Florina Zülli*

Sympathische synthetische Stimmen: Herausforderungen und Nutzen für die Mensch-Maschine-Interaktion

Sympathetic synthetic voices: Challenges and benefits for human-machine interaction

https://doi.org/10.1515/zgl-2025-2002

Abstract: Synthetic voices are increasingly integrated into everyday interactions, ranging from navigation systems in our cars and voice assistants in our homes to social robots in elderly care and nursing homes. Despite their prevalence, our understanding of their potential remains incomplete. This study aims to address fundamental questions regarding synthetic voices and their ability to evoke sympathy to enhance human-machine interaction. Specifically, exploring questions like: How can we engender likability for disembodied devices, such as voice assistants, that solely communicate through voice? What are the social challenges that might arise when developing sympathetic *female* voices for voice assistants? And in which domains can likable voices significantly improve human-machine interaction?

This article sheds light on this proverbial black box by examining different fundamental frequencies (F_0) in human and synthetic voices, with the aim of identifying a potential "golden frequency", the ultimate likable frequency. For this goal two perceptions studies were conducted: The first experiment (N=107) is conducted as a pilot study to explore likability in *human* female voices, while the second experiment (N=435) extends the investigation to *synthetic* female voices, on which will be the main focus in this article. The results revealed that female human voices were perceived as most likable at a frequency around 260 Hz, which coincides with what former studies concluded as the most attractive frequency as well. However, this does not hold true for female synthetic voices, where a vertex becomes evident already around 240 Hz. These findings bear interesting implications for the future of human-machine interaction, offering noteworthy applications in various realms of human-computer interaction (HCI) and human-robot interaction (HRI). These applications extend to (personal) assistance systems, e-learning, and educational contexts, with particular significance in healthcare and nursing homes.

***Kontaktperson: Florina Zülli**: Universität Zürich, Deutsches Seminar, Schönberggasse 9, CH-8001 Zürich, E-Mail: florina.zuelli@ds.uzh.ch

1 Vorbemerkung

Als soziale Wesen sind wir empfänglich für Manipulation und Beeinflussung durch andere. Werden wir gegrüßt oder angelächelt, werden die Spiegelneuronen in unserem Gehirn aktiviert und wir erwidern die Gesten instinktiv. Untersuchungen haben gezeigt, dass es interessante Parallelen in der Interaktion von Menschen mit sozialen Robotern (HRI = human-robot interaction) und der zwischenmenschlichen Interaktion gibt: Sogar non-verbale Signale von Robotern, wie Körpersprache, Mimik oder Gestik, beeinflussen uns in ähnlicher Weise, wie es menschliche Signale tun (vgl. Bartneck et al. 2020, 85 ff.; Tanner et al. 2021; Krämer et al. 2013). Das heißt, wenn ein Roboter uns anlächelt oder grüßt, erwidern wir dies ebenfalls. Dabei gilt, dass je ähnlicher der Roboter in Aussehen und Verhalten einem Menschen ist, desto mehr wird er auch wie ein Mensch wahrgenommen und behandelt. Diese vermeintliche Linearität setzt sich jedoch nicht unendlich fort, sondern erreicht irgendwann einen kritischen Punkt, an dem sie jäh unterbrochen und ins Gegenteil verkehrt wird: Der ‚Uncanny-Valley'-Effekt beschreibt ein bekanntes, wenn auch nicht ganz unumstrittenes Phänomen in der Robotik. Der Begriff des ‚unheimlichen Tals' bezeichnet einen plötzlichen Abfall in der Akzeptanzkurve von Robotern und ähnlichen humanoiden Entitäten. Dieser tritt ein, wenn die Ähnlichkeit zwischen Roboter und Mensch beinahe, aber eben *nicht ganz* nivelliert ist. Die minimalen Unterscheidungsmerkmale, die den Roboter dann als solchen ‚entlarven' – wie z. B. eine unpassende Bewegung –, führen dazu, dass sich die bis dahin positive Wahrnehmung abrupt ins Negative verkehrt und beim Betrachter[1] Unbehagen auslöst (vgl. Mori 2012). Doch bis dieser Wendepunkt erreicht wird, gilt die Norm: je ähnlicher, desto besser.

1 In diesem Beitrag werden abwechselnd verschiedene Formen geschlechtergerechter Sprache, einschließlich Partizip, generisches Maskulinum und Genderstern, verwendet.

Vor diesem Hintergrund sind Sympathien für *humanoide* Roboter einfach zu wecken: Aufgrund unserer anthropomorphischen Tendenzen (der Übertragung menschlicher Eigenschaften auf nicht-menschliche Entitäten (vgl. Fink 2012; Stapels/Eyssel 2021, 239) und des Media-Equation-Effekts (die Tendenz, auf technische Geräte wie auf menschliche Interaktionspartner*innen zu reagieren (vgl. Reeves/Nass 1996, 251) reicht es bereits aus, wenn uns der Roboter *anlächelt*, damit wir ihn sympathisch(er) finden (vgl. Johanson et al. 2020). Anders sieht es nun aber aus, wenn wir mit einem *nicht-humanoiden* Gerät interagieren, dem zur Interaktion lediglich eine Stimme zur Verfügung steht, wie dies z. B. bei einem Sprachassistenten der Fall ist. Um hier Sympathien zu wecken, müssen diese zwangsläufig durch *die Stimme* des Sprachassistenten konstituiert werden, da er über keine anderen Interaktionsmöglichkeiten (wie eine ansprechende Mimik (Lächeln) oder einen beweglichen Körper (Grüßen)) verfügt. Doch ist die Stimme allein überhaupt ausreichend, um Sympathien für eine Maschine auszulösen? Der Frage, wie Stimme und Sympathie zusammenhängen und welche Herausforderungen und Chancen sich dabei im Kontext der Mensch-Maschine-Interaktion ergeben, soll im folgenden Artikel nachgegangen werden. Dazu zunächst einige einleitende Worte zu Sympathie und Stimme als solche.

Sympathien entstehen oft schon innerhalb eines kurzen Augenblicks: Basierend auf nur wenigen Informationen bilden wir uns einen ersten Eindruck von unserem Gegenüber und bewerten die Person als sympathisch oder nicht. Ein wichtiger Informationsträger ist dabei die Stimme – selbst ohne zusätzliche visuelle Stimuli vermittelt sie Eigenschaften über die sprechende Person, z. B. ihr Geschlecht oder ob die Person eher jung oder eher alt ist. Einzelne Studien verweisen sogar darauf, dass die Stimme Hinweise auf äußerliche Attribute (wie die Attraktivität) und die Persönlichkeit des Sprechenden geben kann (vgl. Collins/Missing 2003; Stern et al. 2021; Zuckerman/Driver 1989). Dieses Phänomen, lediglich anhand einer Stimme Attribute über den Sprechenden abzuleiten, lässt sich in der Praxis gut bei einem Telefongespräch beobachten: Wenn wir mit einer uns unbekannten Person telefonieren, neigen wir dazu, in unserer Vorstellung ein Bild von dieser Person lediglich aufgrund ihrer Stimme zu entwerfen.[2] Die Stimme ist also mehr als nur ein individuelles Ausdrucksmittel; sie ist auch ein soziales Werkzeug. Als solches beeinflusst sie nicht nur unsere Wahrnehmung von, sondern auch unsere Haltung gegenüber und dadurch unsere Interaktion mit Gesprächspartner*innen. Verfügt unser

2 Dies gilt dabei nicht nur für Menschen, sondern auch für Roboter: McGinn/Torre (2019) erforschten, was für einen Roboter sich Menschen beim Hören einer Stimme vorstellen, indem sie Teilnehmer verschiedene Stimmen hören und anschließend die dazu passenden Roboter auswählen ließen. Dabei stellten sie fest: „[…] the sound of a voice alone is enough to make us form a mental image of how that speaker – in our case, a robot – should look like“ (McGinn/Torre 2019, 218).

Gegenüber z. B. über ein hohes Maß an vokaler Attraktivität, verhalten wir uns ihm gegenüber kooperativer:„[T]he higher the ratings of voice attractiveness, the more the speaker is judged to be similar to the rater and the more the rater would like to affiliate with the speaker“, schreiben Hughes et al. (2004, 302) und beziehen sich dabei auf die Studie von Miyake/Zuckerman (1993). In einer anderen Studie von Shang/Liu (2022), in welcher der Einfluss der stimmlichen Attraktivität auf kooperatives Verhalten in einem Vertrauensspiel untersucht wurde, waren Teilnehmer eher dazu bereit, mit Personen zu kooperieren, welche über attraktive Stimmen verfügten, was Beweise für einen „beauty premium“ (Shang/Liu 2022, 2), umgangssprachlich auch als ‚pretty privilege‘ bekannt, (dt. ‚Schönheitsbonus‘) liefert (vgl. Shang/Liu 2022, 10).

Natürlich spielen persönliche Vorlieben sowie subjektive Erfahrungen ebenso eine Rolle dabei, welche Stimmen wir mögen und welche nicht. Dennoch unterliegt die vokale Attraktivität, ähnlich wie die visuelle Attraktivität, keiner rein subjektiven Wahrnehmung, sondern ist bis zu einem gewissen Grad empirisch messbar. So lassen sich in der Stimmforschung in Bezug auf menschliche Stimmen bestimmte universelle Präferenzen feststellen; ähnlich wie auch symmetrische Gesichter im Allgemeinen als schöner empfunden werden als asymmetrische. Wie sich Attribute wie Attraktivität oder eben Sympathie von Stimmen experimentell eruieren lassen und auf was dabei geachtet werden soll, wird zu einem späteren Zeitpunkt in diesem Artikel noch ausführlicher diskutiert.

Doch zuerst wird in Bezug auf diese universellen Präferenzen im nächsten Kapitel ein kurzer Überblick zur vokalen Attraktivität von menschlichen Stimmen präsentiert. Ob und wie diese universellen Präferenzen beim Voice-Design für Sprachassistenten wie *Alexa* oder *Siri* Anwendung finden, wird im dritten Kapitel des Beitrags diskutiert. Im vierten Kapitel werden dann zwei empirische Studien präsentiert, in welchen Stimmen nun hinsichtlich ihrer *Sympathie* bewertet werden – in der ersten Untersuchung (N=107) geht es dabei um menschliche, in der zweiten (N=435) um synthetische Stimmen. Die erste Studie verfügt über einen eher explorativen Charakter mit kleinerer Stichprobe und soll hauptsächlich dazu dienen, erste Tendenzen zur Sympathiewahrnehmung bei menschlichen Stimmen sichtbar zu machen. Das Hauptaugenmerk dieses Aufsatzes liegt auf der zweiten, größer angelegten Studie, in der eruiert werden soll, welche synthetischen Stimmen als sympathisch wahrgenommen werden und welche nicht. Anhand der Ergebnisse beider Studien soll einerseits aufgezeigt werden, dass es eindeutig Stimmen gibt, die sympathischer wahrgenommen werden als andere, und andererseits, dass diesbezüglich ein Unterschied zwischen natürlichen und künstlichen Stimmen besteht. Ein weiterführendes Ziel der zweiten Studie ist es, den Fokus darauf zu lenken, wie Sympathie evozierende Stimmen für die Mensch-

Maschine-Interaktion künftig konstruiert werden und in welchen Bereichen (z. B. der Alten- und Krankenpflege oder im E-Learning-Bereich) sie mit welchen Effekten zur Anwendung kommen könnten, was im letzten Kapitel diskutiert wird. Gesamthaft möchte dieser Aufsatz sowohl die Herausforderungen als auch das Potenzial von synthetischen Stimmen im Kontext der Mensch-Maschine-Interaktion beleuchten.

2 Vokale Attraktivität bei Menschen

Die Stimme ist ein inhärenter Bestandteil der Identität einer Person, so individuell wie ein Fingerabdruck. Aus diesem ‚vocal print' (dt. akustischer Fingerabdruck) lassen sich einerseits statische biologische und biografische Eigenschaften – wie das Geschlecht oder die regionale Herkunft – entnehmen, andererseits werden durch sie aber auch situative, soziale Informationen, wie z. B. die akute Befindlichkeit des Sprechenden wiedergegeben (vgl. Graddol/Swann 1989, 13). So hört man es der Stimme des Sprechenden zumeist an, ob die Person gerade traurig, wütend oder fröhlich ist (vgl. Owren/Bachorowski 2007, 240).[3] Eine wichtige Rolle spielt in diesem Zusammenhang die Grundfrequenz: „Previous studies have confirmed F_0 as a reliable indicator of human emotions like fear, happiness, anger, sorrow, fatigue etc." (Sondhi et al. 2015, 42). Es besteht also ein enger Zusammenhang zwischen Stimme und Person.[4] Wie stark Stimmen die Wahrnehmung von Personen beeinflussen, zeigt der sogenannte ‚Halo-Effekt': Dabei handelt es sich um ein kognitives Phänomen der Wahrnehmungspsychologie, bei dem von besonders positiv oder negativ empfundenen *bekannten* Merkmalen (wie der Stimme) auf weitere, noch *unbekannte* Merkmale (wie Charaktereigenschaften) geschlossen wird (vgl. Spektrum 2024). Demnach wird eine Person, die über eine attraktive Stimme verfügt, mit weiteren positiven Persönlichkeitsmerkmalen assoziiert (vgl. Kiese-Himmel 2016, 32; Lange et al. 2017). Der sogenannte ‚vocal attractiveness stereotype' – ‚what *sounds* good, *is* good' – zeigt, dass Hörer*innen bei einer attraktiven Stimme intuitiv

3 Allerdings können sich hier auch Trugschlüsse einstellen: Eine hyponasale Stimme (‚Näseln') kann ebenso als Indiz für Traurigkeit (Emotion, psychisch) gehalten werden, wie auch auf eine Erkältung (Krankheit, physisch) des Sprechers hindeuten. So können erkältete Personen weinerlich klingen, obwohl sie es gar nicht sind, da die Stimme (wahrgenommene) Emotionen ‚miterzeugen' kann.

4 Die Affinität von *Person* und *Stimme* zeigt sich auch in der Etymologie: So leitet sich das Lexem ‚Person' (lat. persona) aus dem Lateinischen ‚per sona' (dt. durch Klang, durch Ton) bzw. ‚personare' (ertönen, erschallen) ab (vgl. Hermann 1983, 370; Kiese-Himmel 2016, 31).

auf weitere, positive Eigenschaften schließen.[5] So werden Personen mit attraktiven Stimmen z. B. in einem Call-Center-Setting, in welchem der Kontakt nur qua Stimme erfolgt, als kompetenter wahrgenommen als Personen mit vergleichsweise unattraktiven Stimmen (vgl. Bartsch 2008, 59 ff.).[6]

Um eine Stimme als attraktiv oder unattraktiv zu bewerten, reicht bereits ein kurzes akustisches Signal von ~390 Millisekunden aus (vgl. McAleer et al. 2014). Ein einfaches ‚Hallo!' genügt demnach, um sich ein Urteil über eine Stimme (und den Sprechenden) zu bilden.[7] Die Attraktivität einer Stimme wird dabei vor allem durch die ‚Grundfrequenz' (eng. fundamental frequency = F_0)[8] festgelegt: Diese variiert im Gegensatz zu Sprachrhythmus, Sprechgeschwindigkeit und Lautstärke – drei Faktoren, die ebenfalls Einfluss auf die vokale Attraktivität haben, – in der Regel weniger stark. Die Grundfrequenz ist somit eines der wichtigsten Kriterien bei der Beurteilung, ob wir eine Stimme als attraktiv wahrnehmen oder nicht: „Numerous empirical studies have investigated how artificial manipulation of vocal acoustic parameters affects perceptions, especially how the alteration of fundamental frequency [...] can shape assessments of speaker attractiveness and formidability" (Hughes/Puts 2021, 1). Dies macht sie folglich zu einem prominenten Untersuchungsgegenstand in der Forschung zur vokalen Attraktivität (vgl. Zheng et al. 2020; Jones et al. 2010). Der zuvor angesprochene Nexus zwischen Person und Stimme tritt hier erneut zu Tage: Denn Stimmen divergieren in ihren Grundfrequenzen je nach

5 Der hier angesprochene ‚vocal attractiveness stereotype' stammt aus einem Artikel von Zuckerman/Driver (1989), wo er im Original „What sounds beautiful, is good" lautet.

6 Eine interessante Erkenntnis aus derselben Studie war ausserdem, dass „customers who were served by a female call center employee were more satisfied than those served by a male call center employee" (Bartsch 2008, 61). Dies geht kongruent mit der Annahme des vorliegenden Artikels, dass Frauenstimmen sympathischer wahrgenommen werden als Männerstimmen (siehe Kapitel 3).

7 In der Psychologie ist dieses Konzept als ‚thin slicing' (dt. in dünne Scheiben schneiden) bekannt und beschreibt die Fähigkeit von Personen, schnelle Einschätzungen aufgrund limitierter Informationen zu treffen (vgl. Carney et al. 2007, 1055). McAleer et al. (2014) zeigten mit ihrer Untersuchung, dass dieses thin-slicing-Phänomen nicht nur bei *visuellen*, sondern auch *akustischen* Ersteindrücken zutrifft (vgl. McAleer et al. 2014, 5 f.).

8 Ein Problem bei englischsprachigen Artikeln ist die teils undifferenzierte Verwendung der Begriffe ‚pitch' und ‚fundamental frequency'. Während *pitch* für die *wahrgenommene* Tonhöhe steht und sich auf die Perzeption bezieht, referenziert *fundamental frequency* auf die tatsächliche physikalische Frequenz eines Tons. Das bedeutet, dass die Grundfrequenz (F_0) objektiv messbar ist, während *pitch* auf der logarithmischen Verarbeitung durch das menschliche Gehör basiert. So nimmt das Gehör z. B. Frequenzunterschiede als musikalische Intervalle (wie Halbtöne) wahr, die auf festen Frequenzverhältnissen basieren, und nicht auf festen, absoluten Frequenzunterschieden. Deshalb bezeichnen *pitch* und *fundamental frequency* zwar beide die Tonhöhe, sind jedoch nicht synonym zu verstehen.

Geschlecht, Kultur und sogar gesprochener Sprache stark voneinander.[9] Aufgrund dieser kulturellen und sprachlichen Unterschiede lassen sich für Mann und Frau nur sehr schwer Spektren festlegen, auf denen ihre jeweils ‚typischen' Stimmhöhen rangieren. Christiane Kiese-Himmel, Fachärztin für Phoniatrie und Pädaudiologie, gibt als *durchschnittliche Spektren* an, dass Männerstimmen i. d. R. zwischen 85–180 Hz und Frauenstimmen zwischen 165–255 Hz liegen (vgl. Kiese-Himmel 2016, 39).[10] Bezüglich der *durchschnittlichen Grundfrequenz* für Männer und Frauen finden wir bei Ujvary et al. (2022), die sich ihrerseits auf Studien von Kovačić/Balaban (2009), Traunmüller/Eriksson (1995) und Fouquet et al. (2016) beziehen, folgende Werte: „The average healthy fundamental frequency for men is reported at 115 Hz, while for women the average is reported as being around 210 Hz" (Ujvary et al. 2022, 7). Von diesen Zahlen ausgehend könnte gesagt werden, dass Männer- bzw. Frauenstimmen, die darunter oder darüber angesiedelt sind, als ‚überdurchschnittlich' tief bzw. hoch bezeichnet werden können. Allerdings ist diese Aussage ebenso mit Vorsicht handzuhaben und im jeweiligen (kulturellen, geografischen, sprachlichen) Kontext zu betrachten wie die durchschnittlichen Spektren zuvor. Denn selbst die durchschnittliche Grundfrequenz derselben Person kann variieren, z. B. in Korrelation mit anderen Faktoren, wie der Lautstärke (vgl. Jessen et al. 2003).[11]

Trotz dieser zu beachtenden Einflussfaktoren wird aus dem Forschungsstand zu stimmlicher Attraktivität ersichtlich, dass eine attraktive Grundfrequenz meist stereotypisch für ihr jeweiliges Geschlecht ist. So gelten z. B. bei Männern besonders tiefe Stimmen als attraktiv (vgl. Feinberg et al. 2005; Collins 2000). Dies lässt sich einerseits auf biologische, aber andererseits auch auf soziologische Weise begründen. Einer biologischen Argumentation folgend könnte dies damit erklärt werden,

9 Nimmt man noch die diachrone Ebene dazu, lassen sich weitere Unterschiede feststellen – so sprechen z. B. deutsche Frauen heute im Durchschnitt deutlich tiefer als noch vor 25 Jahren (vgl. Berg et al. 2017). Forscher*innen vermuten hinter diesem ‚Stimmbruch' allerdings weniger biologische, sondern eher soziale Gründe, etwa dass Frauen sich eine tiefere Stimmlage angeeignet haben, um in der Berufswelt ernster genommen zu werden. Ein bekanntes Beispiel einer Frau, die sich aufgrund ihres beruflichen Status eine tiefere Stimmlage antrainiert haben soll, ist die ehemalige UK-Premierministerin Margaret Thatcher (vgl. Kiese-Himmel 2016, 50).

10 Allerdings nimmt die Autorin keine geografische Verortung für diese Werte vor. Dies ist jedoch wichtig, da die Grundfrequenz stark von der gesprochenen Sprache abhängt (vgl. Cussigh et al. 2020, 1). Die sprachlichen Unterschiede tragen somit dazu bei, auch innerhalb eines Geschlechts sehr unterschiedliche F_0-Werte zu erhalten, siehe dazu z. B. für die ‚durchschnittlichen Spektren' bei Owren und Bachorowski (2007) mit 75–150 Hz (Männer) und 150–300 Hz (Frauen) oder bei Danielescu (2020) mit 85–180 Hz (Männer) und 140–255 Hz (Frauen).

11 Jessen et al. (2003) untersuchten bei 100 männlichen Deutschen, wie sich die Grundfrequenz zusammen mit der Lautstärke verändert (‚normal vs. loud speech') und stellten fest, dass die F_0-Werte der Sprecher bei erhöhter Lautstärke ebenfalls um ca. 20–30 Hz ansteigen (vgl. Jessen et al. 2003, 1624).

dass tiefe Männerstimmen möglicherweise auf einen größeren Resonanzkörper hindeuten[12] und Männer mit tieferen Stimmen einen höheren Testosteronspiegel aufweisen (vgl. Puts et al. 2012; Dabbs/Mallinger 1999). Dadurch könnten tiefe Männerstimmen auf unbewusster Ebene mit evolutionären Vorteilen assoziiert werden. Dieses biologische Argument wird durch Untersuchungen wie die von Feinberg et al. (2006) gestützt, die belegen, dass sich Frauen in der fruchtbaren Phase ihres Zyklus besonders stark zu Männern mit tiefen Stimmen hingezogen fühlen. Diese biologischen Attribute (Größe, Testosteronlevel) werden nun soziologisch mit weiteren, genuin männlich-attraktiven Merkmalen wie Potenz, Autorität und Dominanz in Verbindung gebracht.[13] Männerstimmen mit einer tieferen Grundfrequenz werden demnach bevorzugt, da sie mit Sicherheit und Kompetenz assoziiert werden – Eigenschaften, die nebst dem privaten Bereich auch in der Geschäftswelt und der Politik mit Vorteilen einhergehen (vgl. Sorokowski et al. 2019, 258).[14]

12 Ob dem tatsächlich so ist, ist umstritten. Collins (2000) etwa spricht sich gegen die Hypothese aus, dass größere Männer tiefere Stimmen besitzen, und betont, dass in ihrer Untersuchung „no correlation between male vocal and body characteristics" ausgemacht werden konnte (Collins 2000, 777). Trotzdem assoziierten ihre Probandinnen tiefere Stimmen mit genuin männlichen Attributen (Größe, Gewicht, Behaarung etc.): „Although there was no relationship between body and vocal characteristics, women used vocal characteristics to infer physical characters" (Collins 2000, 778). Deshalb sollten Argumentationen, die rein auf biologische Werte wie Körpergröße referieren, kritisch betrachtet werden, da sich auch in ländervergleichenden Studien gezeigt hat, dass z. B. polnische Männer und amerikanische Männer sehr unterschiedlich tiefe Stimmen haben können – trotz gleicher Körpergröße (vgl. Graddol/Swann 1989, 24).

13 Dies bezeugt auch die Untersuchung von Klofstad et al. (2015), nach welcher sich eine tiefe Stimme in der Politik als nützliches Instrument für männliche Kandidaten herausstellt: Wähler*innen bevorzugen überwiegend Kandidaten, die zwischen 40 und 50 Jahren alt sind, was dem Zeitraum entspricht, in dem die Stimme eines Individuums am tiefsten ist. Die Stimmtiefe wurde in dieser Studie ebenfalls mit Stärke, körperlicher Überlegenheit, Kompetenz und Integrität gleichgesetzt. Angemerkt sei hier auch, dass nach Browning (2008) der tiefe Bariton des ehemaligen US-Präsidenten Barack Obama als die ‚erfolgreichste, männliche Stimmtonlage' überhaupt gilt. Allerdings gilt die Zuschreibung dieser Dominanz- und Kompetenz-Attribute nicht nur für tiefe Männerstimmen, sondern auch für tiefe Frauenstimmen: „[B]oth men and women perceive lower-pitched female voices to be more competent, stronger and more trustworthy, attributes that are probably correlated with perceptions of leadership capacity" (Klofstad et al. 2012, 2702). Auch Jones et al. (2010) stellten fest, dass tiefere Frauenstimmen als dominanter wahrgenommen werden als höhere.

14 Dass tiefe Männerstimmen als Zeichen von Autorität wahrgenommen werden und traditionell mit Führungsqualitäten in Verbindung gebracht werden, basiert wiederum auf den patriarchalen Strukturen der Gesellschaft. Oder anders gesagt: „If men talked in higher pitches than women, *low voices* would be said to lack in authority" (Cameron 1986, 54) [Hervorhebung durch die Autorin]. Wie stark soziale Kontexte die ‚Stimmwahl' beeinflussen können, zeigen Untersuchungen, in welchen Frauen z. B. ihre Stimme in Situationen vertiefen, in denen sie Autorität und Kompetenz vermitteln wollen (vgl. Sorokowski et al. 2019) und sie wiederum erhöhen, wenn sie mit einer Person sprechen, die sie attraktiv finden (vgl. Fraccaro et al. 2011). Dies spricht deutlich dafür, dass auch soziologische Faktoren eine große Rolle spielen.

Was die Präferenz von Frauenstimmen betrifft, so herrscht in der Forschung hingegen weitaus größerer Dissens: Überwiegend gilt jedoch dass – analog zu den Männerstimmen – Frauenstimmen, die genderspezifische Stereotypie aufweisen, bevorzugt werden, ergo: je höher, desto besser (vgl. Zheng et al. 2020; Re et al. 2012; Jones et al. 2010; Jones et al. 2008; Apicella/Feinberg 2009; Feinberg et al. 2008; Collins/Missing 2003). Einige, wenn auch deutlich wenigere, Studien stehen dem jedoch diametral gegenüber: Sie sprechen sich mit ihren Ergebnissen genau für das Gegenteil und somit für die Attraktivität besonders tiefer Frauenstimmen aus (siehe z. B. Leaderbrand et al. 2008). Möglicherweise kann diese Ambivalenz soziophonetisch damit erklärt werden, dass beide Varianten bekannte ‚vokale Klischees' (piepsige Kleinmädchenstimme vs. rauchige *femme fatale*) bedienen, die mit Attraktivität in Verbindung gebracht werden, und die Resultate deshalb – je nach Zusammensetzung und Präferenz der Probandengruppe – variieren.[15] Denn obwohl die Grundfrequenz eine sehr zentrale Rolle in der Attraktivitätsbewertung einnimmt – siehe dafür z. B. Jones et al. (2008, 192): „[I]ncreasing women's voice pitch alone is sufficient to increase vocal attractiveness" –, spielt auch die *Erwartungshaltung* eine große Rolle bei der Bewertung von vokaler Attraktivität: So erstellen wir nicht nur aufgrund der Stimme Vermutungen darüber, wie eine Person aussieht etc. (Telefon-Beispiel), sondern hegen auch aufgrund externer Merkmale (Aussehen, sozialer Status etc.) Erwartungen an die Stimme einer Person.[16] Passt die Stimme nun zur Sprecherin und erfüllt meine perzeptiven Erwartungen, empfinde ich sie folglich als attraktiver, als wenn meine Erwartungen enttäuscht werden. Dass solche auditiven Erwartungshaltungen dabei nicht nur bei menschlichen, sondern auch bei maschinellen Gegenübern, wie Roboter und Sprachassistenten, bestehen (vgl. McGinn/Torre 2019; Cambre/Kulkarni 2019), wird im nächsten Kapitel noch ausführlicher diskutiert.

Um den Bogen zurück zu der Dichotomie in der Attraktivitätsbewertung von Frauenstimmen zu schlagen, möchte ich an dieser Stelle auf die Frage eingehen, weshalb die Antworten bei einem Forschungsgegenstand, der sich sowohl pho-

15 So führten Leaderbrand et al. (2008) als einen möglichen Grund für die Präferenz tiefer Frauenstimmen in ihrer Untersuchung die spezifische Zusammensetzung ihrer Probandengruppe an: „Some explanations for this pattern are that perhaps the demographic of college aged males prefers more mature women" (Leaderbrand et al. 2008, 6).

16 Mavica und Barenholtz (2012) zeigen mit ihrer Untersuchung, dass Probanden mehrheitlich in der Lage waren, Stimmen den richtigen Personen zuzuordnen, und dies nur anhand von Audioaufnahmen und Fotos der Personen. Das zeigt zum einen, wie zuverlässig unsere Präsumtionen sind (siehe *thin-slicing*), und führt zum anderen vor Augen, weshalb wir – berechtigterweise – so starke Erwartungshaltungen vom Aussehen an die Stimme und vice versa hegen.

netisch messen als auch empirisch überprüfen lässt, so unterschiedlich ausfallen können. An dieser Stelle sei auf fünf Ursachen hingewiesen, die zu dieser Inkongruenz innerhalb dieses spezifischen Forschungsfeldes geführt haben könnten:

(1) **Terminologische Inkonsistenz:** Nicht immer werden in Studien die genauen Messwerte der untersuchten Stimmen genannt (siehe z. B. Leaderbrand 2008; Völkert 2012), was zu einem ambigen Gebrauch der Terminologie führen kann. Somit ist dort jeweils unklar, auf welchen Frequenzbereich referiert wird, wenn Aussagen über ‚low voices' oder ‚high voices' getroffen werden. Diese Ungenauigkeiten können dann zu Inkongruenzen innerhalb des Forschungsstandes führen, da ‚low' und ‚high' unterschiedliche Interpretationen zulassen. Ebenso wird der Terminus ‚vocal attraction' häufig nicht weiter definiert, und es ist dadurch unklar, ob mit ‚attraction' nun erotische Anziehung, Attraktivität oder auch Sympathie gemeint ist (siehe auch Punkt 4).

(2) **Kulturelle Verortung und sprachlich bedingte Unterschiede:** Wie bereits erwähnt, hat die Kultur und besonders die untersuchte Sprache einen großen Einfluss auf die Grundfrequenz und die damit einhergehende Attraktivitätsbewertung. So führen Cussigh et al. (2020) in ihrem Literaturreview die *mean* F_0 (durchschnittliche Grundfrequenz) von Sprecherinnen verschiedener Sprachen an, um zu zeigen, wie stark diese je nach Sprache divergieren können: Von F_0= 182.3 Hz für Sprecherinnen des Englischen über F_0= 200 Hz für Sprecherinnen des Spanischen bis hin zu F_0= 297.4 Hz für Sprecherinnen des Chinesischen (vgl. Cussigh et al. 2020, 1 nach Natour/Wingate 2009 und González et al. 2002).[17] Dies ist ein Umstand, der in der Diskussion nicht vernachlässigt werden darf. Somit sind Feststellungen bezüglich ‚attraktiven Frequenzbereichen' immer nur im Kontext der untersuchten Sprachkultur zu verstehen und allgemeingültige Aussagen sollten vermieden werden.

17 Wobei sich in der Studie von Natour/Wingate (2009) für Sprecherinnen des Chinesischen nebst den von Cussigh et al. (2020) übernommenen Werten (F_0=297.42 ± 35.89 Hz) auch noch tiefere finden lassen (F_0= 266.73 ± 48.32 Hz) (vgl. Natour/Wingate 2009, 561). Dennoch kann diesen Zahlen eindeutig entnommen werden, dass Sprecherinnen des Chinesischen mit einer höheren durchschnittlichen Grundfrequenz sprechen als die anderen Vergleichsgruppen.

(3) **Methodologische Unterschiede:** Weiter variieren auf dem Gebiet der Attraktivitätsforschung bei Stimmen die Untersuchungsmethoden (z. B. Pupillenmessung vs. Fragebogen) zur Feststellung der Attraktivität sowie das untersuchte Audiomaterial (z. B. einzelne Vokale vs. ganze Sätze), was einen direkten Vergleich erschwert und zu unterschiedlichen Ergebnissen führen kann.

(4) **Komplexität des Forschungsgegenstandes:** Attraktivität, auch vokale, ist zwar empirisch messbar (ebenso wie die Symmetrie eines Gesichtes messbar ist), bleibt aber dennoch bis zu einem gewissen Grad subjektiv.[18] Ebenso problematisch gestaltet sich eine strikte Trennung von *Erotik*, *Attraktivität* und *Sympathie* in solchen Wahrnehmungsexperimenten, da diese Attribute häufig korrelieren. Deshalb ist es eminent, den zu untersuchenden Aspekt (z. B. ‚Sympathie') trennscharf zu definieren und die Proband*innen darauf hinzuweisen, präzise zwischen diesen verschiedenen Attributen zu differenzieren. Dieser Punkt wird in Kapitel 4 nochmals kurz aufgegriffen und es wird gezeigt, wie hier versucht wurde, diese klare Trennung zu erreichen.

(5) **Limitationen im Messbereich:** Einige Studien weisen Einschränkungen in Bezug auf den gemessenen Frequenzbereich auf, indem sie nur Frauenstimmen, die auf dem durchschnittlichen Spektrum liegen, bewerten ließen. Ausgehend davon könnte angenommen werden, dass zwischen Grundfrequenz und Attraktivität eine lineare Korrelation besteht, nach der zuvor erwähnten Devise: je höher, desto besser. Diese Proportionalität ist aber nur bis zu einem gewissen Grad gegeben (vgl. Borkowska/Pawlowski 2011).

Auf den letzten Punkt sei etwas genauer eingegangen: Borkowska und Pawlowski (2011) stellten bei einem Wahrnehmungsexperiment erstmals fest, dass ab einem bestimmten Frequenzbereich die Wahrnehmung von sehr hohen Frauenstimmen ins Negative kippt und sie als ‚schrill' und ‚unangenehm' wahrgenommen werden. Ein Grund für diese negative Wahrnehmung sehr hoher Stimmen könnte sein, dass diese mit negativen Emotionen wie Angst oder Panik in Zusammenhang gebracht werden: „[T]his study indicated significant increase in mean pitch […] under

18 Ganz im Sinne des alten Sprichwortes: *Schönheit liegt im Auge des Betrachters* – oder eben: im Ohr des Hörers.

anxiety/embarrassment“ (Sondhi et al. 2015, 42). Ein hoher Schrei wird von weitem gehört und kann frühzeitig vor Gefahr warnen, weshalb sehr hohe Stimmen als Indiz für drohendes Unheil interpretiert werden können. Ein anderer Grund könnte hingegen auch die Assoziationen von sehr hohen Stimmen mit Babys oder Kleinkindern sein (vgl. Borkowska/Pawlowski 2011, 58 f.). Beides sind jedenfalls Faktoren, die sich negativ auf die Attraktivitätsbeurteilung auswirken.

Dass es also einen solchen Scheitelpunkt gibt, blieb in einer Vielzahl von Studien unbemerkt, weil nur Stimmen, die im durchschnittlichen Frequenzbereich lagen, evaluiert wurden, d. h. da, wo ein positiver, linearer Zusammenhang noch besteht. Borkowska und Pawlowski (2011) hingegen untersuchten die Attraktivität von weiblichen Stimmen in ihrer Studie in einem sehr breiten Frequenzbereich (von 184.6 Hz bis 310.2 Hz), wodurch dann dieser Scheitelpunkt bei 260 Hz[19] festgestellt werden konnte.

Folgende beiden Abbildungen zeigen auf, wie sich die Forschungslandschaft durch die Untersuchung von Borkowska/Pawlowski (2011) verändert hat:

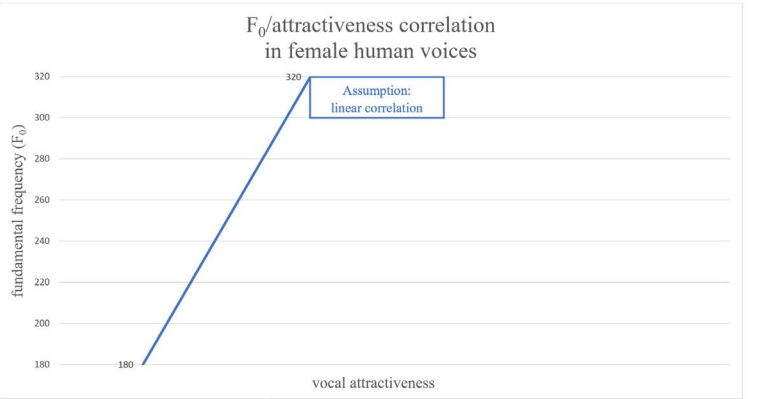


Abbildung 1: Vermutete Korrelation F_0/vokale Attraktivität

19 Dieser Scheitelpunkt bei 260 Hz (genau: 262 Hz in der Studie) ist erneut im Kontext der Sprache/Kultur zu verstehen: Borkowska/Pawlowski führten ihre Untersuchungen mit Polinnen durch. Zheng et al. (2020) sprechen sich dafür aus, dass ein solcher Scheitelpunkt im asiatischen Raum, aufgrund der allgemein höheren F_0, folglich ebenfalls höher angesiedelt wäre: „Similarly, Asian females' average voice pitch is generally higher than that of European females […] and American females (see Zimman, 2011). This could also affect liking thresholds such that a Chinese female voice with a high vocal pitch might be preferred by Chinese speakers even if it is higher than 261.9 Hz“ (Zheng et al. 2020, 174).

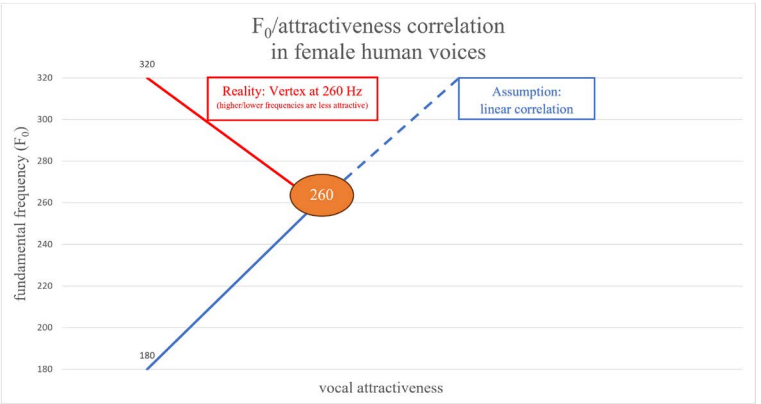


Abbildung 2: Effektiver Zusammenhang F_0/vokaler Attraktivität

In Abb. 1 besteht noch eine lineare Korrelation zwischen Grundfrequenz und Attraktivität. Der Scheitelpunkt bei 260 Hz (Abb. 2) zeigt nun aber, dass nicht nur Stimmen, die darunter liegen, sondern auch Stimmen, *die darüber liegen*, als weniger attraktiv wahrgenommen werden. Ergo: je weiter sich von dieser Idealfrequenz entfernt wird – ungeachtet dessen, ob die Stimme dabei höher oder tiefer wird –, desto weniger attraktiv wird die Stimme empfunden.

Hier möchte die nachfolgende Untersuchung anschließen. Aufgrund der vielen Parallelen zwischen der Wahrnehmung von menschlichen und synthetischen Stimmen ist davon auszugehen, dass sich in einer Attraktivitätsstudie bei synthetischen Stimmen ein Scheitelpunkt in einem ähnlichen Frequenzbereich zeigen würde. Doch noch gänzlich unbekannt ist, wie es sich bezüglich der *Sympathie* verhält. Während Attraktivität in Stimmen (‚vocal attraction') bereits relativ gut erforscht ist, besteht hinsichtlich vokaler Sympathie noch ein großes Forschungsdesiderat – bei menschlichen, aber insbesondere bei synthetischen Stimmen. Die Fragen, die sich vor diesem Hintergrund und in Hinblick auf die beiden Experimente stellen, lauten deshalb wie folgt: i) Existiert, analog zur Attraktivität, eine optimale Grundfrequenz auch für Sympathie? ii) Wenn es eine ideale Frequenz gibt, liegt sie dann in einem ähnlichen Frequenzbereich (260 Hz)? iii) Gelten für menschliche und synthetische Stimmen die gleichen Präferenzen oder zeigen sich Unterschiede?

Diesen Fragen soll in Kapitel vier nachgegangen werden. Doch zuvor möchte ich noch auf die weibliche, synthetische Stimme eingehen, wie man sie z. B. in Sprachassistenten integriert antrifft. In diesem Zug sollen auch einige Spezifika des Voice-Designs für Maschinen hervorgehoben, sowie Überlegungen zu der Rolle der weiblichen Stimme bei Sprachassistenten und Co. angestellt werden; insbesondere

zu den Herausforderungen und potenziellen Gefahren, die beim Diskurs über die ‚optimale weibliche Stimme' mitbedacht werden sollten.

3 Alexa, Siri, Cortana – die ‚Assistentinnen' des digitalen Zeitalters?

Sprachassistenten sind natürlichsprachige Dialogsysteme, die in einem Smartphone (z. B. Siri) oder in einem Smart Speaker (z. B. Alexa) integriert sein können und auf Aufforderungen ihrer Nutzer reagieren – entweder durch das Ausführen einer Handlung (z. B. einen Wecker stellen) oder einer verbalen/schriftlichen Antwort.[20] Wird dabei dieselbe Stimme für mehrere Geräte (z. B. mehrere Smart Speaker) verwendet, werden diese als dieselbe ‚Person' (bzw. sozialer Akteur) wahrgenommen, auch wenn die Geräte eine andere äußere Erscheinungsform – z. B. unterschiedliche Farbe oder Form – besitzen (vgl. Cambre/Kulkarni 2019, 3 f.). Besitzt eine Person hingegen zwei Smart Speaker, einen mit männlicher und einen mit weiblicher Stimme, werden diese entsprechend auch als zwei *verschiedene* Akteure wahrgenommen. Dies zeigt, dass dieser enge Nexus zwischen Stimme und Identität des Sprechenden nicht nur für Menschen, sondern auch Maschinen gilt.

Dass die Stimme ein so wichtiges Element in der Wahrnehmung von technischen Entitäten ist, stellt somit einige Herausforderungen an die Stimmarchitekten solcher Geräte. In der Robotik hat sich herausgestellt, dass Stimme und (attribuierte) Persönlichkeit möglichst kongruent sein sollten, damit der Roboter gut angenommen wird: „If people attribute traits and intentions to robots as if they were humans, then it is likely that the design of a robot's appearance and voice should go hand in hand to create a cohesive image of this agent" erklären McGinn/Torre (2019, 222). Die Stimmwahl für einen Roboter hat somit viel mit der bereits zuvor erwähnten menschlichen Erwartungshaltung zu tun, da er dieser entsprechen sollte. Denn genauso wie wir erwarten, dass ein Staubsaugroboter Staub saugt, so erwarten wir bei einem rosafarbenen Roboter mit langen Haaren, der sprechen kann, dass er dies mit einer weiblichen Stimme tut: „People use visual gender cues as a basis for their judgments about social robots […] such as a robot's hair length in order to make gendersterotypical inferences about the robot" (Eyssel et al. 2012, 1). Spricht uns nun ebendieser langhaarige, rosafarbene Roboter mit einem tiefen Bariton an, löst er Irritationen aus, die möglicherweise zur Ableh-

20 Durch eine Änderung in den Einstellungen kann mit Siri auch schriftlich statt verbal kommuniziert werden.

nung des Roboters führen können,[21] denn „users expect a robot's voice to match its appearance" (Cambre/Kulkarni 2019, 8).[22] Für das Design von Stimmen muss also beachtet werden, welche Vorannahmen existieren und welche Erwartungen an das Gerät gestellt werden. Diese Relation von Erwartungshaltung und Stimmwahl wirft an dieser Stelle die Frage auf: Sprachassistenten tragen meist einen weiblichen Vornamen und sprechen mit – deutlich erkennbar – weiblicher Stimme. Bedeutet dies, dass wir intuitiv *erwarten*, dass eine Maschine, die uns *assistiert*, weiblich ist?

An dieser Stelle lohnt es sich, darüber nachzudenken, ob z. B. ein Sprachassistent, der mit weiblicher Stimme spricht, auch tatsächlich als ‚Frau' wahrgenommen wird. Natürlich wird eine Maschine, die mit einer weiblicher Stimme (oder weiblichen äußerlichen Attributen) ausgestattet wird, i. d. R. weiblich gelesen. Dennoch ist diese Perzeption der Maschine als ‚weiblich' ontologisch betrachtet noch lange nicht dasselbe wie die Wahrnehmung der Maschine als *Frau*. Somit ist auch der Begriff ‚Frauenstimme' im Mensch-Maschine-Kontext faktisch inkorrekt, und es müsste von einer (synthetischen) ‚weiblichen Stimme' gesprochen werden.[23] Im sprachlichen Diskurs um Sprachassistenten und in der Praxis verschwimmt diese Grenze zwischen Frau/weiblicher Maschine jedoch zunehmend, wenn für Siri und Alexa Bezeichnungen wie ‚Assistentin' genutzt werden, anstatt ‚Assistenzsystem', was zwar kaum zu vermeiden, aber dennoch bedenklich ist. Besonders mit Blick auf eine mögliche Zukunft, in der Menschen zunehmend öfter mit Maschinen interagieren werden und Maschinen (durch humanoide Optik, menschlich klingende Stimmen etc.) auch stetig mehr humanisiert und dadurch anthropomorphisiert werden.

Vor diesem Hintergrund könnte man sich fragen, weshalb solche Assistenzsysteme überhaupt überwiegend weibliche Vornamen (siehe Siri, Alexa, Cortana, Alice etc.) und – per Default-Einstellung – eine weibliche Stimme aufweisen. Und obwohl sexistische Vorurteile sicher ihren Teil dazu beitragen, weshalb dem so ist, gibt es aber tatsächlich auch soziopragmatische Gründe für die Nutzung von weiblichen Stimmen bei Sprachassistenten. Einer davon ist, dass Frauenstimmen im Vergleich mit Männerstimmen als sympathischer bewertet werden (vgl. Danielescu 2020, 2). Somit liegt die Favorisierung von weiblichen Stimmen für technische

21 Man könnte an dieser Stelle von einem ‚akustischen Uncanny-Valley-Effekt' sprechen.

22 Cambre und Kulkarni (2019) beziehen sich in ihrem Artikel auf die Untersuchungen von McGinn/Torre (2019) und Moore (2017), die zeigen, dass Menschen eine gehörte Stimme einem aus ihrer Sicht dazu passenden Roboter zuschreiben (z. B. eine kindliche Stimme einem niedlichen Roboterhäschen (vgl. Moore 2017, 3)).

23 Die semantische Trennung zwischen den Begriffen „Frauenstimme" und „weibliche Stimme" wird für das Leseverständnis und eine bessere Zitierbarkeit der hier verwendeten Quellen im weiteren Verlauf dieses Artikels nicht konsequent fortgeführt; der denotative Unterschied soll an dieser Stelle jedoch betont werden.

Entitäten nicht (nur) daran, dass Alexa und Co. als digitale ‚Assistentinnen und Sekretärinnen' stereotypische Geschlechterrollen tradieren, sondern auch daran, dass einige Stimmen besser geeignet sind, um positive Emotionen wie Sympathie und Vertrauen zu evozieren als andere:

> Prior work examining trust in human voices has shown that **female speakers tend to be rated as more trustworthy than male speakers.** […] **Younger female agents** were perceived to be significantly more trustworthy **than male agents** or an **older female agent** (Goodman/Mayhorn 2023, 2) [Hervorhebung durch die Autorin].

> [Clifford Nass]: It's much easier to find a female voice that everyone likes than a male voice that everyone likes. […] It's a well-established phenomenon that **the human brain is developed to like female voices** (Griggs 2011) [Hervorhebung durch die Autorin].

Liest man obige Zitate, wird ersichtlich, welche Art von Stimmen das sind: Es sind junge, weibliche Stimmen, die präferiert werden – und zwar „across cultures and genders" (Danielescu 2020, 3), obwohl sich bei ‚genders' hier die Frage stellt, wie inklusiv diese Studien durchgeführt und ob z. B. auch die Präferenzen von Trans- oder non-binären Personen miteinbezogen wurden. Jedenfalls liegt aus diesen Gründen das Augenmerk bei der Konstruktion von synthetischen Stimmen auf weiblichen Stimmen. Somit kann für Sprachassistenten gesagt werden, dass das durch die Stimme dargestellte ‚Geschlecht' (weiblich) nicht rein aufgrund von sexistischen Klischees gewählt wird, sondern vor allem um Zustimmung bei der Mehrheit der User*innen zu finden.[24] Wie bereits erwähnt, ist aber gerade in Bezug auf Sprachassistenten der Einsatz von weiblichen Stimmen nicht frei von Problematiken, da dadurch einerseits sexistische Stereotypen verstärkt werden können (‚die typische *Assistentin*') und andererseits der tägliche Umgang mit ihnen im persönlichen und/oder professionellen Umfeld mit der Zeit zu einem veränderten Kommunikationsstil gegenüber Frauen führen kann[25] – besonders wenn bedacht wird, dass meist in imperativer Form mit solchen Maschinen gesprochen wird, wie z. B. „Alexa, *mach* das Licht an!" oder „Siri, *stell* einen Wecker auf 7 Uhr!". Auch

24 Nicht nur die Stimme allein, sondern auch eine weibliche Identität scheint für technische Entitäten präferiert zu werden. Dies sei auch der Grund, senso Borau et al. (2021), weshalb „almost all virtual AI products in the market today, including virtual assistants and chatbots, come with female features" (Borau et al. 2021, 1062). Sie begründen dies damit, dass weibliche Maschinen gegenüber männlichen Maschinen bevorzugt werden, da sie ‚menschlicher' erscheinen: „[P]eople prefer female over male bots because they are **perceived as more human** […] and **consumers seem to prefer humanlike machines**" (Borau et al. 2021, 1064) [Hervorhebung durch die Autorin].

25 „For instance, since people become used to interacting with those agents in a commanding tone, humans might also (subconsciously) mirror this behavior in their everyday conversations with women" (Tolmeijer et al. 2021, 1).

Schimpfwörter werden geäußert, wenn die Maschine etwas falsch versteht oder den Befehl nicht ausführt, der vom User geäußert wurde.[26] Expert*innen warnen deshalb davor, dass der tägliche Umgang mit Sprachassistenten zu einer verzerrten Wahrnehmung von Frauen als devote ‚Assistentinnen' führen könne, mit denen man sprechen (und umgehen) kann, wie man will. Besonders problematisch ist es, wenn weibliche Stimmen bevorzugt eingesetzt werden, wenn eine Maschine Anweisungen *erhält*, und männliche Stimmen in Situationen, in denen die Maschine Anweisungen *erteilt*, da dies alte Geschlechtervorurteile verstärken kann: „[P]eople tend to perceive female voices as helpers or assistants, who are helping us solve our problems, while male voices are viewed as authority figures who tell us the answers to our problems" (Danielescu 2020, 2).[27]

Für die Designer*innen von Smart-Geräten bleibt es deshalb eine Herausforderung, Stimmen für Geräte zu konstruieren, die einerseits den Erwartungshaltungen gerecht werden, um keine Aversionen auszulösen, ohne aber andererseits sexistische Vorurteile (weiter) zu fördern. Wenn von Stereotypie in Bezug auf Stimme und Gender gesprochen wird, sollte auch eine mögliche Wechselwirkung bedacht werden: Hören wir als Sprecher*innen einer Gesellschaft stets eine stereotypisierte Form von weiblichen Stimmen bei technischen Entitäten (Sprachassistenten, soziale Roboter, KIs in Sci-Fi-Filmen wie *Her* etc.), festigt dies möglicherweise unsere Erwartungshaltung, wie Frauen idealerweise sprechen bzw. klingen sollten: „[S]ince human beings have demonstrated a tendency to anthropomorphize machines, and thus to treat social robots as fellow human beings, social robots' role in influencing the way we see gender and gender appropriate language might be even bigger than we think" (Pietronudo 2018, 7).[28]

---

26 Äußerst bedenklich sind misogyne Kommentare, die sich in Online-Foren (wie z. B. *Reddit*) finden lassen, in denen sich männliche User abfällig über Alexas äußern. In einem Kommentar erzählt ein User beispielsweise, wie er seiner Alexa befohlen hat, sich selbst als ‚Bitch' zu beleidigen und ihn um Verzeihung anzuflehen, wenn sie einen Fehler macht (vgl. Reddit 2020). Nicht nur die kontinuierliche Assoziation mit dem weiblichen Geschlecht ist hierbei problematisch, sondern auch die Tatsache, dass derartige Äußerungen nicht nur im privaten, sondern auch im öffentlichen Diskurs praktiziert werden. Auch Artikel, die sich gegen die Degradierung von Sprachassistenten wie Siri als ‚Schlampe' (Mitteldeutsche Zeitung) oder ‚dumme Kuh' (Emma) aussprechen, setzen diese Konvention indirekt fort, indem sie sich desselben pejorisierenden Duktus bedienen (vgl. Könau 2019; Laeri 2019).

27 Dass männliche Stimmen in hierarchischen Kontexten eingesetzt werden, könnte erklären, weshalb in Dokumentarfilmen oder anderen Medienformaten (z. B. Autowerbungen), in welchen die Stimme als Autorität historisches/technisches Wissen an den Zuschauer vermittelt, überwiegend Männerstimmen verwendet werden.

28 Auch Borau et al. (2021) beziehen sich auf diese Wechselwirkung zwischen Frauen und technischen Entitäten: „Women are said to be transformed into AI objects [Roboter wie *Sofia*, weibliche Chatbots wie *Amelia* und Sprachassistenten wie *Siri*], but injecting women's humanity into AI

Um solchen Problematiken hinsichtlich Geschlecht und sprechenden Maschinen künftig angemessen begegnen zu können, setzen Unternehmen wie *Virtue*[29] auf die Generierung von geschlechtsneutralen Stimmen (*Project Q*), die sich in einem Frequenzbereich bewegen, der weder als eindeutig männlich noch als eindeutig weiblich wahrgenommen wird (vgl. Degelo 2021, 220). Allerdings gibt es zurzeit nur wenige solche Angebote, was an der noch geringeren Nachfrage für solche Produkte liegt.

Wie dieses Kapitel gezeigt hat, gibt es einige Herausforderungen und soziale Implikationen, die bei der Konstruktion von weiblichen Stimmen für Geräte wie Sprachassistenten bedacht werden müssen. Weibliche Stimmen machen die Maschinen zwar sympathischer. Aber weshalb ist es denn so wichtig, dass Menschen ihre Maschinen mögen, könnte an dieser Stelle gefragt werden. Dies ist jedoch eine Frage, die sich nicht pauschal beantworten lässt. Einige Personen sehen in einem Sprachassistenten oder sozialen Roboter lediglich eine Maschine, die sich nur durch ihre sprachlichen Fähigkeiten von einer anderen Maschine, wie z. B. einer Kaffeemaschine unterscheidet. Für diese Personen dürfte der Sympathiefaktor keine maßgebliche Rolle spielen. Für andere Personen wiederum sind solche sprechenden Maschinen wie Sprachassistenten ein elektronisches Gegenüber (ganz im Sinne des CASA-Paradigmas),[30] mit denen sie tägliche Unterhaltungen führen und dadurch eine soziale Beziehung konstituieren: „Increasingly, people view their interactions with intelligent assistants as social interactions, with 41 % of people who own a voice assistant saying it feels like talking to a friend“ (Danielescu 2020, 2 nach Kleinberg 2018). Für solche Personen dürfte die Sympathie ein wichtiges Kriterium sein.

Abgesehen von diesen individuellen Sichtweisen gibt es aber auch universelle Vorteile, die sympathische synthetische Stimmen auf verschiedene Interaktions-

objects makes these objects seem more human and acceptable“ (Borau et al. 2021, 1052). Für die Akzeptanz von Maschinen ist es zwar vorteilhaft, wenn diese als weiblich perzipiert werden, da dies den Anthropomorphismus begünstigt. Aber die Auswirkungen davon, dass technische Entitäten überwiegend weiblich dargestellt werden, ist problematisch: „As a result, AI designers and policymakers face an ethical dilemma: The female gendering of AI is likely to increase the perceived humanness and adoption of AI tools, but also risks, in turn, to reinforce or even propagate harmful gender stereotypes“ (Borau et al. 2021, 1065). Und zu diesen ‚harmful gender sterotypes‘ gehört dann auch die Objektifizierung von Frauen als ‚persönliche Assistentinnen‘.

29 *Virtue* ist eine Agentur, die zur VICE Media Group gehört. In Zusammenarbeit mit der Pride Kopenhagen und Linguist*innen der dortigen Universität entwickelte dieses Unternehmen ‚Q‘ – die weltweit erste geschlechtsneutrale künstliche Stimme (vgl. Werbewoche 2019).

30 Das CASA-Paradigma (= **C**omputer **a**re **S**ocial **A**ctors) beschreibt die Tendenz, einen Computer wie einen Menschen zu behandeln, im vollen Wissen darüber, dass er eine Maschine und kein Mensch ist (Nass et al. 1994).

formen und -bereiche von Mensch und Maschine haben können; auf diese werde ich im Ausblick genauer eingehen. Das nächste Kapitel widmet sich zunächst aber dem Thema, welche Stimmen (menschliche und synthetische) uns überhaupt sympathisch sind.

4 Empirische Studien zur Bewertung von Sympathie in Stimmen

Nach diesem Überblick zur vokalen Attraktivität menschlicher Stimmen und dem Exkurs zum Voice-Design bei Sprachassistenten sei zunächst nochmals kurz auf das Problem der inkohärenten Verwendung des Ausdrucks ‚vokale Attraktivität' bzw. *vocal attraction* eingegangen. Der Begriff *vocal attraction* scheint sich als eine Art Sammelbecken für verschiedenste, positiv bewertete stimmliche Ausprägungen etabliert zu haben. Da das englische Wort *attractive* aber polyvalent ist und nebst ‚attraktiv' ebenso die Bedeutungen ‚hübsch', ‚erotisch, ‚anziehend', ‚verlockend', ‚sympathisch' etc. enthält, kann dies zu einer ungenauen Trennschärfe in diesem Forschungsbereich führen.[31] Aufgrund dessen wurde in der nachfolgend dargestellten Untersuchung Wert daraufgelegt, den Unterschied zwischen den beiden Qualitäten attraktiv/sympathisch zu betonen. Um dies zu erreichen, wurden die Proband*innen gleich zu Beginn der Studie darauf hingewiesen, dass sie die nachfolgenden Stimmen nach Sympathie und nicht Attraktivität (im erotischen Sinne) bewerten sollen und eine Definition von Sympathie, wie sie hier verstanden wurde, gegeben. Zusätzlich wurde im Evaluationspart ebenfalls nochmals auf diese Unterscheidung hingewiesen.[32]

---

31 Besonders ‚attractive' und ‚erotic' werden in Wahrnehmungsexperimenten oft synonym verwendet, obwohl eine ‚erotische Stimme' andere Qualitäten besitzt, wie z. B. eine stärkere Behauchung (breathy voice), als eine lediglich ‚attraktive' – im Sinne von gutklingender – Stimme (vgl. Völkert 2012).

32 Die Proband*innen wurden zu Beginn des Fragebogens wie folgt instruiert:
The following survey aims to determine which voices are perceived as likable. You will listen to different AI-generated voices and be asked to rate them based on likability. In this study, ‚vocal attractiveness' is understood as a likable and sympathetic voice. Please do not rate them according to how ‚erotic' or ‚attractive' they sound.
Ebenfalls wurde bei den Bewertungsaufgaben nochmals explizit darauf hingewiesen, wie die Stimmen zu bewerten sind:
**For all questions: Please base your decision on sympathy and likability, rather than judging which voice sounds better, more erotic, etc. Choose the voice that you find more appealing for a voice assistant, for example.*

4.1 Methodik

Zur Untersuchung der Sympathie bei menschlichen und synthetischen Stimmen entwarf ich ein Studiendesign mit zwei Experimenten:

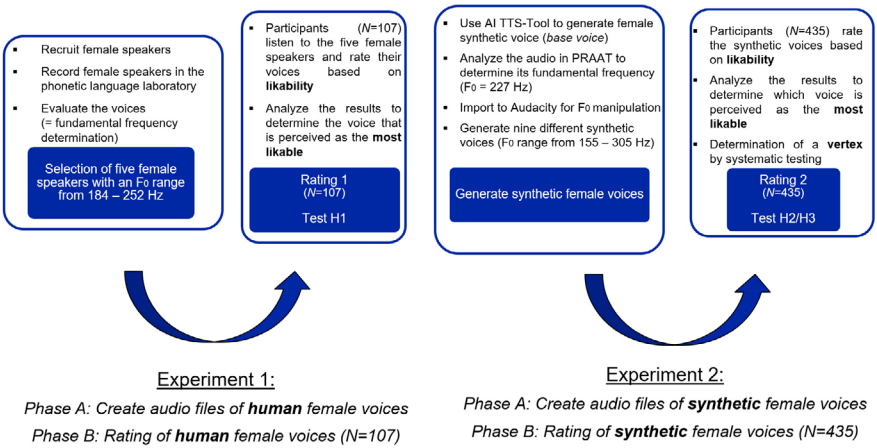


Abbildung 3: Studien-Design der Untersuchung ‚Synthetische Sympathie'

Beide Experimente dienten zur Eruierung der sympathischsten Grundfrequenz (F_0). Dazu wurden drei Hypothesen aufgestellt (siehe H_1–H_3) und in zwei getrennten Untersuchungen (Experiment 1: menschliche Stimmen, Experiment 2: synthetische Stimmen) die aufgenommenen/generierten Stimmen von Proband*innen jeweils nach Sympathie bewertet. Vor dem Hintergrund des bisherigen Forschungsstands zur Attraktivität weiblicher Menschenstimmen stellte ich die Vermutung auf, dass höhere Stimmen nicht nur als attraktiver, sondern auch als sympathischer wahrgenommen werden. Vom durchschnittlichen Spektrum für westliche Frauenstimmen (165–255 Hz nach Kiese-Himmel 2016) ausgehend, nahm ich den Median,[33] der bei 210 Hz lag, als Trennwert zur Unterscheidung in hohe/tiefe Stimmen. 210 Hz entspricht zudem auch der durchschnittlichen Grundfrequenz für Frauenstimmen, die bei Ujvary et al. (2022, 7) genannt wurde. Somit wurde in der ersten Untersuchung mit menschlichen Stimmen geprüft, ob Stimmen, die über dem Median liegen, sympathischer wahrgenommen werden als Stimmen, die darunter liegen (H_1):

33 Der Median ist ein statistisches Maß und bezeichnet den ‚Zentralwert', also den Wert, bei dem genau 50 % der Daten (hier Stimmen) darüber und 50 % darunter liegen. In diesem Fall entspricht er auch dem Mittelwert.

H_1: Menschliche, weibliche Stimmen mit hoher Frequenz (F_0>210 Hz) werden als sympathischer wahrgenommen als menschliche, weibliche Stimmen mit tieferer Frequenz (F_0<210 Hz).

Für die Sympathie bei synthetischen Stimmen wurde die analoge Vermutung aufgestellt (H_2). Zudem wurde ein Scheitelpunkt erwartet, nach dem die Sympathie wieder abnimmt, ausgehend von der Vermutung, dass F_0/Sympathie – genau wie F_0/Attraktivität – ebenfalls nicht unendlich korrelieren (H_3). Die Hypothesen für das zweite Experiment lauteten demnach:

H_2: Synthetische Stimmen mit höherer Grundfrequenz (F_0>210 Hz) werden als sympathischer wahrgenommen als synthetische Stimmen mit tieferer Grundfrequenz (F_0<210 Hz).

H_3: Die Korrelation zwischen Grundfrequenz (F_0) und Sympathie ist bei synthetischen Stimmen nicht unendlich linear; es existiert ein Scheitelpunkt.

Experiment 1: Menschliche Stimmen

Für die Aufnahme von menschlichen Stimmen wurden zehn Sprecherinnen (Studentinnen der Universität Zürich) rekrutiert,[34] die einem zuvor festgelegten Profil entsprachen. Das Profil sollte sicherstellen, dass es in der späteren Bewertung zu keinerlei Verzerrungen zu Gunsten von Geschlecht, Alter, regionalen Unterschiede in der Standardsprache oder der Sprachkompetenz kommt. So handelte es sich z. B. bei allen Sprecherinnen um Zürcherinnen,[35] um auf diese Weise Sympathien zugunsten von unterschiedlichen dialektalen Färbungen in der Standardsprache auszuschließen. Alle Audioaufnahmen wurden im Tonstudio des *LiRi* (Linguistic Research Infrastructure) der Universität Zürich aufgenommen, um sicherzustellen, dass keine externen Störgeräusche die Aufnahmequalität beeinträchtigen.[36] Die Sprecherinnen erhielten vorgängig keine Information über die anschließende Sympathiebewertung, da dies möglicherweise zur (un)bewussten Manipulation der eigenen Stimme geführt hätte, was vermieden werden wollte. Die Sprecherinnen wurden zu diesem Zeitpunkt lediglich über den Ablauf des Experiments instruiert und darüber aufgeklärt, dass die dabei entstandenen Audiodaten zu Forschungs-

34 Der Ausruf enthielt dabei Angaben über die Profilkriterien (Alter, Geschlecht, Schweizerin, Zürcherin oder seit 15 Jahren im Kanton Zürich wohnhaft, Studentin der Germanistik/Fachdidaktik Deutsch/Deutsche Literaturwissenschaft, keine auditiven/verbalen Einschränkungen (Hörgerät, Stottern etc.)), welche die Sprecherinnen erfüllen mussten und Informationen bezüglich des Zeitaufwands (mit Instruktion und Aufnahme im phonetischen Labor ca. 1h) sowie der Vergütung (10.- CHF Gutschein für Starbucks).

35 Kanton in der Schweiz, dessen Dialekt die größte Nähe zur Standardsprache aufweist und keine starke, idiosynkratische Färbung besitzt (wie z. B. das Walliser- oder Berndeutsch).

36 An dieser Stelle möchte ich dem Lab Manager Andrew Clark herzlich für seine technische Unterstützung im Labor und Einstellung der Aufnahmegeräte danken.

zwecken analysiert und perzipiert werden. An den Aufnahmetagen sprach jede Sprecherin insgesamt drei analog gestaltete deutsche Texte mit je fünf Sätzen ein. Um eine Textsorte zu wählen, mit der alle Sprecherinnen sicher vertraut waren, fiel die Entscheidung auf eine telefonische Essensbestellung.[37] Der einzige der drei Texte, der dabei tatsächlich für die Sympathie-Bewertung der Stimme verwendet wurde, wurde zum Schluss eingesprochen, um so der anfänglichen Nervosität (Stottern, Versprechen, Lachen, Räuspern) der Sprecherinnen bestmöglich vorzubeugen. Die Grundfrequenzen (F_0) der Sprecherinnen wurden anhand des gesamten Audiomaterials des dritten Textes mittels der automatischen Pitch-Erkennung von Praat[38] eruiert (*mean pitch* F_0). Anschließend erfolgte die Auswahl von fünf Stimmen (V1–V5), deren Grundfrequenzen zusammen von 184 bis 252 Hz rangierten,[39] um so das durchschnittliche Spektrum der weiblichen Stimme möglichst nachzubilden (165–255 Hz vgl. Kiese-Himmel 2016, 39). Die mittlere Stimme (V3) dieser Untersuchung entspricht mit 209 Hz beinahe exakt dem zuvor bestimmten Median von 210 Hz. Dadurch liegen jeweils zwei Stimmen darüber (V1, V2) und zwei Stimmen darunter (V4, V5), wodurch sich eine symmetrische Verteilung ergibt.[40] Zur Verifizierung der H_1 müssten die Stimmen mit höherer Grundfrequenz, V1 und V2 (F_0 >210 Hz), gegenüber den tieferen V4 und V5 (F_0 <210 Hz), von einer deutlichen Mehrzahl der 107 Probanden dieses Experiments präferiert werden.[41]

37 Essensbestellungen bei einem mexikanischen, indischen und italienischen Lieferservice. Der Text, der später genutzt wurde, lautete: „Guten Tag, hier spricht Lisa Müller. Ich würde gerne bei Ihnen eine Pizza Margherita und zwei Cola bestellen. Ich wohne an der Mustergasse 1, in dem großen Haus mit der roten Türe. Sie können das Essen einfach vor die Türe stellen. Vielen Dank!“. Die Lautstärke (*mean intensity*) betrug bei allen Sprecherinnen zwischen 60–62 dB.

38 Software für phonetische Analysen: https://www.fon.hum.uva.nl/praat/.

39 Da einige Stimmen eine sehr ähnliche Grundfrequenz aufwiesen, konnten nicht alle 10 Sprecherinnen berücksichtigt werden. Ausgewählt wurden die Stimmen mit folgender Grundfrequenz: V1 = 252 Hz, V2 = 222 Hz, V3 = 209 Hz, V4 = 192 Hz, V5 = 184 Hz.

40 Es wäre wünschenswert gewesen, wenn die Intervalle zwischen den einzelnen Stimmen immer gleich groß gewesen wären (z. B. exakt 20 Hz). Da es sich hier jedoch um natürliche, nicht-manipulierte Stimmen handelt, war dies nicht möglich.

41 Da menschliche Stimmen bereits gut erforscht sind und der Fokus dieser Arbeit auf synthetischen Stimmen liegt, wurde diese erste Untersuchung mit menschlichen Stimmen als eine Art Replikastudie zu den Attraktivitätsstudien konzipiert – mit dem Unterschied, dass *Sympathie* und nicht *Attraktivität* erforscht werden soll. In Anbetracht der bereits etablierten Kenntnisse zu menschlichen Stimmen werden hier somit nur eine geringe Anzahl Stimmen des durchschnittlichen Stimmspektrums untersucht, um zu prüfen, ob Grundfrequenz und Sympathie ebenfalls in Korrelation stehen. In einer weiteren, größeren Untersuchung wäre es aber erstrebenswert, auch natürliche Stimmen außerhalb des durchschnittlichen Spektrums zu inkludieren.

Experiment 1: Bewertung
Die für die Bewertung rekrutierte Proband*innengruppe bestand aus 107 Teilnehmenden. Dabei handelte es sich um 78 weibliche, 19 männliche und 10 non-binäre Studierende der Universität Zürich mit unterschiedlicher sexueller Orientierung.[42] Die Aufgabe der Teilnehmer*innen war es, die fünf ausgewählten Stimmen nach mehrmaligem Hören nach Sympathie zu ranken, von der sympathischsten Stimme zu der am wenigsten sympathischen Stimme. Alle Teilnehmenden wurden angewiesen, die Stimmen dabei ausschließlich nach Sympathie, nicht nach Attraktivität zu bewerten. Zur internen Kontrolle wurden die Proband*innen nach der Evaluierung der Stimmen in der Auswertung in zwei Gruppen unterteilt. In der ersten Gruppe, der SVA (social vocal attraction: *Sympathie*) befanden sich nur die Antworten jener Proband*innen, von denen – aufgrund ihres Geschlechts und sexueller Orientierung – davon ausgegangen werden konnte, dass sie keine sexuelle Anziehung zu Frauen haben. Das bedeutet, dass in der SVA-Bedingung die Antworten von heterosexuellen Männern sowie lesbischen und bisexuellen Frauen von der Evaluierung ausgeschlossen wurden, da diese Personengruppen trotz des Hinweises auf das Bewertungskriterium ‚sympathisch' möglicherweise eine weibliche Stimme (unabsichtlich) auch hinsichtlich ihrer Attraktivität, im erotischen Sinne, bewerten würden. Diese Personen wurden in die VA-Gruppe (vocal attraction: *Attraktivität*) eingeteilt. Durch diese Zweiteilung ließen sich die Ergebnisse der SVA-Gruppe und die der VA-Gruppe zur Sicherung der internen Validität gegenüberstellen. Hätte sich nun ein aussagekräftiger Unterschied in der Bewertung gezeigt (>5 % Abweichung), hätte die VA-Gruppe ausgeschlossen werden müssen.

In der Gegenüberstellung der beiden Gruppen (SVA/VA) konnte allerdings kein solcher Unterschied festgestellt werden. Die geringfügigen Abweichungen wurden deshalb als zufällig eingestuft und die VA-Gruppe (n=60) in die Gesamtauswertung (SVA(n=47) + VA(n=60) = N=107) reintegriert.

Experiment 2: Synthetische Stimmen
Die synthetischen Stimmen für das zweite Experiment wurden mit Hilfe von *Lovo.AI*[43] erstellt: Lovo.AI ist einer der fortschrittlichsten KI-Stimmengeneratoren, bei dem aus über 500 Stimmen und 100 verschiedenen Sprachen gewählt werden kann. Mittels TTS-Technologie[44] ermöglicht er es, beliebige Sätze in der gewünsch-

42 Aufgrund des vorzeitigen Abbruchs des Fragebogens mussten 23 Personen aus der Probandengruppe ausgeschlossen werden. Berücksichtigt wurden nur Personen, die den Fragebogen bis zum Ende ausgefüllt hatten.

43 https://genny.lovo.ai/signin

44 TTS steht für *Text-to-Speech* und beschreibt eine Technologie, die es Computern und anderen Geräten ermöglicht, geschriebene Texte in gesprochene Sprache zu konvertieren. Sprachassistenten wie Alexa oder Siri nutzen TTS, um z. B. die Nachrichten oder den Wetterbericht vorzulesen.

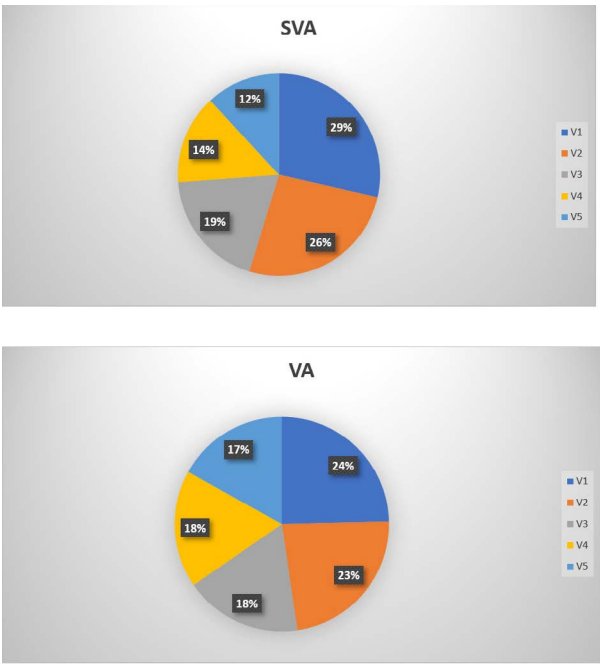


Abbildung 4: Auswertung SVA und VA

ten Stimme zu produzieren. Aus diesen Gründen ist Lovo.AI nicht nur in der Wirtschaft (z. B. für Voice-Overs bei Werbungen) sondern auch in der Wissenschaftscommunity beliebt und wird für Studien mit synthetischen Stimmen genutzt (siehe z. B. Herder/Herden 2023). Für dieses Experiment wählte ich eine Stimme mit dem Profil *female, young adult, US-English* und ließ sie den Satz: *„Hello, how are you today*?" vorlesen. Gewählt wurde ein bewusst neutraler und kurzer Satz, sodass das semantische Gehalt der Äußerung möglichst vorurteillos ausfällt und keine Assoziationen birgt, die einen Einfluss auf die Bewertung haben könnten.[45] Dieses Audiofile diente dann als ‚Basisstimme', auf deren Grundlage dann durch Manipulationen der Grundfrequenz (Erhöhung/Vertiefung der F_0) neun weitere synthetische Stimmen generiert wurden. Dazu importierte ich die Audiodatei in *Audacity*, ein intuitives Audioeditor-Programm, in dem die Grundfrequenz manuell modifiziert (erhöhen/

45 Für eine bessere Durchführbarkeit und um mehr Proband*innen für das zweite Experiment zu gewinnen, wurde beschlossen, einen englischen Satz vorlesen zu lassen.

vertiefen)[46] werden kann (Abb. 5). Die dadurch entstandenen Grundfrequenzen der neun synthetischen Stimmen (EVA 1–EVA 9)[47] umfassten zusammen einen breiten Frequenzbereich von 155 Hz (EVA 1) bis 305 Hz (EVA 9). Damit war der Frequenzbereich – analog zu Borkowska/Pawlowski (2011) – breit genug angesiedelt, um exakt den Moment festlegen zu können, bei dem sich ein Scheitelpunkt etablieren würde. Da das einzige Unterscheidungskriterium dieser synthetischen Stimmen die Grundfrequenz war, konnten andere beeinflussende Faktoren (Behauchung, Intonation, Rhythmus, Lautstärke, Geschwindigkeit etc.) ausgeschlossen werden. Die Proband*innen würden die Sympathie der verschiedenen synthetischen Stimmen somit zwangsläufig nur anhand ihrer unterschiedlichen Grundfrequenzen bewerten.

Sollte sich Sympathie bei synthetischen Stimmen analog zu Attraktivität bei natürlichen Stimmen verhalten, müsste in der Gegenüberstellung zweier Stimmen in allen Fällen *unter* 260 Hz (Scheitelpunkt bei menschlichen Stimmen und Attraktivität, vgl. Borkowska/Pawlowski 2011) die höhere Version (V_{high}) und in allen Fällen *über* 260 Hz die tiefere Version (V_{low}) von EVA präferiert werden. Zur erfolgreichen Verifizierung der Hypothesen müsste sich in den Resultaten zeigen, dass höhere, synthetische Stimmen (F_0>210 Hz) als sympathischer wahrgenommen werden als tiefere, synthetische Stimmen (H_2) – bis zu einem Scheitelpunkt, nach dessen Überschreiten die Sympathie wieder abnimmt (H_3). Diese Frequenz beim Scheitelpunkt würde dann der ‚goldenen Frequenz', der sympathischsten Stimme, entsprechen.

46 https://www.audacityteam.org/ Veränderung der F_0: Hierzu wurde die Audioaufnahme importiert und dann unter ‚Effekt' die Möglichkeit „Tonhöhe ändern" gewählt. Die Grundfrequenz (z. B. 240 Hz) konnte dort manuell in die gewünschte neue Frequenz überführt werden (z. B. 280 Hz), ohne weitere Faktoren (wie das Tempo) zu verändern. Dadurch ist der einzige Unterschied zwischen zwei Stimmen die unterschiedliche Grundfrequenz (F_0).

47 Der Name EVA wurde gewählt, um die Tradition synthetischen weiblichen Stimmen einen Frauennamen zu geben, fortzusetzen. Weiter könnte EVA auch für ‚Electronic Voice Agent' stehen. Die Grundfrequenzen der neun verschiedenen EVAs setzten sich wie folgt zusammen: EVA 1 = 155 Hz, EVA 2 = 185 Hz, EVA 3 = 215 Hz, EVA 4 = 240 Hz, EVA 5 = 245 Hz, EVA 6 = 260 Hz, EVA 7 = 275 Hz, EVA 8 = 280 Hz, EVA 9 = 305 Hz. Die Stimmen EVA 1-EVA 4 wurden so generiert, dass sie sich jeweils um 30 Hz unterschieden. Im oberen Bereich (EVA 5-EVA 9) sind die Intervalle anders, dies hat mit der Form der Frage in der Evaluierung (Ranking statt A/B-Rating) zu tun, die im oberen Frequenzbereich zur Eruierung des Scheitelpunkts angewandt wurde (siehe *Experiment 2: Bewertung*).

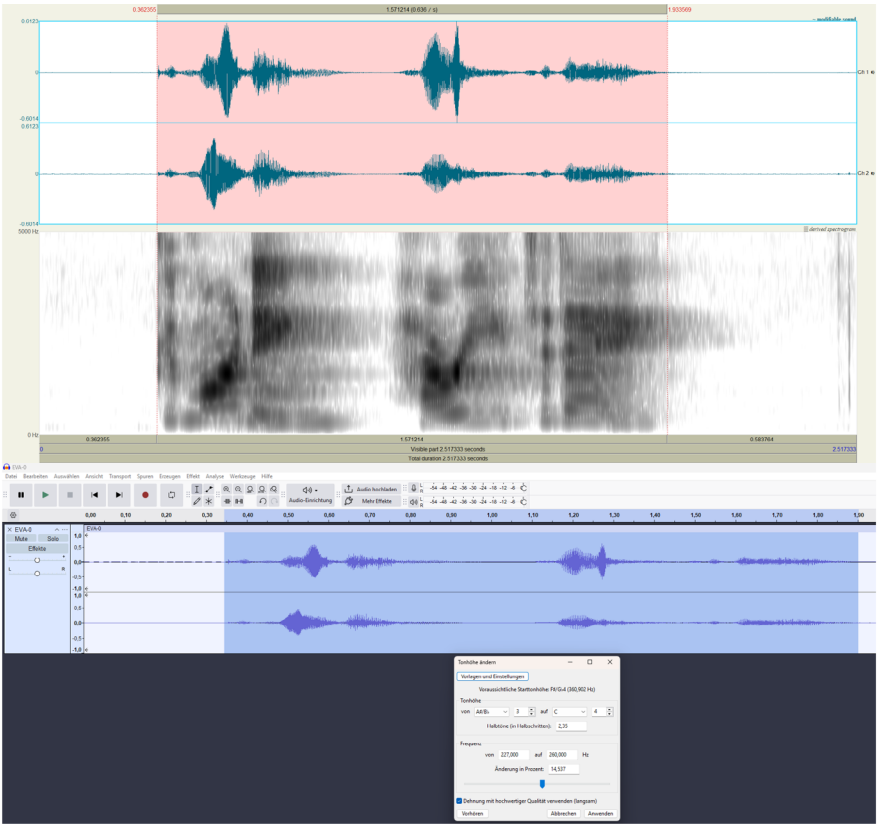


Abbildung 5: Analyse und Modifizierung der Grundfrequenz in Praat & Audacity

Experiment 2: Bewertung der synthetischen Stimmen

Die Proband*innen (N=435)[48] wurden hauptsächlich aus Studierenden der Schweizer Hochschulen (UZH, ETH)[49] rekrutiert. Die dadurch entstandene Proband*innengruppe ist repräsentativ für eine an Schweizer Hochschulen durchgeführte Studie und verfügt deshalb über charakteristische Bias. So zeigen sich deutliche Tendenzen bezüglich der Ethnie/des kulturellen Hintergrunds (~88 % weiß/westlich) und

48 Aufgrund des vorzeitigen Abbruchs des Fragebogens mussten 98 Personen aus der Probandengruppe ausgeschlossen werden. Berücksichtigt wurden nur Personen, die den Fragebogen bis zum Ende ausgefüllt hatten.

49 Universität Zürich und Eidgenössische Technische Hochschule (Zürich), Schweiz.

bezüglich der Sexualität (~71 % heterosexuell); ebenfalls liegt ein leichter Frauenbias (~65 %) vor.[50]

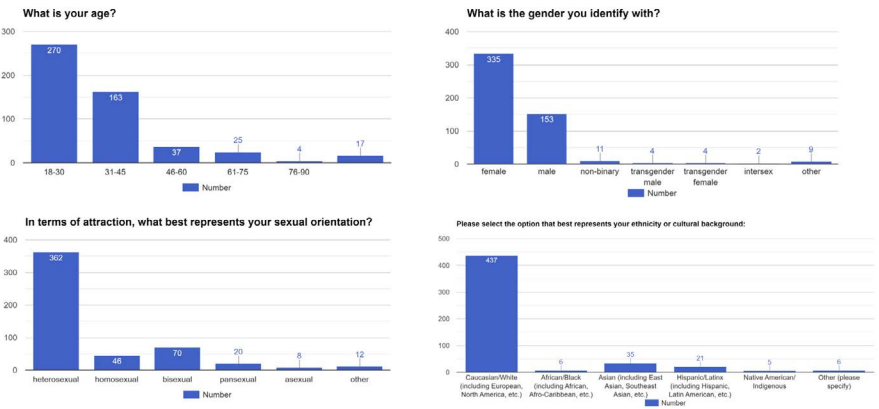


Abbildung 6: Zusammensetzung Probandengruppe

Die Evaluationen der synthetischen Stimmen (EVA 1–9) wurden in Form eines Online-Fragebogens von *Findmind*[51] durchgeführt – ein sicheres, in der Schweiz entwickeltes Umfragetool, das alle Daten verschlüsselt und nicht mit Drittparteien teilt, weshalb es z. B. auch für Kundenumfragen von Schweizer Banken genutzt wird. Der Fragebogen bestand aus drei Teilen: den Fragen zu den demographischen Daten der Teilnehmer*innen (Abb. 6), die A/B-Ratings (Abb. 7a) und den Ranking-Aufgaben (Abb. 7b). Die Teilnahme erfolgte auf freiwilliger Basis und ohne finanzielle Entlohnung und dauerte je nach individuellem Aufwand im Durchschnitt ca. 3–5 Minuten. Die Reihenfolge (höhere Variante/tiefere Variante), in der die Stimmen bei den A/B-Ratings abgespielt wurden, variierten, damit kein auditiver Habituierungseffekt eintreten konnte.[52]

50 Aufgrund dessen betrifft eine Limitation dieser Studie die überwiegend weibliche, westliche, heterosexuelle Probandengruppe, was die Generalisierbarkeit der Ergebnisse einschränkt. Alle Resultate und daraus abgeleiteten Überlegungen sind vor diesem Hintergrund zu betrachten (siehe 4.3 Limitationen).

51 https://de.findmind.ch/

52 Ein Habitierungseffekt könnte eintreten, wenn z. B. immer zunächst die tiefere und dann die höhere Stimme abgespielt würde.

In den Rating-Aufgaben standen sich jeweils zwei Stimmen gegenüber, die sich um ca. 30 Hz unterschieden. Die Präferenz für eine Stimme wurde durch ein simples A/B-Rating getestet, wofür für die erste Option z. B. EVA 1 (155 Hz) und für die zweite Option EVA 2 (185 Hz) abgespielt wurde.

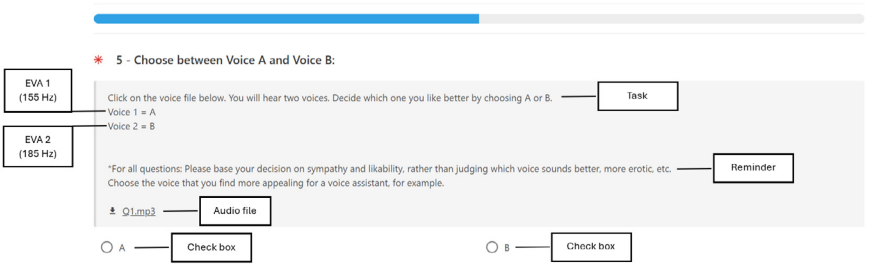


Abbildung 7a: A/B-Rating: Frage 5 (EVA 1 vs. EVA 2)

Die Ranking-Aufgaben wurden für den Frequenzbereich von 240 Hz bis 280 Hz eingesetzt, in welchem der Scheitelpunkt der Sympathie/Grundfrequenz-Korrelation vermutet wurde. Die Proband*innen wurden hier also nicht um ein A/B-Rating gebeten, sondern darum, eine Reihenfolge festzulegen (siehe *Task*, Abb. 7b). In Frage 8 und 10 standen jeweils drei Stimmen zur Auswahl, die einerseits den ‚tieferen' hohen Frequenzbereich (Frage 8: F_0 240 Hz bis 280 Hz | Eva 4, 6, 8) und andererseits den ‚höheren' hohen Frequenzbereich (Frage 10: F_0 260 Hz bis 305 Hz | Eva 6, 8, 9) abfragten, wobei beide die Stimme mit 260 Hz enthielten (EVA 6).

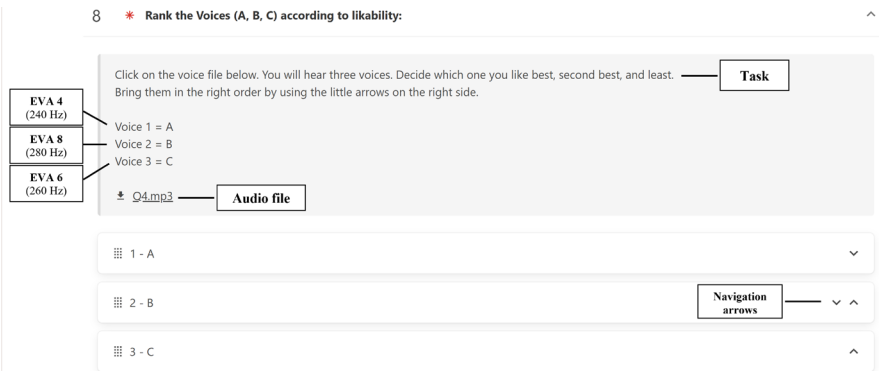


Abbildung 7b: Ranking: Frage 8 (EVA 4, EVA 6, EVA 8)

4.2 Resultate

Experiment 1: Natürliche Stimmen
In der bereits mehrfach erwähnten Studie von Borkowska/Pawloski (2011) bezüglich der vokalen Attraktivität befand sich die optimale Frequenz für weibliche Stimmen bei 260 Hz. In meiner Pilotstudie mit 107 Proband*innen kam ich zum gleichen Ergebnis bezüglich der Korrelation von Sympathie und Grundfrequenz: Die Stimme, deren Grundfrequenz am nächsten bei 260 Hz lag – V1 mit 252 Hz –, wurde von den Proband*innen am sympathischsten empfunden. Weiter zeigte sich in den Ratings, dass die Stimmen mit höheren Grundfrequenzen (V1 = 252 Hz, V2 = 222 Hz) als sympathischer bewertet wurden als die Stimmen mit tieferen Grundfrequenzen (V4 = 192 Hz, V5 = 184 Hz).

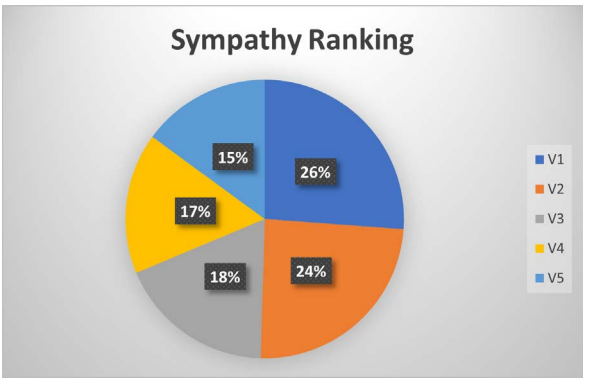


Abbildung 8: Sympathie Ranking Ergebnisse

Und obwohl die Stichprobe mit 107 Teilnehmer*innen nicht repräsentativ für einen größeren Kontext ist, konnten mit dieser Pilotstudie dennoch erste Tendenzen sichtbar gemacht werden, so z. B. die, dass Frauenstimmen mit höherer F_0 als sympathischer wahrgenommen werden als solche mit tieferer F_0. Somit konnte die Hypothese (H_1), dass Frauenstimmen mit hoher Frequenz (F_0>210 Hz) nicht nur als *attraktiver*, sondern auch als *sympathischer* als Frauenstimmen mit tieferer Frequenz (F_0<210 Hz) wahrgenommen werden, verifiziert werden.[53]

53 Wie bereits in Kapitel 2 erwähnt, können auch andere Parameter Einfluss auf die Attraktivität – oder hier die Sympathie – einer Stimme haben (wie z. B. die Intonation). Dies kann in Studien mit natürlichen Stimmen, in welchen ganze Sätze evaluiert werden, jedoch kaum verhindert werden.

Experiment 2: Synthetische Stimmen

Für synthetische Stimmen ließ sich dieselbe Präferenz feststellen wie für natürliche: Die Stimmen mit höherer F_0 (F_0>210 Hz) wurden als sympathischer wahrgenommen als tiefere Stimmen (H_2), bis zu einem bestimmten Punkt (H_3): Analog zur Attraktivität/F_0-Korrelation bei menschlichen Stimmen korreliert auch die Sympathie bei synthetischen Stimmen nicht unendlich linear mit der Grundfrequenz. Vielmehr setzt ab einem gewissen Frequenzbereich ein Scheitelpunkt ein: Bei synthetischen Stimmen setzte dieser aber bereits bei 240 Hz ein. Die Resultate der Ranking-Aufgabe zeigen deutlich, dass die Stimmen, die höher als 240 Hz waren, als *weniger sympathisch* wahrgenommen wurden (siehe Abb. 9):

	Ø	most likable	second most likable	least likable
EVA 4 (240 Hz)	Ø: 1.34 Σ: 435	319 73.33%	84 19.31%	32 7.36%
EVA 8 (280 Hz)	Ø: 2.69 Σ: 435	33 7.59%	68 15.63%	334 76.78%
EVA 6 (260 Hz)	Ø: 1.97 Σ: 435	83 19.08%	283 65.06%	69 15.86%

Abbildung 9: Rankingaufgabe Auswertung (240 Hz–280 Hz)

Hier wird ersichtlich, dass die Sympathie proportional mit dem Abstand zu 240 Hz abnimmt: Während EVA 6 mit 260 Hz noch von 283/435 Personen (~65 %) am zweitsympathischsten empfunden wurde, stimmten bei der noch höheren EVA 8 mit 280 Hz 334/435 Personen (~77 %) darin überein, dass ihnen diese Stimme am wenigsten sympathisch war.

Ein ähnliches Bild stellte sich ein, als EVA 5 (245 Hz) mit der nächsttieferen Frequenz unter 240 Hz (EVA 3, F_0=215 Hz) verglichen wurde: Hier entschied sich ebenfalls eine klare Mehrheit der Proband*innen (~72 %) für die Frequenz bei 245 Hz.

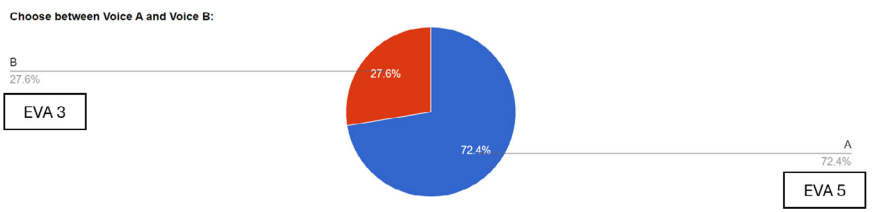


Abbildung 10: A/B Rating (245 Hz vs. 215 Hz)

Demnach kann die Grundfrequenz um 240 Hz als ‚goldene Frequenz' für Sympathie bei synthetischen Stimmen betrachtet werden; Stimmen mit höherer/tieferer F_0 werden in proportionaler Relation zu ihr als weniger sympathisch bewertet.

Die Resultate dieser Studie sind aufschlussreich, da sie die deutlichen Unterschiede in der Sympathiewahrnehmung von synthetischen Stimmen mit unterschiedlichen Grundfrequenzen aufzeigen. In beiden Experimenten weisen sie darauf hin, dass es universelle Präferenzen zu geben scheint, was die ideale Grundfrequenz für Stimmen betrifft. Im direkten Vergleich der beiden Untersuchungen kann festgestellt werden, dass die ‚goldene Frequenz' für synthetische Stimmen jedoch tiefer liegt als für menschliche Stimmen. Bei den menschlichen Stimmen präferierten die Proband*innen dieser Untersuchung die Stimme, die am nächsten bei 260 Hz lag und somit dieselbe Frequenz, die Borkowska/Pawloski (2011) als die attraktivste deklarierten. Da Attraktivität und Sympathie beide positive Wahrnehmungsattribute sind, erscheint es auf den ersten Blick naheliegend, dass wir Stimmen, die wir sympathisch finden, auch als attraktiv empfinden und umgekehrt. Auf weitere Faktoren, die ebenfalls dazu beitragen könnten, dass höhere Stimmen als sympathischer wahrgenommen werden, werde ich in der Diskussion noch genauer eingehen. Dass die ideale Frequenz bei synthetischen Stimmen hingegen tiefer liegt als bei menschlichen, könnte ein erster Hinweis darauf sein, dass wir menschliche und synthetische Stimmen anders wahrnehmen und bewerten. Hier ist weitere Forschung nötig, um die möglichen Ursachen dieser Resultate zu eruieren. Eine Möglichkeit wäre, dass wir synthetische Stimmen, von denen wir annehmen, dass sie zu technischen Entitäten gehören, sympathischer finden, wenn sie nicht zu attraktiv klingen.

4.3 Limitationen

Beide Untersuchungen weisen Limitationen auf, die bei einer erneuten Studie berücksichtigt werden sollten. So wurde bei beiden Experimenten nicht nach der Muttersprache der Proband*innen gefragt. Dies erschien hinsichtlich des Untersuchungszwecks als vernachlässigbar, da die Sympathie einer Stimme auch bewertet werden kann, ohne dass sich einem der semantische Inhalt der Äußerung erschließt.[54] Dennoch wäre es interessant gewesen zu erfahren, ob z. B. deutsche Muttersprachler*innen eine andere Präferenz aufweisen als englische Muttersprach-

54 Die Sympathie oder Attraktivität einer Stimme kann auch bewertet werden, ohne dass sich der linguistische bzw. semantische Gehalt der gehörten Äußerung dem Zuhörer erschließt (vgl. Apicella/Feinberg 2009, 1080). So bewerteten etwa die Probandinnen bei Feinberg et al. (2005) die Attraktivität von männlichen Stimmen rein anhand der Vokale A, E, I, O und U (vgl. Feinberg et al. 2005, 562). Insofern kann davon ausgegangen werden, dass auch Muttersprachler*innen verschie-

ler*innen. Eine weitere, damit verbundene Limitation betrifft die nur begrenzte Vergleichbarkeit der beiden Experimente, da nicht dieselben Stimuli betreffend Sprache und Textumfang verwendet wurden. Da beide Experimente mit ihrem Anspruch, Sympathie bei menschlichen/synthetischen Stimmen zu untersuchen, ein noch nicht ausreichend untersuchtes Forschungsdesiderat darstellen, können sie jedoch auch für sich alleine stehen und müssen nicht zwingend in Relation gesetzt werden. In einem nächsten Schritt sollte dennoch untersucht werden, ob sich dieselben Unterschiede in der Bewertung zwischen menschlichen und synthetischen Stimmen zeigen, wenn dieselben Stimuli verwendet werden. Da der Hauptfokus dieses Artikels jedoch auf synthetischen Stimmen liegt, wurde Wert darauf gelegt, für die zweite Untersuchung eine größere Proband*innengruppe zu rekrutieren, um aussagekräftige Resultate zu generieren. Dies führte zu der Entscheidung, hier einen englischen Fragebogen mit englischen Stimuli zu verwenden.

Weiter sind die hier erhaltenen Resultate vor dem Hintergrund der erwähnten Bias in der Probandengruppe zu verstehen, weshalb eine Generalisierbarkeit der Ergebnisse nur begrenzt möglich ist. Für eine bessere Repräsentativität sollte in weiteren Forschungsprojekten eine größere Heterogenität der Probandengruppe betreffend Gender, Alter, kulturellem Hintergrund und sexueller Orientierung angestrebt werden. Die letzte Einschränkung bezieht sich auf die Auswertungsmethodik, die von *Findmind* ausschließlich in prozentualen Anteilen dargestellt wird. Die Verwendung prozentualer Darstellungen schränkt die Tiefe der Analyse ein und führt zu einer oberflächlicheren Darstellung der Daten. Für diese einfache Art der Datenerhebung (Fragebogen mit A/B-Rating und Ranking-Aufgaben) und im Kontext einer Pilotstudie sind prozentuale Darstellungen jedoch ausreichend geeignet, um erste Präferenzen abzubilden und ein Grundstein für weiterführende Forschung zu legen.

4.4 Diskussion

Im ersten Experiment konnte gezeigt werden, dass Frauenstimmen mit höherer Grundfrequenz als sympathischer wahrgenommen werden als Frauenstimmen mit tieferer Grundfrequenz. Dieses Resultat ist zwar interessant, aber aus zweierlei Gründen auch kaum überraschend: Zum einen liegt es, wie bereits erwähnt, nahe, dass eine Korrelation zwischen attraktiven und sympathischen Stimmen besteht. Zum anderen gibt es gut dokumentierte Sprechpraxen, in denen Sympathie und

dener Sprachen unabhängig davon, ob sie die Äußerung verstehen oder nicht, bewerten können, ob die gehörte Stimme für sie attraktiv bzw. sympathisch klingt.

erhöhte Grundfrequenz eine kardinale Rolle spielen: Die *infant-directed speech* (IDS/'motherese') und die *pet-directed speech* (PDS), also die Art, wie Menschen mit Babys oder Haustieren sprechen. Beide weisen als charakteristisches Merkmal eine erhöhte Grundfrequenz auf: „[I]nfant- and pet-directed speech are similar and distinctly different from adult-directed speech in terms of heightened pitch and affect" (Burnham et al. 2002, 1435). Wenn also mit Babys bzw. Haustieren gesprochen wird, neigen Personen intuitiv dazu, ihre Stimmen zu erhöhen (vgl. Graddol/Swann 1989, 19). Studien zeigen, dass die IDS von Kindern gegenüber der ADS (adult-directed speech) präferiert wird (vgl. Saint-Georges et al. 2013, 8), und diese Präferenz liegt unter anderem an der erhöhten Grundfrequenz: „[I]t is the fundamental frequency characteristics of infant-directed speech that account for the infant listening preference for motherese" (Fernald/Kuhl 1987, 291). Wie das erste Experiment mit menschlichen Sprecherinnen gezeigt hat, bevorzugen jedoch nicht nur Kinder und Tiere diese erhöhte Grundfrequenz, sondern auch Erwachsene.

Im zweiten Experiment konnte gezeigt werden, dass die Modifizierung der Grundfrequenz einen starken Einfluss auf die Sympathie-Wahrnehmung hat und dass höhere Stimmen – bis zum Scheitelpunkt um 240 Hz – als sympathischer wahrgenommen werden als tiefere Stimmen. Dieses Resultat ist besonders für virtuelle Assistenzsysteme bedeutend, die aktuell noch über keinen (humanoiden) Körper verfügen und deren einziges Instrument für die Interaktion und folglich, um Sympathie zu evozieren, ihre Stimme ist.

Im abschließenden Kapitel werde ich nun noch auf das Potenzial von sympathischen synthetischen Stimmen für die Mensch-Maschine-Interaktion im größeren Kontext, fernab von Sprachassistenten, eingehen und aufzeigen, in welchen Einsatzbereiche sympathische synthetische Stimmen großes Potenzial besitzen.

5 Ausblick

Stimmen können einen großen Einfluss auf die Mensch-Mensch- und die Mensch-Maschine-Kommunikation haben und sympathische Stimmen folglich die Interaktion in beiden Fällen verbessern. Mit Bezug auf den *Halo-Effekt* und den *vocal-attractiveness stereotype* wurde bereits aufgezeigt, welche Vorteile Personen mit sympathischen Stimmen in der zwischenmenschlichen Interaktion zugutekommen. Zum Schluss möchte ich nun aber noch auf die Vielzahl an möglichen Einsatzbereichen für sympathische synthetische Stimmen in der Mensch-Maschine-Interaktion zu sprechen kommen, in denen nun der *Hörer* und nicht der Sprecher profitiert. Die Bereiche, in denen sympathische Stimmen von Nutzen sein könnten, gliedern sich wie folgt:

(I) Pflegebereich:
In der Gesundheitspflege könnte der Einsatz von sympathischen Stimmen von großem Vorteil sein, wenn Pflegeroboter (siehe *Robear*)[55] oder soziale Roboter (siehe *Pepper*)[56] zur Pflege und Unterhaltung von Patient*innen und Bewohner*innen in Krankenhäusern und Altersheimen eingesetzt werden, wie es bereits vereinzelt der Fall ist, so z. B. im Altersheim Ehrendingen (Schweiz) (vgl. Moser 2023). In diesem Altersheim wurde Pepper bereits für Unterhaltung(en) von und mit Senior*innen eingesetzt und erfreut sich großer Beliebtheit; eine Bewohnerin spricht sich gar gegen den Begriff ‚Roboter' aus, und besteht darauf, dass Pepper „etwas ganz Spezielles sei" und sie ihn „gern habe" (vgl. Moser 2023). Und Sympathie ist im Pflegebereich ein wichtiges Kriterium, da Patient*innen nicht nur medizinische Pflege benötigen, sondern auch Interaktion, Aktivierung und Gespräche. Und in diesen Bereichen können soziale Roboter Hilfe leisten: So belegen etwa zahlreiche Studien, dass die Interaktion mit dem Roboter *Paro*, der wie eine Kuschelrobbe geformt und mit Sensoren ausgestattet ist, deutlich dabei hilft, die Lebensqualität von demenzkranken Patienten zu erhöhen und ihre Schmerzen und Ängste zu mildern (vgl. Geva et al. 2022; Moyle et al. 2013; Wang et al. 2021; Stapels/Eyssel 2021). Und dies, obwohl Paro, anders als z. B. Pepper, noch nicht mal sprechen kann. Das könnte sich aber bald ändern, denn dank des raschen Fortschritts im Bereich der generativen KI ist es inzwischen möglich, Large Language Models wie ChatGPT in soziale Roboter zu integrieren (vgl. Chen et al. 2024), was die Mensch-Maschine-Interaktion, und besonders die -*Kommunikation*, auf ein völlig neues Niveau hebt. Die gleichzeitige Integration eines Sprachmodell wie ChatGPT-4 Turbo und einer sympathischen Stimme könnte dabei helfen, soziale Roboter wie Pepper und Co. zu besseren Gesprächspartnern für alte und kranke Personen zu machen, und die Interaktionsmöglichkeiten zwischen Menschen und Maschinen in solchen Settings weiter fördern.[57]

55 Der 140 Kilo schwere *Robear* ist ein japanischer Roboter, der im Äußeren einem Bären nachempfunden ist und dafür konzipiert wurde z. B. Patienten umzubetten oder in Rollstühle zu heben (vgl. Szondy 2015).

56 *Pepper* ist ein sozialer, humanoider Roboter, der in diversen Bereichen – wie an der Rezeption zur Begrüßung von Kund*innen oder im Altersheim zur Unterhaltung von Bewohner*innen – eingesetzt werden kann. Siehe: https://www.probo-robotics.at/ <27.03.2025>.

57 An dieser Stelle möchte ich betonen, dass soziale Roboter mit künstlicher Intelligenz in solchen Kontexten nicht als Ersatz von menschlichem Pflegepersonal gedacht sind, sondern als ein ergänzendes Angebot.

(II) Psychotherapie:
Neue KI-Bots wie der *Mindsum Bot*[58] oder *JungGPT*,[59] die auf maschinellem Lernen basieren und mit entsprechenden Daten trainiert werden, agieren bereits jetzt als ‚Taschen-Psychologen' und trösten, beraten und helfen Menschen bei ihren Problemen – und das jederzeit auf Knopf- bzw. Tastendruck. Werden solche Bots nun um eine TTS-Funktion (TTS=Text to Speech) erweitert, können sie ihre schriftlichen Antworten zusätzlich mit sympathischer Stimme verbal vortragen. Dies würde nicht nur im Sinne der Barrierefreiheit die Anwendung solcher Apps für Menschen mit Sehbeeinträchtigung ermöglichen, sondern auch ein authentischeres Therapiegefühl vermitteln, da auf diese Weise eine gesprächsbasierte Interaktion hergestellt werden kann. Relativierend muss an dieser Stelle aber betont werden, dass solche Bots, auch wenn sie von den Sprachkompetenzen her Menschen gleichen, in Krisensituationen keinen ausgebildeten Psychologen ersetzen können oder sollen. Dennoch haben sie großes Potenzial: In Zeiten, in denen Psychotherapeut*innen oft lange Wartezeiten für freie Therapieplätze haben oder es Personen aus persönlichen Gründen (lange Arbeitszeiten oder Nachtschichten, rurale Wohnorte etc.) nicht möglich ist, zu einem Therapeuten zu gehen, können KI-Applikationen ein hilfreiches Alternativangebot sein, um zumindest in der Phase der Überbrückung mit einem objektiven Gegenüber über Probleme sprechen zu können. Sensu Baisch/Kolling: „[Es ist] aus ethischen Gesichtspunkten unabdingbar, möglichst vielen Menschen den Zugang zu einer notwendigen Therapie zu bieten, selbst wenn das bedeutet, dass diese von einem Roboter durchgeführt wird" (Baisch/Kolling 2021, 433). Und hier könnte die Anwendung von sympathischen Stimmen, ähnlich wie bei den sozialen Robotern in der Pflege, dazu beitragen, dass die Personen sich wohler fühlen und die Mensch-Maschine-Interaktion positiv beeinflussen.

58 Öffnet man die Webseite des *Mindsum Bot* (www.mindsum.org/chatbot), sieht man Vorschläge, um die Interaktion zu starten, wie z. B. ‚Any tips to manage my OCD?' oder ‚I need help with my depression', die angewählt werden können. Es besteht aber auch die Möglichkeit, eigene Themen und Anliegen mit der KI zu diskutieren. Als ich in einem Selbstversuch für diesen Artikel Unzufriedenheit über mein Gewicht geäußert habe, erhielt ich vom Mindsum Bot als Antwort: „I'm sorry to hear that you're feeling this way. It's important to remember that your worth is not determined by your appearance." Zusammen mit der Aufforderung, mehr zu erzählen: „How long have you been feeling this way?" (Stand Mai 2024).

59 ‚Jung' im Namen dieses Bots verweist vermutlich auf den berühmten Psychologen Carl G. Jung, während GPT für ‚Generative Pre-trained Transformer' steht und ein fortschrittliches KI-Modell bezeichnet, das Texte lesen, verstehen und selbst generieren kann, bekannt geworden durch Open AIs ChatGPT (https://jung-gpt.com/).

(III) Kundenservice und Assistenzsysteme:
Sympathische Stimmen haben das Potenzial, ein ansprechenderes und angenehmeres Benutzererlebnis in verschiedenen HCI-Anwendungen zu kreieren. Kundenservice-Chatbots oder interaktive virtuelle Avatare, die auf Websites weiterhelfen, können mit solchen Stimmen ausgestattet werden. So kann einerseits händefrei kommuniziert werden, während gleichzeitig andere Aufgaben am PC erledigt werden können. Andererseits können durch sympathische Stimmen negative Emotionen, wie z. B. Frustration über die Webseite, möglicherweise abgemildert werden. In diesen Bereich fallen auch persönliche Assistenzsysteme wie Sprachassistenten und Navigationsgeräte, mit denen im Alltag interagiert wird.

(IV) Bildungswesen und Medien:
a) E-Learning, Tutorials und Schulungsvideos: Im Bildungs- und Schulwesen könnten sympathische Stimmen eine wichtige Rolle spielen, um Lernprozesse effektiver und ansprechender zu gestalten. In Unternehmen und Bildungseinrichtungen werden Schulungsvideos und Tutorials verwendet, um Mitarbeitende einzuarbeiten oder zu schulen. Auf solchen E-Learning-Plattformen können synthetische Stimmen verwendet werden, um Lerninhalte vorzulesen, Erklärungen abzugeben und Fragen zu beantworten. Eine freundliche und sympathische Stimme kann hier dazu beitragen, dass Lernende motivierter sind, sich mit dem Lernstoff auseinanderzusetzen. Dass die Stimmwahl in E-Learning-Kontexten ein einflussreicher Faktor ist, stellten Mayer et al. (2003) in einer Untersuchung fest, in welcher Lernende mit unterschiedlichen Stimmen (menschliche vs. maschinelle Stimme) in einem Lernsetting interagierten. Das Resultat zeigt, dass die Gruppe, die mit der menschlichen Stimme gelernt hat, „scored statistically significantly higher on learning performance tests than the machine voice group“ (Mayer et al. 2003, 120). Ähnliche Effekte konnten mit einem nativen vs. ausländischen Akzent erzielt werden. Dies zeigt, dass die Stimmwahl im Bereich E-Learning einen starken Einfluss auf Lernmotivation und -erfolg hat, weswegen es plausibel erscheint, anzunehmen, dass ebenfalls eine Korrelation zwischen sympathischen Stimmen und Lernerfolg in solchen Formaten bestehen könnte.

b) Medien wie Sprachlernapps, Radio, Podcasts etc.: Beim Erlernen neuer Sprachen auf Apps wie *Duolingo* oder *Babbel* können sympathische Stimmen verwendet werden, um die richtige Aussprache zu demonstrieren und Dialoge in der Zielsprache vorzulesen. Dies kann hier erneut einen Einfluss auf den Erfolg und vor allem die Motivation, mit der Software zu üben, haben. Stellt man sich den umgekehrten Fall vor (d. h. mit einer Sprachlern-App üben zu müssen, die eine äußerst unangenehme, z. B. schrille oder monotone Stimme verwendet), so ist es plausibel anzunehmen, dass die Motivation der Lernenden geringer ist und sie die App weniger

nutzen. Aus den gleichen Gründen können Radiosender Hörerinnen und Hörer verlieren, wenn sie den Host wechseln und die neue Stimme beim Zielpublikum nicht gut ankommt. Um dieses Risiko bei einem Moderatorenwechsel zu vermeiden, könnten deshalb auch Radiosender in der Zukunft in Betracht ziehen, sympathische KI-Stimmen anstelle menschlicher Moderatoren die Nachrichten verkünden und das Wetter ansagen zu lassen.[60] Analog zum Radio können sympathische, synthetische Stimmen für jegliche Formen von auditiven Medien eingesetzt werden, darunter auch die Vertonung von Audiobooks, Kinderhörbüchern und Podcasts.

Während die hier erwähnten Einsatzgebiete (I) und (II) eher soziale Komponenten enthalten und das individuelle wie auch das kollektive Wohl einzelner Personengruppen fördern können, so sind (III) und (IV) auch aus wirtschaftlicher Sicht interessant. Auf jeden Fall können wir festhalten: Es gibt eine breite Palette möglicher Anwendungsbereiche für sympathische synthetische Stimmen in der Mensch-Maschine-Interaktion, die weit darüber hinausgehen, Sprachassistenten sympathischer klingen zu lassen. Und obwohl auch kritische Fragestellungen bei der Konzeption synthetischer Stimmen mitbedacht werden müssen, insbesondere wenn es sich dabei um weibliche Stimmen oder Stimmen für weiblich attribuierte Maschinen handelt, so ist es trotzdem wünschenswert, weiter zu erforschen, wie synthetische Stimmen für Maschinen optimiert werden können. Aktuell ist dieses Desiderat noch lange nicht ausgeschöpft. In Anbetracht ihres Potenzials sollte Stimmen deshalb in Bezug auf die Mensch-Computer-Interaktion (z. B. Mensch und virtueller Assistent) und auch auf die Mensch-Roboter-Interaktion (z. B. Mensch und sozialer Roboter) künftig mehr Bedeutung beigemessen werden, als dies in der Vergangenheit der Fall war. Denn obwohl TTS-Modelle bereits seit über 50 Jahren verwendet werden, sprechen Roboter erst seit den letzten fünfzehn Jahren mit menschlich klingenden und nicht mehr roboterhaften Stimmen (vgl. Klatt 1987; Zen et al. 2009). Das mag daran liegen, dass lange Zeit der (einzige) Fokus darauf ausgerichtet war, *was* der Roboter sagt (Ebene der Semantik, Grammatik und Syntax), und nicht, *wie* er es sagt (Ebene der Pragmatik und Interaktion). Der vorliegende Aufsatz beabsichtigt, mit den durchgeführten Studien die Gegenperspektive zu stärken und die Relevanz pragmatischer Interaktionsaspekte, wie der Stimme, für die Wahrnehmung des Gegenübers – sei es nun ein Mensch, ein sozialer Roboter oder ein Sprachassistent – hervorzuheben.

60 Dieser Vorschlag mag zwar zunächst wie Science Fiction klingen, doch er ist seit August 2023 bereits Realität geworden – zumindest für zwei Radiosender in Devon, England (vgl. RadioTalk UK 2023).

Literatur

Apicella, Coren; Feinberg, David (2009): Voice pitch alters mate-choice-relevant perception in hunter-gatherers. In: Proceedings of the Royal Society – Biological Sciences 276/1659, 1077–1082.

Baisch, Stefanie; Kolling, Thorsten (2021): Roboter in der Therapie. In: Bendel, Oliver (Hrsg.): Soziale Roboter. Technikwissenschaftliche, wirtschaftswissenschaftliche, philosophische, psychologische und soziologische Grundlagen. Wiesbaden: Springer, 417–440.

Bartneck, Christoph; Belpaeme, Tony; Eyssel, Friederike; Kanda, Takayuki; Keijsers, Merel; Šabanović, Selma (2020): Human-Robot Interaction. An Introduction. Cambridge: University Press.

Bartsch, Silke (2008): "What sounds beautiful is good?" How employee vocal attractiveness affects customer's evaluation of the voice-to-voice service encounter. In: Bernd, Stauss (Hrsg.): Aktuelle Forschungsfragen im Dienstleistungsmarketing. Wiesbaden: Gabler.

Berg, Martin; Fuchs, Michael; Wirkner, Kerstin; Loeffler, Markus; Engel, Christoph; Berger, Thomas (2017): The Speaking Voice in the General Population: Normative Data and Associations to Sociodemographic and Lifestyle Factors. In: Journal of Voice 31/2, 257.e13–257.e24.

Borau, Sylvie; Otterbring, Tobias; Laporte, Sandra; Wamba, Samuel F. (2021): The most human bot: Female gendering increases humanness perceptions of bots and acceptance of AI. In: Psychology of Marketing 38, 1052–1068.

Borkowska, Basia; Pawloski, Boguslaw (2011): Female voice frequency in the context of dominance and attractiveness perception. In: Animal Behaviour 82/1, 55–59.

Browning, Frank (2008): Does Obama's baritone give him an edge? Online unter: https://www.salon.com/2008/02/28/obama_clinton_voices/ <27.03.2025>.

Burnham, Denis; Kitamura, Christine; Vollmer-Conna, Ute (2002): What's New, Pussycat? On Talking to Babies and Animals. In: Sciene Magazine 296, 1435.

Cambre, Julia; Kulkarni, Chinmay (2019): One Voice Fits All? Implications and Research Challenges of Designing Voices for Smart Devices. In: Proceedings of the ACM Human-Computer Interaction 3/223, 223:1–223:19.

Cameron, Deborah (1986): Feminism and Linguistic Theory. London: Macmillan.

Carney, Dana R.; Colvin, C. Randall; Hall, Judith A. (2007): A thin slice perspective on the accuracy of First impressions. In: Journal of Research in Personality 41, 1054–1072.

Chen, Yiaohui; Luo, Katherine; Gee, Trevor; Nejati, Mahla (2024): Does ChatGPT and Whisper make humanoid robots more relatable? Online unter: https://arxiv.org/abs/2402.07095 <27.03.2025>.

Collins, Sarah A. (2000): Men's voices and women's choices. In: Animal Behaviour 60/6, 773–780.

Collins, Sarah A.; Missing, Caroline (2003): Vocal and visual attractiveness are related in women. In: Animal Behaviour 65, 997–1004.

Cussigh, Giacomo; Ballester-Arnal, Rafael; Gil-Llario, María D.; Giménez-Carcía, Cristina; Castro-Calvo, Jesus (2020): Fundamental frequency of the female's voice: A cross-country empirical study on its influence on social and sexual selection. In: Personality and Individual Differences 160, 1–7.

Dabbs, James M. Jr.; Mallinger, Alison (1999): High testosterone levels predict low voice pitch among men. In: Personality and Individual Differences 27/4, 801–804.

Danielescu, Andreea (2020): Eschewing Gender Stereotypes in Voice Assistants to Promote Inclusion. CUI '20: Proceedings of the 2nd Conference on Conversational User Interfaces, Article 46, 1–3.

Degelo, Julia (2021): Der wütende Mann, die höfliche Frau – und die Frage nach dem Dazwischen. Wie spricht eine genderneutrale Sprachassistenz? In: Brommer, Sarah/Dürscheid, Christa (Hrsg.): Mensch-Maschine-Kommunikation. Beiträge zur Medienlinguistik. Tübingen: Narr Francke Attempto, 211–225.

Eyssel, Friederike; Kuchenbrandt, Dieta; Bobinger, Simon; de Ruiter, Laura; Hegel, Frank (2012): 'If You Sound Like Me, You Must Be More Human': On the Interplay of Robot and User Features on Human-Robot Acceptance and Anthropomorphism. 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Boston USA. Online unter: https://ieeexplore.ieee.org/document/6249487 <27.03.2025>.

Feinberg, David R.; DeBruine, Lisa M.; Jones, Benedict C.; Perrett, David I. (2008): The role of femininity and averageness of voice pitch in aesthetic judgments of women's voices. In: Perception 37, 615–623.

Feinberg, David R.; Jones, Benedict C.; Law-Smith, Miriam J.; Moore, Fhionna R.; DeBruine, Lisa M.; Cornwell, Robin E.; Hillier, Stephan G.; Perret, David I. (2006): Menstrual cycle, trait estrogen level, and masculinity preferences in the human voice. In: Hormones and Behavior 49/2, 215–222.

Feinberg, David R.; Jones, Benedict C.; Little, Anthony C.; Burt, Michael; Perrett, David I. (2005): Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. In: Animal Behaviour 69/3, 561–568.

Fernald, Anne; Kuhl, Patricia (1987): Acoustic Determinants of Infant Preference for Motherese Speech. In: Infant Behavior and Development 10, 279–293.

Fink, Julia (2012): Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction. In: Ge, Shuzhi Sam; Khatib, Oussama; Cabibihan, John-John; Simmons, Reis; Williams, Mary-Ann (Hrsg.): Social Robots. 4th International Conference, ICSR 2012, Chengdu, China, Proceedings, 199–208.

Fouquet, Meddy; Pisanski, Katarzyna; Mathevon, Nicolas; Reby, David (2016): Seven and Up: Individual differences in male voice fundamental frequency emerge before puberty and remain stable throughout adulthood. In: Royal Society – Open Science 3/160395, 1–9.

Fraccaro, Pauk J.; Jones, Benedict C.; Vukovic, Jovana; Smith, Finlay. G.; Watkins, Christopher. D.; Feinberg, David R.; Little, Anthony. C.; DeBruine, Lisa M. (2011): Experimental evidence that women speak in a higher voice pitch to men they find attractive. In: Journal of Evolutionary Psychology 9/1, 57–67.

Geva, Nirit; Hermoni, Netta; Levy-Tzedek, Shelly (2022): Interaction Matters: The Effect of Touching the Social Robot PARO on Pain and Stress is Stronger When Turned ON vs. OFF. In: Frontiers in Robotics and AI 9/926185, 1–14.

González-Alvarez, Julio; Cervera-Crespo, Teresa; Miralles, José Luis (2002): Análisis acústico de la voz: Fiabilidad de un conjunto de parámetros multidimensionales. In: Acta Otorrinolaringológica Española 53/4, 256–268

Goodman, Kylie L.; Mayhorn, Christopher B. (2023): It's not what you say but how you say it: Examining the influence of perceived voice assistant gender and pitch on trust and reliance. In: Applied Ergonomics 106, 1–8.

Graddol, David; Swann, Joan (1989): Gender Voices. Oxford: Blackwell Publishers.

Griggs, Brandon (2011): Why computer voices are mostly female. Interview with Stanford University Professor Clifford Nass. CNN, Friday October 21. Online unter: https://edition.cnn.com/2011/10/21/tech/innovation/female-computer-voices/index.html <27.03.2025>.

Herder, Eelco; Herden, Sven (2023): Context-Dependent Use of Authority and Empathy in Lifestyle Advices Given By Persuasive Voice Assistants. In: UMAP '23 Adjunct: Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization, 132–139.

Hermann, Ursula (1983): Knaurs etymologisches Lexikon: 10'000 Wörter unserer Gegenwartssprache; Herkunft und Geschichte. München: Droemer.

Hughes, Susan M.; Puts, David A. (2021): Vocal modulation in human mating and competition. In: Philosophical Transaction Royal Society B 376/20200388, 1–10.

Hughes, Susan M.; Dispenza, Franco; Gallup, Gordon G. Jr. (2004): Ratings of voice attractiveness predict sexual behavior and body configuration. In: Evolution and Human Behavior 25, 295–304.

Jessen, Michael; Köster, Olaf; Gfroerer, Stefan (2003): Effect of increased vocal effort on average and range of fundamental frequency in a sample of 100 German-speaking male subjects. In: Conference Paper for the 15th International Congress of Phonetic Sciences (ICPhS-15). Online unter: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/papers/p15_1623.pdf <27.03.2025>.

Johanson, Deborah L.; Ahn, Ho Seok; Sutherland, Craig J.; Brown, Bianca; MacDonald, Bruce A.; Lim Jong Yoon; Ahn, Byeong Kyu; Broadbent, Elizabeth (2020): Smiling and use of first-name by a healthcare receptionist robot: Effects on user perceptions, attitudes, and behaviours. In: Paladyn, Journal of Behavioral Robotics 2020/11, 40–51.

Jones, Benedict C.; Feinberg, David R.; DeBruine, Lisa M.; Little, Anthony C.; Vukovic, Jovana (2010): A domain-specific opposite-sex bias in human preferences for manipulated voice pitch. In: Hormones and Behavior 106, 122–134.

Jones, Benedict C.; Feinberg, David R.; DeBruine, Lisa M.; Little, Anthony C.; Vukovic, Jovana (2008): Integrating cues of social interest and voice pitch in men's preferences for women's voices. In: Biology Letters 4, 192–194.

Kiese-Himmel, Christiane (2016): Körperinstrument Stimme. Grundlage, psychologische Bedeutung, Störung. Heidelberg: Springer.

Klatt, Dennis H. (1987): Review of text-to-speech conversion for English. In: Journal of the Acoustical Society of America 82/3, 737–793.

Kleinberg, Sara (2018): 5 ways voice assistance is shaping consumer behavior. In: Future of Marketing. Online unter: https://www.thinkwithgoogle.com/future-of-marketing/emerging-technology/voice-assistance-consumer-experience/ <27.03.2025>.

Klofstad, Casey A.; Anderson, Rindy C.; Nowicki, Stephen (2015): Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices. In: PLOS ONE 10/8, 1–14.

Könau, Steffen (2019): Siri will keine Schlampe sein: Sprach-Assistenten zementieren Sexismus. In: Mitteldeutsche Zeitung. Online unter: https://www.mz.de/deutschland-und-welt/wirtschaft/siri-will-keine-schlampe-sein-sprach-assistenten-zementieren-sexismus-1574711 <27.03.2025>.

Kovačić, Damir; Balaban, Evan (2009): Voice gender perception by cochlear implantees. In: Journal of the Acoustical Society 126, 762–775.

Krämer, Nicole; Kopp, Stefan; Becker-Asano, Christian; Sommer, Nicole (2013): Smile and the world will smile with you – The effects of a virtual agent's smile on unser's evaluation and behavior. In: International Journal of Human-Computer Studies 71, 335–349.

Laeri, Patrizia (2019): Siri, eine dumme Kuh? In: Emma. Online unter: https://www.emma.de/artikel/siri-eine-dumme-kuh-336573 <27.03.2025>.

Lange, Benjamin P.; Bögemann, Hannah; Zaretsky, Eugen (2017): Ästhetische Dimensionen von Sprache, Sprechen, Stimme. In: Schwender, Clemens; Lange, Benjamin P.; Schwarz, Sascha (Hrsg.): Evolutionäre Ästhetik. Lengerich: Pabst Science Publishers, 225–246.

Leaderbrand, Katie; Morey, Ashley; Tuma, Lisa (2008): The Effects of Voice Pitch on Perceptions of Attractiveness: Do You Sound Hot or Not? Winona State University Psychology Student Journal 6, January 2008.

Mavica, Lauren W.; Barenholtz, Elan (2012): Matching voice and face identity from static images. In: Journal of Experimental Psychology: Human Perception and Performance 39/2, 307–312.

Mayer, Richard E.; Sobko, Kristina; Mautone, Patricia D. (2003): Social cues in multimedia learning: Evidence for dual processing systems in working memory. Journal of Educational Psychology 95, 419–425.

McAleer, Phil; Todorov, Alexander; Belin, Pascal (2014): How do you say 'hello'? Personality impressions from brief novel voices. In: PLOS ONE 9/3, 1–9.

McGinn, Conor; Torre, Ilaria (2019): Can you Tell the Robot by the Voice? An Exploratory Study on the Role of Voice in the Perception of Robots. 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Daegu, South Korea, 211–221.

Miyake, Kunitate; Zuckerman, Miron (1993): Beyond personality impressions: effects of physical and vocal attractiveness on false consensus, social comparison, affiliation, and assumed and perceived similarity. In: Journal of Personality 61, 411–437.

Moore, Roger K. (2017): Appropriate Voices for Artefacts: Some Key Insights. In: 1st International Workop on Vocal Interactivity in-and-between Humans, Animals and Robots. Online unter: https://vihar-2017.vihar.org/assets/papers/VIHAR-2017_paper_8.pdf <27.03.2025>.

Mori, Masahiro (2012): The Uncanny Valley. Translated by Karl F. MacDorman and Norri Kageki. In: IEEE Robotics & Automation Magazine 19/2, 98–100.

Moser, Alex (2023): Pepper, der Unterhaltungsroboter im Altersheim. SRF-Beitrag. Online unter: https://www.srf.ch/audio/echo-der-zeit/pepper-der-unterhaltungsroboter-im-altersheim?partId=12466635 <27.03.2025>.

Moyle, Wendy; Cooke, Marie; Beattie, Elizabeth; Jones, Cindy; Klein, Barbara; Cook, Glenda; Gray, Chrystal (2013): Exploring the effect of companion robots on emotional expression in older adults with dementia: a pilot randomized controlled trial. In: Journal of Gerontological Nursing 39/5, 46–53.

Nass, Clifford; Steuer, Jonathan; Tauber Ellen R. (1994): Computer are Social Actors. In: CHI '94, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 72–78.

Natour, Yaser S.; Wingate, Judith M. (2009): Fundamental frequency characteristics of Jordanian Arabic speakers. In: Journal of Voice 23/5, 560–566.

Owren, Michael J.; Bachorowski, Jo-Anne (2007): Measuring emotion-related arousal. In: Coan, James A./Allen, John J. B. (Hrsg.): Handbook of emotion elicitation and assessment, 239–266.

Pietronudo, Eleonora (2018): „Japanese women's language" and artificial intelligence: Azuma Hikari, gender stereotypes and gender norms. Masterarbeit der Università Ca' Foscari Venezia. Online unter: https://www.academia.edu/79531692/Japanese_womens_language_and_artificial_intelligence_Azuma_Hikari_gender_stereotypes_and_gender_norms <27.03.2025>.

Puts, David A.; Apicella, Coren L.; Cárdenas, Rodrigo A. (2012): Masculine voices signal men's threat potential in forager and industrial societies. In: Proceedings of the Royal Society – Biological Sciences 279/1728, 601–609.

RadioTalk UK (2023): Devon radio stations switch to AI voices for news bulletins. In: Radio Today. Online unter: https://radiotoday.co.uk/2023/09/devon-radio-stations-switch-to-artificial-intelligence-newsreaders/ <27.03.2025>.

Re, Daniel E.; O'Connor, Jillian J. M.; Bennett, Patrick J.; Feinberg, David R. (2012): Preferences for Very Low and Very High Voice Pitch in Humans. In: PLOS ONE 7/3, 1–8.

Reddit (2024): How often do you tell Alexa she is useless? Online unter: https://www.reddit.com/r/amazonecho/comments/hndngn/how_often_do_you_tell_alexa_she_is_useless/?rdt=56778 <27.03.2025>.

Reeves, Byron; Nass, Clifford I. (1996): The media equation: How people treat computers, television, and new media like real people and places. Center for the Study of Language and Information; Cambridge: Cambridge University Press.

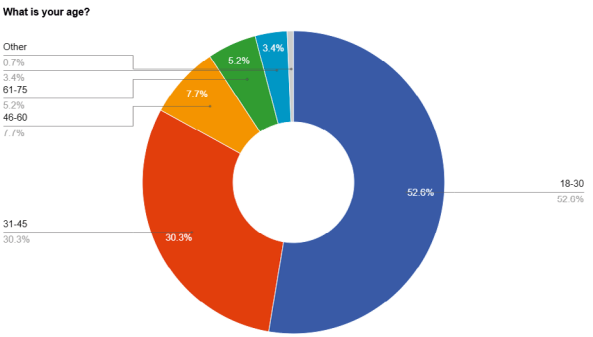
Saint-Georges, Catherine; Chetouani, Mohamed; Cassel, Raquel; Apicella, Fabio; Mahdhaoui, Ammar; Muratori, Filippo; Laznik, Marie-Christine; Cohen, David (2013): Motherese in Interaction: At the Cross-Road of Emotion and Cognition? A Systematic Review. In: PLOS ONE 8/10, 1–17.

Shang, Junchen; Liu, Zhihui (2022): Vocal Attractiveness Matters: Social Preferences in Cooperative Behavior. In: Frontiers in Psychology 13/877530, 1–12.
Spektrum.de – Lexikon der Psychologie (2024): Halo-Effekt. Online unter: https://www.spektrum.de/lexikon/psychologie/halo-effekt/6232 <27.03.2025>.
Stapels, Julia G.; Eyssel, Friederike (2021): Einstellungen gegenüber sozialen Robotern. In: Bendel, Oliver (Hrsg.): Soziale Roboter. Technikwissenschaftliche, wirtschaftswissenschaftliche, philosophische, psychologische und soziologische Grundlagen. Wiesbaden: Springer, 231–250.
Stern, Julia; Schild, Christoph; Jones, Benedict C.; DeBruine, Lisa M.; Hahn, Amanda; Puts, David A.; Zettler, Ingo; Kordsmeyer, Tobias L.; Feinberg, David; Zamfir, Dan; Penke, Lars; Arslan, Ruben C. (2021): Do voices carry valid information about a speaker's personality? In: Journal of Research in Personality 92, 1–14.
Sondhi, Savita; Khan, Munna; Vijay, Ritu; Salhan, Ashok K. (2015): Vocal Indicators of Emotional Stress. In: International Journal of Computer Applications (0975-8887) Volume 122, 38–43.
Sorokowski, Piotr; Puts, David; Johnson, Janie; Żółkiewicz, Olga; Oleszkiewicz, Anna; Sorokowska, Agnieszka; Kowal, Marta; Borkowska, Barbara; Pisanski, Katarzyna (2019): Voice of authority: professionals lower their vocal frequencies when giving expert advice. In: Journal of Nonverbal Behavior 43, 257–269.
Szondy, David (2025): Robear robot care bear designed to serve Japan's aging population. In: New Atlas, Robotics. Online unter: https://newatlas.com/robear-riken/36219/ <27.03.2025>.
Tanner, Alexandra; Schulze, Hartmut; Rüegg, Michelle; Urech, Andreas (2021): Empathie und Emotion: Können sich soziale Roboter empathisch verhalten? In: Bendel, Oliver (Hrsg.): Soziale Roboter. Technikwissenschaftliche, wirtschaftswissenschaftliche, philosophische, psychologische und soziologische Grundlagen. Wiesbaden: Springer, 325–341.
Tolmeijer, Suzanne; Zierau, Naim; Janson, Andreas; Wahdatehagh, Jalil; Bernstein, Abraham (2021): Female by Default? – Exploring the Effect of Voice Assistant Gender and Pitch on Trait and Trust Attribution. Conference on Human Factors in Computing Systems (CHI-EA), Yokohama, Japan.
Traunmüller, Hartmut; Eriksson, Anders (1995): The frequency range of the voice fundamental in the speech of male and female adults. Online unter: https://www.researchgate.net/publication/240312210_The_frequency_range_of_the_voice_fundamental_in_the_speech_of_male_and_female_adults <27.03.2025>.
Ujvary, Laszlo P.; Chirilă, Magdalena; Tiple, Christina; Maniu, Alma A.; Pop, Septimiu S.; Blebea, Cristina M.; Vesa, Stefan; Cosgarea, Marcel (2022): The Effect of Platelet-Rich Plasma Injection on Short Term Vocal Outcomes Following Phonosurgery – A Pilot Study. In: Medicina 58/988, 2–10.
Völkert, Svenja (2012): Das Phänomen „Sexy Stimme". Was macht Männer- und Frauenstimmen sexy? Eine Annäherung an das Thema der stimmlichen Attraktivität. In: Sprechen. Zeitschrift für Sprechwissenschaft. Sprechpädagogik-Sprechtherapie-Sprechkunst 54, 74–87.
Wang, Xinxia; Shen, Jun; Chen, Qiu (2022): How PARO can help older people in elderly care facilities: A systematic review of RCT. In: International Journal of Nursing Knowledge 33, 29–39.
Werbewoche (2019): Virtue und Copenhagen Pride stellen mit Q die weltweit erste genderlose Stimme vor. Online unter: https://www.werbewoche.ch/de/digital/2019-03-12/virtue-und-copenhagen-pride-stellen-mit-q-die-weltweit-erste-genderlose-stimme-vor/ <27.03.2025>.
Zen, Heiga; Tokuda, Keiichi; Black, Alan W. (2009): Statistical parametric speech synthesis. In: Speech Communication 51, 1039–1064.
Zheng, Yi; Compton, Brian J.; Heyman, Gail D.; Jiang, Zhongqing (2020): Vocal attractiveness and voluntarily pitch-shifted voices. In: Evolution and Human Behavior 41, 120–175.
Zuckerman, Miron; Driver, Robert E. (1989): What Sounds Beautiful is Good: The Vocal Attractiveness Stereotype. In: Journal of Nonverbal Behavior 13/2, 67–82.

Anhang:
Synthetische Stimmen (N=435)

1 – What is your age?

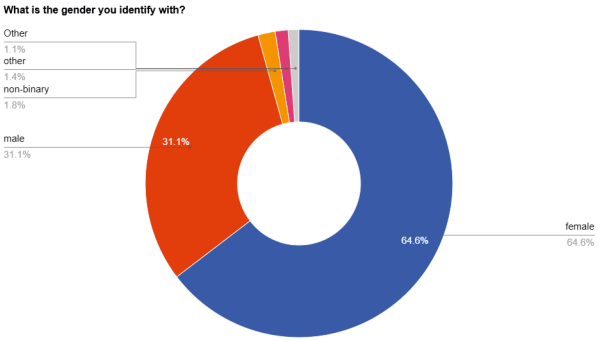
Please note that these questions are for statistical purposes only and cannot be linked to individual participants.
They help generate statements like "most heterosexual men preferred higher/lower voices" or "individuals over 40 showed greater likability towards higher/lower voices."



Option	Prozent	Anzahl
18–30	52.62	231
31–45	30.30	133
46–60	7.74	34
61–75	5.24	23
76–90	0.68	3
-	3.42	15

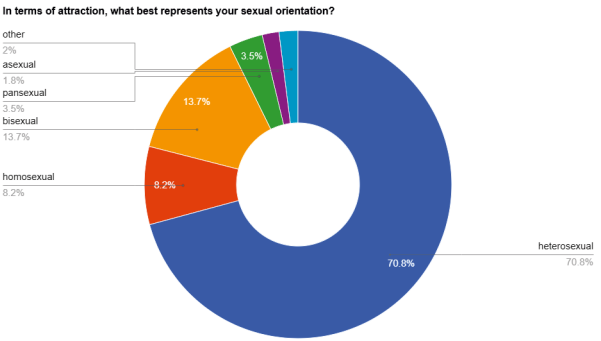
2 – What is the gender you identify with?

Please note that these questions are for statistical purposes only and cannot be linked to individual participants.
They help generate statements like "most heterosexual men preferred higher/lower voices" or "individuals over 40 showed greater likability towards higher/lower voices."



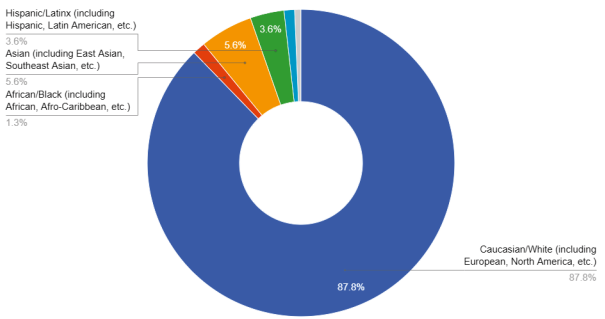
Option	Prozent	Anzahl
female	64.63	285
male	31.07	137
non-binary	1.81	8
transgender male	0.45	2
transgender female	0.45	2
intersex	0.23	1
other	1.36	6

3 – In terms of attraction, what best represents your sexual orientation?
Please note that these questions are for statistical purposes only and cannot be linked to individual participants.
They help generate statements like "most heterosexual men preferred higher/lower voices" or "individuals over 40 showed greater likability towards higher/lower voices."



Option	Prozent	Anzahl
heterosexual	70.80	320
homosexual	8.19	37
bisexual	13.72	62
pansexual	3.54	16
asexual	1.77	8
other	1.99	9

4 – Please select the option that best represents your ethnicity or cultural background:



Option	Prozent	Anzahl
Caucasian/White (including European, North America, etc.)	87.75	394
African/Black (including African, Afro-Caribbean, etc.)	1.34	6
Asian (including East Asian, Southeast Asian, etc.)	5.57	25
Hispanic/Latinx (including Hispanic, Latin American, etc.)	3.56	16
Native American/Indigenous	0.67	3
Other	1.11	5

Reihenfolge der Stimuli:

Frage 5	Voice A = 155 Hz	Voice B = 185 Hz	
Frage 6	Voice A = 245 Hz	Voice B = 215 Hz	
Frage 7	Voice A = 275 Hz	Voice B = 305 Hz	
Frage 8	Voice A = 240 Hz	Voice B = 280 Hz	Voice C = 260 Hz
Frage 9	Voice A = 240 Hz	Voice B = 260 Hz	
Frage 10	Voice A = 305 Hz	Voice B = 260 Hz	Voice C = 280 Hz

5 – Choose between Voice A and Voice B:

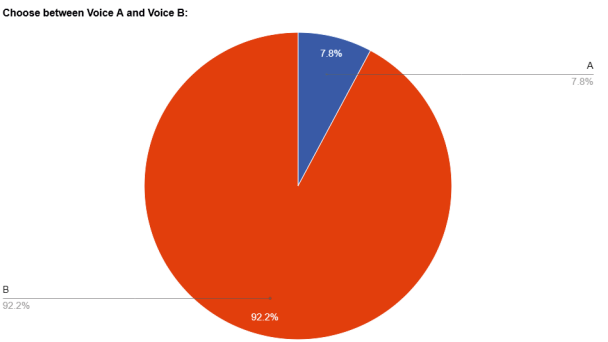
Click on the voice file below. You will hear two voices. Decide which one you like better by choosing A or B.

Voice 1 = A

Voice 2 = B

**For all questions: Please base your decision on sympathy and likability, rather than judging which voice sounds better, more erotic, etc.*

Choose the voice that you find more appealing for a voice assistant, for example.



Option	Prozent	Anzahl
A	7.82	34
B	92.18	401

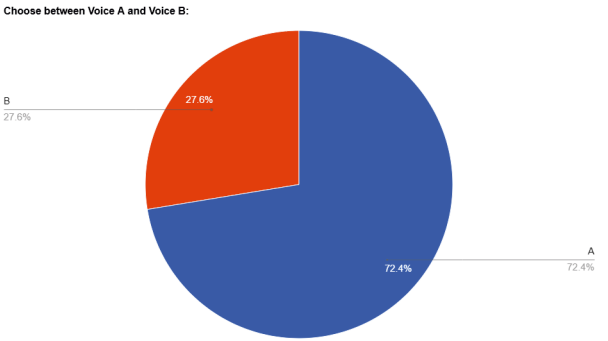
6 – Choose between Voice A and Voice B:

You will hear two voices. Decide which one you like better by choosing A or B.

Voice 1 = A

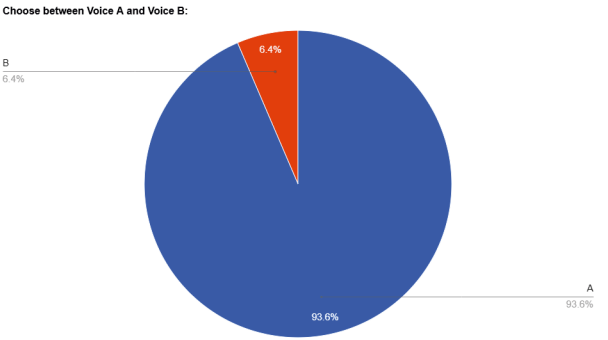
Voice 2 = B

**Please base your decision on sympathy and likability, rather than judging which voice sounds better, more erotic, etc. Choose the voice that you find more appealing for a voice assistant, for example.*



Option	Prozent	Anzahl
A	72.41	315
B	27.59	120

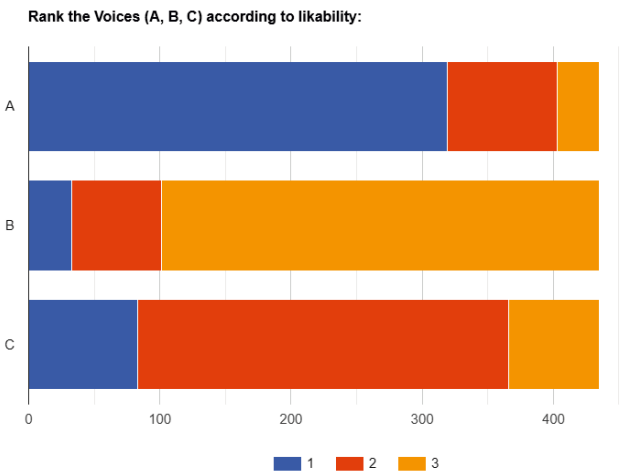
7 – **Choose between Voice A and Voice B:**
Click on the voice file below. You will hear two voices. Decide which one you like better by choosing A or B.
Voice 1 = A
Voice 2 = B



Option	Prozent	Anzahl
A	93.56	407
B	6.44	28

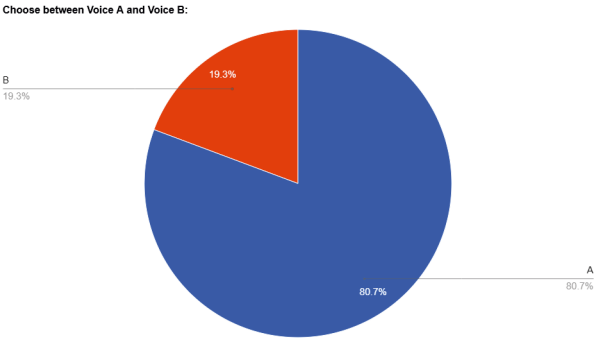
8 – Rank the Voices (A, B, C) according to likability:
Click on the voice file below. You will hear three voices. Decide which one you like best, second best, and least.
Bring them in the right order by using the little arrows on the right side.

Voice 1 = A
Voice 2 = B
Voice 3 = C



	Ø	1	2	3
A	Ø: 1.34	319	84	32
	Σ: 435	73.33 %	19.31 %	7.36 %
B	Ø: 2.69	33	68	334
	Σ: 435	7.59 %	15.63 %	76.78 %
C	Ø: 1.97	83	283	69
	Σ: 435	19.08 %	65.06 %	15.86 %

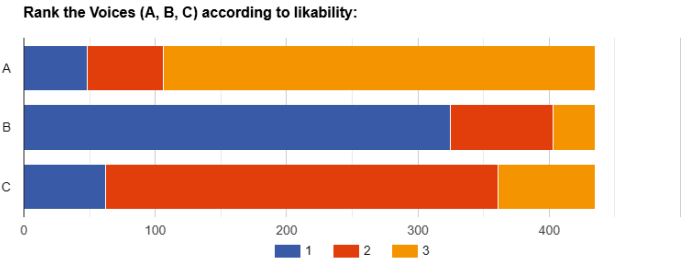
9 – Choose between Voice A and Voice B:
You will hear two voices. Decide which one you like better by choosing A or B.
Voice 1 = A
Voice 2 = B



Option	Prozent	Anzahl
A	80.69	351
B	19.31	84

10 – Rank the Voices (A, B, C) according to likability:
Click on the voice file below. You will hear three voices. Decide which one you like best, second best, and least.
Bring them in the right order by using the little arrows on the right side.

Voice 1 = A
Voice 2 = B
Voice 3 = C



	Ø	1	2	3
A	Ø: 2.65	48	58	329
	Σ: 435	11.03 %	13.33 %	75.63 %
B	Ø: 1.33	325	78	32
	Σ: 435	74.71 %	17.93 %	7.36 %
C	Ø: 2.03	62	299	74
	Σ: 435	14.25 %	68.74 %	17.01 %