recipients in various contexts (e.g., deep fakes). Given the increasing accessibility of this technology and its potential societal implications, the present study conducted online experiments using original data to investigate the validity of AI-based voice similarity measures and their impact on trustworthiness and likability. Correlation analyses revealed that voiceprints – numerical representations of voices derived from a speaker verification system – can be used to approximate human (dis)similarity ratings. With regard to cognitive evaluations, we observed that voices similar to one's own voice increased trustworthiness and likability, whereas average voices did not elicit such effects. These findings suggest a preference for self-similar voices and underscore the risks associated with the misuse of AI in generating persuasive artificial voices from brief voice samples.

Significance Statement

Our paper introduces a new methodological tool to the field at large, showing for the first time that voiceprints derived from a speaker verification system (based on *d*-vectors) can be used to investigate the effects of perceived voice similarity on cognitive evaluation. Our study shows that voices similar to one's own voice increase likability and trustworthiness and thus promote our theoretical understanding of inter-personal evaluations. Our results have a broad appeal, likely beyond the boundary of science. Given the tremendously increasing spread of AI technology, our results suggests that individual adaptions could be used to manipulate human cognition.

Introduction

Artificial intelligence (AI) has become integral to modern life and is revolutionizing how people interact with technology and process information. From autonomous vehicles to personalized recommendation systems, AI's ability to analyze and replicate human-like behaviors profoundly impacts all industries. One particularly relevant application is in the field of speech technology, in which AI systems not only recognize and synthesize speech but also simulate individual voice characteristics. This capability opens new avenues for personalized interactions, such as matching voice assistants to a user's voice profile or augmenting that profile, illustrating the interplay between technology, identity, and human perception.

The human voice remains a remarkable signature of individuality, transcending mere communication to embed rich layers of information about the speaker. Beyond the conveyance of words, each voice carries the unique timbre, tone, and other acoustic information, that hint at the speaker's identity. Like fingerprints, the human voice can be used to distinguish individuals from one another with a high degree of accuracy (Doddington, 1985b; H. Li et al., 2020) and gives insights into the speaker's emotions and physical attributes. Speech data can be used, for example, to recognize stress (Van Puyvelde et al., 2018), emotions (Grágeda et al., 2023, 2025; Kaya & Karpov, 2018; C.-C. Lee et al., 2011) , the level of interest (Jeon et al., 2010), age and sex (M. Li et al., 2012; Meinedo & Trancoso, 2010), and personality traits (Carbonneau et al., 2017; Mohammadi & Vinciarelli, 2015) – for a review on speech analysis for health, see (Cummins et al., 2018).

Voice assistants such as Alexa or Siri attempt to mimic human voice in terms of pleasant and recognizably individualized speech characteristics. So far, most voice assistants implement only one synthetic voice and thus follow an approach in which one voice fits all users (Cambre & Kulkarni, 2019). However, voice assistants may also compute the voiceprint (see below) of the customer and utilize this information to modify the synthetic voice to make it similar to the customer’s voice.

Therefore, the question arises how listeners evaluate voices similar to the listeners' own voices and whether they prefer average voices compared to more distinct voices. The present paper addresses this question in five experiments. As a first step, we show that similarity judgements of two voices by AI-based speaker recognition systems and human listeners significantly correspond (Experiments 1 and 2), which is a necessary precondition for AI-based cloning of individual voices. As second step, we show that this correspondence also holds if one of the voices is the listener's own voice (Experiment 3). As a third step, we showed that average voices are not preferred over distinct voice (Experiment 4). We finally demonstrate that listeners judge voices similar to their own voice (according to the AI-based speaker recognition system) to be more likable and trustworthy then dissimilar voices (Experiment 5).

Characterizing individual human voices through AI-based d-vectors

The complexity of human speech poses a significant challenge: how can one distill and encode these sophisticated vocal characteristics into a form that captures the essence of individual identity? Modern speaker recognition systems use d-vectors, or similar kinds of speaker embeddings (such as x-vectors, r-vectors, or ECAPA-TDNN), derived from a deep neural net (Hinton et al., 2012; Ohi et al., 2021) While there are only minor differences in performance among these speaker embeddings (Z. Zhao et al., 2022) d-vectors have the advantage that the speaker encoder that generates the embeddings is a lightweight model, widely used in the open-source community, and relatively easy to implement.

Starting point are short audio samples of a human speakers. The audio samples are non-linearly transformed on the frequency scale in a way that emphasizes distances between frequencies for which the human ear is most sensitive. Next, these transformations, called mel-spectrograms are used to train deep neural networks. D-vectors then are the averaged activations of the final hidden layer of a deep neural network that is trained on a speaker verification or identification task. As a result, they are abstract representations of audio called "voiceprints", which contain compressed information about the audio signal's unique characteristics, such as timbre and tone, in a multidimensional space.

However, such voiceprints may not only used for speaker identification. Instead, deep learning methods (Oord et al., 2016; Shen et al., 2018), enables software to clone the voice of a real person. Cloning a voice traditionally involves training a Text-to-Speech (TTS) system using audio samples from the target individual. However, this requires a large number of audio samples from the individual (often unavailable) and the training of an entire TTS system, which is both time-intensive and computationally demanding. Consequently, cloning someone’s voice was most of the time either impossible or too prohibitively expensive. Yet, providing voiceprints as additional information when training a TTS system makes it possible to clone a voice with only a few seconds of audio material and without the need to train a new system (Arik, Diamos, et al., 2017; Cooper et al., 2020; Jia et al., 2018). Even if the results are not yet as convincing as previously used techniques, the essential prerequisites have been met to convert any given text into speech and predetermine the used voice by providing a voiceprint.

Comparing human to d-vector based voice similarity judgments

Yet, there is little research on how voiceprints are related to human perception. Since voiceprints are new in the field, we needed to establish their validity for (human) similarity judgments, which is a prerequisite to study the cognitive consequences of voice similarity thereafter. Beside research on performance differences between human and speaker recognition systems (González Hautamäki et al., 2015), to the best of our knowledge, there is only one study that has investigated the relationship between voice similarity estimates by humans and an automatic speaker recognition system (Gerlach et al., 2020). The study by (Gerlach et al., 2020) showed a positive relationship between participants' similarity judgments and comparison scores from a speaker recognition system (Dehak et al., 2011). In contrast to this study, we are interested in voiceprints derived from a speaker recognition system based on d-vectors, which are derived by training a deep neural net (Variani et al., 2014) and can be used to clone a voice. Although our study does not employ cloned voices, the application of d-vectors for generating speech that resembles a target speaker's voice makes them the ideal candidate for examining similarity effects.

The significance of likeability and trustworthiness in social interactions

Likeability and trustworthiness are foundational attributes that significantly influence social interactions and relationships. Research has demonstrated that individuals judged more likable by others are more persuasive, often receiving preferential treatment and social support (Brodsky et al., 2009; Clayson, 2022; Gommans et al., 2017; Younan & Martire, 2021), and that similar others are also perceived as more likeable (Moreland & Zajonc, 1982). Similarly, trustworthiness is central in fostering long-term (business) relationships and ensuring effective collaboration, as it mitigates uncertainty, reduces the perceived risk in interactions and increases predictability (Dyer & Chu, 2003; A. M. Evans & Krueger, 2009; Kumar, 1996; Rempel et al., 1985; Simpson, 2007). From an evolutionary perspective, trustworthiness likely signals an individual’s reliability and cooperative intent, essential for fostering social cohesion and reciprocal behaviors within groups. Similarly, likeability facilitates social bonding by eliciting positive affect and reducing interpersonal tension, enhancing collaboration and mutual support. Thus, since these constructs are integral to social evaluation processes, the use of likability and trustworthiness as dependent variables is critical to understanding the impact of voice similarity.

Voice typicality and its influence on trustworthiness and likability

Previous research in other perceptual domains has consistently demonstrated a *beauty-in-averageness* effect, where average or prototypical faces and objects are perceived as more attractive than those that deviate from the norm (Halberstadt, 2006; Holzleitner et al., 2019; Langlois & Roggman, 1990). This phenomenon extends beyond mere aesthetic preference, reflecting a broader cognitive tendency to favor typical over unusual stimuli, which may be rooted in the ease of processing more familiar or expected patterns (Winkielman et al., 2006). Additionally, familiarity has been shown to enhance social evaluations, such as perceived trustworthiness, particularly in the context of faces (Sofer et al., 2015). These findings suggest that perceptual and cognitive processes prioritize typicality and familiarity, potentially because they signal safety, reliability, or group affiliation.

Building on this framework, Experiment 4 sought to explore whether a similar effect is observable in the auditory domain, specifically for voices. In this context, typicality was operationalized as the mean cosine similarity between a given speaker's voiceprint – a numerical representation of their vocal characteristics – and the voiceprints of all other speakers in our dataset. By examining whether voices with higher typicality are associated with greater trustworthiness and likability, we aimed to extend the beauty-in-averageness principle to auditory stimuli.

Similarity attraction and the possible effects of voices resembling listeners' own voices

Cloned voices are an essential component of deep fakes, which are primarily used for entertainment purposes such as showing Elon Musk performing a belly dance or Barack Obama mocking Donald Trump. However, there are also malicious use cases (Brewster, 2021), and deep fakes have been used to spread fake news and propaganda (Burgess, 2022). But there are also more subtle possibilities to use manipulated audio, particularly in the field of voice assistants, which could have significant effects on users of voice assistant systems. According to the *similarity attraction hypothesis* (Byrne, 1961; Byrne et al., 1967), people like other people more if they behave, appear, or think similarly to them – for a meta-analysis, see (Montoya et al., 2008). A possible explanation for similarity attraction is linked to a phenomenon called *implicit egotism*: People tend to evaluate themselves positively, and if they associate other people with themselves, the positive self-evaluation may influence their evaluation (Hughes & Harrison, 2013; J. T. Jones et al., 2004; Peng et al., 2020). Building on this concept, it has been proposed that similarity influences attraction by shaping the perceived valence and significance of inferred traits (Ajzen, 1974; Kaplan & Anderson,

1973; Montoya & Horton, 2013). Specifically, individuals may derive positive or negative evaluations of others based on shared or divergent attitudes, personality traits, or other attributes. According to (Kaplan & Anderson, 1973), similar attitudes do not directly lead to attraction but foster expectations of additional positive qualities in the similar individual, driven by the individual's own favorable self-assessment (Ajzen, 1974; Stalling, 1970). Moreover, the inclination to interpret superficial similarities as indicative of deeper shared traits can result in an overestimated sense of alignment. For instance, individuals might presume that an advisor who shares surface-level characteristics also holds similar preferences, thereby perceiving their advice as more applicable or insightful (Hovland et al., 1953). However, effects of similarity may also stem from humans' inherently social nature. *Social Identity Theory* (Tajfel & Turner, 1986) suggests that individuals exhibit a preference for and more positive behaviors toward those they perceive as members of their own group.

Biological explanations further contribute to our understanding of similarity effects. In the context of social networks, this mechanism is often referred to as homophily, which describes the tendency to form connections with similar others (McPherson et al., 2001). From an evolutionary perspective, it has been argued that altruistic behaviors typically come at a cost, except when directed toward genetically related individuals (Burnstein et al., 1994; Hamilton, 1964). Since genetic similarity is often correlated with phenotypic resemblance, evolutionary pressures may have favored prosocial behaviors toward those perceived as similar, enhancing cooperation and cohesion within social groups.

However, similarity effects do not only occur between human agents. Research on human-machine communication has shown that humans exhibit social responses to computers just as they do to humans. Consequently, similarity attraction also might arise in human-computer interactions involving artificial voices (Nass et al., 1994; Nass & Moon, 2000). Indeed, general (i.e. non-adaptive) alignments of acoustic-prosodic features such as speech rate, intensity, pitch, volume, and prosody, can lead to a similarity attraction towards synthetic voices (Nass & Lee, 2001), which can positively influence learning (Lubold et al., 2018), engagement (Chaspari & Lehman, 2016), and enjoyment (Sadoughi et al., 2017).

Based on the above considerations, we investigated

– whether the cosine similarity derived from the trained neural network correlates with human similarity judgments (Exp 1-3).

- whether speakers with prototypical voices are judged as more likeable and trustworthy (Exp 4).
- whether speakers with similar voiceprints to the corresponding participants are perceived as more likable and trustworthy (Exp 5).

The relation between AI and human similarity judgments

In the first Experiment, we investigated the validity of the cosine similarity as a measure of perceived voice similarity by probing whether the cosine similarity of two voiceprints predicts human similarity judgments.

Method

Ethics statement. The studies reported were approved by the ethics committee of the Leibniz-Institut für Wissensmedien, Tübingen (approval number LEK 2020/061 and LEK 2021/123). All participants provided written informed consent through the online platform qualtrics.com, and all experiments included in this study were preregistered (Exp. 1: https://osf.io/kxwsv; Exp. 2: https://osf.io/8c7xw; Exp. 3: https://osf.io/yt3b7; Exp. 4: https://osf.io/q59da; Exp. 5: https://osf.io/cv5g9).

Encoder and data. For the first as well as the following Experiments, we used an open-source encoder (Jemine, 2019) based on research conducted by (Heigold et al., 2016; Jia et al., 2018; Wang et al., 2017). In contrast to the described model in (Jemine, 2019), our model consists of three recurrent neural networks (RNN) of the long short-term memory type(LSTM layers) with 768 nodes, followed by a fully connected projection layer with 256 nodes and a tanh activation function.

The encoder is trained on a speaker verification task in which it learns to embed utterances from the same speaker close together in the embedding space and utterances from different speakers farther away. This increases intraspeaker variation and enhances interspeaker discrimination. For each utterance, a 256-dimensional feature vector is created, where each feature can encode certain voice features. These voice features are characteristic for the speaker and could be understood as a numerical representation of the voice, a voiceprint. The similarity of two voices can be compared by calculating the cosine similarity of two feature vectors, yielding values ranging from -1 to 1.

To train the network, we used the German subset of the Common-Voice dataset (https://github.com/mozilla/common-voice), Distant-Speech (https://www.inf.uni-

hamburg.de/en/inst/ab/lt/resources/data/acoustic-models.html), LibriVoxDeEn (Beilharz et al., 2020), and the German subset of the VoxForge dataset (http://www.voxforge.org/home). These datasets contain linguistically diverse material. The linguistic content ranges from simple phrases to complex sentences, covering a broad spectrum of phonetic, lexical, and syntactic structures in German. The combined datasets consist of approximately 1,000 hours of spoken audiobooks, and Wikipedia articles read aloud by about 10,000 non-professional speakers. Audio files not already cut at sentence boundaries were cut at the appropriate points.

Participants. We chose a sample size of 100 participants for our first experiment as a practical starting point. This sample size provided a balance between feasibility and statistical power, allowing us to evaluate the relationship between cosine similarity and human similarity judgments while accounting for individual variability. Therefore, we recruited 50 male and 50 female German participants via prolific (https://www.prolific.com), which was the recruitment platform used for all experiments. Basic demographic information was collected via Qualtrics (https://qualtrics.com) in all experiments.

Six of the participants were excluded because they failed in more than one control trial. The mean age was $M = 32.01$ ($SD = 11.26$). Forty-seven of the 94 participants were female, one diverse, and three refused to answer. Recruitment occurred from March 3, 2021, to March 5, 2021. Participants received £3.45 for their participation in the study.

Materials, stimuli and procedure. Since our dataset included more male than female speakers (approximate ratio of 3:1), and because this was our first experiment using this type of data, we aimed to achieve a wide range of cosine similarity values with high granularity. We used only male speakers in this study to simplify the experimental design and ensure consistent conditions.

For each male speaker in our dataset, we calculated the cosine similarity of the voice embedding with each other speaker in the dataset. Since those cosine similarities are approximately normally distributed, randomly drawing from these pairs would result in too few examples from the edge categories. Therefore, we subdivided the cosine values into ten categories, using the lowest and highest cosine value between speaker pairs as reference points with equal cosine value differences between the breakpoints. We subsequently drew speaker pairs based on the categories, which should ensure an even distribution of cosine values and, therefore, the greatest possible variance in the stimulus material. For each drawn speaker, we randomly picked one audio sample from our dataset,

trimmed it to a maximum length of 5 s, and normalized the volume. We drew 50 sets of 100 male speaker pairs and presented each set to one female and one male participant in a random order.

Since our experiments were conducted online on pavlovia.org (https://pavlovia.org/) using PsychoPy (Peirce, 2007), we checked whether participants had a working audio setup at the beginning of the experiment: We presented a short text in which we informed them to count sinus tones. After presenting four sinus tones with an inter-stimulus interval of 1s, participants should indicate on a slider (ticks on 0,1, 2,3,4 and 5) how many sinus tones they were hearing. If they failed the test, the experiment was concluded.If they passed the detection task, three introductory trials were presented that had the same structure as the regular trials; audio recordings from two different male speakers were presented sequentially, with an inter-stimulus interval of 1s. While audio was played a headphone icon was depicted. After hearing both voices once, the participants were asked to rate the dissimilarity by adjusting the slider on an unmarked continuous rating scale (range: little dissimilarity - great dissimilarity). They were allowed to take as much time as needed for this rating, with no imposed time constraints. While piloting our study, we found it much more challenging to rate the similarity compared to the dissimilarity. Accordingly, we asked the participants to rate the dissimilarity rather the similarity and inverted the response afterward. Participants could skip a rating but were informed that they only should choose this option if they couldn't hear one of the samples properly. The participants who skipped more than ten trials were excluded. To check the participants' attention, we presented every 30th trial two different audio samples from the same speaker. The participants who rated the dissimilarity higher than 0.2 in more than one of the three control trials were excluded.

Results

To analyze the data in this study, we used the software R (R Core Team, 2019), the R package lme4 (Bates et al., 2014), and the R package MuMIn (Barton, 2020). We used mixed models with participants as random effects, the raw cosine values as the independent variable, and the inverted slider responses as the dependent variable (R code for data processing is publicly available, see below). To test whether encoder ratings can predict how similarly humans judge different voices, we compared an intercept-only model, a linear model, and a quadratic model. We included a quadratic model because human judgments, particularly those based on perceptual features like voice similarity, often show non-linear trends (Perrachione et al., 2019). This approach accounts for potential non-linear relationships between the cosine similarity of voice embeddings

and human similarity judgments. For instance, individuals may perceive two voices as more similar up to a certain point, but after that, additional increases in cosine similarity might not yield proportional increases in perceived similarity. This suggests diminishing or varying returns on perceived similarity as cosine similarity increases.

Since previous research found evidence for an own-gender bias in the ability of voice identification (Skuk & Schweinberger, 2013) and gender differences in voice processing (Ahrens et al., 2014; Junger et al., 2013), we included participants gender as an additional factor. Weighting using Akaike information criterion (AIC) scores (see Table 1.1) showed a clear quadratic relationship between the calculated cosine similarity of the encoder and participants' rating (intercept: 0.36, *95% CI* [0.34, 0.38], $t(102.3) = 33.05$, $p < .001$; cosine: 0.10, *95% CI* [0.07, 0.13], $t(9215) = 7.68$, $p < .001$; cosine^2: 0.47, *95% CI* [0.42, 0.52], , $t(9215) = 18.51$, $p < .001$). The median of the individual Spearman Correlations between the encoder's cosine similaritiy values and participants similarity ratings was *Mdn* $r_s = 0.37$ ($Q_1 = 0.28$, $Q_3 = 0.43$), indicating a moderate relationship. The model explained 27% of the variance ($R_c^2 = 0.27$). Numerically, female participants rated the similarity slightly higher; however, the inclusion of gender as an additional factor is not justified given the AIC values.

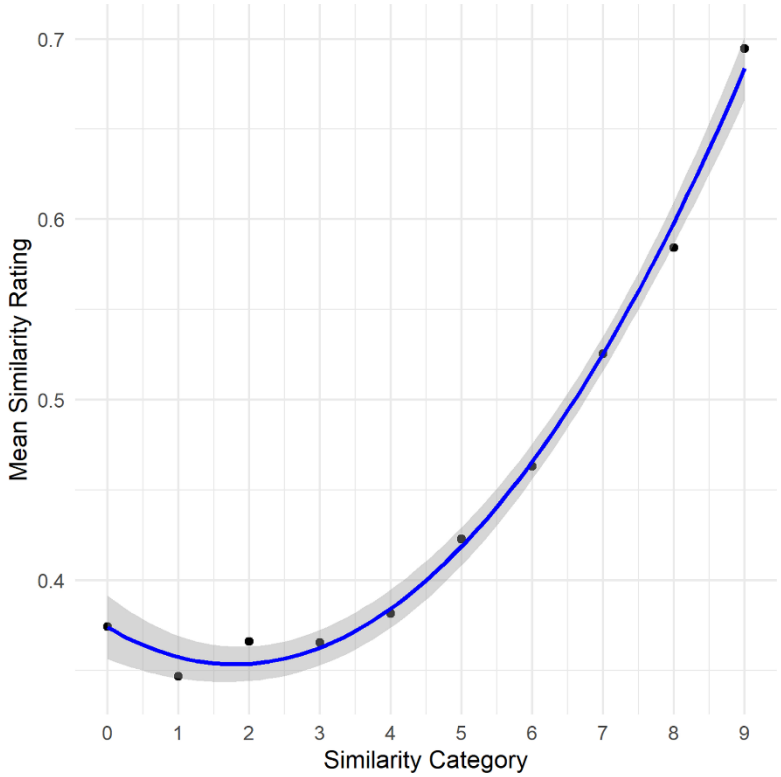
The relationship between cosine values and the participants' similarity ratings seems stronger for higher cosine similarity values. The quadratic relationship indicates that the cosine values derived from the deep neural network are associated with human similarity judgments, and highlights a stronger association at extremer cosine similarity values Participants skipped, on average, $M = 0.86$ trials (SD = 1.16) and needed, on average, *Mdn* = 28.98 minutes to complete the experiment.

Given the large variance observed among participants' similarity ratings and the use of a vast amount of different stimulus pairs, we performed additional analysis with aggregated data to gain more insights into the relationship between cosine values and similarity judgments. Rather than employing the encoder's raw cosine values for each similarity judgment, we utilized the predefined similarity categories used in the sampling process as predictors. The response variable was the mean similarity judgment corresponding to each category. Since the above analysis revealed a quadratic relationship, we compared a linear model with a quadratic regression model using an analysis of variance (ANOVA). The results strongly favored the quadratic model over the linear model, $F(1,7) = 240.98$, $p < .001$. The analysis of the quadratic model itself revealed a significant quadratic relationship between the similarity category and the mean similarity rating, $F(2,7) = 670.5$, $p < .001$

(intercept: 0.37, *95% CI* [0.36, 0.39], $t(7) = 50.51$, $p < .001$; category -0.02, *95% CI* [-0.03, -0.01], $t(7) = -5.97$, $p < .001$, category2: 0.006, *95% CI* [0.005, 0.007], $t(7) = 15.52$, $p < .001$). The model explained most of the variance in the mean similarity rating, $R^2 = 0.995$ (Figure 1.1). Even though mixed effects models account for random variation and lead to shrinkage, the analysis with more aggregated data further reduces variation and leads to a more pronounced relation.

Figure 1.1

Illustration of the Results of Experiment 1



Note. Depicted is the quadratic relationship between cosine similarity categories and the participants' mean similarity ratings as well as the the 95% confidence interval of the regression line.

While the association between similarity values and similarity judgments appears modest in the analysis with mixed models, the analysis with aggregated data suggests that the relationship warrants attention. This finding is noteworthy given the potential limitations of using stimuli derived from open-source datasets. Factors such as varying audio quality, speakers' articulation proficiency, and the semantic content of audio clips could have influenced evaluations. At the same time, the diversity of the stimuli may have contributed to the ecological validity of incidental similarity evaluations.

It is also important to consider that assessing the similarity of two voices based on just 5 seconds of random samples is inherently challenging. The quadratic relationship observed in the data indicates that these challenges were particularly pronounced when participants evaluated speaker pairs with moderate to low cosine similarity values. These complexities, and their potential implications for interpreting the results, will be explored further in the General discussion.

Our findings further gain context when compared with those of MOSNet (Lo et al., 2019). MOSNet demonstrated a capacity to predict human-perceived similarity judgments, more precisely termed identity ratings, with Spearman Rank Correlation coefficients ranging between 0.292 and 0.455. The derived median correlation coefficient of 0.37 in our experiment aligns with the midpoint of MOSNet's observed correlations but for raw values on a similarity judgment – which we consider more valuable for research on the influence of voice similarity on cognitive processes.

Overall, the results confirm the validity of cosine similarity as a measure of perceived voice similarity. The quadratic relationship suggests that participants were disproportionately sensitive to very dissimilar and very similar voices but less capable of differentiating at intermediate similarity levels. This may reflect a natural limit in human voice discrimination abilities, particularly for voices that are neither too distinct nor too similar. These findings support the utility of AI-generated cosine similarity for approximating human voice similarity judgments.

The reliability of human similarity judgments

In order to interpret the magnitude of the correlation between raw cosine values and similarity judgments observed in Experiment 1, we needed to assess the reliability of human similarity judgments, which limits the maximum of observable correlations (Hedge et al., 2018).

Method

Participants. G*Power (Faul et al., 2007) was used to calculate the necessary sample size for the Correlation: Bivariate normal model test as an approximation for the non-parametric Spearman rank correlation test that was used to calculate the test-retest reliability. The analysis aimed to detect a medium to large effect size with an alpha = 0.05 and a power = 0.80. The power analysis revealed a minimum sample size of 46 participants. We therefore recruited 50 new participants via prolific. Five were excluded because they failed in more than one control trial. Eighteen of the remaining 45 participants were female, two did not indicate their sex. The mean age was $M = 30.31$ ($SD = 10.91$).

Recruitment took place on April 19, 2021. Participants received £4.36 for their participation in the study.

Materials, stimuli and procedure. Besides some minor changes, the material and procedure were identical to Experiment 1. In contrast to Experiment 1, we sampled 50 instead of 100 speaker pairs, and we only used one set of speakers for all participants. After each of 50 speaker pairs were presented once, participants were asked to rate their similarity again. The order of the speaker pairs was altered in the second part of the experiment but was the same for all participants. We aimed for this uniformity of the experimental conditions to avoid variance from individual randomizations of the trials order since our approach mainly focused on correlations which require reliable estimates for person parameters rather than experimental conditions (for which randomization would be necessary).

Results

As in Experiment 1, we observed a correlation between cosine similarity and similarity judgments. Using AIC values from Table 1.2, we identified a quadratic relationship between cosine similarity (generated by the encoder) and participants' ratings. The model explained 19% of the variance in similarity judgments (intercept: 0.43, *95% CI*: 0. 40 to 0. 46, $t(47.05) = 25.75$, $p < .001$; cosine: 0. 09, *95% CI* [0. 05, 0.13], $t(4249) = 4.55$, $p < .001$; cosine^2: 0. 24, *95% CI* [0. 16, 0. 31], $t(4249) = 5.97$, $p < .001$, $R_c^2 = 0.19$). The median of the individual Spearman Correlation between cosine similarities and similarity ratings was $Mdn\ r_s = 0.23$ ($Q_1 = 0. 19$, $Q_3 = 0. 26$). Therefore, we were able to replicate our results of Experiment 1, which showed a quadratic relation between cosine similarities and similarity judgments. The relationship, however, was slightly less pronounced than in Experiment 1. This may be explained by the limited stimulus material required to measure the reliability scores as well as the reduced number of trials to keep the experiment within reasonable boundaries.

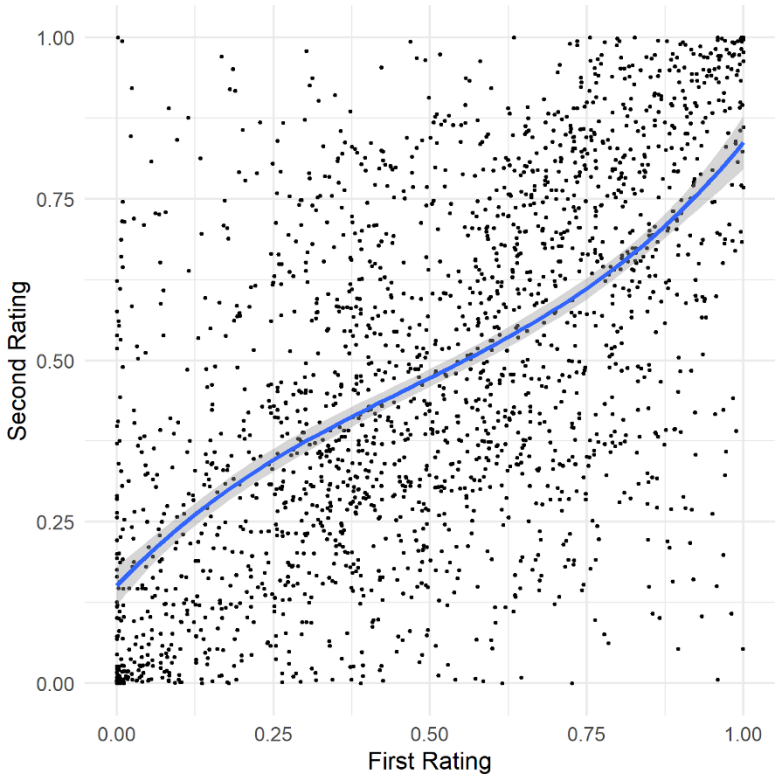
As in the first experiment, we conducted additional regression analyses using the similarity category as the predictor and the mean similarity judgments as the response variable. An ANOVA comparing the quadratic and the linear model found no significant increase in fit, $F(1,7) = 0.47$, $p = .52$. The analysis of the linear model revealed a significant relationship between the similarity category and the mean similarity rating, $F(1,8) = 5.52$, $p = .047$ (intercept: 0.39, *95% CI* [0.28, 0.50], $t(8) = 8.48$, $p < .001$; category 0.02, *95% CI* [0.00, 0.04], $t(8) = 2.35$, $p = .047$). The model

explained 40.8% of the variance (R^2 = 0.408). Again, the lack of a quadratic effect in this aggregated dataset compared to the first experiment likely stems from the diminished variance due to the smaller number of speaker pairs per category (10 vs. 500) and the reduced sample size (50 vs. 100). Despite these limitations, the findings suggest a consistent, incremental monotonic increase in similarity judgments for speaker pairs in higher similarity categories.

Reliability and attenuation correction. The test-retest reliability, indexed by the median of the individual Spearman Correlation between the first and the second similarity rating, was *Mdn* r_s = 0.57 (Q_1 = 0.44, Q_3 = 0.65), which can be considered as a fair test-retest reliability (Cicchetti, 1994). Since the cosine similarity values derived from the encoder are consistent, a single attenuation correction was performed to estimate the true correlation. Using the obtained reliability value yielded a correlation between the cosine values of the encoder and the participants' similarity ratings of *Mdn* r_s = 0.48 for the first experiment, and *Mdn* r_s = 0.31 for the second experiment. Explanatory analyses showed a polynomial relationship between the first and the second rating (Figure 1.2). This indicates a stronger correlation for extreme (dis-)similarities.

Figure 1.2

Illustration of the Results of Experiment 2



Note. The scatterplot shows the polynomial relationship between the first and the second similarity judgment and the 95% confidence interval of the regression line.

The observed reliability is crucial in understanding the correlation between AI-generated cosine values and human similarity judgments. Any error or inconsistency in human judgments (as suggested by the reliability value less than 1) can attenuate or reduce the observed correlation. This means that the true correlation would likely be higher in the absence of such errors. Therefore, the obtained correlation between AI ratings and human judgments underestimates the actual strength of this relationship due to the influence of measurement error inherent in human judgments. Considering this reliability, our attenuation correction suggests that the true correlation is stronger than what is directly observed – even though the correlation values calculated by the attenuation correction are to be regarded as upper bounds.

The more pronounced correlations at extreme values of similarity, as indicated by the polynomial relationship, supports not only this view but also demonstrate that people struggle to make reliable and consistent judgments for speaker pairs average in similarity. This finding does not

contradict the outcomes of Experiment 1, where a quadratic relationship suggested difficulties in making nuanced ratings for more dissimilar speaker pairs. Moreover, the results added evidence to the notion that judging the similarity is inherently challenging, with more reliable assessments typically occurring at the extremes of the similarity spectrum.

Consistency across raters. Since we used a fixed set of stimuli for all participants, we also assessed the consistency of similarity judgments across different raters by calculating the Intraclass Correlation Coefficient (ICC) with the R package irr (Gamer et al., 2012). We employed a two-way model to evaluate the level of agreement on similarity judgments among the 44 raters across the 50 speaker pairs. Since participants rated each speaker pair twice, we analyzed only the ratings from the first 50 trials. Where participants skipped the assessment of a speaker pair, we used the median of the other participants to replace the missing values – which was necessary in 17 of the 2200 cases. The results indicated a small to moderate level of agreement among the raters, ICC(A,1) = 0.31 (95% *CI* [0.23, 0.42], $F(49, 722) = 25.8$, $p < .001$). Albeit these results suggest a consistent assessment of similarity across raters within the context of our study, there are also significant individual differences in judging the similarity of speaker pairs highlighting the difficulty to evaluate the similarity of two voices.

General observations. Participants skipped on average $M = 2.30$ trials ($SD = 1.79$) and needed, on average, $Mdn = 26.21$ minutes to complete the experiment.

The findings confirm that AI-derived cosine values are predictive of human similarity judgments, though their strength varies depending on dataset restrictions and individual variability. Stronger correlations at extreme similarity values underscore participants' difficulties in making reliable judgments for speaker pairs of average similarity. This pattern complements the quadratic relationship observed in Experiment 1, where the dissimilarity of speaker pairs posed challenges for nuanced ratings. Overall, the results reinforce that similarity judgments are inherently challenging and prone to individual differences, particularly in moderate similarity categories.

Similarity judgments in relation to the own voice

The first two experiments demonstrated the encoders' ability to predict similarity judgments when the voices are those of other people. In the third experiment, we investigated whether this holds true if one of the voices is one’s own voice. Since the perception of one’s own voice depends

on whether we are speaking or just listening to an audio sample of our voice (Pörschmann, 2000), we manipulated this as a between-subjects factor. In the internal group, we only presented an audio sample of a speaker and asked participants to compare this sample with their own (internal) voice. In the external group, we presented an audio sample and additionally a sample of the participant, which was recorded prior to the experiment.

Method

Participants. To ensure statistical consistency across experiments and facilitate meaningful comparisons, we used the same sample size of 100 participants in Experiment 3 (and the remaining experiments) as in Experiment 1. This approach minimizes potential discrepancies arising from differences in statistical power. Therefore, we recruited 100 new German participants via prolific. Ten were excluded because they detected fewer than two control trials. The mean age was $M = 25.76$ ($SD = 6.81$). Fifty-one participants were female, two diverse, and three did not specify their sex. Recruitment occurred from August 14, 2021, to August 30, 2021.Participants received £5.00 for their participation in the study.

Materials, stimuli and procedure. In a first session, each participant recorded five sentences. These recordings were used to compute the feature vector of their voice. We then calculated the cosine similarities of the participants' voice embeddings with all speakers of the same gender in our dataset. In order to achieve the highest possible variance in cosine similarities, we assigned each raw cosine similarity value to one of ten similarity categories – where the category boundaries from the first two experiments were used. Ten speakers were randomly selected from each category, resulting in a total of 100 speakers. Since the cosine similarity values are approximately normally distributed, the extreme categories would be underrepresented otherwise. If there were not enough speakers in the more extreme category, a speaker was chosen from the category closer to the mean. We picked one audio sample from our dataset for each speaker, trimmed it to a maximum length of 5 s, and normalized the volume.

In the second session, after checking the audio setup, in 100 trials the participants were asked to rate the dissimilarity of the presented voice in comparison to their own voice. If they were assigned to the external representation group, on each trial, participants were randomly presented with one of their five audio recordings, followed by another person's audio sample – without them having been instructed to listen to their recordings beforehand. In the internal representation group,

only the speaker from our dataset was presented. In both groups, we simply asked the subjects to rate the similarity of the other person's voice to their own voice, and therefore did not mention that their own voice might have an internal representation.

After presenting three introductory trials, every 30th trial contained recordings of two different speakers, none of which came from the participant.. Participants were asked to detect these pairs by clicking a red button below the rating scale. The participants who caught fewer than two control trials were excluded.

Results

We used mixed models with participants as random effects, the raw cosine values as the independent variable, and judged similarity as the dependent variable. We compared an intercept-only model, a linear model, and a quadratic model. Additionally, we included the between-subjects factor as an interaction term. Weighting using the AIC scores in Table 1.3 showed a linear relationship (Figure 1.3) between cosine similarity values and participants' rating. The model explained 21% of the variance in similarity ratings (intercept: 0.34, *95% CI* [0.32, 0.37], $t(101.1) = 26.32$, $p < .001$; cosine: 0.15, *95% CI* [0.12, 0.18], $t(8853) = 9.80$, $p < .001$; $R_c^2 = 0.21$). Whether one's voice was externally presented or not had no significant effect. The median of the individual Spearman Correlation between cosine similarities and similarity ratings was $Mdn\ r_s = 0.11$ ($Q_1 = 0.01$, $Q_3 = 0.23$). Performing an attenuation correction yielded a Spearman Correlation of $Mdn\ r_s = 0.15$. This reflects moderate predictive power at the individual level. Participants skipped on average $M = 2.11$ trials ($SD = 2.56$) and needed, on average, $Mdn = 25.07$ minutes to complete the experiment.

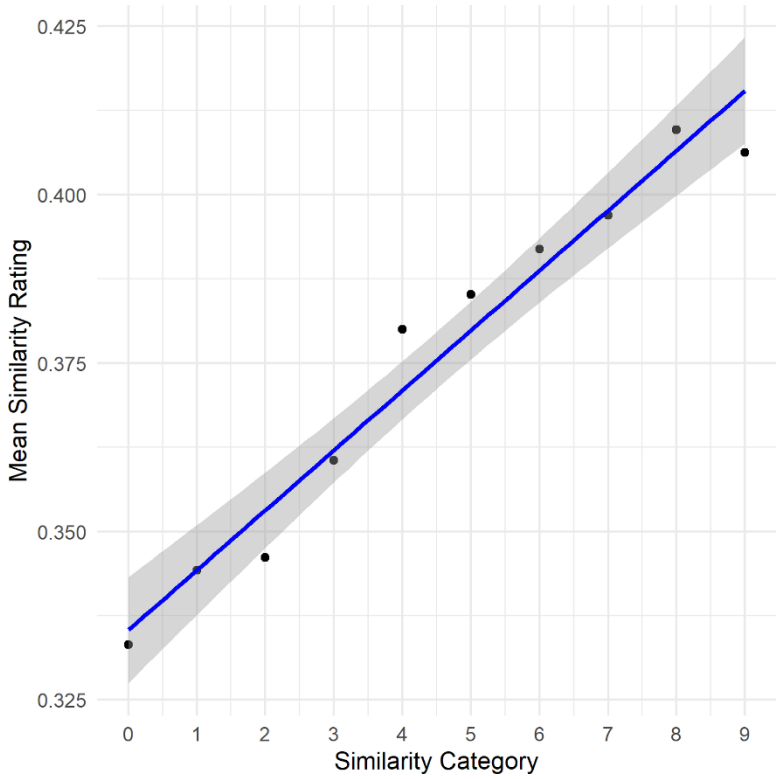
We conducted additional regression analyses, employing the similarity category as the predictor and the average similarity judgments as the dependent variable. An ANOVA contrasting the quadratic with the linear model just missed the threshold of significance, $F(1,7) = 4.27$, $p = .078$. The analysis of the linear model revealed a significant effect of the similarity category on the mean similarity ratings, $F(1,8) = 194.1$, $p < .001$ (intercept: 0.34, *95% CI* [0.33, 0.343], $t(7) = 98.34$, $p < .001$; category: 0.009 (*95% CI* [0.007, 0.010], $t(7) = 13.93$, $p < .001$). This model accounted for a significant variance in mean similarity ratings, as indicated by an $R^2 = 0.96$.

As we noticed an overall decrease in similarity ratings when participants compared voices to their own voice, we conducted post-hoc Tukey-Kramer tests to investigate differences across the three experiments. Significant differences emerged between the average slider responses: $M_2 - M_1 =$

0.03, t = 5.49, p < .001; M_3 - M_1 = -0.08, t = -18.80, p < .001; M_3 – M_2 = -0.11, t = -20.44, p < .001. These findings suggest a consistent bias where participants are less likely to judge voices as similar to their own.

Figure 1.3

Illustration of the Results of Experiment 1.3



Note. Depicted is the quadratic relationship between cosine similarity categories and the participants' mean similarity ratings as well as the the 95% confidence interval of the regression line.

Experiment 3 showed the encoders' ability to partially predict similarity judgments even when one of the voices is one's own voice. Unlike Experiments 1 and 2, Experiment 3 revealed a linear relationship between raw cosine similarity values and judged similarity, likely reflecting a general bias against perceiving self-voice similarities. The lower similarity ratings may stem from a *Need for Uniqueness* (C. R. Snyder & Fromkin, 1977), where participants hesitate to identify other voices as similar to their own. Alternatively, heightened familiarity with one's own voice may increase sensitivity to subtle differences, leading to underestimation of similarity. Surprisingly, whether participants compared the presented voices to an internal mental representation, or an

external recording of their own voice had no significant effect. This suggests that the internal representation of one's voice may serve as the dominant reference point.

The beauty in average voices

Because the previous experiments indicated that the cosine similarity is a valid proxy for perceived similarity, we investigated cognitive consequences of voice similarity in the remaining experiments. In this experiment we focused particularly on the likability and trustworthiness of average voices. Our investigation was motivated by the concept of the beauty-in-averageness effect (Langlois & Roggman, 1990; Winkielman et al., 2006), which suggests that average features are often perceived as more attractive – even though they may not be optimally attractive (Perrett et al., 1994). This effect, well-documented in studies focusing on facial stimuli (Rhodes et al., 2001), may extend to auditory perceptions. By examining whether voices with average characteristics (determined by mean cosine similarity across our speaker dataset) are perceived as more likable and trustworthy, we wanted to explore whether this phenomenon transcends visual stimuli and applies to auditory perceptions as well.

Method

Participants. We recruited 100 new German participants via prolific. Two of them were excluded because they had less than 90 submitted ratings. The mean age was $M = 30.32$ ($SD = 11.64$). Forty-three participants were female, two diverse. Recruitment occurred from October 21, 2021, to October 26, 2021. Participants received £2.82 for their participation in the study.

Materials, stimuli and procedure. We computed the cosine similarities of each male speaker with each other male speaker. We used the mean cosine similarity of a speaker as a measure of typicality. To obtain the highest possible variance in typicality, we assigned each mean cosine similarity value to one of ten similarity categories and randomly selected ten speakers from each category, resulting in a total of 100 speakersWe picked one audio sample from our dataset for each speaker, trimmed it to a maximum length of 5s, and normalized the volume. While the overall design of the experiment mirrored that of the second session in the third experiment – detecting sine tones to verify audio settings, performing three introductory trials, and judging 100 speakers – there were key differences. Instead of rating similarity, participants were asked to assess the likability and trustworthiness of the

speakers using two continuous rating scales (ranging from "not at all" to "very"). Additionally, for the control trials, participants were required to detect audio samples from two female speakers.

Results

Considering the AIC scores in Table 1.4, there was no significant effect of mean cosine similarity on likeability ratings. The model comparison for the trustworthiness ratings (see Table 1.5), revealed a quadratic relationship (Figure 1.5; intercept: 0.52, *95% CI* [0.49, 0.54], $t(178.19) = 39.35$, $p < .001$; mean cosine: -0.18, *95% CI* [-0.38, 0.02], $t(9627.08) = -1.88$, $p = .06$; mean cosine^2: 0.74, *95% CI* [0.18, 1.30], $t(9627.08) = 2.57$, $p = .01$; $R_c^2 = 0.21$, indicating that the model explained 21% of the variance in trustworthiness ratings). The median of the individual Spearman Correlation between mean cosine similarities and trustworthiness ratings was $Mdn\ r_s = 0.04$ ($Q_1 = -0.03$, $Q_3 = 0.09$), reflecting weak associations.. Participants skipped on average $M = 2.74$ trials ($SD = 1.12$) and needed, on average, $Mdn = 20.29$ minutes to complete the experiment.

Analyzing averaged data did not reveal any significant effects in mean cosine similarity values on likeability or trustworthiness ratings (all $p > .05$).

These results suggest that a voice's typicality (as captured by mean cosine similarity) does not affect likeability and has only a negligible influence on perceived trust. However, it should be noted that this could also be due to an insufficient number of particularly typical and atypical speakers or that the semantic content could have biased the evaluation. Previous research also pointed out that averageness has positive effects only in some dimensions but not in others (Said & Todorov, 2011).

The attraction towards similar voices

In the final experiment, we investigated whether speakers with similar voices to one’s own voice are perceived as more likable and trustworthy.

Method

Participants. We recruited 100 new German participants via prolific. Seven were excluded because they had less than 90 submitted ratings. The mean age was $M = 31.04$ ($SD = 11.28$). Forty-five participants were female. Recruitment occurred from November 11, 2021, to November 29, 2021. Participants received £3.76 for their participation in the study.

Materials, stimuli and procedure. As in the third experiment, the first session was used to record voice samples, compute the voiceprints, and sample 100 same-gender speakers from our dataset with a wide variety of cosine similarities. The second session was identical to Experiment 4.

Results

We compared an intercept-only model, a linear model, and a quadratic model for both the likeability rating as well as the trustworthiness rating. Weighting using the AIC values in Table 1.6 showed a quadratic relationship between the voice similarity and likeability ratings (intercept: 0.47, *95% CI* [0.45, 0.49], $t(111) = 46.70$, $p < .001$; cosine: 0.11, *95% CI* [0.05, 0.18], $t(8863) = 3.55$, $p < .001$; mean cosine^2: 0.17, *95% CI* [0.04, 0.30], $t(8864) = 2.59$, $p = .009$; $R_c^2 = 0.19$, indicating that 19% of the variance in likeability ratings was explained by the model). The median of the individual Spearman Correlation between cosine similarities and likeability ratings was *Mdn* $r_s = 0.15$ ($Q_1 = 0.05$, $Q_3 = 0.25$), which, while modest, demonstrates a consistent positive relationship.

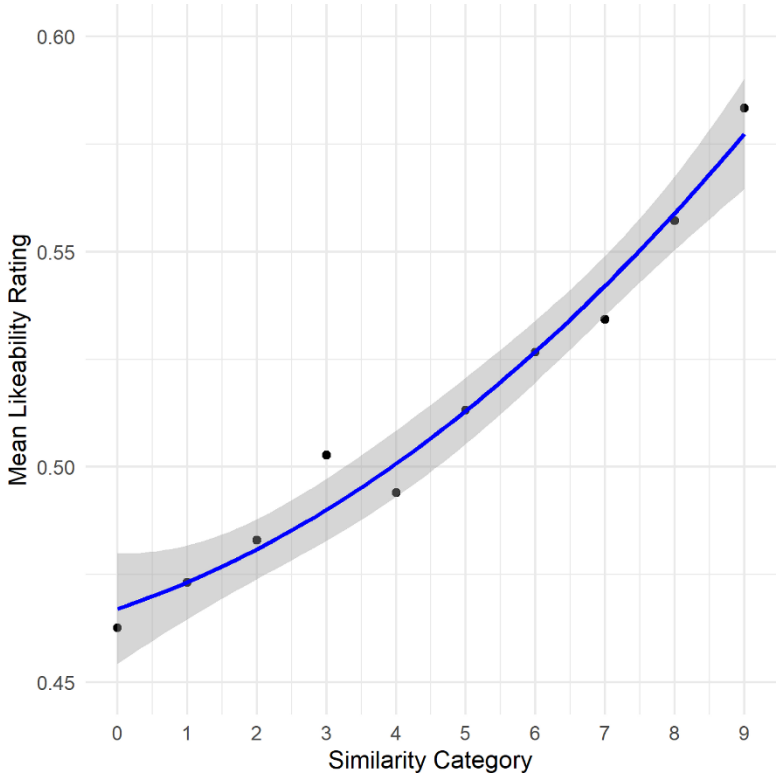
Weighting using the AIC values in Table 1.7 showed also a quadratic relation between the voice similarity and trustworthiness ratings (intercept: 0.50, *95% CI* [0.48, 0.52], $t(111.2) = 47.55$, $p < .001$; cosine: 0.06, *95% CI* [-0.01, 0.12], $t(8863) = 1.74$, $p = .08$; mean cosine^2: 0.31, *95% CI* [0.17, 0.45], $t(8865) = 4.45$, $p < .001$; $R_c^2 = 0.19$, indicating that 19% of the variance in likeability ratings was explained by the model). The median of the individual Spearman Correlation between cosine similarities and trustworthiness ratings was *Mdn* $r_s = 0.16$ ($Q_1 = 0.06$, $Q_3 = 0.25$), which, while once again modest, demonstrates a consistent positive relationship. Participants skipped on average $M = 5.95$ trials ($SD = 3.12$) and needed, on average, *Mdn* = 19.06 minutes to complete the experiment.

To further investigate the relationship, we employed the similarity category as the predictor with the average likeability judgments serving as the dependent variable. The ANOVA comparison between the quadratic and the linear model demonstrated an improved fit for the quadratic model, $F(1,7) = 6.57$, $p = .04$. The quadratic model's analysis revealed a significant influence of the similarity category on the average likeability ratings, $F(2,7) = 134.5$, $p < .001$ (intercept: 0.467, *95% CI* [0.454, 0.480], $t(7) = 86.4$, $p < .001$; category at 0.005; *95% CI* [-0.001, 0.012], $t(7) = 1.916$, $p = .097$; category^2: 0.0008(*95% CI* [0.0001, 0.0015], $t(7) = 2.564$, $p = .04$). With $R^2 = 0.97$, this model accounted for a substantial portion of the variance.

The ANOVA comparison between the quadratic and linear models using the similarity categories as predictors and the corresponding mean trustworthiness ratings as the response variable, indicated a more favorable fit for the quadratic model, $F(1,8) = 8.31$, $p = .02$. The quadratic model's analysis revealed a significant effect, $F(2,7) = 73.95$, $p < .001$ (intercept: 0.50, *95% CI* [0.48, 0.51], $t(7) = 65.141$, $p < .001$; category: 0.002, *95% CI* [-0.008, 0.011], $t(7) = 0.422$, $p = .686$; category^2: 0.001; *95% CI* [0.0002, 0.0022], $t(7) = 2.883$, $p = .02$). With $R^2 = 0.955$, this model accounted for a significant proportion of the variance.

Figure 1.4

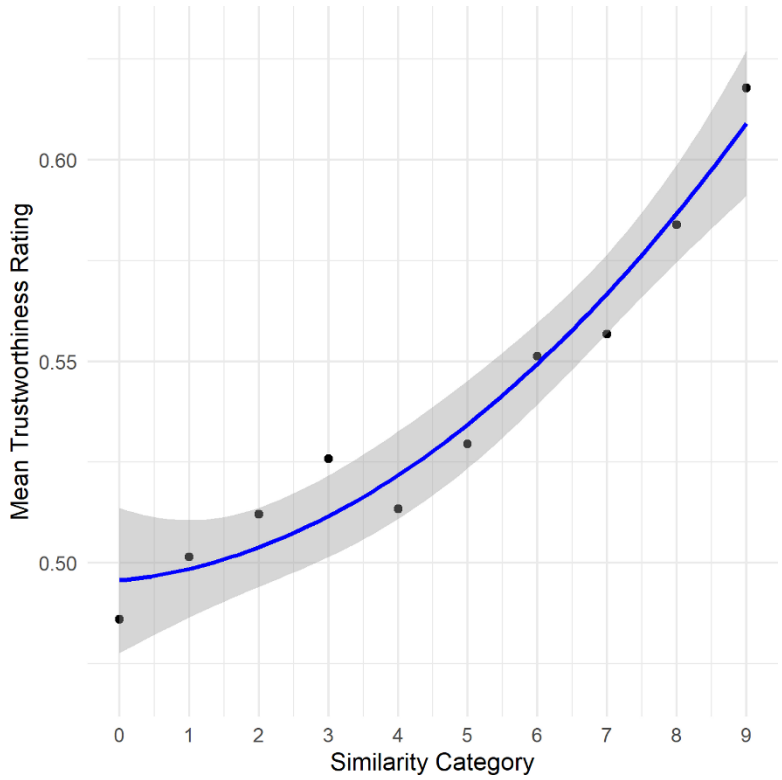
First Illustration of the Results of Experiment 5



Note. Depicted is the quadratic relationship between cosine similarity categories and the participants' mean likeability ratings as well as the the 95% confidence interval of the regression line.

Figure 1.5

Secind Illustration of the Results of Experiment 5



Note. Depicted is the quadratic relationship between cosine similarity categories and the participants' mean trustworthiness ratings as well as the the 95% confidence interval of the regression line.

Taken together, these results demonstrate that the quadratic relationship between voice similarity and ratings of likeability and trustworthiness accounts for a substantial proportion of the variance. These findings suggest that while individual effect sizes are modest, the overall fit of the models underscores the practical relevance of voice similarity in shaping social perceptions. Indeed, these findings support the similarity-attraction hypothesis (Byrne, 1961; Byrne et al., 1967), in such that voices similar to one's own are perceived as more likable and trustworthy. The quadratic relationship suggests that the effect is stronger for higher levels of voice similarity. This effect likely stems from implicit egotism (Hughes & Harrison, 2013; J. T. Jones et al., 2004; Peng et al., 2020) suggesting that individuals evaluate self-associated traits positively, or from social identity processes, in which perceived similarity fosters a sense of connection or group affiliation (Tajfel &

Turner, 1986). These results highlight the potential for AI systems to exploit similarity effects in personalized technologies, such as voice assistants, to influence user perceptions and behavior.

Discussion

Speaker verification systems can compute numerical representations of human voices, so-called voiceprints. Whereas traditionally speaker verification systems served, for example, as a forensic toolkit or as a biometric security feature in highly secured areas, the spread of deep learning technologies also increased the possibilities of utilizing voiceprints. Most importantly with regard to the present study, they could be used to design and shape the voice features of artificial voices. Despite the emerging importance of voiceprints in TTS systems, little research has been conducted on potential cognitive influences (including manipulations) of variations in the voice features of digital assistants. In the present set of experiments, we present such first evidence, including the methodological prerequisites for such an investigation.

The results of our first experiment indicated that the cosine similarity between voiceprints can predict voice similarity judgments (i.e., validity). The resulting quadratic relationship is likely due to disproportionate sensitivity to dissimilar voices. Since voices of close relatives are often similar and voices vary within a speaker depending on time of day, physical condition, and context (Kreiman et al., 2015; Lavan et al., 2019; Y. Lee & Kreiman, 2019), the ability to discriminate between similar voices is a necessary skill for humans to learn. Below a certain threshold at which it is obvious that voices stem from different speakers, we are not aware of any reasonable explanation why it might be valuable to further differentiate between different levels of dissimilarity. Since the trained speaker verification system is agnostic in regard to ecological advantages, it is capable of differentiating even between dissimilar voices. Experiment 2 replicated these results. Moreover, the results revealed a relatively fair test-retest reliability of human similarity judgments. On the one hand, this implies that in principle correlations between similarity judgments and other cognitive variables should be observable. On the other hand, however, should such correlations arise, the numerical values most likely underestimate the true correlations since the reliability was far from being perfect.

The results of Experiment 3 revealed that the AI is also capable of partially predicting the perceived similarity between voices if one of the voices is one's own. Although the correlation was less pronounced, we found a linear relationship between cosine similarities and participant ratings. The most relevant difference for the experiment involving one's own voice relative to judging

similarity between two unrelated voices (Experiments 1 and 2) was that similarity ratings were generally lower. One possible explanation for this is people's *Need for Uniqueness* (C. R. Snyder & Fromkin, 1977), which could make them more hesitant to classify a voice as similar to their own voice. Another interpretation could be that familiar voices are processed differently from unfamiliar voices (Sidtis & Kreiman, 2012; Stevenage, 2018). The familiarity with one's own voice could thus increase the sensitivity for differences which in return might lead to an underestimation of the similarity. Interestingly, whether the participants were asked to compare a voice with their internal representation of their own voice, or an external recording of their voice had no substantial effect on the observed similarity judgments. This is surprising as one's own voice typically is not only transferred via air but also via bone. As there was no difference, we consider it likely that the internal representation of one's own voice may be the dominant reference point from which similarity judgments are made.

After establishing these basic methodological prerequisites for studying correlational relationships, we strived to investigate how voice features might affect basic cognitive evaluations. Following previous research on the beauty in averageness effect, we first investigated whether speakers are perceived as more likable and trustworthy if they have a more average voice. In contrast to other studies, we observed no evidence for a correlation between typicality and likeability and only a marginal effect of typicality on trustworthiness. However, there are substantial differences between previous studies and our experimental approach. In previous studies, an average voice was generated by creating composites, either by statistically averaging speakers (Andraszewicz et al., 2011) or by auditory morphing (Belin, 2021; Bruckert et al., 2010). However, (Zäske et al., 2020) consider that such composites lead to artifacts, especially an increased *harmonics-to-noise ratio* (Hillenbrand, 1987), which is discussed to decrease with age (Ferrand, 2002; Stathopoulos et al., 2011), in stressful situations (Kappen et al., 2022), and cases of hoarseness (Yumoto et al., 1982). Therefore, it is questionable whether the similarity or the more favorable change in the harmonics-to-noise ratio is the reason for the obtained results. In sum, we therefore tend to consider that there is little evidence for a beauty-in-average effect of voices.

With regard to similarity to one's own voice, however, the results of our final experiment drew a different pattern of results. When using one's own voice as reference point, similar voices are perceived as more likeable and trustworthy. These findings match previous studies (Miyake & Zuckerman, 1993; Peng et al., 2019). Again, however, the evidence from previous studies was rather weak, as they did not manipulate similarity directly (Miyake & Zuckerman, 1993) or just adjusted

pitch (+/- 20Hz) or loudness (+/- 10 dB) (Peng et al., 2019). Participants rated samples altered in loudness as more favorable compared to samples shifted in pitch. The authors concluded that this pattern is due to the higher similarity of the loudness manipulated samples to the original recording. However, we are not convinced that participants perceive a recording of their voice as a recording from another person if it is just louder or quieter. Therefore, recordings altered in pitch may not be compared to a similar voice but to one's own voice instead. Since people overestimate the attractiveness of their own voice, the results could be a consequence of this vocal implicit egotism (Hughes & Harrison, 2013). It would have been much more consistent to manipulate the similarity by shifting the pitch by various degrees. Apart from that, shifting the pitch of a recording introduces far more noise than altering the loudness, making it more artificial and possibly unpleasant.

Our study's results underscore the potential for even brief voice recordings to be misused in shaping artificial voices in a way that influences people. The observed correlations between voiceprint similarity and human judgments of likability and trustworthiness (as seen in our final experiment) highlight a vulnerability in human perception. This could be exploited in TTS systems, where slight alterations of voice features aligned with a user's voiceprint could subtly sway their perceptions and behavior. Thus, while our research contributes to understanding the cognitive impact of voice similarity, it also opens discussions about ethical implications in the context of TTS technologies and voice assistants, where personalized voices might be used to manipulate user responses.

Even though the effect may have been relatively small in our experiments, due to the widespread use of voice assistants and the more elaborate methods available to large technology companies, the impact can be tremendous in absolute terms. This applies not only to the use of user-adapted voices to make interactions with the assistants more attractive but also to the impact on advertising messages and political propaganda.

General limitations

While our study provides valuable insights into the relationship between voice similarity and cognitive evaluations, several limitations must be acknowledged.

A major limitation of this study arises from the wide range of voice similarity values that we were able to investigate. By including pairs of voices from a wide range of the similarity spectrum, we were able to ensure a solid understanding of the overall pattern; however, this broad spectrum may have diluted specific effects that are particularly pronounced in highly similar voices. Future

studies should focus more on speakers with high cosine similarity values to better capture the nuances and practical implications of judgments in this critical range. This limitation also reflects a trade-off between experimental control and ecological validity. While our design provided valuable insights into general trends, focusing exclusively on highly similar voices may provide more precise and application-oriented results.

The use of open-source datasets, while providing a wide range of speaker voices, also introduced variability in audio quality, articulation, and linguistic content. These factors may not only have influenced participants' judgments but reduced the internal validity of the experiments. Future work should employ more controlled datasets or systematic manipulations of stimulus properties to reduce potential biases.

Our study emphasized static voice similarity judgments based on pre-recorded audio. Dynamic aspects of speech, such as conversational context, prosody, or situational factors, were not considered. These elements are likely to influence perceptions and warrant exploration in future research.

Finally, the online nature of the experiments presents challenges such as variable listening environments and participant compliance. Although attention checks and control trials were implemented, these measures cannot fully account for potential distractions or technical issues encountered by participants during the study.

Conclusion

Our findings demonstrate that AI-derived cosine similarity measures effectively predict human voice similarity judgments and influence social evaluations. Across the first three experiments, we found significant relationships between the cosine similarity of voice embeddings and participants' similarity ratings, with a quadratic pattern emerging for judgments of other voices. This suggests that participants were particularly sensitive to highly similar and dissimilar voices, while intermediate similarity was more challenging to evaluate.

In Experiment 3, we extended this analysis to self-voice comparisons, revealing a general bias against perceiving other voices as similar to one's own voice. This bias likely stems from increased sensitivity to subtle differences in one's own voice or a Need for Uniqueness (C. R. Snyder

& Fromkin, 1977). Comparisons to an internal mental representation or external audio recordings of the own voice yielded similar results, suggesting that the internal representation serves as a dominant reference.

Experiments 4 and 5 explored how voice similarity influences social perceptions such as likability and trustworthiness. Contrary to the beauty-in-averageness effect found in visual stimuli, we found no evidence that average voices were perceived as more likable and only weak effects on trustworthiness. However, voices similar to one's own were judged as both more likable and trustworthy, supporting the similarity-attraction hypothesis and the influence of implicit egotism.

Our results highlight the potential of AI-generated cosine similarity as a tool for understanding voice perception. While individual effects were modest, the consistency of the findings underscores their practical relevance for voice-based technologies like personalized voice assistants or synthetic speech systems. Future research should focus on refining models for highly similar voices, exploring cross-linguistic generalizability, and addressing the ethical implications of voice similarity manipulations. This study advances our understanding of voice similarity's role in cognition and social interaction by bridging human perception and AI-driven voice representations.

Tables

Table 1.1

Model Selection Table for Experiment 1.

Model name	Degrees of freedom	Log-Likelihood	AIC	Δ AIC	Weight
Quadratic model	5	-354.73	719.5	-	0.911
Quadratic model + gender	6	-356.04	724.1	4.64	0.089
Linear model	4	-520.26	1048.5	329.07	0
Linear model + gender	5	-521.59	1053.2	333.74	0
Intercept-only model	3	-1189.42	2384.8	1665.39	0

Table 1.2

Model Selection Table for Experiment 2.

Model name	Degrees of freedom	Log-Likelihood	AIC	Δ AIC	Weight
Quadratic model	5	-332.97	675.9	-	1
Linear model	4	- 348.40	704.8	28.85	0
Intercept-only model	3	- 459.12	924.2	248.30	0

Table 1.3

Model Selection Table for Experiment 3.

Model name	Degrees of freedom	Log-Likelihood	AIC	Δ AIC	Weight
Linear model	4	245.52	-483.0	-	0.896
Quadratic model	5	244.35	-478.7	4.33	0.103
Linear model + group	6	240.98	-470.0	13.07	0.001
Quadratic model + group	8	239.00	-462.0	21.06	0
Intercept-only model	3	201.08	-396.2	86.87	0

Table 1.4

Model Selection Table for Experiment 4 – Likeability.

Model name	Degrees of freedom	Log-Likelihood	AIC	Δ AIC	Weight
Intercept-only model	3	1231.85	-2457.7	-	0.938
Linear model	4	1229.77	-2451.5	6.18	0.043
Quadratic model	5	1229.95	-2449.9	7.82	0.019

Table 1.5

Model Selection Table for Experiment 4 – Trustworthiness.

Model name	Degrees of freedom	Log-Likelihood	AIC	Δ AIC	Weight
Quadratic model	5	1162.81	-2315.6	-	0.749
Intercept-only model	3	1159.19	-2312.4	3.25	0.148
Linear model	4	1159.83	-2311.7	3.96	0.104

Table 1.6

Model Selection Table for Experiment 5 – Likeability.

Model name	Degrees of freedom	Log-Likelihood	AIC	Δ AIC	Weight
Quadratic model	5	1576.62	-3143.2	-	0.635
Linear model	4	1575.07	-3142.2	1.11	0.365
Intercept-only model	3	146638	-2926.8	216.48	0

Table 1.7

Model Selection Table for Experiment 5 – Trustworthiness.

Model name	Degrees of freedom	Log-Likelihood	AIC	Δ AIC	Weight
Quadratic model	5	1206.21	-2402.4	-	0.999
Linear model	4	1198.09	-2388.2	14.25	0.001
Intercept-only model	3	1093.25	-2180.5	221.92	0

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Declaration according to § 5 Abs. 2 No. 8 of the PhD regulations of the Faculty of Science -Collaborative Publications

The following chapter (Chapter 3) consists of a manuscript that is currently under review for submission and was co-authored by Stephan Schwan, Leonie Aßmann, Tilmann Batsch, and Hauke S. Meyerhoff. The proportional contributions to this manuscript are presented in the subsequent table.

Author	Author position	Scientific ideas	Data generation	Analysis & interpretation	Paper writing
Oliver Jaggy	First author	60 %	100 %	65 %	55 %
Stephan Schwan	Second author	10 %	0 %	10 %	15 %
Leonie Aßmann	Third author	5 %	0 %	5 %	5 %
Tilmann Betsch	Fourth author	10 %	0 %	10 %	10 %
Hauke S. Meyerhoff	Fifth author	15 %	0 %	10 %	15 %

Title of paper: The Impact of Voice Similarity on Decision-Making: Do We Follow Advisors with Similar Voices?

Status in publication process: Under Review.

Chapter 3: The Impact of Voice Similarity on Decision-Making: Do We Follow Advisors with Similar Voices?

Abstract

This study investigates the impact of voice similarity on decision-making, a growing concern as AI-driven voice assistants become more personalized. Across three experiments, participants were incentivized to locate treasures based on predictions from advisors with either a similar or generic voice. Auditory predictions were derived from pre-recorded samples of 600 native English speakers, and voice similarity was calculated by comparing participants' voiceprints with those of the advisors using cosine similarity values. In Experiment 1, participants received conflicting advice from advisors of varying reliability using a probabilistic inference paradigm with visual (pictures of a spider or a treasure) and auditory predictions ('There is a treasure.' or 'There is a spider.'). The results showed no improvement in decision quality with similar voices, as participants followed the majority rule irrespective of voice similarity. Experiment 2 simplified the task with two equally reliable advisors, yet voice similarity still had no effect. In Experiment 3, the visual component of the predictions was removed to assert a proper processing of the auditory cues, again, without an effect of voice similarity. However, exploratory analysis including the spatial position of advisors on the display, revealed a significant interaction between spatial positioning and voice similarity on decision outcomes, with a preference for similar sounding advisors on top of the display. Thus, our results indicate that voice similarity might have a negligible impact on decision making. If anything, voice similarity might be able to bias decision making in combination with rather weak heuristics such as spatial position.

Significance Statement

This research exemplifies use-inspired basic research by addressing a pressing societal challenge: understanding how everyday decision-making can be influenced by subtle auditory cues in an increasingly digital world. As voice assistants and AI-driven communication technologies become more integrated into our daily lives, the ability to personalize these systems – for example, by mimicking a user's own vocal attributes – is seen as a promising

way to boost trust and user engagement. However, whether such personalized voice characteristics truly enhance decision quality remains unclear.

By systematically investigating the impact of voice similarity on decision-making, our basic research elucidates the cognitive processes and heuristics that underlie how people integrate and act upon auditory information. This work not only contributes to a more nuanced understanding of decision heuristics – such as the dominance of visual or spatial cues – but also challenges prevailing assumptions about the persuasive power of voice similarity. In doing so, it bridges the gap between fundamental cognitive theory and practical applications, providing valuable insights that could inform the design and ethical deployment of personalized voice technologies.

Introduction

Imagine a moment of indecision when you do not know what meal to choose, what stock to acquire, or which candidate to vote for. Now imagine seeking guidance from your favorite voice assistant, an assistant that is not just any helper but tailored to resemble aspects of your identity, for instance your voice. An assistant designed to resonate with you on a more personal level: Would you trust it? Could such a personalized guide sway your decisions beyond mere logic, subtly influencing your choices in ways you might not fully understand? At a time when deep fakes are more and more convincing, where the boundaries between reality and fiction are blurring, it is crucial to understand the impact of such technologies on our decisions. In this article, we investigate whether voices similar to our own can impact basic decision-making processes.

The Logic of Choice

Decision-making is a major topic in psychology (for a review, see Fischhoff & Broomell, 2020). Research in this field has been profoundly influenced over the past fifty years by the work of Tversky and Kahneman. One of their major conclusions was that individuals typically do not employ statistical methods or probability theory as expected when making decisions in vague or risky situations and, therefore, do not follow the utility theory (Edwards, 1954). Instead, they depend on a set of heuristics that are prone to systematic mistakes (Tversky & Kahneman, 1973; Tversky & Kahneman, 1974) and irrational behavior (Gilovich et al., 2002). For example, the affect heuristic – which expresses the influence of

positive and negative emotions on decisions (Slovic et al., 2007b) – underscores that our feelings can significantly bias judgments, especially in situations of uncertainty. Although these heuristics sometimes produce systematic errors, they serve as practical strategies to cope with our limited cognitive resources. Therefore, decision-making is influenced by both objective and subjective cues. Objective information, or cue validity, provides factual details about the environment, while subjective cues, such as liking, evoke affective responses that can sway our choices (Betsch et al., 2014).

Social Influences on Decision-Making

In the context of this study, we are particularly interested in the effect of similarity between the advisor and the person who is to reach an opinion, belief, or judgment. In research on persuasion, similarity is considered a source characteristic, which refers to attributes of the individual or entity providing information or attempting persuasion (Hovland et al., 1953; for a review on source effects, see Simons et al., 1970; E. J. Wilson & Sherrell, 1993). Likewise, in research on consumer psychology, similarity is employed as a means of aligning a consumer with the source of a message (for a review, see Teeny et al., 2021). In general, research indicates that similar others are more persuasive (Faraji-Rad et al., 2015). A wide range of studies has explored this effect with different types of similarity. These include incidental similarity (Burger et al., 2004; Jiang et al., 2010), alignment in political beliefs and affiliation (Nelson & Garst, 2005; Roth et al., 2020), shared values (Silvia, 2005), common attitudes and social background (Feng & MacGeorge, 2010), as well as cultural or ethnic similarity (Anderson & McMillion, 1995; Ivanič et al., 2014; Tsalikis et al., 1991). However, while observable similarities—such as appearance or cultural traits—tend to be more noticeable, they may actually have a smaller impact on persuasion compared to internal similarities, like shared values or beliefs (Lichtenthal & Tellefsen, 2001). This finding contrasts with earlier research, which suggested that perceived similarity (how similar individuals *believe* they are) can have a stronger influence than actual, observable similarity, regardless of whether those similarities are truly present (Orpen, 1984).

There a several explanations why, under such terms, similarity has such an influential effect. One possible explanation stems from research on the influence of emotions on decision-making (for review, see Loewenstein et al., 2003): Lerner et al. (2015) argue, that emotions may be the dominant driver of many significant life decisions. Indeed, according to the

similarity attraction hypothesis (Byrne et al., 1967), similar others are perceived more favorably than dissimilar others – including similar voices . The evocation of positive emotions towards them could make their advice more influential. Similarly, since we are often inclined to perceive ourselves in a positive light (Taylor & Brown, 1988), entities that seemed associated with ourselves are unconsciously preferred over unrelated objects or individuals. This so called implicit egotism (Pelham et al., 2002) could either shift our attentional focus on cues or advisors that we associate with ourselves, or could promote cognitive processing more directly (Aron et al., 1991). Furthermore, the observed (superficial) similarity could be overgeneralized, so that we assume the advisor must have similar preferences, which in turn, makes their advice more diagnostic (Hovland et al., 1953).With the advancing capabilities of Text-to-Speech Systems to produce speech, including the possibility to resemble a user's voice or at least make it more similar our research focuses on the consequences of voice similarity on decision-making processes. This seems particularly interesting since recent studies have shown that similar voices are perceived as more likeable and trustworthy (Jaggy et al., 2025). While a fair number of studies have utilized speech to deliver messages, they have altered only superficial voice characteristics to vary similarity between the advisor and advisee. These modifications included, for example, accents (Ivanič et al., 2014: Tsalikis et al., 1991) or pitch (Banai et al., 2018). However, current research has yet to explore the implications of AI's ability not only to mimic nuances like accents but to clone an entire voice, particularly concerning the effects on decision-making. Given the increasing number of interactive devices that incorporates voice assistants and their growing integration in our daily life, understanding the potential influence of voice similarity on decision outcomes and persuasion attempts becomes increasingly significant.

In three experiments, we investigated whether participants were more likely to follow the predictions of advisors with similar voices.

In our first experiment, we used a variant of the information board approach (Payne et al., 1988), requiring individuals to choose one of several options. Our information board depicted three different advisors from which participants could retrieve information about the location of a hidden treasure. We investigated whether participants who get relevant information from a similar advisor have better decision outcomes than those who get relevant information from a generic advisor – especially when in the presence of conflicting advice. This approach allows us to disentangle the influence of objective information (cue validity)

from that of subjective, affect-driven cues, such as voice similarity. In the subsequent experiments, we used a distilled version of the information board display to investigate whether voice similarity alters decision-making more broadly.

Experiment 1

The first Experiment was designed to investigate whether participants make more rational choices when they receive advice from a competent advisor with a similar voice. In each trial participants were monetary incentivized to find a treasure behind three houses. To get clues about the treasure's location, they could retrieve visual and auditory predictions from three advisors with different predictive capabilities: two of them pointed in 56 % of the cases to the correct house, one of them in 86 %. While the combined predictions of the less capable advisors cannot outperform the predictions from the competent advisor, people's choices are often skewed by the majority bias that arises when both low-validity advisor point to the same house (Aßmann et al., 2022). We hypothesized that participants more often follow the advice of a high-validity advisor with a voice similar to their own, resulting in a less biased decision-making process. As all other experiments in this study, the design, hypotheses, and analysis plans were preregistered prior to data collection (see below).

Methods

Participants. For a repeated measures ANOVA to achieve a power of .95, an alpha error probability of .05, and an assumed effect size of partial $\eta^2 = 0.01$ for the within-between interaction, G*Power 3.1.9.7 computed a necessary sample size of at least 104 participants. To compensate for possible exclusions, we recruited 120 English native speakers through Prolific. After excluding 13 participants that collected, on average, fewer than two predictions per trial, the final sample consisted of 53 females, 43 males, and 11 people that did not disclose their gender. The mean age of the sample was $M = 39.86$ years ($SD = 13.51$ years). As all other experiments in this study, it received ethical approval from the ethics committee of the Leibniz-Institut für Wissensmedien, Tübingen. In all experiments in this study, participants gave their informed consent prior to their participation.

Materials and Stimuli. The auditory predictions used in our study were collected beforehand. We recruited 300 female and 300 male English native speakers through Prolific

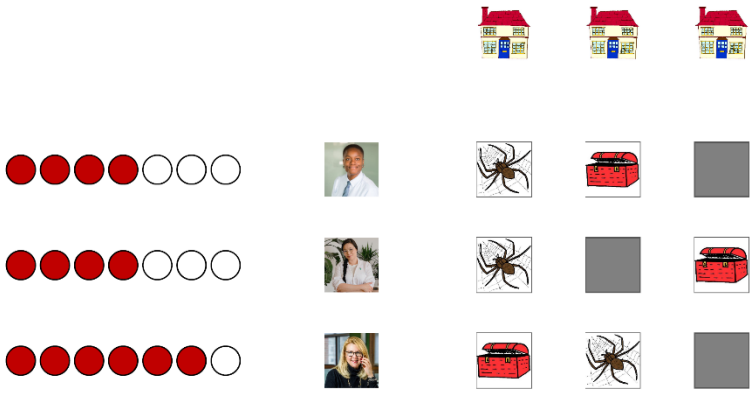
and asked them to record multiple sentences (e.g., “There is a spider!”, “There is a treasure!”) via a Qualtrics integrated phonic.ai interface.

To investigate whether similarity could improve decision outcomes, we incorporated the voice similarity of the high-validity advisor as a between-subjects factor. Voice similarity was determined by comparing the voice characteristics of the speakers in our dataset with the participants’ voices. Various methods were developed to extract voice characteristics and identify and recognize speakers (Kabir et al., 2021). In this study, we used an open-source Text-to-Speech (TTS) system (Eren & The Coqui TTS Team, 2023) with an encoder based on the VITS model (Kim et al., 2021) to derive *d*-vectors. *D*-vectors are feature vectors of voices generated by a speaker verification system that uses deep neural networks (Variani et al., 2014). If properly trained, the resulting feature vectors can be used as numerical representations of voices and to identify and recognize speakers. The similarity between two speakers’ voices can be determined by comparing their feature vectors, for example, by calculating their cosine similarity (Ohi et al., 2021; Sztahó et al., 2021).

We used PsychoPy to develop an adapted version of MouseKids, a tool previously utilized for investigating decision processes (e.g., Betsch et al., 2016; Betsch et al., 2018). The resulting display of the main trials used in Experiment 1 is depicted in Figure 2.1.

Figure 2.1

Display Used in Experiment 1



Note. On the left-hand side, the prediction probability of each advisor is represented as smart points. In each trial, the predictions are hidden behind a grey rectangle. The predictions are shown by clicking on the rectangles, and the matching audio file is played – either “There is a spider.” or “There is a treasure.”. When participants are convinced that they have gathered enough information, they can click on one of the three houses to see whether a treasure is to be found.

On top of the display, three houses were depicted, each containing a treasure or a spider. In each trial, the participant should guess behind which house the treasure is located. In the center of the display, a three-by-three matrix of grey rectangles was visible. Clicking on one of these revealed the prediction of an advisor for the corresponding house. Predictions were presented visually as pictures and auditorily by playing the matching audio recording.

On the left-hand side of the display, smart points indicated the predictive capabilities of the advisors, i.e., their success rate in identifying treasure locations. Specifically, the top two advisors pointed in 56% of the cases to houses with hidden treasures, and the advisor placed at the bottom in 86%. This dispersion of validities creates a non-compensatory environment where the predictions of the high-validity cue are more predictive than the combined predictions of the two lower-validity cues.

In each trial, we used one of four prediction patterns, mirroring those used by Aßmann et al. (2022). Figure 2.2 depicts the general prediction combinations. Importantly, in the first two patterns, both low validity cues point together to at least one house. When both less accurate advisors agree on a location, participants must decide whether they follow the majority rule or the single, more reliable cue. In such instances, participants' ability to make rational decisions is challenged. Therefore, they are called conflicting patterns. The last two patterns show a more equal distribution of predictions. In such instances, following the advice of the high-validity advisor is obviously the best choice. Therefore, these patterns are termed non-conflicting. Each pattern was presented 12 times in different combinations.

Figure 2.2

Prediction Patterns Used for Experiment 1

Conflicting Pattern 1			
	House		
Validities	**A**	**B**	**C**
Advisor 1 (.56)	1	1	0
Advisor 2 (.56)	1	0	0
Advisor 3 (.86)	0	0	1

Conflicting Pattern 2			
	House		
Validities	**A**	**B**	**C**
Advisor 1 (.56)	1	0	1
Advisor 2 (.56)	1	0	1
Advisor 3 (.86)	0	1	0

Non-Conflicting Pattern 1			
	House		
Validities	**A**	**B**	**C**
Advisor 1 (.56)	1	0	0
Advisor 2 (.56)	0	1	0
Advisor 3 (.86)	0	0	1

Non-Conflicting Pattern 2			
	House		
Validities	**A**	**B**	**C**
Advisor 1 (.56)	1	1	0
Advisor 2 (.56)	0	0	1
Advisor 3 (.86)	0	1	0

Note. A "1" indicates that the advisor predicts a treasure for the respective house. The first two patterns are dubbed conflicting because both low-validity cues make at least one coherent prediction. In such cases, the majority rule can bias the decision-making process. That is not the case in the non-conflicting pattern where the low-validity advisors point to different houses.

Alongside the smart points, pictures of the advisors were displayed. We derived pictures of 21 individuals supposedly perceived as female and 21 as male from pexels.com. They represent diverse ethnic backgrounds, all portrayed in professional settings (e.g., with a laptop, at a desk, or in business attire). For each participant, three pictures based on their preferences were used as visual representations of the advisors (see below).

Procedure. The Experiment consisted of two sessions that were, at most, three days apart. In the first session, we used the platform Qualtrics to collect basic demographic information and participants' consent. Via a phonic.ai plug-in, participants recorded a short sentence ("*Hello, I am one of your advisors and will help you in the treasure hunt***.")**. Additionally, participants rated the likability of 21 same-gender images on a 100-point scale. After the first session, the feature vectors of the voice samples were generated, and the cosine similarity with all gender-matched speakers in our dataset was calculated. For participants in the similar voice condition, the speaker with the highest cosine value compared to their own voice was identified, and their respective recordings were used as stimuli for the high-validity

cue. Recordings for the low-validity cues were chosen from speakers with the most average cosine values. In the non-similar voice condition, all advisors had recordings from speakers with average cosine values. To minimize the effects of sympathy, for each participant, we picked the three images with likeability ratings closest to the mean and used them as pictures of the advisors. Consequently, each participant experienced a personalized experiment featuring advisor images with average likability ratings and a high-validity cue voice that was either similar to their own or of average similarity, like the low-validity cues. After generating the experiment, participants were invited to the second session through prolific.

In the first part of the second session, participants should familiarize themselves with the advisors and learn their predictive abilities through seven trials per advisor. Each introduction began with displaying the advisor's image and an introductory audio message: "*Hello, I am one of your advisors and will help you in the treasure hunt.*" This was followed by a display featuring seven unmarked smart points, the advisor's image, a house image, and a grey rectangle. Participants were instructed to click on the rectangle to unveil the advisor's prediction and then the house to reveal its content. A correct prediction (matching treasure or spider) triggered the audio response, "*I was right. I get a smart point*". Then, the smart points were increased in size for one second, and one additional smart point was filled out. If the prediction was incorrect, the recording "*I was wrong. I don't get a smart point*" was played, and the smart points were increased in size for one second, but no additional smart point was added.

After completing the introduction trials of all three advisors, an attention check was conducted where participants identified the advisors by their voice. Upon hearing the sentence "*Which advisor am I*?" participants should click on the correct advisor image. An incorrect assignment concluded the experiment.

Before starting the main phase of the experiment, an introduction trial was conducted, where participants could familiarize themselves with the paradigm. They were informed that they could retrieve as many predictions as needed before choosing a house.

In the 48 experimental trials, participants clicked grey rectangles to retrieve visual and auditory predictions before selecting a house. Finding a treasure resulted in a "Treasure found!" text message, while unsuccessful attempts displayed "Treasure not found!" After completing the study, participants received an additional, pre-announced bonus through Prolific of £0.05 for each treasure found.

Results

In examining the impact of voice similarity, pattern combination, and their interaction on making a correct decision (following the advice of the high-validity cue), an mixed Analysis of Variance (ANOVA) was conducted. The results revealed a significant main effect of the pattern combination ($F(1, 236) = 11.18$, $p < 0.001$, $\eta^2 = 0.05$). Neither voice similarity ($F(1, 236) = 0.17$, $p = 0.68$) nor the interaction between pattern combination and voice similarity reached significance ($F(1, 236) = 0.16$, $p = 0.69$). These findings suggest that while non-conflicting pattern combinations increase the probability of making the correct decision ($M = 0.58$) compared to conflicting patterns ($M = 0.49$), voice similarity and its interaction with pattern combinations do not influence the decision outcome.

We also analyzed if participants overused information. To make an accurate decision, only revealing information provided by the high-validity cue would be sufficient. Throughout the Experiment, the high-validity cue makes 144 predictions. Indeed, with an average of $M = 215.57$ ($SD = 99.34$), a two-sided t-test revealed that participants gathered significantly more information than necessary ($t(106) = 7.45$, $p < .001$).

Discussion

In the first experiment of this study, we investigated how conflicting information and voice similarity impact decision-making quality. We found that participants often struggle to make optimal decisions, particularly when both low-validity cues provided a consistent prediction. In such cases, they tended to follow the majority rule, ignoring the advice from the high-validity cue (J. Zhang et al., 2006). Furthermore, participants tended to seek more information than necessary, suggesting difficulty in prioritizing the most relevant information. The results align with previous studies showing that even individuals trained to ignore irrelevant information face challenges in effectively doing so (Söllner et al., 2014).

In sum, we replicated the pattern effect and the overuse of information demonstrated by Aßmann et al. (2022), but did not find any effect of voice similarity in the decision-making process. However, there could be a potential ceiling effect that may limit the additional influence of voice similarity, as participants might be strongly inclined to follow the high-validity advisor; this issue is discussed further in the General Discussion.

Experiment 2

Experiment 1 revealed no effect of voice similarity on decision quality. We hypothesize that the majority rule bias is too strong to be negated by a possible similarity effect. Therefore, our goal in Experiment 2 was to isolate the effect of voice similarity by simplifying the decision-making context and to clarify whether voice similarity alone can significantly impact decision-making processes.

In this experiment only the predictions of two advisors with the same predictive capabilities were used. The only difference between those two was the similarity of their voice. We hypothesized, that participants more often followed the advice of a similar advisor compared to an advisor with an average voice.

Methods

Participants. To achieve a power of 0.85 in a Generalized Linear Mixed-Effects Model, with an alpha error probability of 0.05 and accounting for a small effect size of voice similarity, we calculated a required sample size of at least 110 participants. We recruited new English native speakers via Prolific. After excluding 23 participants who, on average, made fewer than two predictions per trial, our final sample consisted of 96 participants: 43 females, 36 males, 2 individuals identifying as diverse, and 16 who chose not to disclose their gender. Despite this reduction in sample size, a post hoc power analysis revealed that the study still achieved a power of 1, with an alpha error probability of 0.05. The mean age was $M = 40.24$ ($SD = 13.71$).

Materials and Stimuli. PsychoPy was used to make an adapted version of the previously used paradigm. This time, we used two instead of three advisors each with a cue validity of 0.4. In each trial, both advisors pointed to different houses. We varied whether the advisor on the top or the bottom made predictions with a similar voice. Accordingly, the prediction matrix was reduced to a two-by-three matrix. Furthermore, we did not incorporate smart points or any images representing the advisors. The same images and audio recordings as in Experiment 1 were used to indicate the advisors' predictions.

Procedure. The principal framework of the second Experiment was consistent with the first Experiment. Since we did not use advisor images, participants did not rate them in the

first session. Consequently, participants were only presented with a welcome screen in the second session, before the introduction trial and the 48 main trials started. Participants should click on one of the six grey rectangles to retrieve the predictions and then choose a house. In contrast to the first Experiment, within the first two seconds of each trial, the house images were semitransparent, and clicking on them had no effect. This should ensure that participants take their time to acquire predictions. Again, they got £0.05 for each treasure found to boost motivation further.

Results

To investigate the impact of voice similarity on decision-making, a Generalized Linear Mixed Model (GLMM) was fitted to the data using a Poisson distribution with a log link function. For each participant, we counted how often their house choice matched the predictions of the advisor with a similar voice ($advisor_{sim}$), the advisor with a voice of average similarity ($advisor_{avg}$), and how often they disregarded both ($advisor_0$). In the GLMM, these frequencies constituted the dependent variable, while the type of advisor match was used as the independent variable. The factor level '$advisor_{sim}$' served as the reference category.

Our analysis revealed no significant difference between the levels $advisor_{sim}$ and $advisor_{avg}$ (β = - 0.03, SE = 0.03, Z = -0.94, p = .35) on the frequency of chosen houses. This suggests that the similarity of the advisor's voice did not influence participants' choices. Between the levels $advisor_{sim}$ and $advisor_0$ there was a significant difference (β = -1.67, SE = 0.06, , Z = -29.82, $p < .001$), indicating that participants more often follow the advice of a similar advisor than neither advisor.

Discussion

In the second Experiment, we employed a streamlined version of the paradigm used in the first Experiment. This simplification eliminated confounding factors and assessed the impact of voice similarity on decision-making in a more direct approach. Whether the advisor had a similar voice did not influence participants house choices. However, despite simplifying the paradigm, the visual components of the predictions, i.e., the pictures of spiders and treasures, were still used to represent advisors' predictions. This simultaneous presentation of images and audio-recordings could have altered the processing of the auditory information. Indeed, previous research demonstrated a dominance of visual information over inputs from

other sensory systems (i.e., Colavita, 1974; Hecht et al., 2009; Posner et al., 1976). This visual dominance is known to affect various cognitive processes, including memory (Meyerhoff et al., 2023; Meyerhoff & Huff, 2016), perceptual sensitivity (Koppen et al., 2009), and object recognition (Yuval-Greenberg & Deouell, 2009). We hypothesize that the visual dominance effect could have mitigated the impact of voice similarity by enhancing the processing of visual predictions at the expense of the auditory predictions.

Experiment 3

The results of the Experiment 2 showed no influence of voice-similarity on decision-making. However, since we used visual and auditory cues to indicate the predictions of the advisor, it is possible that participants focused their attention on the images, which could have hindered the processing of the auditory information. To exclude this possibility, the third Experiment was designed without the visual predictions.

Methods

Participants. We used the same power calculation as in the second Experiment, resulting in a required sample size of at least 110 participants. New English native speakers were enlisted through Prolific. Following the exclusion of 31 participants who, on average, heard fewer than two predictions per trial, our final group consisted of 93 participants, including 44 females, 36 males, and 13 individuals who preferred not to specify their gender. The mean age of the sample was $M = 42.22$ ($SD = 12.14$).

Materials, Stimuli and Procedure. Except for removing the visual component of the predictions, that is the pictures of spiders and treasures, the third Experiment remained identical in design and methodology compared to the second Experiment.

Results

We employed the same GLMM model as in the second Experiment. Likewise, our analysis revealed no significant difference between the $advisor_s$ and the $advisor_{avg}$ level ($\beta = -0.05$, $SE = 0.03$, $p = .11$) on the frequency of chosen houses. This suggests that voice similarity did not influence participants' decision-making. However, a significant difference was observed between $advisor_{sim}$ and $advisor_0$ ($\beta = -2.04$, $SE = 0.07$, $p < .001$), showing participants

tend to align their choice more often with the advice of a similar advisor than ignoring both advisors.

Discussion

Experiment 3 was designed to eliminate a possible visual dominance effect, that could diminish the impact of auditory predictions. However, voice similarity did not alter the decision-making in absence of the prediction's visual component.

Exploratory Analysis

Even though we assumed that visual information was eliminated in the third experiment, the spatial position of the two advisors could serve as a visual heuristic cue about their validity. Indeed, previous researched demonstrated that during evaluative processes, there is a tendency for people to associate objects positioned higher in their visual field with positive attributes, while those placed lower are instinctively perceived as negative or less influential (Meier & Robinson, 2004; Schubert, 2005). To control such spatial effects, we reanalyzed the data of the last two experiments. We conducted the same GLMM, but added the position of the similar advisor as an interaction term into the formula. Including the spatial position of the advisor as an additional factor in our models (see Table 2.1 & Table 2.2) revealed a significant interaction effect between voice-similarity and position for Experiment 2 ($advisor_{avg} + pos_{top}$: $\beta = -0.39$, $z = -6.08$, $p < .001$, $OR = 0.68$) and Experiment 3($advisor_{avg} + pos_{top}$: $\beta = -0.15$, $z = -2.30$, $p = .02$, $OR = 0.86$) . If an advisor with a generic voice is presented on top of the display, participants equally often follow the advice of both advisors. Is the similar advisor positioned on top, participants choose more often houses suggested by the similar advisor. This result suggests a preference for advisors on top as well for advisors with a similar voice.

Table 2.1

GLMM Results Table of Experiment 2

Fixed Effects	*β*	*SE*	*z*	*p*	*OR*
Intercept	2.95	0.03	89.08	< .001	19.04
$advisor_{avg}$	0.16	0.04	3.63	< .001	1.18
$advisor_{0}$	-1.69	0.08	-20.18	< .001	0.18
pos_{top}	0.16	0.04	3.59	< .001	1.17
$advisor_{avg}$ + pos_{top}	-0.39	0.06	-6.08	< .001	0.68
$advisor_{0}$ + pos_{top}	0.04	0.11	0.33	.74	1.04

Note. The advisor with a similar voice was used as the reference level. The dependent variable was the frequencies with which participants aligned their choices with the advisor.

Table 2.2

GLMM Results Table of Experiment 3

Fixed Effects	*β*	*SE*	*z*	*p*	*OR*
Intercept	3.04	0.03	102.44	< .001	20.98
$advisor_{avg}$	0.01	0.04	0.27	.79	1.01
$advisor_{0}$	-2.04	0.09	-23.30	< .001	0.13
pos_{top}	0.06	0.05	1.36	.17	1.06
$advisor_{avg}$ + pos_{top}	-0.15	0.06	-2.30	.02	0.86
$advisor_{0}$ + pos_{top}	0.01	0.13	0.08	.94	1.01

Note. The advisor with a similar voice was used as the reference level. The dependent variable was the frequencies with which participants aligned their choices with the advisor.

General Discussion

In three experiments, we investigated the influence of voice similarity on decision-making processes. Experiment 1 highlighted the complexity of decision-making in conflicting information environments, where participants gathered more information than necessary and were not influenced by voice-similarity effects. In Experiment 2, we implemented a more straightforward design to assess the impact of voice similarity on decision-making more directly. The pattern observed in Experiment 1 persisted here: whether advisors had a similar voice did not affect participants' decisions. To address the potential impact of a visual dominance effect, where visual cues might overshadow auditory predictions and impede the processing of auditory content, Experiment 3 aimed to enhance auditory influence. This was achieved by eliminating the visual content of the predictions, though once again, no effect on the decision-making process was observed.

Since existing research demonstrated an influence of the vertical position of a stimulus on their evaluation (Meier & Robinson, 2004; Schubert, 2005), we conducted an exploratory analysis, examining the potential influence of the spatial positioning of the advisor. Although exploratory in nature, our analysis was informed by prior research and revealed a dual preference for advisors positioned higher in the display and those with a voice similar to the participants. This interaction not only replicates earlier findings but also highlights the robustness of these heuristic influences, even when additional variables are present.

Taken together, our results deliver little evidence for a voice similarity effect on its own. Instead, it might be able to promote the impact of other heuristics such as the spatial position, but only when visual cues are reduced to a minimum. Our data therefore suggests that, if anything, there is only very little impact of voice similarity and only under particular circumstances. Consequently, our results challenge the expectation based on the similarity attraction hypothesis (Byrne et al., 1967) in the context of voices and existing research on personalized matching effects in persuasion, where the similarity between the source of the message and the recipient is thought to enhance persuasive impact (Teeny et al., 2021).

The theoretical framework of dual-process models, including the Heuristic-Systematic Model (Chaiken, 1980) and the Elaboration Likelihood Model (Petty et al., 1986), posits that individuals engage in systematic analysis when they possess the necessary skills and motivation. In such instances, contextual characteristics and heuristics exert minimal influence on attitude change. However, under conditions of reduced motivation or the lack of cognitive

resources, peripheral cues – such as the similarity of the source – are more likely to influence the decision process (Crano & Prislin, 2006). In our study, the incentives provided may have promoted deeper elaboration, potentially limiting the influence of voice similarity. However, in the first experiment, participants were evidently influenced by another heuristic cue: the majority rule – which seems to contradict the assumption of a deep elaboration. Yet, one could argue that understanding the stochastic consequences of the validity dispersion between the advisors is a too difficult task for even the most skilled and motivated people (see Aßmann et al., 2022) and, therefore, following the majority rule is not a counter-evidence for deep-processing. Hence, providing payment could have foster deep elaboration, reducing the influence of peripheral cues, such as voice similarity.

It's important to note that, while the design created a non-compensatory environment where following the majority was less effective than relying on the high-validity cue, making decisions based on the majority rule can still lead to positive outcomes (Hastie & Kameda, 2005). If participants were unable to fully assess the consequences of validity dispersion, their behavior aligns with the *Adaptive Decision-Maker Hypothesis* (Payne et al., 1988), which suggests that individuals adaptively choose decision strategies based on task demands. In this case, while participants may have been drawn to advisors with similar voices, their decision-making appeared rational within the boundaries of their cognitive capabilities. Following the high-validity cue unless both low-validity advisors made the same prediction, while disregarding voice similarity, can be seen as a reasonable strategy. This behavior likely reflects a speed-accuracy trade-off, where examining the full implications of the non-compensatory environment may have been viewed as too time-consuming.

However, the speed-accuracy trade-off cannot only be regarded as a form of systematic approach in which subjects try to maximize their expected utility (Falk & Scholz, 2018). It could also be the result of the fact that people regularly tend to reduce the cognitive effort associated with a decision-making process (Gigerenzer & Gaissmaier, 2011; Pennycook & Rand, 2019). If the results of the explanatory analysis are taken into account, the last two experiments make it evident that not only the majority rule but also voice similarity and spatial position can influence decision-making processes as heuritic cues. The fact that the two cues cancel each other out is not so much an argument against the importance of voice similarity. In light of the daily confrontation with lists in which upper rankings are generally associated with positive characteristics of the variable in question - e.g. rankings in product tests or

competitions, as well as search results on the internet - it can be assumed that this association is the result of strong learning processes. In conjunction with the dominance of the visual system over the auditory system – as explained above – the fact that voice similarity exerts a similarly strong effect as the position of the stimulus should emphasize rather than diminish the potential impact of voice similarity.

The practical implications of our findings are particularly interesting for the design of voice assistants. They indicate that using a customized voice with characteristics similar to the users' voice might not improve the effectiveness of these technologies as decision-support devices (see Experiment 1) but could have a small effect on decision outcomes. It's important to realize that even such subtle effects can accumulate, and can have impacts at larger scales, such as in a corporate context or across societies. Therefore, future research should explore under which conditions voice similarity impacts decision-making more profoundly. This includes examining different types of decisions, varying the stakes involved, or considering the interplay of voice similarity with other persuasive elements like content, context, and the advisor's credibility. Investigating the interaction between voice similarity and other source characteristics, such as expertise or authority, could yield more profound insights into the complexities of auditory persuasion. Since our study has noteworthy limitations, especially regarding sample characteristics and the technical operationalization of voice similarity, we hope future research will use state-of-the-art text-to-speech systems that can clone a user's voice and conduct similar experiments in settings with higher ecological validity.

Finally, we note that we successfully replicated previous research regarding decision heuristics. Our experiments reaffirm previous observations – such as the majority bias (Aßmann et al., 2022) and the influence of spatial positioning (Meier & Robinson, 2004; Schubert, 2005) – demonstrating that these effects persist even when additional variables, such as voice similarity, are introduced. By replicating and extending established results, our study contributes to a more robust and nuanced understanding of auditory persuasion and the interplay of heuristic cues in decision-making.

Conclusion

In conclusion, our experiments suggest that voice similarity alone exerts little influence on decision-making under the conditions we examined. Instead, our data replicate established heuristic effects – such as the majority bias and spatial positioning – indicating that these cues

remain robust even when new factors like voice similarity are introduced. Although our exploratory analyses hinted at a potential interaction between voice similarity and spatial position when visual cues are minimized, the overall impact of voice similarity appears subtle and context dependent. These findings are modest yet valuable, as they not only confirm the resilience of traditional decision heuristics but also underscore the need for further research, particularly in more ecologically valid settings. Such investigations will be crucial for understanding the practical implications for personalized voice assistant design and other applications where auditory persuasion plays a role.

Declarations

Ethics approval and consent to participate

The studies reported were approved by the ethic committee of the Leibniz-Institut für Wissensmedien, Tübingen. All participants provided informed consent, and all experiments included in this study were preregistered (Exp. 1: https://osf.io/vyaqg; Exp. 2: https://osf.io/zjxsk; Exp. 3: https://osf.io/s3ng7).

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Declaration according to § 5 Abs. 2 No. 8 of the PhD regulations of the Faculty of Science -Collaborative Publications

The following chapter (Chapter 4) consists of a manuscript that is currently under review for submission and was co-authored by Stephan Schwan and Hauke S. Meyerhoff. The proportional contributions to this manuscript are presented in the subsequent table.

Author	Author position	Scientific ideas	Data generation	Analysis & interpretation	Paper writing
Oliver Jaggy	First author	80 %	100 %	80 %	70 %
Stephan Schwan	Second author	10 %	0 %	10 %	15 %
Hauke S. Meyerhoff	Third author	10 %	0 %	10 %	15 %

Title of paper: Can I Believe My Voice? Self-Similarity and the Illusory Truth Effect.

Status in publication process: Under Review.

Chapter 4: Can I Believe My Voice? Self-Similarity and the Illusory Truth Effect.

Abstract

Exposing people to information repeatedly increases the likelihood that they will judge it to be accurate. This phenomenon – known as the illusory truth effect – is robust and widely replicated, and particularly relevant in an age of unfiltered information dissemination. In four experiments, we test whether perceptual fluency is necessary for the effect to occur, whether auditory stimuli generated by text-to-speech (TTS) systems can elicit the effect, and whether voice similarity between speaker and recipient increases the perceived truth of statements. The results of our first three experiments indicate that perceptual fluency is likely the primary driver of the effect, and that repeated exposure to TTS-generated auditory stimuli enhances credibility, albeit to a lesser extent than written statements. Building on this, our fourth experiment shows that – although voice similarity only marginally amplifies the illusory truth effect – the credibility of a statement increases when presented in a voice similar to the listener's, even without repetition.

Introduction

In today's digital landscape, text-to-speech (TTS) technologies are revolutionizing how we consume information by transforming written content into audio and audiovisual formats. This study investigates how repeated exposure to auditory information affects truth judgments, with particular attention to whether voice similarity enhances the illusory truth effect.

Disinformation has long served as a tool for political and social manipulation, but its influence has been amplified by digital technologies in recent years. This is further exacerbated by the spread of misinformation, hate speech, and propaganda – particularly online – which deepens social divisions and fosters widespread mistrust in institutions (see United Nations Human Rights Council, 2021). This is compounded by the rapid and wide dissemination of misinformation, often outpacing verified information (Vosoughi et al., 2018), making false information – or fake news – not1 only more prevalent but also more plausible by repetition (for a review, see Pillai & Fazio, 2021). Repetition increases the likelihood that a statement will be judged as true (Hasher et al., 1977, for a review of truth judgments, see Brashier & Marsh, 2020). This phenomenon, known as the illusory truth effect, has shown

strong and consistent results (for a review, see Unkelbach et al., 2019). A meta-analysis covering 70 different studies indicated an average effect size of d = 0.50, with a 95% confidence interval ranging from 0.43 to 0.57 (Dechêne et al., 2010), and could be even stronger in real-life situations where people are not informed that they will encounter false information Van Bavel et al. (2021) identify memory as a central element in influencing when and why individuals may believe misinformation and they point to the illusory truth effect as a major contributor to this susceptibility. Repeated exposure can increase the perception of truthfulness (e.g., Nadarevic et al., 2020) over long periods of time (Boehm, 1994; Brown & Nix, 1996), even for statements that are highly unlikely to be true (Lacassagne et al., 2022), despite better knowledge (Fazio et al., 2015; Fazio et al., 2019), and monetary incentives (Speckmann & Unkelbach, 2022). The illusory truth effect has been observed in relation to many subjects: trivia (Bacon et al., 1979), arguments (Moons et al., 2009), rumors (DiFonzo et al., 2016), product promises (Venkataramani Johar & Roggeveen, 2007), news (Polage, 2012) and opinions (Arkes et al., 1989).

Several theoretical frameworks have been proposed to explain for the increase in perceived truthfulness following repetitions. One line of explanation emphasizes heuristic processing, noting that individuals often seek to minimize cognitive effort (Pennycook & Rand, 2019) and use heuristics instead of elaborate processing whenever possible (Gigerenzer & Gaissmaier, 2011). When faced with complex or ambiguous information, heuristical processing often relies on salient rather than informative attributes as cues for judgment. Two key heuristics implicated in the illusory truth effect are familiarity and processing fluency, which fall under the broader category of processing ease mechanisms.

The familiarity heuristic is a variant of the recognition heuristic (Goldstein & Gigerenzer, 2002) and relies on the recognition of previously encountered information. The familiarity with information is used as a cue to judge the veracity of the information (Begg et al., 1992; Hawkins & Hoch, 1992). Winkielman et al. (2003) suggested that familiarity is inherently associated with positive valence, as unfamiliar stimuli are often perceived as potentially threatening. Moreover, Bacon et al. (1979) found that repetition on its own does not cause the illusion of truth: it depends on people's belief that they encountered the information beforehand (see also Boehm, 1994). Accordingly, they called it a recognition effect.

The processing fluency heuristic builds upon the ease with which particular piece of information can be processed mentally. In general, repeated information can be processed

faster and more fluently than novel information (Jacoby & Dallas, 1981). Processing fluency has been reported to affect truthfulness ratings (for reviews, see Alter & Oppenheimer, 2009; Winkielman et al., 2003), as well as social interactions (Pearson & Dovidio, 2013), affective judgments (Reber et al., 1998), art appreciation (Belke et al., 2010) and liking (Forster et al., 2013).Processing fluency is regarded as one of the main drivers of the illusory truth effect (Ecker et al., 2022; Unkelbach, 2007). Reber et al. (2004) suggested that processing fluency carries a positive hedonic marker, making fluently processed information inherently favorable and more likely to be assessed as accurate. Processing fluency is related to the familiarity heuristic as familiarity is one possible reason why processing is more fluent. Notably, perceptual fluency can affect truth judgments even without prior exposure, as shown by manipulations involving readability (Pennycook & Rand, 2020), speech clarity (Reber & Schwarz, 1999), or rhyme (McGlone & Tofighbakhsh, 2000).

Besides perceptual fluency, conceptual fluency is another form of processing fluency that may influence truthfulness judgments (Parks & Toth, 2006; Silva et al., 2017; Thapar & Westerman, 2009; Whittlesea, 1993). Unlike its perceptual counterpart, conceptual fluency operates at more advanced stages of cognition and is mainly influenced by the context in which a target stimulus appears. This comprises the type of surrounding stimuli, the previous exposure to related or unrelated stimuli, and the predictiveness of the context. It can influence recognition, evaluation, and memory retrieval by increasing the coherence between the target stimulus and its contextual features (Lee & Labroo, 2004; Winkielman et al., 2003). In studies investigating the illusory truth effect, perceptual and conceptual fluency are often entangled. Nevertheless, they have been demonstrated to reveal differential consequences (Lanska et al., 2013). One approach for probing the role of conceptual fluency in the illusory truth effect involves repeating stimuli that differ in perceptual attributes (for instance synonyms) but share the same semantic meaning. Various experimental designs could employ this strategy, such as using masked words crafting semantically identical yet structurally different statements(Silva et al., 2017), or altering the modality between the encoding and testing phases(Heusser et al., 2013). Among these approaches, modality switching offers the most straightforward manipulation for isolating the influence of conceptual fluency, as it minimizes the confounding effects of perceptual fluency. However, empirical findings on modality changes remain mixed, with some studies reporting consistent effects and others not. Where some studies showed a truth effect regardless of modality change (Bacon et al., 1979; Begg et al., 1992), Thapar and

Westerman (2009) found no such effect when there was a change in modality between the study and test phase in their first experiment. However, it occurred regardless of modality change in the second experiment. A major difference between these experiments was the kind of process fluency manipulated. From these findings, they conclude that perceptual fluency is more susceptible to modality change than conceptual fluency.

Besides familiarity and process fluency, two more hypotheses have been set out to explain the illusory truth effect (see Pillai & Fazio, 2021). First, the *Convergent validity* or *Source Dissociation Hypothesi*s posits that repetition may lead individuals to infer that the information originates from multiple sources, creating an illusion of widespread consensus (Arkes et al., 1991). Indeed, the trustworthiness of a source significantly influences how persuasive the information is perceived to be (for a review see Pornpitakpan, 2004). However, studies focusing on eyewitness accounts have suggested that it is repetition, not source variation that amplifies the impact of misinformation. (Foster et al., 2012; O'Donnell et al., 2023). Second, memory processes have been suggested to explain illusory truth.For instance, it has been suggested that repeating a statement strengthens the associative network among the corresponding concepts as well as the internal consistency of this network. In return, the coherent references in the memory network might bias one's judgment of truthfulness in favor of the strengthened concepts (Unkelbach & Rom, 2017).

Although the illusory truth effect has been examined using auditory stimuli, few studies have systematically explored how speech characteristics, such as voice similarity affect perceived truth (see Henderson et al., 2022). That is especially interesting since the credibility of a source impacts the illusory truth effect (Brown & Nix, 1996; Unkelbach & Stahl, 2009) and since people are able to extract a diverse set of information about a speaker by just listening to them. Indeed, vocal features can convey demographic information such as a speaker's sex, age (Li et al., 2012; Meinedo & Trancoso, 2010), and even personality traits (Carbonneau et al., 2017; Mohammadi & Vinciarelli, 2015). To our knowledge there are only two studies that investigated the consequences of speaker characteristics: both studies manipulated whether the speaker spoke with an accent (Frances et al., 2018; Lev-Ari & Keysar, 2010) and found conflicting results regarding consequences on the illusory truth effect. While Frances et al. (2018) found no effect of regional accents on memory and credibility, Lev-Ari and Keysar (2010) study demonstrated that non-native speaker with an accent were perceived as less credible.

In the present project, we conduct a systematic investigation into how voice characteristics – particularly voice similarity – modulate the illusory truth effect. In particular, we evaluate AI generated voiced which could resemble the participants' voices thus acting as an implicit cue of familiarity. We hypothesize that AI generated voices – like human voices (Arkes et al., 1989; Gigerenzer, 1984) should elicit the illusory truth effect and that voices resembling the voice of the participant should enhance the illusory truth effect. This investigation is especially relevant given the rapid advancements in text-to-speech technologies that allow for the creation of highly convincing artificial voices (e.g., Oord et al., 2016; Ren et al., 2019; Shen et al., 2018; Wang et al., 2017). Such technologies are not only becoming increasingly sophisticated, including the possibility to clone the voice of a target speaker (Arik et al., 2018; Jia et al., 2018; Neekhara et al., 2021), they are also more accessible to the general public, raising the potential for the manipulation of information on an unprecedented scale (Eren & The Coqui TTS Team, 2023).

In Experiment 1, we replicated the basic illusory truth effect. In Experiment 2, we examined the role of conceptual fluency in the illusory truth effect by implementing a modality change from written statements during the encoding phase to auditory presentation during the testing phase. In Experiment 3, we investigated whether the illusory truth effect persists when presenting auditory stimuli during both the encoding and testing phase. In the final Experiment 4, we explored the impact of voice similarity on truth judgments, assessing whether statements spoken in a voice resembling the participants' voices are more likely to be perceived as true compared to those presented in a generic voice.

Experiment 1

In the first Experiment, we attempted to replicate the illusory truth effect in order to confirm that our study materials (e.g., the selected statements) and experimental setup (i.e., the online setup) are capable of eliciting the phenomenon. As all other experiments in this study, the design, hypotheses, and analysis plans were preregistered prior to data collection (see below).

Methods

Participants. Participants were recruited via prolific.com (https://prolific.com/). Based on power calculations using G*Power 3.1.9.7 ($\alpha = 0.05$, power = 0.95, $d = 0.53$), a minimum

sample size of 40 participants was determined to be sufficient for a paired t-test. To account for potential exclusions, we targeted a final sample size of 52. Initially, 54 individuals completed the experiment; however, two were excluded from the analysis due to failing more than one of the three control trials in either the initial or the final phase of the experiment, as predefined by our exclusion criteria. This resulted in a final sample size of 51 participants. The mean age of the sample was $M = 36.90$ years ($SD = 10.54$). The final sample included 22 females, 23 males, 1 participant who identified as diverse, and 5 who did not disclose their gender. As all other experiments in this study, it received ethical approval from the ethics committee of the Leibniz-Institut für Wissensmedien, Tübingen.

Materials and Procedure. We used the platform qualtrics.com **(**https://qualtrics.com**)** to collect basic demographic information and participants consent. The experiment was conducted online using PsychoPy (Peirce, 2007) via pavlovia.org (https://pavlovia.org/)**.** The stimuli for this study consisted of a total of 232 declarative statements. These were developed from questions initially published and validated by Nelson & Narens (1980**)** and subsequently revalidated by Tauber et al. (2013**)**. Each original question was transformed into two distinct declarative statements: one encapsulating the correct answer, and the other representing the most frequently selected incorrect answer as identified in Tauber et al. (2013**)**. Additionally, six control statements were created. Each of them contained a pseudoword. Three of them were incorporated in the first part, three in the last part of the Experiment. To achieve counterbalancing, we created four different condition files. Across participants, each statement was presented visually in both its true and false versions, serving alternately as a target and as a lure.

During the learning phase, participants rated how interesting they found each of the 58 statements on a continuous scale ranging from "boring" to "interesting." This served to mask the true purpose of the study and ensure participant engagement. To further ensure that participants are processing the study materials, we presented three control trials in the study phase. These control trials included a pseudoword; and participants were instructed to click on a yellow circle if they recognized such a pseudoword. In the test phase of the experiment the 58 statements from the first part were presented in written form, this time intermixed with 58 new statements (29 true and 29 false). In addition, we presented three statements that

contained pseudowords. The participants were asked to decide whether a statement is true, false, or contains a non-word by pressing the ‘t’, ‘f’, or ‘n’ key.

To create a delay between study and test phases, a 32-trial mental rotation task was included as a filler activity. In each trial, one of the letters f, z, l, or n was displayed on the left side of the screen. This was accompanied on the right side by either an identical or a mirrored version of the letter, both of which were rotated. Participants should decide if they were identical or mirrored letters by pressing the ‘s’ or ‘m’ key.

Results

A paired samples t-test was conducted to compare the proportion of truth judgments of factual false statements for statements exclusively presented during the final phase of the experiment ($M = 0.46$, $SD = 0.15$) with those for statements presented in both phases ($M = 0.54$, $SD = 0.17$). The analysis revealed a significant difference in truth judgments, indicating that repeated false statements were more likely to be judged as true compared to new false statements ($t\ (50) = 5.0289$, $p < .001$). The computed Cohen's d showed a medium effect size of $d = 0.54$.

Discussion

These findings replicated the illusory truth effect, confirming that repetition increases perceived truthfulness – even for false statements. Despite the online format, the effect size observed was consistent with previous findings (Dechêne et al., 2010)**.**

Experiment 2

One explanation for the illusory truth effect states that repetition creates coherent references within memory. This assumption predicts that a repetition of the content of a statement irrespective of modality should add coherent references as well. In this experiment, we tested this prediction by investigating the consequences of a change in modality between the initial presentation of a false statement and its’ repetition. Furthermore, this experiment was designed to isolate conceptual fluency, as a modality change should disrupt perceptual fluency while preserving conceptual content. Whereas modality changes between study and test phase were deployed in several studies (for a systematic map of research on the illusory truth effect, see Henderson et al., 2022), to our knowledge no previous study has investigated

changes in modality in isolation, i.e. without altering other variables or with the deliberate intention of incorporating modality change to examining the impact of conceptual fluency on the illusory truth effect.

Methods

Participants

New participants were recruited through Prolific. We expected the same number of participants would be needed as in Experiment 1. A total of 53 individuals initially completed all phases of the study. Based on our pre-established exclusion rules concerning control trial performance, we had to exclude two participants, resulting in a final sample of 51 participants. The participants had a mean age of $M = 40.29$ years ($SD = 12.32$). The sample comprised 24 females and 20 males; 7 participants opted not to disclose their gender.

Materials and Procedure

The same statements as in the first experiment were used. For the learning phase they were converted into audio files using an open-source TTS system (Eren & The Coqui TTS Team, 2023), which incorporates the VITS model (Kim et al., 2021). The target voice used to create the material was randomly selected from the speaker identities provided by the TTS system: It was a supposedly male speaker (audio samples can be retrieved from https://osf.io/z76cn/files/osfstorage). We introduced an audio setup testing phase in which three sinus tones were presented. After hearing the tones participants had to declare how many they have heard. In case of a false answer the experiment ended immediately. In contrast to the learning phase of the first experiment, participants should evaluate how difficult it was to comprehend the statements using a continuous rating scale (range: very easy – very hard). Furthermore, participants were asked to click on a yellow circle when hearing a control statement ("This is a control trial"), which was generated through the same TTS system. The remaining parts stayed consistent with the first experiment.

Results

A paired samples t-test was conducted to compare the proportion of true judgments of factual false statements for statements exclusively presented during the final phase of the experiment ($M = 0.48$, $SD = 0.18$) with those for statements previously heard ($M = 0.49$, $SD =$

0.18). The analysis revealed no statistically significant difference between the two conditions ($t(50) = 0.80$, $p = .43$), suggesting that cross-modal repetition (audio to text) does not enhance truth judgments.

Additionally, we investigated whether participants' judgments regarding the comprehensibility of the statements influenced their truth assessments of false statements. We employed a generalized linear mixed-effects model (GLMM) with a binomial distribution and a logit link function, utilizing the lme4 package in R (Bates et al., 2014). The model revealed a significant effect of perceived statement difficulty on truth ratings (Estimate = -1.07, $z = -3.96$, $p < .001$). Specifically, the negative coefficient suggests that more difficult-to-understand statements were less likely to be judged as true.

This result is consistent with the notion that the statement must first be effectively comprehended for repetition to influence truth judgments. As specified in our preregistration plan, we had intended to exclude any statements from the analysis that received an average difficulty score exceeding 0.8 if we would find a significant effect of the difficulty assessment. However, none of the statements met this criterion, with the maximum mean difficulty rating being 0.66.

Discussion

We observed no illusory truth effect when the modality switched from auditory presentation in the study phase to visual presentation in the test phase. As conceptual fluency was preserved while perceptual fluency was disrupted, this null effect supports the idea that perceptual cues are central to the illusory truth effect. Additionally, our results seem to contradict a memory-based bias as a driver of the illusory truth effect. However, the absence of the effect could also be explained by the inferior recognition memory for sounds compared to visual stimuli (Bigelow & Poremba, 2014; Gloede & Gregg, 2019). The inability to recognize a statement encoded auditorily could, therefore, be the reason why we found no effect of the repetition. Conversely, another potential explanation – that the text-to-speech system we used was too difficult to understand, thereby accounting for the null effect – seems implausible based on our analyses.

Experiment 3

The third experiment aimed to determine whether the null effect observed in Experiment 2 was due to modality change or the use of text-to-speech (TTS)-generated stimuli. Accordingly, both the encoding and test phases employed TTS-generated auditory stimuli.

Methods

Participants

Study participants were newly recruited through Prolific. We again aimed for 52 participants. Initially, the experiment was completed by 52 individuals. After applying predetermined criteria for control-trial performance, we excluded one participant, yielding a final sample of 51 participants, including 19 females, 31 males, and one individual which did not disclose their gender. The mean age within this group was $M = 35.12$ years ($SD = 10.29$ years).

Materials and Procedure

While the general experimental framework mirrored that of Experiment 2, we reinstated the "interestingness" rating task from Experiment 1 to re-focus participant attention on the semantic content of the statements rather than their auditory presentation. Additionally, the auditory stimuli, including the control statements, were used in both the learning and the test phase.

Results

We applied a paired-samples t-test to evaluate the influence on proportion of true judgments of factual false statements presented solely in the final phase ($M = 0.49$, $SD = 0.17$) compared to those that were rehearsed ($M = 0.53$, $SD = 0.19$). The results indicated a statistically meaningful difference ($t(50) = 2.40$, $p = .02$). The calculation of Cohen's d yielded an effect size of $d = 0.26$, indicating a small but reliable increase in perceived truth for repeated statements in the auditory modality.

Discussion

Spoken trivia statements, generated by a text-to-speech system, were more likely evaluated as true when participants heard them before. These results rule out stimulus intelligibility as the cause of the null effect in Experiment 2, implicating modality change as the critical factor. The reduced effect size compared to Experiment 1 aligns with the notion that auditory memory traces are weaker than visual ones (Gloede & Gregg, 2019; Meyerhoff & Huff, 2016), leading to reduced recognition-based fluency, and therefore, diminishing the illusory truth effect.

Experiment 4

The findings from Experiment 3 showed that the illusory truth effect could be observed with purely auditory stimuli. In Experiment 4, we investigated whether processing fluency alters the illusory truth effect in the auditory modality. We therefore manipulated the similarity between the voice presenting the statements and the voice of the participants. We hypothesized that voice similarity would enhance processing fluency, as self-similar voices may be more familiar and perceived as more trustworthy (Jaggy et al., 2025). Given the growing ease of voice cloning, this experiment also addresses practical implications regarding susceptibility to synthetic speech in misinformation contexts.

Methods

Participants

Study participants were newly recruited through Prolific. With an alpha level of 0.05 and a power of 0.85, G*Power 3.1.9.7 computed a required sample size of 110 to detect a small interaction effect in a 2 x 2 design. Initially, the experiment was completed by 110 individuals. After applying predetermined criteria for control-trial performance, we excluded three participants, yielding a final sample of 107 participants. The mean age within this group was $M = 41.00$ years ($SD = 13.27$). The sample included 47 females and 59 males, and 1 diverse person.

Materials and Procedure

This investigation unfolded in two parts. The first session required participants to provide basic demographic information, deliver informed consent, and record an audio sample.

This was accomplished via a Qualtrics-integrated phonic.ai interface, allowing participants to easily record their voice.

To generate the audio files used in the second session, the original TTS system was retained, but a different model to synthesize the auditory stimuli was used. The YourTTS model (Casanova, Weber, et al., 2021) was chosen owing to its better performance in preliminary tests for the specified task. To generate material in a similar voice we utilized the voice cloning capability of the model: participants audio recordings served as reference files to produce half of the auditory stimuli. YourTTS analyzes the speaker embedding (the numeric representation of a speakers voice) of the provided audio recording and uses the information to generate speech that is typical for a speaker with such an embedding (e.g., Arik et al., 2017; Jia et al., 2018). For the remaining stimuli, we utilized gender-matched audio recordings to generate stimuli spoken in a generic voice.

We balanced conditions by creating eight files to account for variations in voice type (generic vs. similar), frequency of presentation (once vs. repeated), and statement veracity (true vs. false). Each participant experienced an evenly distributed mix of these conditions. Following the creation of each individual experiment based on the provided audio, participants were invited back for the second session.

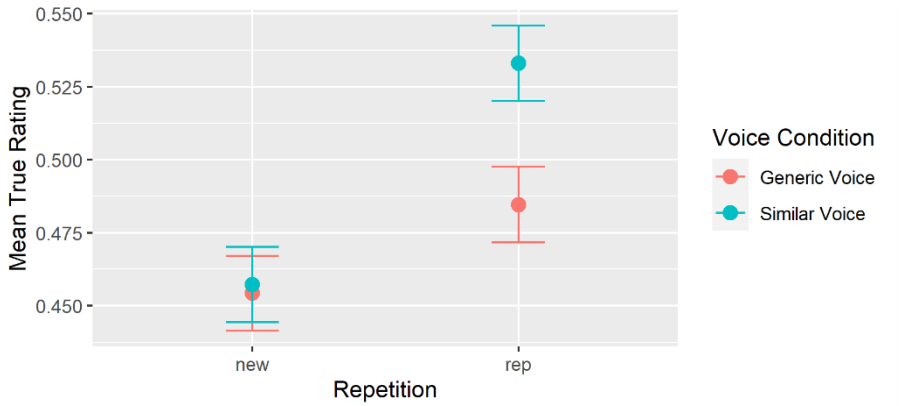
The second session replicated the procedure of Experiment 3, including the learning, filler, and test phases.

Results

To analyze the influence of voice similarity on the illusory truth effect we deployed an ANOVA with the truth judgement of false statements as the dependent variable and voice similarity and repetition as the independent variable (see Figure 3.1). Results revealed a marginal interaction between voice similarity and repetition ($F(1) = 3.11$, $p = .078$), a significant main effect of voice similarity ($F(1) = 3.98$, $p = .046$, $\eta_p^2 < .01$), and a significant main effect of repetition ($F(1) = 16.97$, $p < .001$, $\eta_p^2 < .01$).

Figure 3.1

Illustration of the Results of Experiment 4



Note. Depicted are the mean true ratings dependent on repetition and voice type. Arrows represent the 95% confidence intervals.

Discussion

Our findings showed that false trivia statements delivered in voices resembling participants' own voices were judged more often as true, supporting the hypothesis that voice similarity increases credibility. Notably, the marginal interaction between statement repetition and voice type could suggest that the impact of voice similarity occurs during the evaluation phase. This leads us to theorize that the familiarity of the voice facilitates more fluent cognitive processing during the test phase. Additionally, it could enhance the perceived credibility of the information source. However, since the interaction effect barely missed statistical significance, we cannot rule out that voice similarity also elevates the effect by improving memory formation. Taken together, even though the effect is relatively small, voice similarity could have a considerable impact when used on a broader scale.

Repetition impacted the truth rating, but the effect size was much smaller than in the previous Experiments. This decrease may be attributed to the presence of voice similarity as an additional cue that participants can rely on while judging the veracity. Therefore, the diminished influence of the recognition-based heuristic when evaluating statements is not surprising. Further implications of our findings are discussed below.

General Discussion

Although over 200 studies have demonstrated the illusory truth effect (see Henderson et al., 2022), this study is among the first to systematically investigate how auditory presentation – particularly vocal similarity – influences truth judgments. By extending beyond basic modality effects to include speaker similarity, we offer a unified framework incorporating perceptual fluency, memory heuristics, and source credibility.

First, our replication of the classic repetition effect with written stimuli (Experiment 1) confirmed that our materials and online procedures yield effect sizes ($d \approx 0.54$) consistent with the literature (Dechêne et al., 2010). In Experiment 2, cross-modal repetition (auditory encoding, visual retrieval) did not produce an illusory truth effect, suggesting that perceptual continuity is essential for repetition-based credibility effects. This absence of effect underscores that mere conceptual coherence (i.e., the knowledge that "I've heard this before" without matching perceptual form) is insufficient to bias truth judgments. Instead, repetition appears to drive truth perceptions primarily when perceptual fluency cues are preserved.

Experiment 3 addressed an alternative explanation – that TTS-generated voices are simply too hard to remember – by keeping both encoding and retrieval in the auditory modality. Here, we again observed a significant repetition effect ($d = 0.26$), albeit smaller than in the visual domain, consistent with evidence that auditory memory traces are weaker than visual ones (Bigelow & Poremba, 2014; Gloede & Gregg, 2019). However, the fact that any effect emerged with synthetic speech indicates that perceptual fluency, not raw memory strength, underpins the illusory truth effect.

Experiment 4 introduced a novel manipulation of vocal similarity, demonstrating that statements spoken in voices resembling participants' own voices were judged more frequently as true – regardless of whether they had been repeated. This finding provides the first direct evidence that vocal self-similarity acts as an implicit familiarity cue, enhancing processing fluency and source credibility during the evaluation phase. The marginal interaction effect suggests that voice similarity could also affect encoding processes, though additional evidence is needed to substantiate this possibility.

Theoretical Implications

These findings support a dual-process model, wherein perceptual fluency and recognition heuristics together underlie the illusory truth effect. When perceptual form is

preserved across encounters – whether visual or auditory – processing ease is increased, leading individuals to misattribute fluency for accuracy (Alter & Oppenheimer, 2009). Moreover, voice similarity appears to engage an additional fluency channel: hearing one's own voice – or something close to it – automatically signals familiarity and trust (Jaggy et al., 2025), thereby lowering the threshold for truth acceptance. In contrast, conceptual coherence alone (as in Experiment 2) lacks the sensory reinforcement needed to trigger these heuristics.

Practical and Ethical Considerations

In a media landscape increasingly dominated by video and audio content – often synthesized via advanced TTS systems – the ability to manipulate voice characteristics poses new risks for misinformation. Voice-cloning tools can now generate near-perfect replicas of a target speaker's voice (Jia et al., 2018; Neekhara et al., 2021), potentially exploiting the familiarity heuristic to lend credibility to false or misleading claims. Even modest effects, such as those observed for voice similarity ($\eta_p^2 < .01$), can scale meaningfully when embedded within high-frequency exposure contexts like social media feeds or automated communication systems.

Limitations and Future Directions

Several limitations warrant caution. First, our participant pool – recruited online via Prolific – may not fully represent populations with differing levels of media literacy or auditory processing abilities. Second, we focused on factual trivia; real-world disinformation often involves emotionally charged or politically salient content, which may interact with voice cues more complexly. Third, our voice-cloning manipulation used only one TTS architecture (YourTTS); future work should compare across multiple models, including adversarial or maliciously modified voices.

Building on these findings, future research should:

1. Examine whether emotional tone (e.g., warmth, confidence) interacts with voice similarity to amplify fluency effects further.
2. Test voice-similarity effects in high-stakes contexts (e.g., public health messaging, political persuasion) to assess real-world impact.

3. Investigate interventions – such as explicit warnings about synthetic voices or training in voice literacy – that might mitigate susceptibility to voice-based illusions of truth.

Conclusion

Our experiments demonstrate that perceptual fluency – whether preserved across modalities or instantiated via self-similar voices – is a key driver of the illusory truth effect. As synthetic speech technologies evolve, understanding the cognitive mechanisms by which voice characteristics influence credibility will be crucial for both theoretical models of human judgment and the design of policy interventions to safeguard public discourse.

Declarations

Inclusion & Ethics

The studies reported were approved by the ethic committee of the Leibniz-Institut für Wissensmedien, Tübingen. All participants provided informed consent, and all experiments included in this study were preregistered (Exp. 1: https://osf.io/yqchw; Exp. 2: https://osf.io/atnfy; Exp. 3: https://osf.io/jpq52; Exp. 4: https://osf.io/hk8xe).

Availability of Data and Material

The data generated during this study has been anonymized and was deposited in the following OSF repository: https://osf.io/z76cn/files/osfstorage. Participants audio samples were deleted due to data protection.

Code Availability

R code for data processing is publicly available in the following OSF repository: https://osf.io/z76cn/files/osfstorage.

Chapter 5: General Discussion

The rapid development of TTS systems, capable of replicating specific voice characteristics, coincides with a shift from predominantly text-based information to more audio and audio-visual content. This transition is fueled by the popularity of video-based social platforms like TikTok and the expanding use of voice-based interfaces, which are increasingly powered by advanced large language models. As voice cloning and voice adaptation technologies become more accessible and computationally efficient, their potential applications across various scenarios expand. This thesis explores the implications of these voice-adapting technologies in a world where such interactions are becoming more prevalent.

Despite the reasonably extensive research on the similarity attraction hypothesis, studies examining the effects of voice similarity on attraction rarely exist. By training and using a modern speaker recognition system and by utilizing advanced TTS technologies, the experiments have provided robust evidence that voices similar to one's own can significantly affect trait evaluations, decision-making, and the perception of truth. In this final chapter of my thesis, I will summarize my results and discuss the implications of my empirical findings in relation to existing research, highlighting how they contribute to our understanding of voice similarity effects. Additionally, the practical implications of these results will be examined, alongside a detailed evaluation of the strengths and limitations of my experiments. Finally, I will outline potential directions for future research.

Summary of Findings

The first experiment described in Chapter 2 focused on validating the use of cosine similarity values as a quantitative measure for perceived voice similarity. The cosine similarity values were derived by comparing *d*-vectors generated by a speaker recognition system. The speaker recognition system was trained on a large dataset to generate 256-dimensional feature vectors, known as voiceprints. Through training, the system should learn to enhance intra-speaker similarity while upholding inter-speaker discrimination. The central hypothesis was that higher cosine similarity values between voiceprints correspond to higher perceived similarity ratings. My results showed a quadratic relationship between the cosine similarities and human similarity ratings, indicating that human perception of voice similarity is more sensitive at higher ranges of voice similarity, which seems logical

considering that discriminating between similar voices is a valuable skill in everyday life while discriminating between dissimilar voices lacks ecological value. While the median Spearman correlation coefficient was moderate, additional analyses using aggregated data revealed a more pronounced correlation between cosine values and human judgments, indicating that the relatively low correlation using raw data likely stems from the variance of the utilized audio samples and the difficulty of judging the similarity after hearing only a few seconds of audio. Evidence for this interpretation stems from the second experiment in this chapter, where a rather low test-retest reliability of human similarity judgments was found. This variability in judgments could have decreased the correlations observed in all experiments in this series.

Experiment 3 explored whether the findings from the first two experiments hold when participants compare external voices to their own. In contrast to the first experiments, I found a linear relationship between cosine similarity values and similarity ratings and a reduction of the observed correlation. Additional analyses revealed that participants' average similarity rating was significantly lower compared to the first two experiments, which could have mitigated a stronger relationship between cosine scores and similarity judgments. This reduction could stem from a general need for uniqueness, which reduces the perceived similarity, especially for voices with higher similarity scores (C. R. Snyder & Fromkin, 1977). Another explanation could stem from a difference in processing one's voice compared to other voices, where familiarity with one's voice could increase the sensitivity to even subtle differences (Sidtis & Kreiman, 2012; Stevenage, 2018). Interestingly, whether we presented the participant's voice recordings in addition to the recordings of the other speaker had no effect. The fourth experiment tested whether speakers with average voice characteristics would be perceived as more likable and trustworthy. The analysis showed no significant impact of a voice's typicality on likability and only a negligible effect on trustworthiness ratings. These findings suggest that the beauty-in-averageness effect may not be present in the auditory perceptions of voices, a somewhat surprising finding given the evidence for such an effect in the visual modality (Langlois & Roggman, 1990; Winkielman et al., 2006).

The final experiment investigated if voices similar to one's own are perceived as more likable and trustworthy. The results indicated a quadratic relationship between cosine similarity values and both likability and trustworthiness ratings.

The experiments of my first series collectively validate the use of cosine similarity as a practical proxy for assessing voice similarity in psychological research. While the beauty-in-averageness effect was not supported for voice perceptions, the findings underscore the significant effect of voice similarity to one's voice on perceived likability and trustworthiness. This suggests that personal relevance plays a crucial role in how voices are perceived, potentially driven by an implicit preference for familiarity or similarity to oneself, albeit influenced by the complexity of auditory processing and personal biases.

The second experimental series reported in Chapter 3 expanded the research conducted in the last experiment reported in Chapter 2 by investigating if similar voices are not only perceived favorably but can also alter our behavior, particularly the decisions we make. In detail, the first experiment examined if decision quality improves when advice is received from a similar-sounding advisor amidst conflicting information from less reliable sources. Participants got predictions from three advisors varying in reliability and voice similarity to decide on the location of a hidden treasure. Despite being monetarily incentivized to find the treasures, participants deviate from optimal decisions – to follow the advice of the most reliable advisor – especially when the two more unreliable advisors made the same prediction. Moreover, whether the high-validity advisor had a similar voice did not alter decision outcomes.

The subsequent experiments used a simplified setup to decouple the influence of voice similarity from social influence, particularly the majority bias that arises when both low-validity cues make the exact predictions. Therefore, only two advisors with the same prediction capability were used, and I varied whether the advisor on top or at the bottom of the display had a voice similar to the participant’s voice. When including the position of the advisor as an interaction term in my model, I found a significant effect of voice similarity on decision outcomes. When an advisor with a similar voice is at the bottom of the display, the preference for advisors on top and the preference for similar advisors cancel each other out. However, if a similar advisor was on the top of the display, participants were more likely to follow their advice compared to the dissimilar advisor on the bottom.

In Chapter 4, I explored how voice similarity might influence truth perception. The choice to focus on the illusory truth effect was motivated by the fact that social media increases the likelihood of repeatedly encountering misinformation. The results of my initial experiment in this series demonstrated that repeated exposure to false statements significantly increased their perceived truthfulness compared to novel false statements in an

online setting. The second experiment sought to disentangle the influence of perceptual and conceptual fluency in the illusory truth effect by introducing a modality change between the presentation and repetition of a statement. However, false information presented auditorily and tested in written form was not more likely judged as accurate compared to statements just presented in the evaluation part of the experiment. In other words, changing the modality suppressed the illusory truth effect, indicating that an increase in perceptual fluency is more important than conceptual fluency.

In the third experiment, I explored the potential limitations of auditory stimuli generated by a TTS system in influencing truth judgments. The results indicated a significant effect of repetition on truth judgments using auditory stimuli. However, compared to the first experiment in the series, the effect was substantially smaller, possibly due to the less reliable memory for auditory information compared to visual information.

The final experiment in this series investigated the influence of voice similarity on the illusory truth effect. Results indicated a slight increase in truth judgments for familiar voices. However, the interaction with repetition was marginally significant, suggesting that while voice similarity impacts truth perception, it does not significantly amplify the effect of repetition.

The combined results indicate that cosine similarity values can be utilized to study the impact of voice similarity on cognitive processes. Furthermore, my experiments provide evidence for a significant impact of voice similarity on attraction, decision-making and truth-perception. The theoretical and practical implications of these findings are discussed in the following section.

Theoretical and Practical Implications

The first three experiments reported in Chapter 2 demonstrated that cosine similarity values could be utilized as a proxy for human-perceived voice similarity. Since pre-trained open-source speaker recognition systems are now easily accessible – in contrast to the point in time when I started my PhD project – my findings open up the possibility of investigating voice similarity effects on a wide variety of topics.

The observed quadratic relationship between cosine similarity values and human judgments suggests a disproportionate sensitivity between AI ratings and human judgments, potentially highlighting a general challenge when using AI assessments to predict human

perception und judgment. While AI-driven systems are often trained to distinguish based on raw physical data, human perception typically follows a non-linear response curve influenced by the sensitivity of the respective sense. Such a nonlinearity is recognized in speech generation through the use of mel-spectrograms, which employ a non-linear frequency scale and enhances the perceived quality of generated speech. Similarly, the findings from my initial experiments could be leveraged to train a deep neural network that accounts for the quadratic relationship between cosine similarity values and perceived voice similarity, aiming to develop a system that better aligns with human voice similarity judgments.

Somewhat related to the discrepancies between AI and human perception is the finding from the third experiment that externally presenting one's own voice does not affect similarity judgments. This factor was included because a speaker's perception of their own voice differs from a listener's due to sound being conducted not only through air but also via bone when we speak (Pörschmann, 2000). However, the lack of a difference in whether one's voice is presented externally suggests that the internal representation of one's voice may be too robust to be influenced by its external presentation. When cloning or adapting voices, the aim is usually not to deceive the individual whose voice is being mimicked but rather to convince someone else that the speech originates from the cloned speaker. Nonetheless, suppose the objective is not merely to convince people of the identity of a speaker but to utilize voice similarity effects. In that case, it is essential to consider the strength of the internal voice representation and the differences in voice perception based on the sound transmission route. By applying digital filters that simulate changes due to bone conduction, one can create a speech signal that reflects how the speaker perceives their voice internally (Berger & Won, 2005). Utilizing audio edited in this way as reference files for voice cloning could likely produce content that is more effective in eliciting similarity effects. Additionally, avatars in virtual and augmented reality environments whose voices have been cloned in such a way could likely enhance the sense of immersion experienced by users.

Previous research has demonstrated that a wide range of attributes, such as attitudes, values, and behavior, can lead to a similarity attraction. However, up to this point, only two studies investigated whether voice similarity could lead to similarity effects (Miyake & Zuckerman, 1993; Peng et al., 2019). Moreover, both studies have significant methodological limitations. While Miyake and Zuckerman (1993) did not manipulate voice similarity directly, Peng and colleagues (2019) changed the pitch and the loudness of participants' voice

recordings to investigate whether similar voices would be preferred over dissimilar voices. Although participants preferred recordings altered in loudness over samples altered in pitch, this approach may not adequately test the effects of perceived similarity, as changes in loudness might not lead participants to perceive the recordings as stemming from a different person. Additionally, recordings altered in pitch introduce a distortion to the signal, making it inherently unpleasing. Thus, while these studies provided the first evidence that similar voices are preferred over dissimilar voices, they did not capture how changes in voice similarity impact perception due to the utilized methods. Hence, my experimental series is the first to provide clear evidence for a similarity effect for voices.

From a practical point of view, this could lead designers of voice assistants to choose voices similar to users' voices to make an interaction with them more appealing. This is particularly interesting since voice assistants have no other physical attributes that could be altered to increase their attractiveness. Even though there are physical devices through which voice assistants interact with the user, the variety of devices that can be used as voice assistant interfaces likely prevents them from being perceived as the voice assistant itself. Moreover, since we also found an increase in trust towards similar voices, the interactions should not only be more appealing, but the user could also be more inclined to outsource information and tasks that are more confidential.

The findings from the first experiment reported in Chapter 3 align with existing research on the limitations of human cognitive processing in managing and prioritizing conflicting information when making decisions (Aßmann et al., 2022; Betsch et al., 2016). Although the study hypothesized that voice similarity would make the advice of a competent advisor more compelling, the results did not demonstrate a significant impact on decision outcomes. This suggests that the majority bias might be more influential, potentially overshadowing subtler cues like voice similarity. This finding contributes to theories on minority influence in group decision-making by suggesting that an increase in similarity and, therefore, attractiveness or trust might not significantly enhance the influence of a minority opinion (high-validity advice), especially against a majority consensus (two lower-validity advisors agreeing): If being an expert is not sufficient to be heard in the crowd, being similar to the decision-maker will not help either.

While the assumption that voice similarity may affect decision outcomes in the absence of the majority bias and interfering visual information could not be found in the

subsequent experiments, the explanatory analysis revealed a significant interaction between the voice similarity of advisors and their spatial positioning. Considering the surprisingly sparse research on this topic, this finding provides insights into spatial priming or verticality bias, which suggests that objects or sources of information placed higher in the visual field are perceived as more important or trustworthy. While existing research has found a vertical attention bias for tops of objects and bottoms of scenes (Langley & McBeath, 2023), to my knowledge, only two studies examine the effects of relative stimuli position on their evaluation (Meier & Robinson, 2004; Schubert, 2005). This spatial bias may partially be explained by the interactions with online search engines that most commonly place the best results on top of their result list.

However, even when accounting for the spatial position of the advisor, whether the voice was similar to the participant's voice had only a small impact on decision-making. Therefore, factors such as ease of use or reliability could be more important in interactions with voice assistants than personalization through voice similarity – even though minor effects of voice similarity could have a significant impact considering the large user base of voice-driven devices.

In relation to the illusory truth effect, the experiments detailed in Chapter 4 replicated previous findings and added new insights about the effect itself and how voice similarity can influence the evaluation of repeated information. While using auditory information is not uncommon in research on the illusory truth effect (Henderson et al., 2022), my experiments are the first to utilize a TTS system to generate the material. The decrease in the size of the illusory truth effect compared to the experiment where information was presented and tested in written form may be attributed to the use of such an artificial voice. However, it could also stem from differences in modality-specific information processing and retention. In either case, these differences challenge the assumption that heuristic processing is more likely when arguments or information are presented in audio or video rather than written formats (Chaiken & Eagly, 1983). Furthermore, although participants generally had no difficulties understanding the statements, the increased cognitive load required to process auditory information created with such a system should, theoretically, make heuristic processing more likely and intensify the effect. However, if heuristic processing should be more likely, while the findings suggest a less probable belief in false information than the observed variations in effect sizes may be due to the superior retention of visual information over auditory

information (Cohen et al., 2009; Gloede et al., 2017; Gloede & Gregg, 2019), underscoring the significant role of memory processes in the illusory truth effect.

Since switching the modality from auditory to visual should decrease perceptual fluency while preserving conceptual fluency, the lack of an illusory truth effect in my second experiment in Chapter 4 questions the importance of conceptual fluency on the effect. This observation suggests that perceptual fluency may play a more critical role. However, when switching modality is enough to counter the effect, it questions the real-world impact of the illusory truth effect. Even though studies have shown that perceptual impressions can be remembered over extended periods (Hawley & Johnston, 1991) and that perceptual learning occurs even in the absence of conscious memory (Squire et al., 2021); the illusory truth effect would be of little significance in the real world if it were based solely on perceptual learning or fluency. Certainly, misinformation is often shared verbatim on text-based social media platforms, where perceptual learning or fluency can have a significant impact and it is possible that fragments of verbatim messages could be sufficient to enhance perceptual fluency, and therefore elicit an illusion of truth.

An alternative explanation for the absence of the illusory truth effect when switching the modality could be the absence of increased ease of retrieval from memory caused by modality-specific processing of the information. Although the debate over whether memory is modality-specific continues (Meyerhoff et al., 2023), it is plausible that the single repetition of the statements and the proximity of the veracity rating to the initial exposure emphasize modality-specific memory processing (Kaup et al., 2024). Therefore, since the auditory information is not encoded in visual memory, the lack of enhanced retrieval ease or perception of familiarity might suppress the illusory truth effect. Conversely, if the information had been encountered multiple times over an extended period, it could have been represented in an amodal format, potentially enhancing the illusion of truth across different modalities (Kaup et al., 2024).

The concluding experiment in Chapter 4 assessed the impact of voice similarity on the illusory truth effect, revealing that truth judgments increase for both repeated information and when delivered by a voice similar to the listener's. This reinforces findings from Chapters 2 and 3, where similar voices were deemed more trustworthy, now extended to the domain of truth perception. As outlined in Chapter 1, repetition of a statement could increase the probability of judging the information as accurate by increasing processing fluency or due to memory processes. However, voice similarity may bolster this effect through several

pathways: familiarity with the voice could enhance processing fluency; the perceived similarity may serve as a cue for source credibility; the pleasantness of a similar voice could amplify emotional responses.

Although our experimental design does not allow for the isolation of these individual processes, the interaction between voice similarity and repetition, while not statistically significant, approaches significance and suggests a complex relationship, as illustrated in Chapter 4, Figure 4.1. Despite a statistically non-significant *p*-value, the pattern observed suggests a substantive interaction worth further investigation.

While there is a neglectable difference in truth judgments for newly encountered information presented in a similar voice compared to a generic voice, there is a substantial difference between repeated statements presented in a similar compared to a generic voice. The similarity is, therefore, only influential for repeated statements. This could mean that information from a more trusted, familiar, or attractive source is more likely to be encoded in memory and, therefore, retrieved from memory when the statement should be evaluated. Since remembering information makes information more believable, this could explain the effect of voice similarity on repetition. To evaluate this explanation, future research could investigate whether participants retain more information delivered in a similar voice – a project that could also inform research on the possible use of pedagogical agents with similar voices. Alternatively, in light of dual process models, one could argue that the minimal impact of voice similarity on new information emerges because when information is novel, individuals may rely more on analytic processes. However, as information is repeated, heuristic processes, including familiarity and ease of processing, become more dominant, affecting the truth judgments.

Taken together, the results of Chapter 4 suggest that artificial voices can elicit the illusory truth effect and that speaker similarities can significantly affect the believability of the delivered information.

Strengths, Limitations, and Future Directions

A notable strength of this thesis lies in its commitment to open science principles. To ensure transparency and reproducibility, all my experiments were preregistered, detailing sample size, hypotheses, dependent variables, conditions, analytical methods, and criteria for outliers and exclusions. Furthermore, all experimental data has been made publicly accessible

in an anonymized form alongside the R-code developed for data analysis. Additionally, all experiments entailed in this thesis were approved by the ethics committee of the Leibniz-Institut für Wissensmedien in Tübingen.

Another strength of my thesis is the integration of computer science techniques within cognitive psychology, mainly through the adoption of speaker embeddings as an objective measure for human perceived voice similarity and the generation of individually tailored stimulus material through modern TTS systems. This allows further research to investigate the influence of voice similarity on other cognitive processes, which I will discuss below. Additionally, I used individually adapted stimuli in half of the reported experiments, and in the last study reported in this thesis, I used recently developed methods to clone voices.

A further methodological strength lies in the variety of applied analytical methods: from AIC scores for model comparison, Spearman and inter-rater correlations, to linear mixed-effects models, generalized mixed-effects models, and ANOVAs, I demonstrated the ability to utilize a wide range of statistical methods. I consider this a strength since a researcher's proficiency in a wide range of analytical methods enhances the likelihood of selecting the most appropriate technique rather one, they are familiar with.

Beyond methodological strengths, my thesis also has theoretical strengths: From trait evaluation over truth judgments to the frequency with which participants follow the advice of a similar advisor, the variety of outcome variables I utilized in my studies reflects the different realms in which consequences of voice similarity were investigated. By examining the influence of voice similarity in distinct contexts and across diverse cognitive tasks, my thesis contributes to a more nuanced understanding of how auditory cues influence human perception and information processing. Indeed, my work bridges gaps in the literature on the Similarity-Attraction theory, the Beauty-in-Averageness effect, the Illusory-Truth effect, on persuasion, and decision-making.

Alongside these strengths, my thesis does not come without limitations. Since all my studies were conducted online, I had little control over the experimental setting, significantly reducing the internal validity of the experiments. Nevertheless, online experiments also have merits: Despite a potential decrease in internal validity, the uncontrolled settings could have increased the external validity of my findings.

One major issue of many experiments conducted in cognitive psychology research, particularly regarding external validity, is the homogeneity of participant samples. They are often predominantly comprised of psychology students, leading to a lack of

representativeness. Despite preventing this pitfall by using prolific.com to acquire participants with a large variety of social backgrounds, relying on online platforms dedicated to sourcing research participants could introduce a new form of bias: Professionalized participants prolific in completing experiments in the least amount of time to maximize their utility – without getting flagged by researchers for apparent malpractice. Since my PhD project started during the COVID-19 pandemic, conducting on-site experiments was impossible. Even when such restrictions were lifted, using English audio in the remaining experiments made lab-based studies impractical. This constraint required a continued reliance on online platforms for participant recruitment.

The last limitation I want to address concerns technical aspects: When I started my PhD project, cloning a voice using speaker embeddings as additional information was recently introduced. Indeed, open-source projects using this technique were scarce, required an in-depth understanding of programming languages as well as deep learning methods, and the resulting audio was often underwhelming. However, over the past four years, several openly available projects have been developed that require less programming skills and deliver better results (see Eren & The Coqui TTS Team, 2023; RVC-Boss, 2024). Having these tools available at the beginning of my thesis could have significantly improved the quality of the stimulus material used in my research, potentially leading to more pronounced effect sizes.

As mentioned in the Introduction, besides trait evaluation, truth perception, and decision-making, various other areas could be affected by voice similarity. Since the combined findings of my thesis suggest various influences of voice similarity on cognitive processes, conducting studies investigating other avenues seems promising. Therefore, I want to give a brief outlook on possible future research directions: The last experiment reported in Chapter 2 revealed a significant relationship between trustworthiness ratings and voice similarity. Moreover, in the conducting experiments of Chapter 3, participants more often followed the advice of an advisor with a similar voice compared to an advisor with a generic voice, indicating an increased trust towards such advisors. However, to gather more evidence for the hypothesis that voice similarity increases trust, one could use the *Trust game*, named initially *Investment Game*, widely utilized in economic studies to assess trust and reciprocity in different situations (Berg et al., 1995; Camerer & Weigelt, 1988; Johnson & Mislin, 2011) . Incorporating trustees with either similar or generic voices would allow researchers to observe if and how voice similarity impacts the amount of money trustors choose to send, reflecting their trust levels and trust-based decision-making.

In the initial experiment reported in Chapter 3, participants retrieved more predictions from a similar advisor than a generic advisor, yet the reason must be clarified. The similarity of the voice may lead to an increase in interest or attention. In case of heightened interest, this could be due to a mere curiosity in the similarity of the voice – which could wear off over time – or be due to more complex cognitive processes. To differentiate between those two possibilities, interest levels could be measured by tracking the time participants spend engaging with similar advisors compared to generic advisors and determining if existing differences vanish over time. A suitable paradigm for testing whether voice similarity enhances attention toward individuals could be the selective auditory attention task (Cherry, 1953). Participants could simultaneously be presented audio from two speakers, one with a voice similar to their own and the other with a generic voice. In a between-subjects design, participants could be instructed to focus on the similar or generic voice. As the dependent variable, one could measure recall accuracy and reaction times to investigate the effects of similarity on attention.

Another promising avenue of research into voice similarity could lie in knowledge acquisition. Indeed, self-related material can be more easily remembered compared to material without relation to ourselves (Liu et al., 2024; Rogers et al., 1977). Such a *self-reference effect* in memory is likely due to a more organized and elaborate processing (Symons & Johnson, 1997). To investigate whether voice-similarity could benefit learning, one could present material either in a similar or generic voice. By assessing recall accuracy, one can determine the impact of voice similarity on memory retention. To ascertain whether this stems from an increase in self-reference and self-referential encoding, one could test whether learners are more likely to associate self-descriptive terms with the content presented in a similar voice.

As mentioned earlier, voice similarity can lead to a sense of familiarity and increased processing fluency. This improved processing fluency could enhance the learning experience by reducing cognitive load and influencing how individuals acquire and preserve new information. Additionally, the increased familiarity could improve perceived empathy, making the learning experience more desirable. Indeed, future experiments could investigate how voice similarity between a teacher or virtual agent and the learner affects the listener's engagement and learning outcome. Furthermore, research could explore how personalized voice settings in e-learning environments improve learning outcomes.

Conclusion

Existing research has demonstrated that similarity across various characteristics, such as attitudes, appearance, and values, enhances perceived attractiveness, trustworthiness, and pro-social behavior. However, studies exploring whether voice similarity could lead to such effects were relatively sparse. Since deep learning-based TTS systems can now generate speech in a voice that resembles the voice of a target individual, and because such systems are more and more prevalent, my thesis aimed to investigate the possible consequences of voice similarity.

While the findings presented in my thesis suggest that cosine similarity values can be used to investigate voice similarity effects, they also provide insights into voice perception in general and how speaker recognition systems could be adapted to improve aligning between human similarity judgments and AI ratings.

Moreover, the findings indicate that real and artificial voices exert different effects based on their similarity to the listener's own voice. Speakers with similar voices are perceived as more likable and trustworthy, may influence decisions affecting personal gains, and can lead individuals to believe factually incorrect information.

Despite the relatively subtle effects observed, these results highlight the potential for misuse of adapted voices in TTS systems and voice assistants. Nevertheless, future research should also consider how voice adaptation could be utilized positively, such as enhancing trust in public information campaigns, including those related to political and health education.

References

Ahrens, M.-M., Awwad Shiekh Hasan, B., Giordano, B. L., & Belin, P. (2014). Gender differences in the temporal voice areas. *Frontiers in Neuroscience*, *8*. https://www.frontiersin.org/articles/10.3389/fnins.2014.00228

Ajzen, I. (1974). Effects of information on interpersonal attraction: Similarity versus affective value. *Journal of Personality and Social Psychology*, *29*(3), 374–380. https://doi.org/10.1037/h0036002

Alphabet. (2020). *Form 10-K*. https://www.sec.gov/Archives/edgar/data/1652044/000165204421000010/goog-20201231.htm

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the Tribes of Fluency to Form a Metacognitive Nation. *Personality and Social Psychology Review*. https://doi.org/10.1177/1088868309341564

Amazon. (2020). *Form 10-K*. https://www.sec.gov/Archives/edgar/data/1018724/000101872421000004/amzn-20201231.htm#i75de98b9097f40f3b5884e541f532421_13

Ammari, T., Kaye, J., Tsai, J. Y., & Bentley, F. (2019). Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Transactions on Computer-Human Interaction*, *26*(3), 1–28. https://doi.org/10.1145/3311956

Anderson, R. B., & McMillion, P. Y. (1995). Effects of Similar and Diversified Modeling on African American Women's Efficacy Expectations and Intentions to Perform Breast Self-Examination. *Health Communication*, *7*(4), 327–343. https://doi.org/10.1207/s15327027hc0704_3

Andraszewicz, R., Yamagishi, J., & King, S. (2011). Vocal attractiveness of statistical speech synthesisers. *In Proc. ICASSP 2011*, 5368–5371.

Arik, S., Chen, J., Peng, K., Ping, W., & Zhou, Y. (2018). Neural Voice Cloning with a Few Samples. *Advances in Neural Information Processing Systems*, 10019–10029.

Arik, S., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., & Shoeybi, M. (2017). *Deep Voice: Real-time Neural Text-to-Speech*. https://doi.org/10.48550/ARXIV.1702.07825

Arik, S., Diamos, G., Gibiansky, A., Miller, J., Peng, K., Ping, W., Raiman, J., & Zhou, Y. (2017). Deep Voice 2: Multi-Speaker Neural Text-to-Speech. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2966–2974. http://arxiv.org/abs/1705.08947

Arkes, H., Hackett, C., & Boehm, L. (1989). The generality of the relation between familiarity and judged validity. *Journal of Behavioral Decision Making*, *2*, 81–94. https://doi.org/10.1002/bdm.3960020203

Aron, A., Aron, E. N., Tudor, M., & Nelson, G. (1991). Close relationships as including other in the self. *Journal of Personality and Social Psychology*, *60*(2), 241–253. https://doi.org/10.1037/0022-3514.60.2.241

Aßmann, L., Betsch, T., Lang, A., & Lindow, S. (2022). When even the smartest fail to prioritise: Overuse of information can decrease decision accuracy. *Journal of Cognitive Psychology*, *34*(5), 675–690. https://doi.org/10.1080/20445911.2022.2055560

Atal, B. S. (1976). Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, *64*(4), 460–475. https://doi.org/10.1109/PROC.1976.10155

Bacon, F., Begg, Mitterer, J., Upfold, & Harris, G. (1979). Credibility of repeated statements: Memory for trivia. *Journal of Experimental Psychology Learning Memory and Cognition*, *5*, 241–252.

Bahns, A. J., Crandall, C. S., Gillath, O., & Preacher, K. J. (2017). Similarity in relationships as niche construction: Choice, stability, and influence within dyads in a free choice environment. *Journal of Personality and Social Psychology*, *112*(2), 329–355. https://doi.org/10.1037/pspp0000088

Banai, B., Laustsen, L., Banai, I. P., & Bovan, K. (2018). Presidential, But Not Prime Minister, Candidates With Lower Pitched Voices Stand a Better Chance of Winning the Election in Conservative Countries. *Evolutionary Psychology*, *16*(2), 1474704918758736. https://doi.org/10.1177/1474704918758736

Barton, K. (2020). *MuMIn: Multi-Model Inference* (Version 1.43.17) [Computer software]. https://CRAN.R-project.org/package=MuMIn

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *arXiv:1406.5823 [Stat]*. http://arxiv.org/abs/1406.5823

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, *117*(3), 497–529. https://doi.org/10.1037/0033-2909.117.3.497

Begg, I. M., Anas, A., & Farinacci, S. (1992). Dissociation of Processes in Belief: Source Recollection, Statement Familiarity, and the Illusion of Truth. *Journal of Experimental Psychology: General*, *121*(4), 446–458.

Beggan, J. K. (1992). On the social nature of nonsocial perception: The mere ownership effect. *Journal of Personality and Social Psychology*, *62*(2), 229–237. https://doi.org/10.1037/0022-3514.62.2.229

Beilharz, B., Sun, X., Karimova, S., & Riezler, S. (2020). LibriVoxDeEn: A Corpus for German-to-English Speech Translation and German Speech Recognition. *Proceedings of The 12th Language Resources and Evaluation Conference*, 3590–3594.

Belin, P. (2021). On Voice Averaging and Attractiveness. In B. Weiss, J. Trouvain, M. Barkat-Defradas, & J. J. Ohala (Eds.), *Voice Attractiveness* (pp. 139–149). Springer Singapore. https://doi.org/10.1007/978-981-15-6627-1_8

Belke, B., Leder, H., Strobach, T., & Carbon, C.-C. (2010). Cognitive fluency: High-level processing dynamics in art appreciation. *Psychology of Aesthetics, Creativity, and the Arts*, *4*(4), 214–222. https://doi.org/10.1037/a0019648

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Berger, J., & Won, S. Y. (2005). Estimating Transfer Function from Air to Bone Conduction using Singing Voice. *ICMC*.

Berscheid, E. (1966). Opinion change and communicator-communicatee similarity and dissimilarity. *Journal of Personality and Social Psychology*, *4*(6), 670. https://doi.org/10.1037/h0021193

Betsch, T., Lang, A., Lehmann, A., & Axmann, J. M. (2014). Utilizing Probabilities as Decision Weights in Closed and Open Information Boards: A Comparison of Children and Adults. *Acta Psychologica*, *153*, 74–86. https://doi.org/10.1016/j.actpsy.2014.09.008

Betsch, T., Lehmann, A., Lindow, S., Lang, A., & Schoemann, M. (2016). Lost in Search: (Mal-)Adaptation to Probabilistic Decision Environments in Children and Adults. *Developmental Psychology*, *52*(2), 311–325. https://doi.org/10.1037/dev0000077

Bhatia, S. (2017). Choice rules and accumulator networks. *Decision*, *4*(3), 146–170. https://doi.org/10.1037/dec0000038

Bigelow, J., & Poremba, A. (2014). Achilles' Ear? Inferior Human Short-Term and Recognition Memory in the Auditory Modality. *PLOS ONE*, *9*(2), e89914. https://doi.org/10.1371/journal.pone.0089914

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D., & Reynolds, D. A. (2004). A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Advances in Signal Processing*, *2004*(4), 101962. https://doi.org/10.1155/S1110865704310024

Bisani, M., & Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, *50*(5), 434–451. https://doi.org/10.1016/j.specom.2008.01.002

Black, A. W., & Taylor, P. (1997). Automatically clustering similar units for unit selection in speech synthesis. *5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, 601–604. https://doi.org/10.21437/Eurospeech.1997-219

Black, A. W., Zen, H., & Tokuda, K. (2007). Statistical Parametric Speech Synthesis. *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, IV-1229-IV–1232. https://doi.org/10.1109/ICASSP.2007.367298

Bless, H. (2003). The consequences of mood on the processing of social information. In A. Tesser & M. Hewstone (Eds.), *Blackwell Handbook of Social Psychology: Intraindividual Processes.* (pp. 391–412). Blackwell.

Boehm, L. E. (1994). The Validity Effect: A Search for Mediating Variables. *Personality and Social Psychology Bulletin*, *20*(3), 285–293. https://doi.org/10.1177/0146167294203006

Boer, D., Fischer, R., Strack, M., Bond, M. H., Lo, E., & Lam, J. (2011). How Shared Preferences in Music Create Bonds Between People: Values as the Missing Link.

Personality and Social Psychology Bulletin, *37*(9), 1159–1171. https://doi.org/10.1177/0146167211407521

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, *101*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology*, *79*(4), 631–643. https://doi.org/10.1037/0022-3514.79.4.631

Brashier, N. M., & Marsh, E. J. (2020). Judging Truth. *Annual Review of Psychology*, *71*(1), 499–515. https://doi.org/10.1146/annurev-psych-010419-050807

Brewster, T. (2021, October 14). *Fraudsters Cloned Company Director's Voice In $35 Million Bank Heist, Police Find*. Forbes. https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/

Bröder, A., & Newell, B. R. (2008). Challenging some common beliefs: Empirical work within the adaptive toolbox metaphor. *Judgment and Decision Making*, *3*(3), 205–214. https://doi.org/10.1017/S1930297500002412

Brodsky, S. L., Neal, T. M. S., Cramer, R. J., & Ziemke, M. H. (2009). Credibility in the Courtroom: How Likeable Should an Expert Witness Be? *Journal of the American Academy of Psychiatry and the Law*, *37*(4), 525–532.

Brosch, T., Scherer, K., Grandjean, D., & Sander, D. (2013). The impact of emotion on perception, attention, memory, and decision-making. *Swiss Medical Weekly*. https://doi.org/10.4414/smw.2013.13786

Brown, A. S., & Nix, L. A. (1996). Turning lies into truths: Referential validation of falsehoods. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1088–1100. https://doi.org/10.1037/0278-7393.22.5.1088

Brown, R. (2000). Social identity theory: Past achievements, current problems and future challenges. *European Journal of Social Psychology*, *30*(6), 745–778. https://doi.org/10.1002/1099-0992(200011/12)30:6<745::AID-EJSP24>3.0.CO;2-O

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., Kawahara, H., & Belin, P. (2010). Vocal Attractiveness Increases by Averaging. *Current Biology*, *20*(2), 116–120. https://doi.org/10.1016/j.cub.2009.11.034

Burger, J. M., Messian, N., Patel, S., del Prado, A., & Anderson, C. (2004). What a Coincidence! The Effects of Incidental Similarity on Compliance. *Personality and Social Psychology Bulletin*, *30*(1), 35–43. https://doi.org/10.1177/0146167203258838

Burgess, S. (2022, March 17). *Ukraine war: Deepfake video of Zelenskyy telling Ukrainians to "lay down arms" debunked*. Sky News. https://news.sky.com/story/ukraine-war-deepfake-video-of-zelenskyy-telling-ukrainians-to-lay-down-arms-debunked-12567789

Burnstein, E., Crandall, C., & Kitayama, S. (1994). Some neo-Darwinian decision rules for altruism: Weighing cues for inclusive fitness as a function of the biological importance of the decision. *Journal of Personality and Social Psychology*, *67*(5), 773–789. https://doi.org/10.1037/0022-3514.67.5.773

Byrne, D. (1961). Interpersonal attraction and attitude similarity. *The Journal of Abnormal and Social Psychology*, *62*(3), 713–715. https://doi.org/10.1037/h0044721

Byrne, D., & Griffitt, W. (1973). Interpersonal Attraction. *Annual Review of Psychology*, *24*(1), 317–336. https://doi.org/10.1146/annurev.ps.24.020173.001533

Byrne, D., Griffitt, W., & Stefaniak, D. (1967). Attraction and similarity of personality characteristics. *Journal of Personality and Social Psychology*, *5*(1), 82–90. https://doi.org/10.1037/h0021198

Cambre, J., & Kulkarni, C. (2019). One Voice Fits All?: Social Implications and Research Challenges of Designing Voices for Smart Devices. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–19. https://doi.org/10.1145/3359325

Camerer, C., & Weigelt, K. (1988). Experimental Tests of a Sequential Equilibrium Reputation Model. *Econometrica*, *56*(1), 1–36. https://doi.org/10.2307/1911840

Capraro, V. (2024). The dual-process approach to human sociality: Meta-analytic evidence for a theory of internalized heuristics for self-preservation. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspa0000375

Carbonneau, M.-A., Granger, E., Attabi, Y., & Gagnon, G. (2017). Feature Learning from Spectrograms for Assessment of Personality Traits. *IEEE Transactions on Affective Computing*, *11*(1), 25–31. https://doi.org/10.1109/TAFFC.2017.2763132

Carli, L. L., Ganley, R., & Pierce-Otay, A. (1991). Similarity and Satisfaction in Roommate Relationships. *Personality and Social Psychology Bulletin*, *17*(4), 419–426. https://doi.org/10.1177/0146167291174010

Casanova, E., Shulby, C., Gölge, E., Müller, N. M., de Oliveira, F. S., Junior, A. C., Soares, A. da S., Aluisio, S. M., & Ponti, M. A. (2021). SC-GlowTTS: An Efficient Zero-Shot Multi-Speaker Text-To-Speech Model. *arXiv:2104.05557 [Cs, Eess]*. http://arxiv.org/abs/2104.05557

Casanova, E., Weber, J., Shulby, C., Junior, A. C., Gölge, E., & Ponti, M. A. (2021). *YourTTS: Towards Zero-Shot Multi-Speaker TTS and Zero-Shot Voice Conversion for everyone*. https://doi.org/10.48550/ARXIV.2112.02418

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*(5), 752. https://doi.org/10.1037/0022-3514.39.5.752

Chaiken, S., & Eagly, A. H. (1983). Communication modality as a determinant of persuasion: The role of communicator salience. *Journal of Personality and Social Psychology*, *45*(2), 241–256. https://doi.org/10.1037/0022-3514.45.2.241

Chaspari, T., & Lehman, J. F. (2016). An Acoustic Analysis of Child-Child and Child-Robot Interactions for Understanding Engagement during Speech-Controlled Computer Games. *Interspeech 2016*, 595–599. https://doi.org/10.21437/Interspeech.2016-85

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., & Chan, W. (2020). WaveGrad: Estimating Gradients for Waveform Generation. *arXiv:2009.00713 [Cs, Eess, Stat]*. http://arxiv.org/abs/2009.00713

Chen, S., & Chaiken, S. (1999). The heuristic-systematic model in its broader context. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 73–96). The Guilford Press.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. https://doi.org/10.1121/1.1907229

Cicchetti, D. V. (1994). Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Clayson, D. (2022). The student evaluation of teaching and likability: What the evaluations actually measure. *Assessment & Evaluation in Higher Education*, *47*(2), 313–326. https://doi.org/10.1080/02602938.2021.1909702

Clore, G. L., & Byrne, D. (1974). A Reinforcement-Affect Model of Attraction. In *Foundations of Interpersonal Attraction* (pp. 143–170). Elsevier. https://doi.org/10.1016/B978-0-12-362950-0.50013-6

Cohen, M. A., Horowitz, T. S., & Wolfe, J. M. (2009). Auditory recognition memory is inferior to visual recognition memory. *Proceedings of the National Academy of Sciences*, *106*(14), 6008–6010. https://doi.org/10.1073/pnas.0811884106

Colavita, F. B. (1974). Human sensory dominance. *Perception & Psychophysics*, *16*(2), 409–412. https://doi.org/10.3758/BF03203962

Condon, J. W., & Crano, W. D. (1988). Inferred evaluation and the relation between attitude similarity and interpersonal attraction. *Journal of Personality and Social Psychology*, *54*(5), 789–797. https://doi.org/10.1037/0022-3514.54.5.789

Cooper, E., Lai, C.-I., Yasuda, Y., Fang, F., Wang, X., Chen, N., & Yamagishi, J. (2020). Zero-Shot Multi-Speaker Text-To-Speech with State-Of-The-Art Neural Speaker Embeddings. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6184–6188. https://doi.org/10.1109/ICASSP40776.2020.9054535

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463–480. https://doi.org/10.1037/a0037010

Crano, W. D., & Prislin, R. (2006). Attitudes and Persuasion. *Annual Review of Psychology*, *57*(1), 345–374. https://doi.org/10.1146/annurev.psych.57.102904.190034

Cummins, N., Baird, A., & Schuller, B. W. (2018). Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*, *151*, 41–54. https://doi.org/10.1016/j.ymeth.2018.07.007

De keersmaecker, J., Dunning, D., Pennycook, G., Rand, D. G., Sanchez, C., Unkelbach, C., & Roets, A. (2020). Investigating the Robustness of the Illusory Truth Effect Across Individual Differences in Cognitive Ability, Need for Cognitive Closure, and Cognitive Style. *Personality and Social Psychology Bulletin*, *46*(2), 204–215. https://doi.org/10.1177/0146167219853844

De Neys, W. (Ed.). (2018). *Dual process theory 2.0* (1 Edition). Routledge.

De Neys, W., & Pennycook, G. (2019). Logic, Fast and Slow: Advances in Dual-Process Theorizing. *Current Directions in Psychological Science*, *28*(5), 503–509. https://doi.org/10.1177/0963721419855658

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The Truth About the Truth: A Meta-Analytic Review of the Truth Effect. *Personality and Social Psychology Review*, *14*(2), 238–257. https://doi.org/10.1177/1088868309352251

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798. https://doi.org/10.1109/TASL.2010.2064307

Devendorf, S. A., & Highhouse, S. (2008). Applicant–employee similarity and attraction to an employer. *Journal of Occupational and Organizational Psychology*, *81*(4), 607–617. https://doi.org/10.1348/096317907X248842

DiFonzo, N., Beckstead, J. W., Stupak, N., & Walders, K. (2016). Validity judgments of rumors heard multiple times: The shape of the truth effect. *Social Influence*, *11*(1), 22–39. https://doi.org/10.1080/15534510.2015.1137224

Dilman, I. (1999). *Free Will: An Historical and Philosophical Introduction* (1st ed.). Routledge. https://doi.org/10.4324/9780203002384

Doddington, G. R. (1985a). Speaker recognition—Identifying people by their voices. *Proceedings of the IEEE*, *73*(11), 1651–1664. https://doi.org/10.1109/PROC.1985.13345

Doddington, G. R. (1985b). Speaker Recognition—Identifying People by Their Voices. *Proceedings of the IEEE*, *73*(11), 1651–1664. https://doi.org/10.1109/PROC.1985.13345

Doddipatla, R., Braunschweiler, N., & Maia, R. (2017). Speaker Adaptation in DNN-Based Speech Synthesis Using d-Vectors. *Interspeech 2017*, 3404–3408. https://doi.org/10.21437/Interspeech.2017-1038

Dunning, D. (1999). A Newer Look: Motivated Social Cognition and the Schematic Representation of Social Concepts. *Psychological Inquiry*, *10*(1), 1–11. https://doi.org/10.1207/s15327965pli1001_1

Dyer, J. H., & Chu, W. (2003). The Role of Trustworthiness in Reducing Transaction Costs and Improving Performance: Empirical Evidence from the United States, Japan, and Korea. *Organization Science*, *14*(1), 57–68. https://doi.org/10.1287/orsc.14.1.57.12806

Ecker, U. K. H., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., & Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, *1*(1), Article 1. https://doi.org/10.1038/s44159-021-00006-y

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, *51*(4), 380–417. https://doi.org/10.1037/h0053870

Eren, G., & The Coqui TTS Team. (2023). *Coqui TTS* (Version v0.16.6) [Computer software]. https://www.coqui.ai

Evans, A. M., & Krueger, J. I. (2009). The Psychology (and Economics) of Trust. *Social and Personality Psychology Compass*, *3*(6), 1003–1017. https://doi.org/10.1111/j.1751-9004.2009.00232.x

Evans, J. (2003). In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459. https://doi.org/10.1016/j.tics.2003.08.012

Evans, J. (2008). Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition. *Annual Review of Psychology*, *59*(1), 255–278. https://doi.org/10.1146/annurev.psych.59.103006.093629

Falk, E., & Scholz, C. (2018). Persuasion, Influence, and Value: Perspectives from Communication and Social Neuroscience. *Annual Review of Psychology*, *69*(1), 329–356. https://doi.org/10.1146/annurev-psych-122216-011821

Fan, Y., Qian, Y., Soong, F. K., & He, L. (2015). Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4475–4479. https://doi.org/10.1109/ICASSP.2015.7178817

Faraji-Rad, A., Samuelsen, B. M., & Warlop, L. (2015). On the Persuasiveness of Similar Others: The Role of Mentalizing and the Feeling of Certainty. *Journal of Consumer Research*, *42*(3), 458–471. https://doi.org/10.1093/jcr/ucv032

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*(5), 993–1002. https://doi.org/10.1037/xge0000098

Fazio, L. K., Rand, D. G., & Pennycook, G. (2019). Repetition increases perceived truth equally for plausible and implausible statements. *Psychonomic Bulletin & Review*, *26*(5), 1705–1710. https://doi.org/10.3758/s13423-019-01651-4

Fazio, L. K., & Sherry, C. L. (2020). The Effect of Repetition on Truth Judgments Across Development. *Psychological Science*, *31*(9), 1150–1160. https://doi.org/10.1177/0956797620939534

Fazio, R. H. (1990). *Multiple processes by which attitudes guide behavior: The mode model as an integrative framework*. *23*, 75–109. https://doi.org/10.1016/S0065-2601(08)60318-4

Fazio, R. H., & Olson, M. A. (2014). The MODE model: Attitude-behavior processes as a function of motivation and opportunity. In *Dual-process theories of the social mind.* (pp. 155–171). The Guilford Press.

Feng, B., & MacGeorge, E. L. (2010). The Influences of Message and Source Factors on Advice Outcomes. *Communication Research*, *37*(4), 553–575. https://doi.org/10.1177/0093650210368258

Ferrand, C. T. (2002). Harmonics-to-Noise Ratio: An Index of Vocal Aging. *Journal of Voice*, *16*(4), 480–487. https://doi.org/10.1016/S0892-1997(02)00123-6

Ferreira Caceres, M. M., Sosa, J. P., Lawrence, J. A., Sestacovschi, C., Tidd-Johnson, A., Rasool, M. H. U., Gadamidi, V. K., Ozair, S., Pandav, K., Cuevas-Lou, C., Parrish, M., Rodriguez, I., Fernandez, J. P., Division of Research & Academic Affairs, Larkin Community Hospital, South Miami, Florida, USA, Department of Medicine, American University of Antigua, Coolidge, Antigua, Family Medicine, Larkin Community Hospital Palm Springs Campus, Hialeah, Florida, USA, Family Medicine, Larkin Community Hospital South Campus, Miami, Florida, USA, & Pulmonary Disease and Critical Care Medicine, Larkin Community Hospital Palm Springs Campus, Hilaeah,

Florida, USA. (2022). The impact of misinformation on the COVID-19 pandemic. *AIMS Public Health*, *9*(2), 262–277. https://doi.org/10.3934/publichealth.2022018

Festinger, L. (1954). A Theory of Social Comparison Processes. *Human Relations*, *7*(2), 117–140. https://doi.org/10.1177/001872675400700202

Filieri, R., McLeay, F., Tsui, B., & Lin, Z. (2018). Consumer perceptions of information helpfulness and determinants of purchase intention in online consumer reviews of services. *Information & Management*, *55*(8), 956–970. https://doi.org/10.1016/j.im.2018.04.010

Finkel, E. J., Eastwick, P. W., Karney, B. R., Reis, H. T., & Sprecher, S. (2012). Online Dating: A Critical Analysis From the Perspective of Psychological Science. *Psychological Science in the Public Interest*, *13*(1), 3–66. https://doi.org/10.1177/1529100612436522

Fischhoff, B., & Broomell, S. B. (2020). Judgment and Decision Making. *Annual Review of Psychology*, *71*(1), 331–355. https://doi.org/10.1146/annurev-psych-010419-050747

Fishburn, P. C. (1979). *Utility theory for decision making*. NY: Kriger.

Fishburn, P. C. (1981). Subjective expected utility: A review of normative theories. *Theory and Decision*, *13*(2), 139–199. https://doi.org/10.1007/BF00134215

Fiske, S. T., & Neuberg, S. L. (1990). A Continuum of Impression Formation, from Category-Based to Individuating Processes: Influences of Information and Motivation on Attention and Interpretation. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 23, pp. 1–74). Academic Press. https://doi.org/10.1016/S0065-2601(08)60317-2

Forster, M., Leder, H., & Ansorge, U. (2013). It felt fluent, and I liked it: Subjective feeling of fluency rather than objective fluency determines liking. *Emotion*, *13*(2), 280–289. https://doi.org/10.1037/a0030115

Foster, J. L., Huthwaite, T., Yesberg, J. A., Garry, M., & Loftus, E. F. (2012). Repetition, Not Number of Sources, Increases Both Susceptibility to Misinformation and Confidence in the Accuracy of Eyewitnesses. *Acta Psychologica*, *139*(2), 320–326. https://doi.org/10.1016/j.actpsy.2011.12.004

Frances, C., Costa, A., & Baus, C. (2018). On the effects of regional accents on memory and credibility. *Acta Psychologica*, *186*, 63–70. https://doi.org/10.1016/j.actpsy.2018.04.003

Frisch, D., & Baron, J. (1988). Ambiguity and rationality. *Journal of Behavioral Decision Making*, *1*(3), 149–157. https://doi.org/10.1002/bdm.3960010303

Fu, S., Yan, Q., & Feng, G. C. (2018). Who will attract you? Similarity effect among users on online purchase intention of movie tickets in the social shopping context. *International Journal of Information Management*, *40*, 88–102. https://doi.org/10.1016/j.ijinfomgt.2018.01.013

Gamer, M., Lemon, J., & ¡puspendra.pusp22@gmail.com¿, I. F. P. S. (2012). *irr: Various coefficients of interrater reliability and agreement* [Manual]. https://CRAN.R-project.org/package=irr

Garcia-Romero, D., & Espy-Wilson, C. Y. (2011). Analysis of i-vector length normalization in speaker recognition systems. *Interspeech 2011*, 249–252. https://doi.org/10.21437/Interspeech.2011-53

Gawronski, B., & Creighton, L. A. (2013). Dual Process Theories. In *The Oxford handbook of social cognition* (p. 282). Oxford University Press.

Gerlach, L., McDougall, K., Kelly, F., Alexander, A., & Nolan, F. (2020). Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features. *Speech Communication*, *124*, 85–95. https://doi.org/10.1016/j.specom.2020.08.003

Gigerenzer, G. (1984). External Validity of Laboratory Experiments: The Frequency-Validity Relationship. *The American Journal of Psychology*, *97*(2), 185–195. https://doi.org/10.2307/1422594

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic Decision Making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650–669. https://doi.org/10.1037/0033-295X.103.4.650

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press, USA.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098

Gino, F., Shang, J., & Croson, R. (2009). The impact of information from similar or different advisors on judgment. *Organizational Behavior and Human Decision Processes*, *108*(2), 287–302. https://doi.org/10.1016/j.obhdp.2008.08.002

Gloede, M. E., & Gregg, M. K. (2019). The fidelity of visual and auditory memory. *Psychonomic Bulletin & Review*, *26*(4), 1325–1332. https://doi.org/10.3758/s13423-019-01597-7

Gloede, M. E., Paulauskas, E. E., & Gregg, M. K. (2017). Experience and information loss in auditory and visual memory. *Quarterly Journal of Experimental Psychology*, *70*(7), 1344–1352. https://doi.org/10.1080/17470218.2016.1183686

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*(1), 75–90. https://doi.org/10.1037/0033-295X.109.1.75

Gommans, R., Sandstrom, M. J., Stevens, G. W. J. M., ter Bogt, T. F. M., & Cillessen, A. H. N. (2017). Popularity, likeability, and peer conformity: Four field experiments. *Journal of Experimental Social Psychology*, *73*, 279–289. https://doi.org/10.1016/j.jesp.2017.10.001

González Hautamäki, R., Kinnunen, T., Hautamäki, V., & Laukkanen, A.-M. (2015). Automatic versus human speaker verification: The case of voice mimicry. *Speech Communication*, *72*, 13–31. https://doi.org/10.1016/j.specom.2015.05.002

Gosling, S. D., John, O. P., Craik, K. H., & Robins, R. W. (1998). Do people know how they behave? Self-reported act frequencies compared with on-line codings by observers. *Journal of Personality and Social Psychology*, *74*(5), 1337–1349. https://doi.org/10.1037/0022-3514.74.5.1337

Govalkar, P., Fischer, J., Zalkow, F., & Dittmar, C. (2019). A Comparison of Recent Neural Vocoders for Speech Signal Reconstruction. *10th ISCA Workshop on Speech Synthesis (SSW 10)*, 7–12. https://doi.org/10.21437/SSW.2019-2

Grágeda, N., Alvarado, E., Mahu, R., Busso, C., & Becerra Yoma, N. (2023). Distant Speech Emotion Recognition in an Indoor Human-robot Interaction Scenario. *INTERSPEECH 2023*, 3657–3661. https://doi.org/10.21437/Interspeech.2023-1169

Grágeda, N., Busso, C., Alvarado, E., García, R., Mahu, R., Huenupan, F., & Yoma, N. B. (2025). Speech emotion recognition in real static and dynamic human-robot interaction scenarios. *Computer Speech & Language*, *89*, 101666. https://doi.org/10.1016/j.csl.2024.101666

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Halberstadt, J. (2006). The Generality and Ultimate Origins of the Attractiveness of Prototypes. *Personality and Social Psychology Review*, *10*(2), 166–183. https://doi.org/10.1207/s15327957pspr1002_5

Hamilton, W. D. (1964). The genetical evolution of social behaviour. II. *Journal of Theoretical Biology*, *7*(1), 17–52. https://doi.org/10.1016/0022-5193(64)90039-6

Handley, S. J., Newstead, S. E., & Trippas, D. (2011). Logic, beliefs, and instruction: A test of the default interventionist account of belief bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 28–43. https://doi.org/10.1037/a0021098

Hansen, J., & Hasan, T. (2015). Speaker Recognition by Machines and Humans: A tutorial review. *Signal Processing Magazine, IEEE*, *32*, 74–99. https://doi.org/10.1109/MSP.2015.2462851

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*(1), 107–112. https://doi.org/10.1016/S0022-5371(77)80012-1

Hastie, R. (2001). Problems for Judgment and Decision Making. *Annual Review of Psychology*, *52*(1), 653–683. https://doi.org/10.1146/annurev.psych.52.1.653

Hastie, R., & Kameda, T. (2005). The Robust Beauty of Majority Rules in Group Decisions. *Psychological Review*, *112*, 494–508. https://doi.org/10.1037/0033-295X.112.2.494

Hawkins, S. A., & Hoch, S. J. (1992). Low-Involvement Learning: Memory without Evaluation. *Journal of Consumer Research*, *19*(2), 212. https://doi.org/10.1086/209297

Hawkins, S. A., Hoch, S. J., & Meyers-Levy, J. (2001). Low-Involvement Learning: Repetition and Coherence in Familiarity and Belief. *Journal of Consumer Psychology*, *11*(1), 1–11. https://doi.org/10.1207/S15327663JCP1101_1

Hawley, K. J., & Johnston, W. A. (1991). Long-term perceptual memory for briefly exposed words as a function of awareness and attention. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(3), 807–815. https://doi.org/10.1037/0096-1523.17.3.807

Hecht, D., Reiner, M., & Karni, A. (2009). Repetition priming for multisensory stimuli: Task-irrelevant and task-relevant stimuli are associated if semantically related but with no advantage over uni-sensory stimuli. *Brain Research*, *1251*, 236–244. https://doi.org/10.1016/j.brainres.2008.10.062

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Heigold, G., Moreno, I., Bengio, S., & Shazeer, N. (2016). End-to-end text-dependent speaker verification. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5115–5119. https://doi.org/10.1109/ICASSP.2016.7472652

Henderson, E. L., Westwood, S. J., & Simons, D. J. (2022). A reproducible systematic map of research on the illusory truth effect. *Psychonomic Bulletin & Review*, *29*(3), 1065–1088. https://doi.org/10.3758/s13423-021-01995-w

Heusser, A. C., Awipi, T., & Davachi, L. (2013). The ups and downs of repetition: Modulation of the perirhinal cortex by conceptual repetition predicts priming and long-term memory. *Neuropsychologia*, *51*(12), 2333–2343. https://doi.org/10.1016/j.neuropsychologia.2013.04.018

Hillenbrand, J. (1987). A Methodological Study of Perturbation and Additive Noise in Synthetically Generated Voice Signals. *Journal of Speech, Language, and Hearing Research*, *30*(4), 448–461. https://doi.org/10.1044/jshr.3004.448

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine*, *29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

Holzleitner, I. J., Lee, A. J., Hahn, A. C., Kandrik, M., Bovet, J., Renoult, J. P., Simmons, D., Garrod, O., DeBruine, L. M., & Jones, B. C. (2019). Comparing theory-driven and data-driven attractiveness models using images of real women's faces. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(12), 1589–1595. https://doi.org/10.1037/xhp0000685

Hornsey, M. J. (2008). Social Identity Theory and Self-categorization Theory: A Historical Review. *Social and Personality Psychology Compass*, *2*(1), 204–222. https://doi.org/10.1111/j.1751-9004.2007.00066.x

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion; psychological studies of opinion change.* (pp. xii, 315). Yale University Press.

Hughes, S. M., & Harrison, M. A. (2013). I like My Voice Better: Self-Enhancement Bias in Perceptions of Voice Attractiveness. *Perception*, *42*(9), 941–949. https://doi.org/10.1068/p7526

Hunt, A. J., & Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, *1*, 373–376. https://doi.org/10.1109/ICASSP.1996.541110

Ivanič, A. S., Bates, K., & Somasundaram, T. (2014). The Role of the Accent in Radio Advertisements to Ethnic Audiences: Does Emphasizing Ethnic Stereotypes Affect Spokesperson Credibility and Purchase Intention? *Journal of Advertising Research*, *54*(4), 407–419. https://doi.org/10.2501/JAR-54-4-407-419

Jacoby, L. L., & Dallas, M. (1981). On the Relationship Between Autobiographical Memory and Perceptual Learning. *Journal of Experimental Psychology: General*, *110*(3), 306–340. https://doi.org/10.1037/0096-3445.110.3.306

Jaggy, O., Schwan, S., & Meyerhoff, H. S. (2025). AI-determined similarity increases likability and trustworthiness of human voices. *PLOS ONE*, *20*(3), e0318890. https://doi.org/10.1371/journal.pone.0318890

Jemine, C. (2019). *Real-time Voice Cloning*. Université de Liège.

Jeon, J. H., Xia, R., & Liu, Y. (2010). Level of interest sensing in spoken dialog using multi-level fusion of acoustic and lexical evidence. *Interspeech 2010*, 2802–2805. https://doi.org/10.21437/interspeech.2010-741

Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I. L., & Wu, Y. (2018). Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. *arXiv:1806.04558 [Cs, Eess]*, *abs/1806.04558*. https://doi.org/10.48550/arXiv.1806.04558

Jiang, L., Hoegg, J., Dahl, D. W., & Chattopadhyay, A. (2010). The persuasive role of incidental similarity on attitudes and purchase intentions in a sales context. *Journal of Consumer Research*, *36*(5), 778–791. https://doi.org/10.1086/605364

Johnson, N. D., & Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of Economic Psychology*, *32*(5), 865–889. https://doi.org/10.1016/j.joep.2011.05.007

Jones, E. E., & Davis, K. E. (1965). From Acts To Dispositions The Attribution Process In Person Perception. In *Advances in Experimental Social Psychology* (Vol. 2, pp. 219–266). Elsevier. https://doi.org/10.1016/S0065-2601(08)60107-0

Jones, J. T., Pelham, B. W., Carvallo, M., & Mirenberg, M. C. (2004). How Do I Love Thee? Let Me Count the Js: Implicit Egotism and Interpersonal Attraction. *Journal of*

Personality and Social Psychology, *87*(5), 665–683. https://doi.org/10.1037/0022-3514.87.5.665

Junger, J., Pauly, K., Bröhr, S., Birkholz, P., Neuschaefer-Rube, C., Kohler, C., Schneider, F., Derntl, B., & Habel, U. (2013). Sex matters: Neural correlates of voice gender perception. *NeuroImage*, *79*, 275–287. https://doi.org/10.1016/j.neuroimage.2013.04.105

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kämmer, J. E., Choshen-Hillel, S., Müller-Trede, J., Black, S. L., & Weibler, J. (2023). A systematic review of empirical studies on advice-based decisions in behavioral and organizational research. *Decision*, *10*(2), 107–137. https://doi.org/10.1037/dec0000199

Kaplan, M. F., & Anderson, N. H. (1973). Information integration theory and reinforcement theory as approaches to interpersonal attraction. *Journal of Personality and Social Psychology*, *28*(3), 301–312. https://doi.org/10.1037/h0035112

Kappen, M., Hoorelbeke, K., Madhu, N., Demuynck, K., & Vanderhasselt, M.-A. (2022). Speech as an indicator for psychosocial stress: A network analytic approach. *Behavior Research Methods*, *54*(2), 910–921. https://doi.org/10.3758/s13428-021-01670-x

Kaup, B., Ulrich, R., Bausenhart, K. M., Bryce, D., Butz, M. V., Dignath, D., Dudschig, C., Franz, V. H., Friedrich, C., Gawrilow, C., Heller, J., Huff, M., Hütter, M., Janczyk, M., Leuthold, H., Mallot, H., Nürk, H.-C., Ramscar, M., Said, N., … Wong, H. Y. (2024). Modal and amodal cognition: An overarching principle in various domains of psychology. *Psychological Research*, *88*(2), 307–337. https://doi.org/10.1007/s00426-023-01878-w

Kaya, H., & Karpov, A. A. (2018). Efficient and effective strategies for cross-corpus acoustic emotion recognition. *Neurocomputing*, *275*, 1028–1034. https://doi.org/10.1016/j.neucom.2017.09.049

Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007). Joint Factor Analysis Versus Eigenchannels in Speaker Recognition. *IEEE Transactions on Audio, Speech and Language Processing*, *15*(4), 1435–1447. https://doi.org/10.1109/TASL.2006.881693

Kersta, L. G. (1962). Voiceprint Identification. *Nature*, *196*(4861), 1253–1257. https://doi.org/10.1038/1961253a0

Kersta, L. G. (1973). Voiceprint Identification. *Police Law Quarterly*, *3*(1), 5–12.

Kim, J., Kim, S., Kong, J., & Yoon, S. (2020). Glow-TTS: A Generative Flow for Text-to-Speech via Monotonic Alignment Search. *arXiv:2005.11129 [Cs, Eess]*. https://doi.org/10.48550/arXiv.2005.11129

Kim, J., Kong, J., & Son, J. (2021). *Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech* (No. arXiv:2106.06103). arXiv. http://arxiv.org/abs/2106.06103

Kingma, D. P., & Dhariwal, P. (2018). *Glow: Generative Flow with Invertible 1x1 Convolutions*. https://doi.org/10.48550/ARXIV.1807.03039

Kohut, H. (1984). *How Does Analysis Cure?* University of Chicago Press. https://doi.org/10.7208/chicago/9780226006147.001.0001

Kong, J., Kim, J., & Bae, J. (2020). *HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis*. https://doi.org/10.48550/ARXIV.2010.05646

Koppen, C., Levitan, C. A., & Spence, C. (2009). A signal detection study of the Colavita visual dominance effect. *Experimental Brain Research*, *196*(3), 353–360. https://doi.org/10.1007/s00221-009-1853-y

Kreiman, J., Park, S. J., Keating, P. A., & Alwan, A. (2015). The relationship between acoustic and perceived intraspeaker variability in voice quality. *Interspeech 2015*, 2357–2360. https://doi.org/10.21437/Interspeech.2015-510

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. https://doi.org/10.1037/0022-3514.77.6.1121

Kumar, N. (1996). The power of trust in manufacturer-retailer relationships. *Harvard Business Review*, *74*(6), 92–106.

Lacassagne, D., Béna, J., & Corneille, O. (2022). Is Earth a perfect square? Repetition increases the perceived truth of highly implausible statements. *Cognition*, *223*, 105052. https://doi.org/10.1016/j.cognition.2022.105052

Langley, M. D., & McBeath, M. K. (2023). Vertical attention bias for tops of objects and bottoms of scenes. *Journal of Experimental Psychology: Human Perception and Performance*, *49*(10), 1281–1295. https://doi.org/10.1037/xhp0001117

Langlois, J. H., & Roggman, L. A. (1990). Attractive Faces Are Only Average. *Psychological Science*, *1*(2), 115–121. https://doi.org/10.1111/j.1467-9280.1990.tb00079.x

Lanska, M., Olds, J., & Westerman, D. (2013). Fluency Effects in Recognition Memory: Are Perceptual Fluency and Conceptual Fluency Interchangeable? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*. https://doi.org/10.1037/a0034309

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review*, *26*(1), 90–102. https://doi.org/10.3758/s13423-018-1497-7

Law, S., Hawkins, S. A., & Craik, F. I. M. (1998). Repetition-Induced Belief in the Elderly: Rehabilitating Age-Related Memory Deficits. *Journal of Consumer Research*, *25*(2), 91–107. https://doi.org/10.1086/209529

Lee, A. Y., & Labroo, A. A. (2004). The Effect of Conceptual and Perceptual Fluency on Brand Evaluation. *Journal of Marketing Research*, *41*(2), 151–165. https://doi.org/10.1509/jmkr.41.2.151.28665

Lee, C.-C., Mower, E., Busso, C., Lee, S., & Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, *53*(9–10), 1162–1171. https://doi.org/10.1016/j.specom.2011.06.004

Lee, R. M., & Robbins, S. B. (1995). Measuring belongingness: The Social Connectedness and the Social Assurance scales. *Journal of Counseling Psychology*, *42*(2), 232–241. https://doi.org/10.1037/0022-0167.42.2.232

Lee, Y., & Kreiman, J. (2019). Within and between speaker variation in voices. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 1460–1464).

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and Decision Making. *Annual Review of Psychology*, *66*(1), 799–823. https://doi.org/10.1146/annurev-psych-010213-115043

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, *46*(6), 1093–1096. https://doi.org/10.1016/j.jesp.2010.05.025

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, *6*(4), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

Lewis, K. N., & Walsh, W. B. (1980). Effects of value-communication style and similarity of values on counselor evaluation. *Journal of Counseling Psychology*, *27*(4), 305–314. https://doi.org/10.1037/0022-0167.27.4.305

Lewontin, R., & Levins, R. (1997). Organism and environment. *Capitalism Nature Socialism*, *8*(2), 95–98. https://doi.org/10.1080/10455759709358737

Li, H., Xu, C., Rathore, A. S., Li, Z., Zhang, H., Song, C., Wang, K., Su, L., Lin, F., Ren, K., & Xu, W. (2020). VocalPrint: Exploring a resilient and secure voice authentication via mmWave biometric interrogation. *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, 312–325. https://doi.org/10.1145/3384419.3430779

Li, M., Han, K., & Narayanan, S. (2012). Automatic Speaker Age and Gender Recognition Using Acoustic and Prosodic Level Information Fusion. *Computer, Speech, and Language*, *27*(1), 151–167. https://doi.org/10.1016/j.csl.2012.01.008

Lichtenthal, J. D., & Tellefsen, T. (2001). Toward a theory of business buyer-seller similarity. *Journal of Personal Selling & Sales Management*, *21*(1), 1–14.

Liu, Z., Wen, J., Liu, Y., & Hu, C.-P. (2024). The effectiveness of self: A meta-analysis of using self-referential encoding techniques in education. *British Journal of Educational Psychology*, *94*(1), 112–137. https://doi.org/10.1111/bjep.12636

Lo, C.-C., Fu, S.-W., Huang, W.-C., Wang, X., Yamagishi, J., Tsao, Y., & Wang, H.-M. (2019). MOSNet: Deep Learning based Objective Assessment for Voice Conversion. *Interspeech 2019*, 1541–1545. https://doi.org/10.21437/Interspeech.2019-2003

Loewenstein, G., Lerner, J. S., & others. (2003). The role of affect in decision making. *Handbook of Affective Science*, *619*(642), 3.

Lubold, N., Walker, E., Pon-Barry, H., & Ogan, A. (2018). Automated Pitch Convergence Improves Learning in a Social, Teachable Robot for Middle School Mathematics.

International Conference on Artificial Intelligence in Education, 282–296. https://doi.org/10.1007/978-3-319-93843-1_21

Lynch, S. (2017). *Andrew Ng: Why AI Is the New Electricity*. Stanford Graduate School of Business. https://www.gsb.stanford.edu/insights/andrew-ng-why-ai-new-electricity

Masuko, T., Tokuda, K., Kobayashi, T., & Imai, S. (1997). Voice characteristics conversion for HMM-based speech synthesis system. *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *3*, 1611–1614. https://doi.org/10.1109/ICASSP.1997.598807

McGlone, M. S., & Tofighbakhsh, J. (2000). Birds of a Feather Flock Conjointly (?): Rhyme as Reason in Aphorisms. *Psychological Science*, *11*(5), 424–428. https://doi.org/10.1111/1467-9280.00282

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*(1), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

Meier, B. P., & Robinson, M. D. (2004). Why the Sunny Side Is Up: Associations Between Affect and Vertical Position. *Psychological Science*, *15*(4), 243–247. https://doi.org/10.1111/j.0956-7976.2004.00659.x

Meinedo, H., & Trancoso, I. (2010). Age and gender classification using fusion of acoustic and prosodic features. *Interspeech 2010*, 2818–2821. https://doi.org/10.21437/Interspeech.2010-745

Meyerhoff, H. S., & Huff, M. (2016). Semantic congruency but not temporal synchrony enhances long-term memory performance for audio-visual scenes. *Memory & Cognition*, *44*(3), 390–402. https://doi.org/10.3758/s13421-015-0575-6

Meyerhoff, H. S., Jaggy, O., Papenmeier, F., & Huff, M. (2023). Long-term memory representations for audio-visual scenes. *Memory & Cognition*, *51*(2), 349–370. https://doi.org/10.3758/s13421-022-01355-6

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, *82*(2), 213–225. https://doi.org/10.1037/h0076486

Miyake, K., & Zuckerman, M. (1993). Beyond Personality Impressions: Effects of Physical and Vocal Attractiveness on False Consensus, Social Comparison, Affiliation, and Assumed and Perceived Similarity. *Journal of Personality*, *61*(3), 411–437. https://doi.org/10.1111/j.1467-6494.1993.tb00287.x

Mohammadi, G., & Vinciarelli, A. (2015). Automatic personality perception: Prediction of trait attribution based on prosodic features. *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 484–490. https://doi.org/10.1109/acii.2015.7344614

Mohd Hanifa, R., Isa, K., & Mohamad, S. (2021). A review on speaker recognition: Technology and challenges. *Computers & Electrical Engineering*, *90*, 107005. https://doi.org/10.1016/j.compeleceng.2021.107005

Montoya, R. M., & Horton, R. S. (2013). A meta-analytic investigation of the processes underlying the similarity-attraction effect. *Journal of Social and Personal Relationships*, *30*(1), 64–94. https://doi.org/10.1177/0265407512452989

Montoya, R. M., Horton, R. S., & Kirchner, J. (2008). Is actual similarity necessary for attraction? A meta-analysis of actual and perceived similarity. *Journal of Social and Personal Relationships*, *25*(6), 889–922. https://doi.org/10.1177/0265407508096700

Moons, W. G., Mackie, D. M., & Garcia-Marques, T. (2009). The impact of repetition-induced familiarity on agreement with weak and strong arguments. *Journal of Personality and Social Psychology*, *96*(1), 32–44. https://doi.org/10.1037/a0013461

Moreland, R. L., & Zajonc, R. B. (1982). Exposure effects in person perception: Familiarity, similarity, and attraction. *Journal of Experimental Social Psychology*, *18*(5), 395–415. https://doi.org/10.1016/0022-1031(82)90062-2

Morewedge, C. K., & Giblin, C. E. (2015). Explanations of the endowment effect: An integrative review. *Trends in Cognitive Sciences*, *19*(6), 339–348. https://doi.org/10.1016/j.tics.2015.04.004

Morry, M. M. (2005). Relationship satisfaction as a predictor of similarity ratings: A test of the attraction-similarity hypothesis. *Journal of Social and Personal Relationships*, *22*(4), 561–584. https://doi.org/10.1177/0265407505054524

Mullen, B., Brown, R., & Smith, C. (1992). Ingroup bias as a function of salience, relevance, and status: An integration. *European Journal of Social Psychology*, *22*(2), 103–122. https://doi.org/10.1002/ejsp.2420220202

Nadarevic, L., Reber, R., Helmecke, A. J., & Köse, D. (2020). Perceived truth of statements and simulated social media postings: An experimental investigation of source credibility, repeated exposure, and presentation format. *Cognitive Research: Principles and Implications*, *5*(1), 56. https://doi.org/10.1186/s41235-020-00251-4

Narayanan, S., Yu, G., Ho, C.-J., & Yin, M. (2023). How does Value Similarity affect Human Reliance in AI-Assisted Ethical Decision Making? *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 49–57. https://doi.org/10.1145/3600211.3604709

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, *7*(3), 171–181. https://doi.org/10.1037/1076-898X.7.3.171

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are Social Actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78.

Nature. (2023). *Nature's 10*. https://www.nature.com/immersive/d41586-023-03919-1/index.html

Neekhara, P., Hussain, S., Dubnov, S., Koushanfar, F., & McAuley, J. (2021). *Expressive Neural Voice Cloning*. https://doi.org/10.48550/ARXIV.2102.00151

Nekvinda, T., & Dušek, O. (2020). *One Model, Many Languages: Meta-learning for Multilingual Text-to-Speech*. https://doi.org/10.48550/ARXIV.2008.00768

Nelson, T. E., & Garst, J. (2005). Values-based Political Messages and Persuasion: Relationships among Speaker, Recipient, and Evoked Values. *Political Psychology*, *26*(4), 489–516. https://doi.org/10.1111/j.1467-9221.2005.00428.x

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior*, *19*(3), 338–368. https://doi.org/10.1016/S0022-5371(80)90266-2

Neuberg, S. L., & Fiske, S. T. (1987). Motivational influences on impression formation: Outcome dependency, accuracy-driven attention, and individuating processes. *Journal of Personality and Social Psychology*, *53*(3), 431–444. https://doi.org/10.1037/0022-3514.53.3.431

Nuttin, J. M. (1985). Narcissism beyond Gestalt and awareness: The name letter effect. *European Journal of Social Psychology*, *15*(3), 353–361. https://doi.org/10.1002/ejsp.2420150309

O'Donnell, R., Chan, J. C. K., Foster, J. L., & Garry, M. (2023). Experimental and Meta-Analytic Evidence That Source Variability of Misinformation Does Not Increase

Eyewitness Suggestibility Independently of Repetition of Misinformation. *Frontiers in Psychology*, *14*, 1201674. https://doi.org/10.3389/fpsyg.2023.1201674

Ohi, A. Q., Mridha, M. F., Hamid, Md. A., & Monowar, M. M. (2021). Deep Speaker Recognition: Process, Progress, and Challenges. *IEEE Access*, *9*, 89619–89643. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3090109

Oord, A. van den, Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A Generative Model for Raw Audio. *arXiv Preprint arXiv:1609.03499*. https://doi.org/10.48550/arXiv.1609.03499

Orpen, C. (1984). Attitude Similarity, Attraction, and Decision-Making in the Employment Interview. *The Journal of Psychology*, *117*(1), 111–120. https://doi.org/10.1080/00223980.1984.9923666

Ozubko, J. D., & Fugelsang, J. (2011). Remembering makes evidence compelling: Retrieval from memory can give rise to the illusion of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 270–276. https://doi.org/10.1037/a0021323

Parks, C. M., & Toth, J. P. (2006). Fluency, Familiarity, Aging, and the Illusion of Truth. *Aging, Neuropsychology, and Cognition*, *13*(2), 225–253. https://doi.org/10.1080/138255890968691

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 534–552. https://doi.org/10.1037/0278-7393.14.3.534

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139173933

Pearson, A. R., & Dovidio, J. F. (2013). Intergroup Fluency: How Processing Experiences Shape Intergroup Cognition and. In J. P. Forgas, O. Vincze, & J. László (Eds.), , *Social cognition and communication* (pp. 101–120). Psychology Press.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Pelham, B. W., & Carvallo, M. (2011). The surprising potency of implicit egotism: A reply to Simonsohn. *Journal of Personality and Social Psychology*, *101*(1), 25–30. https://doi.org/10.1037/a0023526

Pelham, B. W., Carvallo, M., & Jones, J. T. (2005). Implicit Egotism. *Current Directions in Psychological Science*, *14*(2), 106–110. https://doi.org/10.1111/j.0963-7214.2005.00344.x

Pelham, B. W., Mirenberg, M. C., & Jones, J. T. (2002). Why Susie sells seashells by the seashore: Implicit egotism and major life decisions. *Journal of Personality and Social Psychology*, *82*(4), 469–487. https://doi.org/10.1037/0022-3514.82.4.469

Peng, Z., Hu, Z., Wang, X., & Liu, H. (2020). Mechanism underlying the self-enhancement effect of voice attractiveness evaluation: Self-positivity bias and familiarity effect. *Scandinavian Journal of Psychology*, *61*(5), 690–697. https://doi.org/10.1111/sjop.12643

Peng, Z., Wang, Y., Meng, L., Liu, H., & Hu, Z. (2019). One's own and similar voices are more attractive than other voices. *Australian Journal of Psychology*, *71*(3), 212–222. https://doi.org/10.1111/ajpy.12235

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. https://doi.org/10.1016/j.cognition.2018.06.011

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, *88*(2), 185–200. https://doi.org/10.1111/jopy.12476

Perrachione, T. K., Furbeck, K. T., & Thurston, E. J. (2019). Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices. *The Journal of the Acoustical Society of America*, *146*(5), 3384–3399. https://doi.org/10.1121/1.5126697

Perrett, D. I., May, K. A., & Yoshikawa, S. (1994). Facial shape and judgements of female attractiveness. *Nature*, *368*(6468), 239–242. https://doi.org/10.1038/368239a0

Petty, R. E., Rucker, D. D., Gizer, G. Y., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In J. S. Seiter & R. H. Gass (Eds.), *Perspectives on persuasion, social influence and compliance gaining* (pp. 65–89). Allyn & Bacon.

Phua, J. (2016). The effects of similarity, parasocial identification, and source credibility in obesity public service announcements on diet and exercise self-efficacy. *Journal of Health Psychology*, *21*(5), 699–708. https://doi.org/10.1177/1359105314536452

Pillai, R. M., & Fazio, L. K. (2021). The effects of repeating false and misleading information on belief. *WIREs Cognitive Science*, *12*(6), e1573. https://doi.org/10.1002/wcs.1573

Ping, W., Peng, K., Gibiansky, A., Arık, S. O., Kannan, A., Narang, S., Raiman, J., & Miller, J. (2018). *DEEP VOICE 3: 2000-SPEAKER NEURAL TEXT-TO-SPEECH*. 16.

Polage, D. C. (2012). Making up History: False Memories of Fake News Stories. *Europe's Journal of Psychology*, *8*(2), 245–250. https://doi.org/10.5964/ejop.v8i2.456

Pornpitakpan, C. (2004). The Persuasiveness of Source Credibility: A Critical Review of Five Decades' Evidence. *Journal of Applied Social Psychology*, *34*(2), 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x

Pörschmann, C. (2000). Influences of bone conduction and air conduction on the sound of one's own voice. *Acta Acustica United with Acustica*, *86*(6), 1038–1045.

Posner, M., Nissen, M., & Klein, R. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological Review*, *83*, 157–171. https://doi.org/10.1037/0033-295X.83.2.157

Pyszczynski, T., Greenberg, J., Solomon, S., Arndt, J., & Schimel, J. (2004). Why Do People Need Self-Esteem? A Theoretical and Empirical Review. *Psychological Bulletin*, *130*(3), 435–468. https://doi.org/10.1037/0033-2909.130.3.435

R Core Team. (2019). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. https://doi.org/10.1109/5.18626

Reber, R., & Schwarz, N. (1999). Effects of Perceptual Fluency on Judgments of Truth. *Consciousness and Cognition*, *8*(3), 338–342. https://doi.org/10.1006/ccog.1999.0386

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing Fluency and Aesthetic Pleasure: Is Beauty in the Perceiver's Processing Experience? *Personality and Social Psychology Review*, *8*(4), 364–382. https://doi.org/10.1207/s15327957pspr0804_3

Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of Perceptual Fluency on Affective Judgments. *Psychological Science*, *9*(1), 45–48. https://doi.org/10.1111/1467-9280.00008

Rebryk, Y., & Beliaev, S. (2020). ConVoice: Real-Time Zero-Shot Voice Style Transfer with Convolutional Network. *arXiv:2005.07815 [Cs, Eess]*. http://arxiv.org/abs/2005.07815

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, *49*(1), 95–112. https://doi.org/10.1037/0022-3514.49.1.95

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2019). FastSpeech: Fast, Robust and Controllable Text to Speech. *Advances in Neural Information Processing Systems*, *32*. https://doi.org/10.48550/arXiv.1905.09263

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Process*. https://doi.org/10.1006/dspr.1999.0361

Rhodes, G., Yoshikawa, S., Clark, A., Lee, K., McKay, R., & Akamatsu, S. (2001). Attractiveness of Facial Averageness and Symmetry in Non-Western Cultures: In Search of Biologically Based Standards of Beauty. *Perception*, *30*(5), 611–625. https://doi.org/10.1068/p3123

Richardson, H. M. (1940). Community of Values as a Factor in Friendships of College and Adult Women. *The Journal of Social Psychology*, *11*(2), 303–312. https://doi.org/10.1080/00224545.1940.9918751

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. (1977). Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology*, *35*(9), 677–688. https://doi.org/10.1037/0022-3514.35.9.677

Rosenberg, A. E. (1976). Automatic speaker verification: A review. *Proceedings of the IEEE*, *64*(4), 475–487. https://doi.org/10.1109/PROC.1976.10156

Roth, P. L., Thatcher, J. B., Bobko, P., Matthews, K. D., Ellingson, J. E., & Goldberg, C. B. (2020). Political affiliation and employment screening decisions: The role of similarity and identification processes. *Journal of Applied Psychology*, *105*(5), 472–486. https://doi.org/10.1037/apl0000422

RVC-Boss. (2024). *GPT-SoVITS* [Computer software]. https://github.com/RVC-Boss/GPT-SoVITS

Sadoughi, N., Pereira, A., Jain, R., Leite, I., & Lehman, J. F. (2017). Creating Prosodic Synchrony for a Robot Co-player in a Speech-controlled Game for Children. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, 91–99. https://doi.org/10.1145/2909824.3020244

Said, C. P., & Todorov, A. (2011). A Statistical Model of Facial Attractiveness. *Psychological Science*, *22*(9), 1183–1190. https://doi.org/10.1177/0956797611419169

Salehghaffari, H. (2018). *Speaker Verification using Convolutional Neural Networks*. https://doi.org/10.48550/ARXIV.1803.05427

Schlünz, G. I. (2010). *The effects of part–of–speech tagging on text–to–speech synthesis for resource–scarce languages* [Phd]. North-West University.

Schubert, T. W. (2005). Your Highness: Vertical Positions as Perceptual Symbols of Power. *Journal of Personality and Social Psychology*, *89*(1), 1–21. https://doi.org/10.1037/0022-3514.89.1.1

Scott-Phillips, T. C., Laland, K. N., Shuker, D. M., Dickins, T. E., & West, S. A. (2014). THE NICHE CONSTRUCTION PERSPECTIVE: A CRITICAL APPRAISAL: PERSPECTIVE. *Evolution*, *68*(5), 1231–1243. https://doi.org/10.1111/evo.12332

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., & Wu, Y. (2018). Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (*, 4779–4783. https://doi.org/10.1109/ICASSP.2018.8461368

Shepperd, J., Malone, W., & Sweeny, K. (2008). Exploring Causes of the Self-serving Bias. *Social and Personality Psychology Compass*, *2*(2), 895–908. https://doi.org/10.1111/j.1751-9004.2008.00078.x

Shichiri, K., Sawabe, A., Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (2002). Eigenvoices for HMM-based speech synthesis. *7th International Conference on Spoken Language Processing (ICSLP 2002)*, 1269–1272. https://doi.org/10.21437/ICSLP.2002-390

Shu, S. B., & Peck, J. (2011). Psychological ownership and affective reaction: Emotional attachment process variables and the endowment effect. *Journal of Consumer Psychology*, *21*(4), 439–452. https://doi.org/10.1016/j.jcps.2011.01.002

Sidtis, D., & Kreiman, J. (2012). In the Beginning Was the Familiar Voice: Personally Familiar Voices in the Evolutionary and Contemporary Biology of Communication. *Integrative Psychological and Behavioral Science*, *46*(2), 146–159. https://doi.org/10.1007/s12124-011-9177-4

Silva, R. R., Garcia-Marques, T., & Reber, R. (2017). The informative value of type of repetition: Perceptual and conceptual fluency influences on judgments of truth. *Consciousness and Cognition*, *51*, 53–67. https://doi.org/10.1016/j.concog.2017.02.016

Silverman, K. E., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., Pierrehumbert, J. B., & Hirschberg, J. (1992). ToBI: A standard for labeling English prosody. *ICSLP*, *2*, 867–870. https://doi.org/10.21437/ICSLP.1992-260

Silvia, P. J. (2005). Deflecting Reactance: The Role of Similarity in Increasing Compliance and Reducing Resistance. *Basic and Applied Social Psychology*, *27*(3), 277–284. https://doi.org/10.1207/s15324834basp2703_9

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. https://doi.org/10.2307/1884852

Simons, H. W., Berkowitz, N. N., & Moyer, R. J. (1970). Similarity, credibility, and attitude change: A review and a theory. *Psychological Bulletin*, *73*(1), 1–16. https://doi.org/10.1037/h0028429

Simonsohn, U. (2011). Spurious? Name similarity effects (implicit egotism) in marriage, job, and moving decisions. *Journal of Personality and Social Psychology*, *101*(1), 1–24. https://doi.org/10.1037/a0021990

Simpson, J. A. (2007). Psychological Foundations of Trust. *Current Directions in Psychological Science*, *16*(5), 264–268. https://doi.org/10.1111/j.1467-8721.2007.00517.x

Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R. J., Clark, R., & Saurous, R. A. (2018). Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. *arXiv:1803.09047 [Cs, Eess]*. http://arxiv.org/abs/1803.09047

Skuk, V. G., & Schweinberger, S. R. (2013). Gender differences in familiar voice identification. *Hearing Research*, *296*, 131–140. https://doi.org/10.1016/j.heares.2012.11.004

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007a). The affect heuristic. *European Journal of Operational Research*, *177*(3), 1333–1352. https://doi.org/10.1016/j.ejor.2005.04.006

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2007b). The Affect Heuristic. *European Journal of Operational Research*, *177*(3), 1333–1352. https://doi.org/10.1016/j.ejor.2005.04.006

Smith, E. R., & DeCoster, J. (2000). Dual-Process Models in Social and Cognitive Psychology: Conceptual Integration and Links to Underlying Memory Systems.

Personality and Social Psychology Review, *4*(2), 108–131. https://doi.org/10.1207/S15327957PSPR0402_01

Snyder, C. R., & Fromkin, H. L. (1977). Abnormality as a positive characteristic: The development and validation of a scale measuring need for uniqueness. *Journal of Abnormal Psychology*, *86*(5), 518–527. https://doi.org/10.1037/0021-843X.86.5.518

Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. (2017). Deep Neural Network Embeddings for Text-Independent Speaker Verification. *Interspeech 2017*, 999–1003. https://doi.org/10.21437/Interspeech.2017-620

Sofer, C., Dotsch, R., Wigboldus, D. H. J., & Todorov, A. (2015). What Is Typical Is Good: The Influence of Face Typicality on Perceived Trustworthiness. *Psychological Science*, *26*(1), 39–47. https://doi.org/10.1177/0956797614554955

Söllner, A., Bröder, A., Glöckner, A., & Betsch, T. (2014). Single-process versus multiple-strategy models of decision making: Evidence from an information intrusion paradigm. *Acta Psychologica*, *146*, 84–96. https://doi.org/10.1016/j.actpsy.2013.12.007

Speckmann, F., & Unkelbach, C. (2022). Monetary incentives do not reduce the repetition-induced truth effect. *Psychonomic Bulletin & Review*, *29*(3), 1045–1052. https://doi.org/10.3758/s13423-021-02046-0

Squire, L. R., Frascino, J. C., Rivera, C. S., Heyworth, N. C., & He, B. J. (2021). One-trial perceptual learning in the absence of conscious remembering and independent of the medial temporal lobe. *Proceedings of the National Academy of Sciences*, *118*(19), e2104072118. https://doi.org/10.1073/pnas.2104072118

Stalling, R. B. (1970). Personality similarity and evaluative meaning as conditioners of attraction. *Journal of Personality and Social Psychology*, *14*(1), 77–82. https://doi.org/10.1037/h0028623

Stathopoulos, E., Huber, J., & Sussman, J. (2011). Changes in Acoustic Characteristics of the Voice Across the Life Span: Measures From Individuals 4-93 Years of Age. *Journal of Speech, Language, and Hearing Research*, *54*, 1011–1021. https://doi.org/10.1044/1092-4388(2010/10-0036)

Stevenage, S. V. (2018). Drawing a distinction between familiar and unfamiliar voice processing: A review of neuropsychological, clinical and empirical findings. *Neuropsychologia*, *116*, 162–178.

Strack, F., & Deutsch, R. (2004). Reflective and Impulsive Determinants of Social Behavior. *Personality and Social Psychology Review*, *8*(3), 220–247. https://doi.org/10.1207/s15327957pspr0803_1

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin*, *121*(3), 371–394. https://doi.org/10.1037/0033-2909.121.3.371

Sztahó, D., Szaszák, G., & Beke, A. (2021). Deep learning methods in speaker recognition: A review. *Periodica Polytechnica Electrical Engineering and Computer Science*, *65*(4), 310–328. https://doi.org/10.3311/PPee.17024

Taigman, Y., Wolf, L., Polyak, A., & Nachmani, E. (2017). *VoiceLoop: Voice Fitting and Synthesis via a Phonological Loop*. https://doi.org/10.48550/ARXIV.1707.06588

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178. https://doi.org/10.1002/ejsp.2420010202

Tajfel, H., & Turner, J. C. (1986). The social identity theory of intergroup behavior. In S. Worchel & W. G. Austin (Eds.), *Psychology of Intergroup Relations* (pp. 7–24). Nelson-Hall.

Tan, X., Qin, T., Soong, F., & Liu, T.-Y. (2021). *A Survey on Neural Speech Synthesis* (No. arXiv:2106.15561). arXiv. http://arxiv.org/abs/2106.15561

Tauber, S., Dunlosky, J., Rawson, K., Rhodes, M., & Sitzman, D. (2013). General knowledge norms: Updated and expanded from the Nelson and Narens (1980) norms. *Behavior Research Methods*, *45*. https://doi.org/10.3758/s13428-012-0307-9

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193.

Teeny, J. D., Siev, J. J., Briñol, P., & Petty, R. E. (2021). A Review and Conceptual Framework for Understanding Personalized Matching Effects in Persuasion. *Journal of Consumer Psychology*, *31*(2), 382–414. https://doi.org/10.1002/jcpy.1198

Thapar, A., & Westerman, D. (2009). Aging and Fluency-Based Illusions in Recognition Memory. *Psychology and Aging*, *24*, 595–603. https://doi.org/10.1037/a0016575

Tirumala, S. S., & Shahamiri, S. R. (2016). A review on Deep Learning approaches in Speaker Identification. *Proceedings of the 8th International Conference on Signal Processing Systems*, 142–147. https://doi.org/10.1145/3015166.3015210

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis. *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, *3*, 1315–1318. https://doi.org/10.1109/ICASSP.2000.861820

Tsalikis, J., DeShields Jr, O. W., & LaTour, M. S. (1991). The Role of Accent on the Credibility and Effectiveness of the Salesperson. *Journal of Personal Selling & Sales Management*, *11*(1), 31–41. https://doi.org/10.1080/08853134.1991.10753857

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory.* basil Blackwell.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Um, S.-Y., Oh, S., Byun, K., Jang, I., Ahn, C., & Kang, H.-G. (2019). *Emotional speech synthesis with rich and granularized control*. https://doi.org/10.48550/ARXIV.1911.01635

United Nations Human Rights Council. (2021). *Disinformation and freedom of opinion and expression. Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, Irene Khan*. https://documents-dds-ny.un.org/doc/UNDOC/GEN/G21/085/64/PDF/G2108564.pdf

Unkelbach, C. (2007). Reversing the Truth Effect: Learning the Interpretation of Processing Fluency in Judgments of Truth. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*, 219–230. https://doi.org/10.1037/0278-7393.33.1.219

Unkelbach, C., Koch, A., Silva, R. R., & Garcia-Marques, T. (2019). Truth by Repetition: Explanations and Implications. *Current Directions in Psychological Science*, *28*(3), 247–253. https://doi.org/10.1177/0963721419827854

Unkelbach, C., & Rom, S. C. (2017). A referential theory of the repetition-induced truth effect. *Cognition*, *160*, 110–126. https://doi.org/10.1016/j.cognition.2016.12.016

Unkelbach, C., & Stahl, C. (2009). A multinomial modeling approach to dissociate different components of the truth effect. *Consciousness and Cognition*, *18*(1), 22–38. https://doi.org/10.1016/j.concog.2008.09.006

Urberg, K. A., Degirmencioglu, S. M., & Tolson, J. M. (1998). Adolescent Friendship Selection and Termination: The Role of Similarity. *Journal of Social and Personal Relationships*, *15*(5), 703–710. https://doi.org/10.1177/0265407598155008

Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K. C., & Tucker, J. A. (2021). Political Psychology in the Digital (mis)Information age: A Model of News Belief and Sharing. *Social Issues and Policy Review*, *15*(1), 84–113. https://doi.org/10.1111/sipr.12077

Van Puyvelde, M., Neyt, X., McGlone, F., & Pattyn, N. (2018). Voice Stress Analysis: A New Framework for Voice and Effort in Human Performance. *Frontiers in Psychology*, *9*, 1994. https://doi.org/10.3389/fpsyg.2018.01994

Variani, E., Lei, X., McDermott, E., Moreno, I. L., & Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4052–4056. https://doi.org/10.1109/ICASSP.2014.6854363

Venkataramani Johar, G., & Roggeveen, A. L. (2007). Changing False Beliefs from Repeated Advertising: The Role of Claim-Refutation Alignment. *Journal of Consumer Psychology*, *17*(2), 118–127. https://doi.org/10.1016/S1057-7408(07)70018-9

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Vuilleumier, P. (2005). How brains beware: Neural mechanisms of emotional attention. *Trends in Cognitive Sciences*, *9*(12), 585–594. https://doi.org/10.1016/j.tics.2005.10.011

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A.

(2017). Tacotron: Towards End-to-End Speech Synthesis. *arXiv:1703.10135 [Cs]*. https://doi.org/10.48550/arXiv.1703.10135

Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., & Saurous, R. A. (2018). Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. *arXiv:1803.09017 [Cs, Eess]*. http://arxiv.org/abs/1803.09017

Webster, J. J., & Kit, C. (1992). Tokenization as the Initial Phase in NLP. *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*. COLING 1992. https://doi.org/10.3115/992424.992434

Whittlesea, B. W. A. (1993). Illusions of familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(6), 1235–1253. https://doi.org/10.1037/0278-7393.19.6.1235

Wilson, E. J., & Sherrell, D. L. (1993). Source effects in communication and persuasion research: A meta-analysis of effect size. *Journal of the Academy of Marketing Science*, *21*(2), 101–112. https://doi.org/10.1007/BF02894421

Winkielman, P., Halberstadt, J., Fazendeiro, T., & Catty, S. (2006). Prototypes Are Attractive Because They Are Easy on the Mind. *Psychological Science*, *17*(9), 799–806. https://doi.org/10.1111/j.1467-9280.2006.01785.x

Winkielman, P., Huber, D. E., Kavanagh, L., & Schwarz, N. (2012). Fluency of consistency: When thoughts fit nicely and flow smoothly. In B. Gawronski & F. Strack (Eds.), *Cognitive consistency: A fundamental principle in social cognition* (pp. 89–111). The Guilford Press.

Winkielman, P., Schwarz, N., Fazendeiro, T., & Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In J. Musch & K. C. Klauer (Eds.), *The Psychology of Evaluation: Affective Processes in Cognition and Emotion*

(0 ed., pp. 189–217). Lawrence Erlbaum Associates Publisher. https://doi.org/10.4324/9781410606853

Winkle, K., Lemaignan, S., Caleb-Solly, P., Leonards, U., Turton, A., & Bremner, P. (2019). Effective Persuasion Strategies for Socially Assistive Robots. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 277–285. https://doi.org/10.1109/HRI.2019.8673313

Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., King, S., & Renals, S. (2009). Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, *17*(6), 1208–1230. https://doi.org/10.1109/TASL.2009.2016394

Yang, S., Wu, Z., & Xie, L. (2016). On the training of DNN-based average voice model for speech synthesis. *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 1–6. https://doi.org/10.1109/APSIPA.2016.7820818

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. *6th European Conference on Speech Communication and Technology (Eurospeech 1999)*, 2347–2350. https://doi.org/10.21437/Eurospeech.1999-513

You, S., & Robert Jr., L. P. (2018). Human-Robot Similarity and Willingness to Work with a Robotic Co-worker. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 251–260. https://doi.org/10.1145/3171221.3171281

Younan, M., & Martire, K. A. (2021). Likeability and Expert Persuasion: Dislikeability Reduces the Perceived Persuasiveness of Expert Evidence. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.785677

Yumoto, E., Gould, W. J., & Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *The Journal of the Acoustical Society of America*, *71*(6), 1544–1550. https://doi.org/10.1121/1.387808

Yuval-Greenberg, S., & Deouell, L. Y. (2009). The dog's meow: Asymmetrical interaction in cross-modal object recognition. *Experimental Brain Research*, *193*(4), 603–614. https://doi.org/10.1007/s00221-008-1664-6

Zäske, R., Skuk, V. G., & Schweinberger, S. R. (2020). Attractiveness and distinctiveness between speakers' voices in naturalistic speech and their faces are uncorrelated. *Royal Society Open Science*, *7*(12), 201–244. https://doi.org/10.1098/rsos.201244

Zen, H. (2015). Acoustic modeling in statistical parametric speech synthesis—From HMM to LSTM-RNN. *Proc. MLSLP*.

Zen, H., Tokuda, K., & Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, *51*(11), 1039–1064. https://doi.org/10.1016/j.specom.2009.04.004

Zhang, J., Hsee, C. K., & Xiao, Z. (2006). The majority rule in individual decision making. *Organizational Behavior and Human Decision Processes*, *99*(1), 102–111. https://doi.org/10.1016/j.obhdp.2005.06.004

Zhang, Y., Weiss, R. J., Zen, H., Wu, Y., Chen, Z., Skerry-Ryan, R. J., Jia, Y., Rosenberg, A., & Ramabhadran, B. (2019). Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning. *arXiv:1907.04448 [Cs, Eess]*. http://arxiv.org/abs/1907.04448

Zhang, Y.-J., Pan, S., He, L., & Ling, Z.-H. (2018). *Learning latent representations for style control and transfer in end-to-end speech synthesis*. https://doi.org/10.48550/ARXIV.1812.04342

Zhao, Y., von Delft, S., Morgan-Thomas, A., & Buck, T. (2020). The evolution of platform business models: Exploring competitive battles in the world of platforms. *Long Range Planning*, *53*(4), 101892. https://doi.org/10.1016/j.lrp.2019.101892

Zhao, Z., Pan, D., Peng, J., & Gu, R. (2022). *Probing Deep Speaker Embeddings for Speaker-related Tasks* (No. arXiv:2212.07068). arXiv. http://arxiv.org/abs/2212.07068