AI vs. Real Voice

Dr. Manuel D. S. Hopp, manuel.d.s.hopp@gmail.com

in Zusammenarbeit mit Daniel Crucius

Table of contents

Einleitung 2
Import und Datenbereinigung 2
Samplebeschreibung 2
Alter 3
Geschlecht, Bildung, Audio-Erfahrung und Wirtschaftssektor 3
Zusammenfassung der Samplebeschreibung 7
Präferenzanalyse: Modell 2 vs. Modell 1 vs. Real Voice 7
Modell 2 vs. Modell 1 8
Modell 1 vs. Real Voice 9
Modell 2 vs. Real Voice 10
Zusammenfassung der Präferenzanalyse 10
Qualitätskriterien der Sprecher-Typen 11
Vergleich - Klangqualität 12
Vergleich - Professionalität 13
Vergleich - Natürlichkeit 14
Vergleich - Verständlichkeit 15
Vergleich - Gesamteindruck 16
Zusammenfassung und Schlussfolgerung der Ergebnisse 17

Einleitung

Dieser Bericht präsentiert die Ergebnisse einer Studie, die darauf abzielt, die wahrgenommene Qualität und Präferenz von verschiedenen Sprachgenerierungsmodellen mit einer menschlichen Sprecherin zu vergleichen. Im Fokus stehen dabei zwei KI-basierte Sprachklon-Modelle sowie eine professionelle menschliche Stimme.

Die Studie untersucht die Präferenzen der Probanden in direkten Vergleichen zwischen diesen Stimmen sowie deren Bewertungen hinsichtlich verschiedener Qualitätskriterien wie Klangqualität, Professionalität, Natürlichkeit, Verständlichkeit und Gesamteindruck. Die Analyse basiert auf einem Datensatz von 150 Probanden.

Die Struktur des Berichts gliedert sich wie folgt: Zunächst werden die importierten Daten bereinigt. Anschließend folgt eine Beschreibung des Samples. Danach werden die Präferenzen der Probanden in direkten Vergleichen analysiert. Abschließend werden die Bewertungen der einzelnen Qualitätskriterien detailliert betrachtet und die Ergebnisse zusammenfassend interpretiert.

Import und Datenbereinigung

In diesem Abschnitt wird der Datensatz importiert und für die weitere Analyse vorbereitet.

Samplebeschreibung

Dieser Abschnitt präsentiert deskriptive Statistiken über das Sample der Studie.

Alter

In Table 1 sind die deskriptiven Statistiken zum Alter der Probanden aufgeführt. Die Verteilung des Alters ist in Figure 1 dargestellt.

Table 1: Deskriptive Statistik des Alters der Probanden

	n	mean	sd	min	max	median
year	150	41.53	11.46	15	74	40

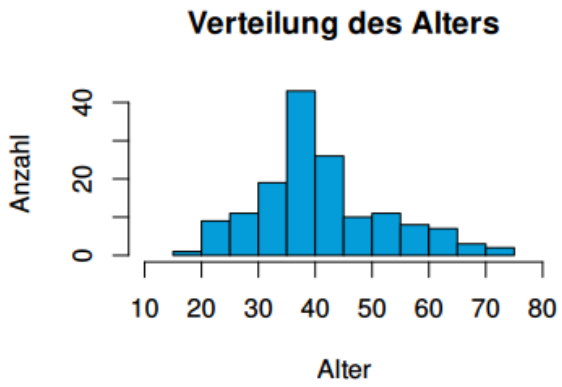

Figure 1: Verteilung des Alters der Probanden

Die Personen sind durchschnittlich um 40 Jahre alt.

Geschlecht, Bildung, Audio-Erfahrung und Wirtschaftssektor

Die folgenden Tabellen und Plots zeigen die Verteilungen der demografischen Merkmale und der Audio-Erfahrung der Probanden.

Table 2: Verteilung des Geschlechts der Probanden (Prozentsätze)

Var1	Freq
Frau	56
Mann	43
Diverse	1

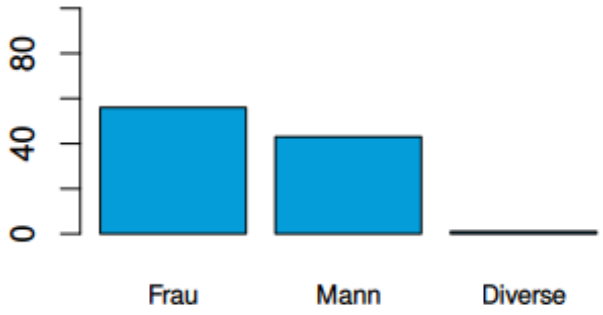

Figure 2: Verteilung des Geschlechts der Probanden

Table 3: Verteilung des Bildungsstands der Probanden (Prozentsätze)

Var1	Freq
Obligatorische Schule / Hauptschulabschluss	1
Sekundarschule / Realschulabschluss	6
Berufslehre / Berufsausbildung	22
Maturität / Abitur	9
Fachhochschulabschluss	31
Universitätsabschluss	24
Promotion / Doktorat	3
Sonstiger Abschluss	3
Keine Angabe	1

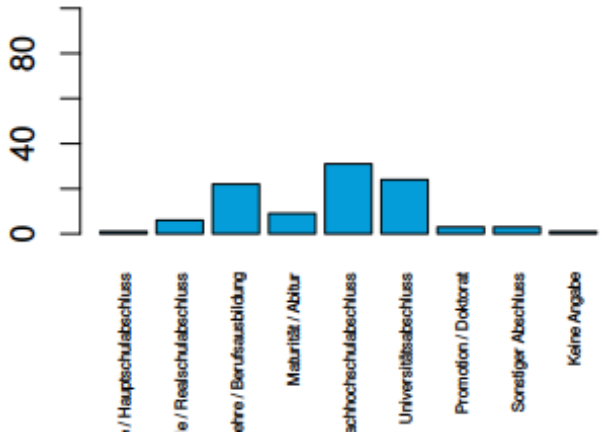

Figure 3: Verteilung des Bildungsstands der Probanden

Table 4: Verteilung der Audio-Erfahrung der Probanden (Prozentsätze)

Var1	Freq
Ja	27
Nein	73

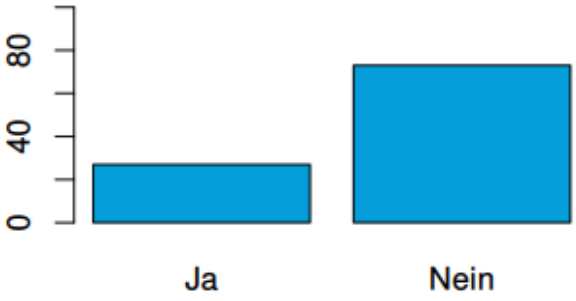

Figure 4: Verteilung der Audio-Erfahrung der Probanden

Table 5: Verteilung des Wirtschaftssektors der Probanden (Prozentsätze)

Var1	Freq
Primärer Sektor (Landwirtschaft, Rohstoffgewinnung)	0
Sekundärer Sektor (Industrie, Handwerk, Produktion)	18
Tertiärer Sektor (Dienstleistungen, Handel)	61
Quartärer Sektor (Information, Forschung, Entwicklung)	11
Derzeit nicht berufstätig	5
Keine Angabe	5

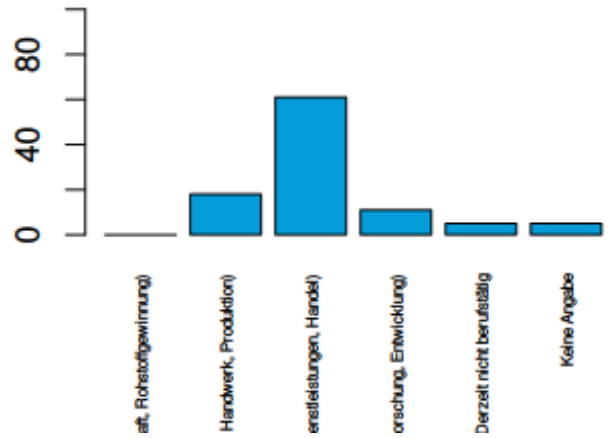

Figure 5: Verteilung des Wirtschaftssektors der Probanden

Wie aus Table 2 und Figure 2 ersichtlich, sind 56% Frauen, 43 % Männer und 1% Divers. Gemäss Table 3 und Figure 3 haben mehr als die Hälfte einen Fachhochschul- oder Universitätsabschluss. Table 4 und Figure 4 zeigen, dass 3/4 keine Erfahrung mit Audio haben. Die meisten (ca. 60%) arbeiten im Tertiären Sektor, wie in Table 5 und Figure 5 dargestellt.

Zusammenfassung der Samplebeschreibung

Die Gruppe besteht hauptsächlich aus Personen im Alter von etwa 40 Jahren, wobei Frauen mit 56 % die Mehrheit bilden. Ein Großteil der Gruppe (über 50 %) verfügt über einen Fachhochschul- oder Universitätsabschluss. Auffällig ist, dass die überwiegende Mehrheit (75 %) keine Erfahrung mit Audio hat. Die meisten Personen (ungefähr 60 %) sind im tertiären Sektor tätig.

Präferenzanalyse: Modell 2 vs. Modell 1 vs. Real Voice

Im ersten Teil des Fragebogens wurde die Präferenz der Probanden gegenüber “Modell 2 vs. Modell 1 Voice Clone vs. Real Voice” abgefragt. Hierzu wurden ihnen mehrfach je zwei Audiosamples präsentiert und nach der Präferenz gefragt. Die Reihenfolge der Samples war hierbei zufällig gewählt.

Die Analyse betrachtet nun die folgenden drei Vergleiche:

1. Modell 2 vs. Modell 1
2. Real Voice vs. Modell 2
3. Real Voice vs. Modell 1

Um verlässliche Ergebnisse zu erzielen, werden die einzelnen zusammengehörenden Vergleiche gemittelt und anschließend ein statistischer Signifikanztest durchgeführt. Für die statistische Auswertung der Präferenzen wurde der Wilcoxon-Vorzeichen-Rang-Test verwendet, ein nicht-parametrischer Test, der Unterschiede zwischen gepaarten Stichproben untersucht. Dieser Test überprüft, wie wahrscheinlich das beobachtete Ergebnis (oder ein extremeres) ist, unter der Annahme, dass keine wirkliche Präferenz für das eine oder andere vorliegt. Falls die Signifikanzwerte kleiner als 0.05 sind, können wir davon ausgehen, dass eine Präferenz für das eine oder andere vorliegt.

Modell 2 vs. Modell 1

Fünf aufeinanderfolgende Präferenz-Fragen wurden gestellt. Die Ergebnisse sind in Figure 6 dargestellt.

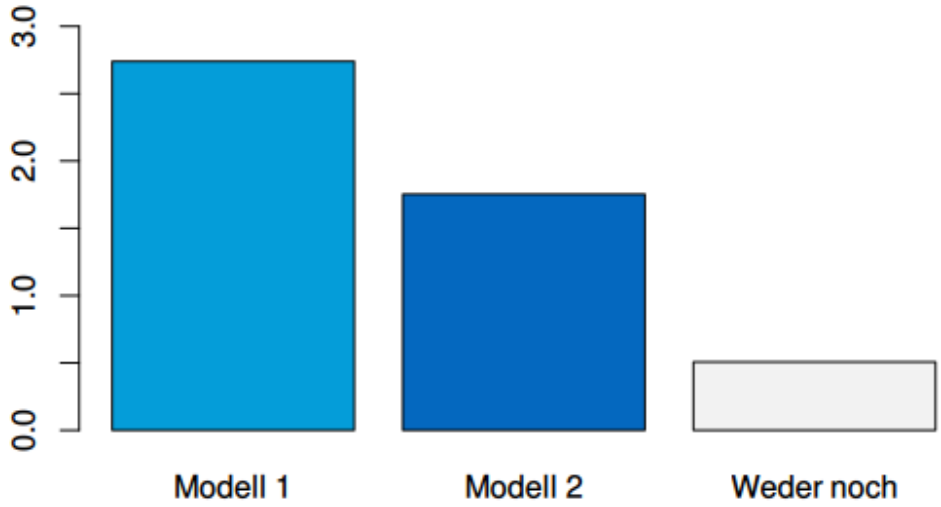

Figure 6: Durchschnittliche Präferenz für Modell 1, Modell 2 oder 'Weder noch' im direkten Vergleich (Modell 2 vs. Modell 1)

Im Durchschnitt bevorzugten die Probanden bei fünf Fragen Modell 1 ($M = 2.74$), Modell 2 ($M = 1.75$) oder wählten "weder noch" ($M = 0.51$). Der Unterschied in der Präferenz zwischen Modell 1 und Modell 2 war statistisch signifikant ($p < .001$). Dies deutet darauf hin, dass Modell 1 signifikant gegenüber Modell 2 bevorzugt wurde.

Modell 1 vs. Real Voice

Zwei Präferenzfragen wurden gestellt. Die Ergebnisse sind in Figure 7 dargestellt.

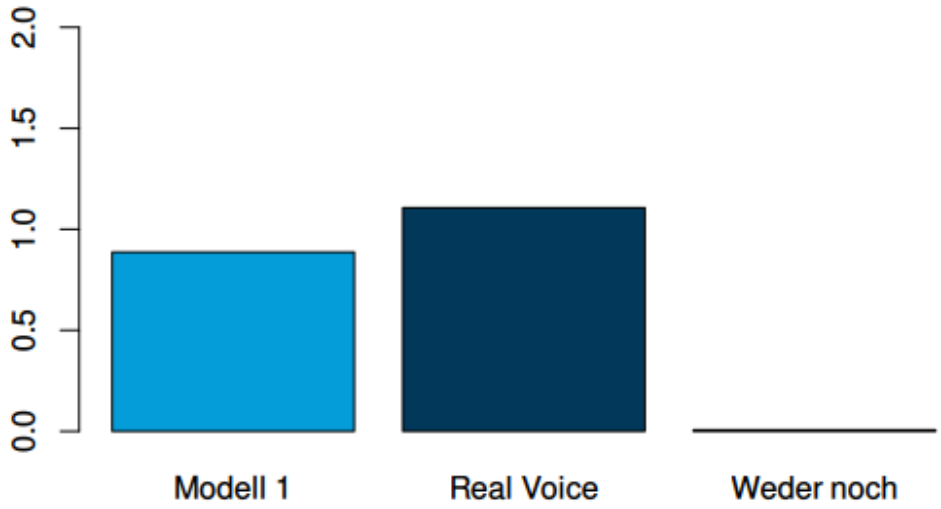

Figure 7: Durchschnittliche Präferenz für Modell 1, Real Voice oder 'Weder noch' im direkten Vergleich (Modell 1 vs. Real Voice)

Im Durchschnitt bevorzugten die Probanden bei zwei Fragen Modell 1 ($M = 0.89$), die Real Voice ($M = 1.11$) oder wählten "weder noch" ($M = 0.01$). Obwohl die Real Voice im Durchschnitt etwas häufiger bevorzugt wurde, war der Unterschied in der Präferenz zwischen Modell 1 und der Real Voice statistisch **nicht** signifikant ($p = .10$). Daraus folgt, dass keine statistisch signifikante Präferenz zwischen Modell 1 und der Real Voice nachgewiesen werden konnte.

Modell 2 vs. Real Voice

Wie zuvor wurden auch hier zwei Präferenzfragen gestellt. Die Ergebnisse sind in Figure 8 dargestellt.

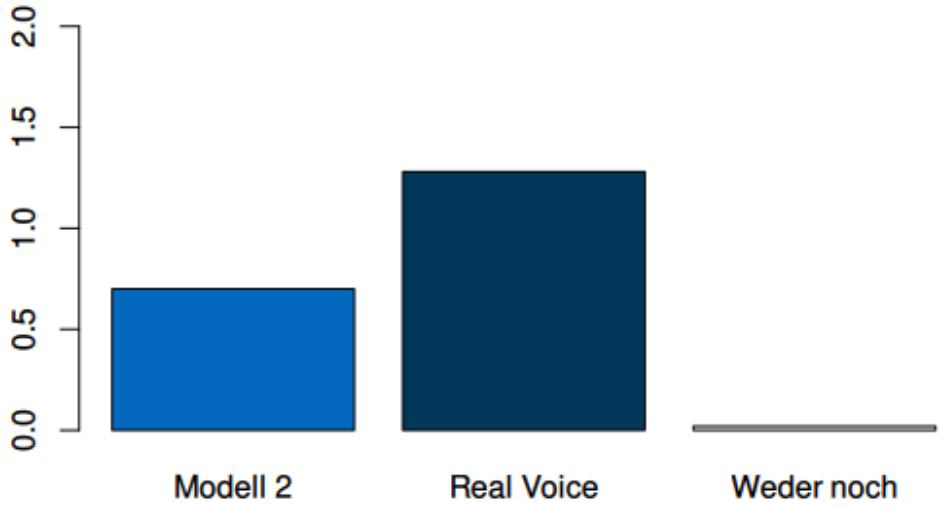

Figure 8: Durchschnittliche Präferenz für Modell 2, Real Voice oder 'Weder noch' im direkten Vergleich (Modell 2 vs. Real Voice)

Im Durchschnitt bevorzugten die Probanden bei zwei Fragen Modell 2 ($M = 0.7$), die Real Voice ($M = 1.28$) oder wählten "weder noch" ($M = 0.02$). Die Real Voice wurde statistisch signifikant häufiger bevorzugt ($p < .001$). Dies deutet darauf hin, dass die Probanden die Real Voice signifikant gegenüber Modell 2 bevorzugten.

Zusammenfassung der Präferenzanalyse

Die Präferenzanalyse der Probanden ergab unterschiedliche Ergebnisse je nach verglichenen Sprechern. Bei der Gegenüberstellung von Modell 1 und Modell 2 wurde Modell 1 signifikant bevorzugt ($p < .001$), wie in Figure 6 gezeigt. Im Vergleich zwischen Modell 2 und Real Voice zeigten die Probanden eine signifikante Präferenz für die Real Voice ($p < .001$), dargestellt in Figure 8. Interessanterweise gab es keine statistisch signifikante Präferenz zwischen Modell 1 und der Real Voice ($p = .10$), obwohl die Real Voice im Durchschnitt leicht häufiger gewählt wurde (siehe Figure 7).

Basierend auf den vorliegenden statistisch signifikanten Präferenzunterschieden (und dem Fehlen eines solchen Unterschieds in einem Vergleich) lässt sich folgende Rangfolge der Präferenz ableiten:

1. **Real Voice / Modell 1** (Es gab keinen statistisch signifikanten Unterschied in der Präferenz zwischen diesen beiden, obwohl die Real Voice im Durchschnitt leicht höher bewertet wurde. Sie können daher nicht eindeutig voneinander unterschieden werden.)
2. **Modell 2** (Wurde sowohl gegenüber Modell 1 als auch gegenüber der Real Voice signifikant weniger bevorzugt.)

Zusammenfassend: Modell 2 wurde signifikant weniger bevorzugt als Modell 1 und die Real Voice. Zwischen Modell 1 und der Real Voice konnte kein statistisch signifikanter Präferenzunterschied festgestellt werden.

Qualitätskriterien der Sprecher-Typen

Es wurden verschiedene Qualitätskriterien zwischen Modell 1, Modell 2 und Real Voice erhoben, und zwar: Klangqualität, Professionalität, Natürlichkeit, Verständlichkeit und Gesamteindruck. Um Unterschiede in den Bewertungen der Qualitätskriterien zwischen den Sprecher-Typen zu analysieren, wurde eine einfaktorielle Varianzanalyse (ANOVA) durchgeführt. Bei signifikanten Ergebnissen folgten Post-hoc-Tests nach Tukey (Tukey HSD) zur Bestimmung spezifischer Gruppenunterschiede.

Vergleich - Klangqualität

Die durchschnittlichen Bewertungen für die Klangqualität der verschiedenen Sprecher-Typen sind in Figure 9 dargestellt.

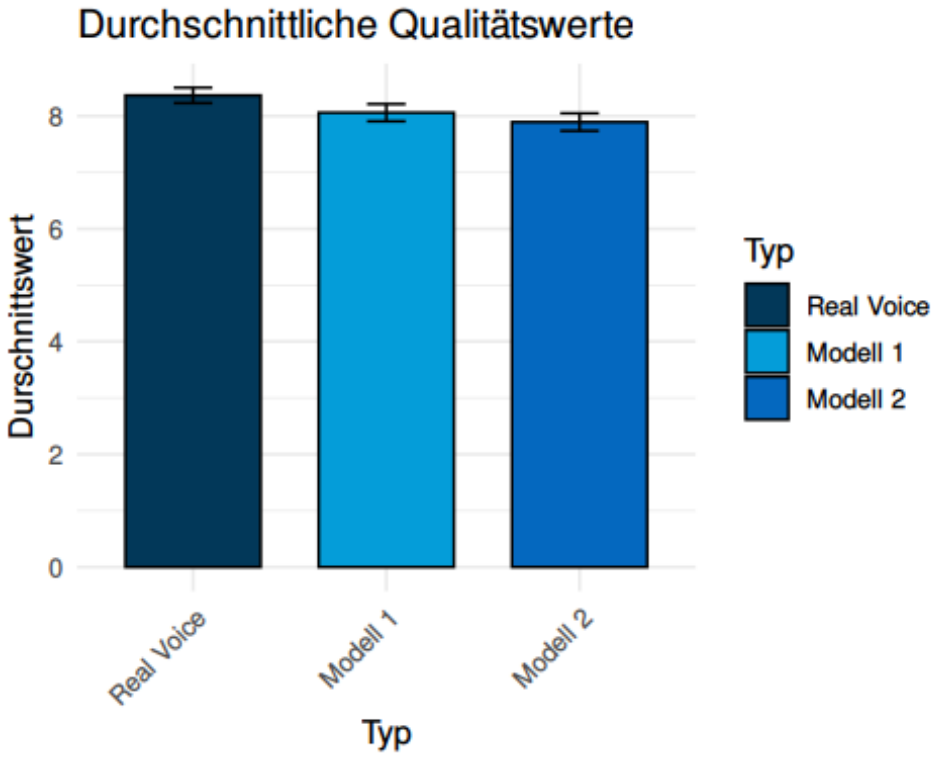

Figure 9: Durchschnittliche Bewertung der Klangqualität nach Sprecher-Typ

Wir haben untersucht, ob sich die durchschnittlichen Bewertungen für die drei Qualitätsarten (Real Voice, Modell 1, Modell 2) statistisch bedeutsam unterscheiden. Die Durchschnittswerte lagen bei 8.37 für die Real Voice, 8.06 für Modell 1 und 7.89 für Modell 2. Ein statistischer Test (ANOVA) zeigte jedoch, dass die gefundenen Unterschiede zwischen diesen Mittelwerten nicht statistisch signifikant sind (p = .076). Das bedeutet, wir haben keinen ausreichenden statistischen Nachweis dafür, dass die drei Methoden tatsächlich zu unterschiedlichen durchschnittlichen Qualitätsbewertungen führen, auch wenn die Mittelwerte leicht variieren.

Vergleich - Professionalität

Die durchschnittlichen Bewertungen für die Professionalität der verschiedenen Sprecher-Typen sind in Figure 10 dargestellt.

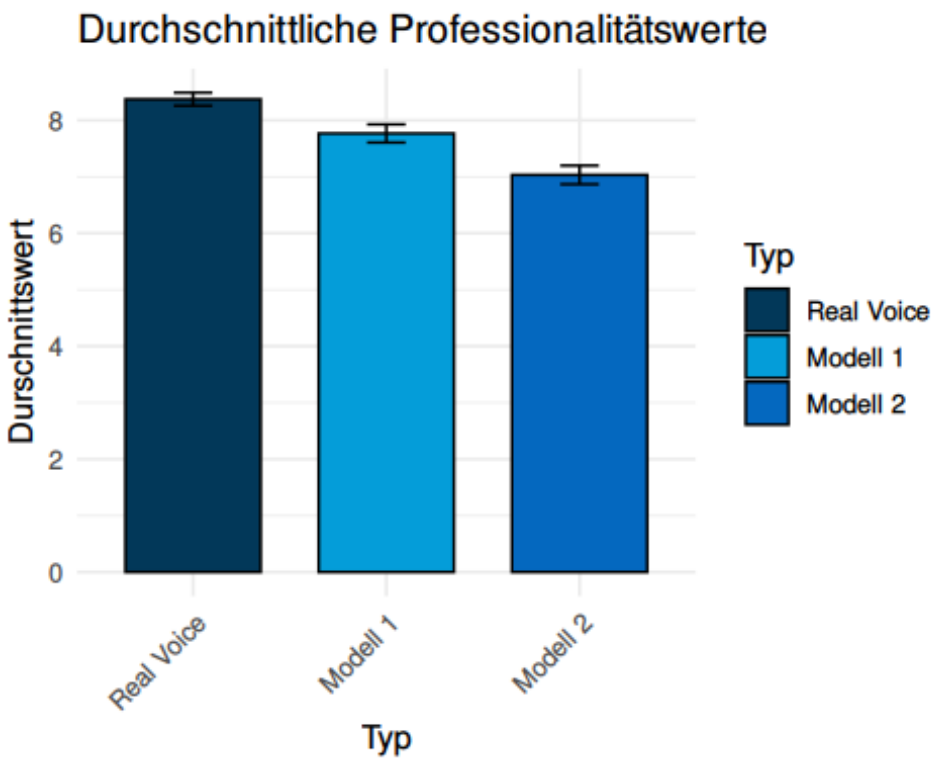

Figure 10: Durchschnittliche Bewertung der Professionalität nach Sprecher-Typ

Wir haben untersucht, ob sich die Durchschnittswerte der drei Typen (Real Voice, Modell 1, Modell 2) statistisch unterscheiden. Die Mittelwerte waren 8.37 für Real Voice, 7.77 für Modell 1 und 7.03 für Modell 2. Ein statistischer Test (ANOVA) zeigte, dass es insgesamt einen hoch signifikanten Unterschied zwischen diesen Gruppen gibt ($p < .001$). Anschließende Tests (Tukey HSD) ergaben, dass sich alle drei Gruppen signifikant voneinander unterscheiden: Real Voice hatte signifikant höhere Werte als Modell 1 ($p = .012$) und Modell 2 ($p < .001$), und auch Modell 1 hatte signifikant höhere Werte als Modell 2 ($p = .002$).

Vergleich - Natürlichkeit

Die durchschnittlichen Bewertungen für die Natürlichkeit der verschiedenen Sprecher-Typen sind in Figure 11 dargestellt.

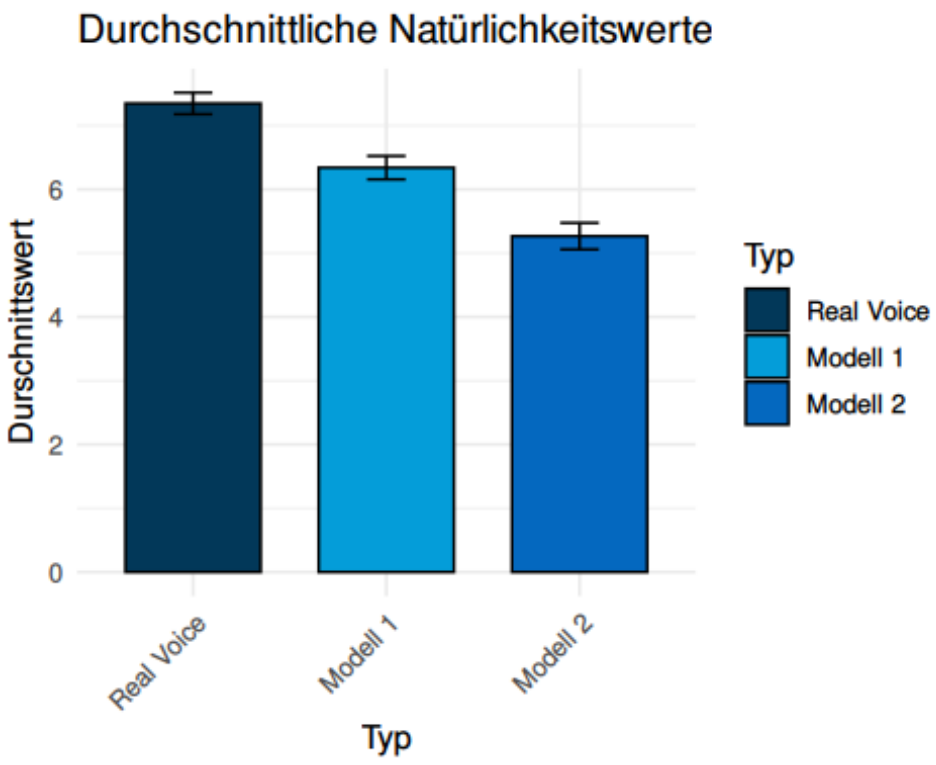

Figure 11: Durchschnittliche Bewertung der Natürlichkeit nach Sprecher-Typ

Wir haben die durchschnittliche Bewertung der “Natürlichkeit” für drei Typen (Real Voice, Modell 1, Modell 2) verglichen. Die Mittelwerte lagen bei 7.35 für Real Voice, 6.34 für Modell 1 und 5.27 für Modell 2. Ein statistischer Test (ANOVA) zeigte einen hoch signifikanten Unterschied zwischen den Gruppen insgesamt ($p < 0.001$). Die nachfolgenden Vergleiche (Tukey HSD) bestätigten, dass sich alle drei Gruppen signifikant voneinander unterscheiden: Real Voice wurde signifikant natürlicher bewertet als Modell 1 ($p < 0.001$) und Modell 2 ($p < 0.001$), und auch Modell 1 wurde signifikant natürlicher bewertet als Modell 2 ($p < 0.001$).

Vergleich - Verständlichkeit

Die durchschnittlichen Bewertungen für die Verständlichkeit der verschiedenen Sprecher-Typen sind in Figure 12 dargestellt.

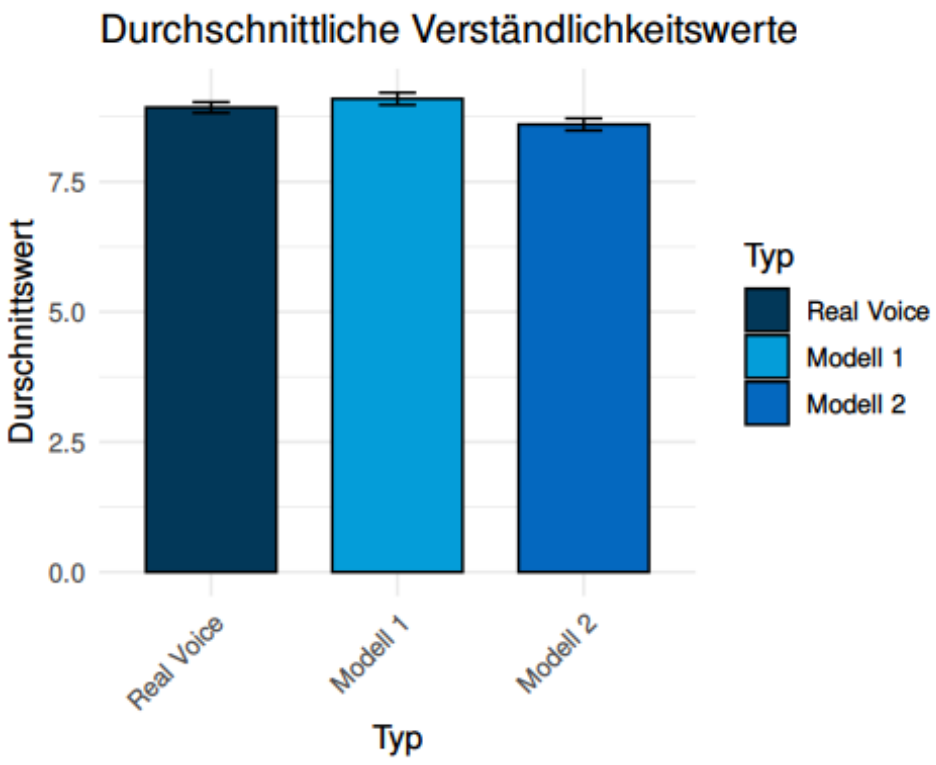

Figure 12: Durchschnittliche Bewertung der Verständlichkeit nach Sprecher-Typ

Wir haben die durchschnittliche Bewertung der “Verständlichkeit” für drei Typen (Real Voice, Modell 1, Modell 2) verglichen. Die Mittelwerte lagen bei 8.93 für Real Voice, 9.09 für Modell 1 und 8.60 für Modell 2. Ein statistischer Test (ANOVA) zeigte einen signifikanten Unterschied zwischen den Gruppen insgesamt (p = .007). Die genaueren Vergleiche (Tukey HSD) ergaben jedoch nur einen signifikanten Unterschied: Modell 1 wurde signifikant verständlicher bewertet als Modell 2 (p = .006). Die Unterschiede zwischen Real Voice und den beiden anderen Gruppen waren nicht statistisch signifikant (p > .10).

Vergleich - Gesamteindruck

Die durchschnittlichen Bewertungen für den Gesamteindruck der verschiedenen Sprecher-Typen sind in Figure 13 dargestellt.

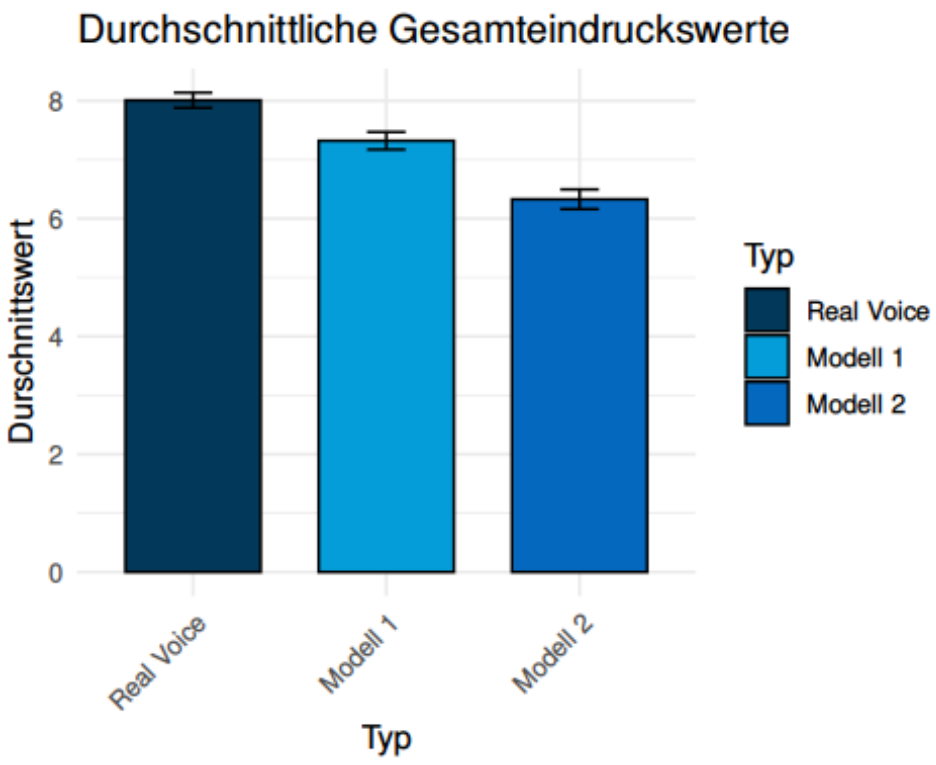

Figure 13: Durchschnittliche Bewertung des Gesamteindrucks nach Sprecher-Typ

Wir haben den durchschnittlichen “Gesamteindruck” für drei Typen (Real Voice, Modell 1, Modell 2) verglichen. Die Mittelwerte lagen bei 8.01 für Real Voice, 7.32 für Modell 1 und 6.33 für Modell 2. Ein statistischer Test (ANOVA) zeigte einen hoch signifikanten Unterschied zwischen den Gruppen insgesamt ($p < .001$). Die genaueren Vergleiche (Tukey HSD) bestätigten, dass sich alle drei Gruppen signifikant voneinander unterscheiden: Real Voice erzielte einen signifikant besseren Gesamteindruck als Modell 1 ($p = .004$) und Modell 2 ($p < .001$), und auch Modell 1 wurde signifikant besser bewertet als Modell 2 ($p < .001$).

Zusammenfassung und Schlussfolgerung der Ergebnisse

Die vorliegende Analyse untersuchte die Präferenzen und Qualitätsbewertungen von drei verschiedenen Sprecher-Typen: einer professionellen menschlichen Sprecherin (“Real Voice”), Modell 1 Voice Clone (“Modell 1”) und Modell 2 Voice Clone (“Modell 2”) anhand eines Samples von 150 Probanden.

Stichprobenbeschreibung: Die Stichprobe bestand mehrheitlich aus Frauen (56%) mit einem Durchschnittsalter von etwa 40 Jahren, wie in den Section 3 Details zu finden ist. Über die Hälfte der Teilnehmenden verfügte über einen Fachhochschul- oder Universitätsabschluss. Auffällig war, dass drei Viertel der Probanden keine Vorerfahrung mit Audio hatten. Die Mehrheit (ca. 60%) war im tertiären Sektor tätig.

Präferenzanalyse: Die Analyse der direkten Präferenzvergleiche zwischen den Sprecher-Typen, detailliert in Section 4, ergab klare Unterschiede:

- **Modell 1 vs. Modell 2:** Modell 1 wurde signifikant häufiger bevorzugt als Modell 2 ($p < 0.001$), wie in Figure 6 dargestellt.
- **Modell 2 vs. Real Voice:** Die Real Voice wurde signifikant häufiger bevorzugt als Modell 2 ($p < 0.001$), siehe Figure 8.
- **Modell 1 vs. Real Voice:** Es gab *keinen* statistisch signifikanten Unterschied in der Präferenz zwischen Modell 1 und der Real Voice ($p = 0.10$), obwohl die Real Voice im Durchschnitt leicht häufiger gewählt wurde (siehe Figure 7).

Aus diesen Ergebnissen lässt sich die folgende Präferenz-Rangfolge ableiten: **Real Voice ≈ Modell 1 > Modell 2**. Modell 2 wurde eindeutig am wenigsten bevorzugt, während zwischen Modell 1 und der Real Voice keine klare statistische Präferenz festgestellt werden konnte.

Bewertung von Qualitätskriterien: Die Probanden bewerteten die Sprecher-Typen anhand verschiedener Qualitätskriterien, wie in Section 5 detailliert:

- **Klangqualität:** Es gab *keine* statistisch signifikanten Unterschiede in der wahrgenommenen Klangqualität zwischen den drei Typen ($p = 0.076$), wie Figure 9 zeigt.
- **Professionalität:** Die Real Voice wurde als signifikant professioneller bewertet als Modell 1 ($p = 0.012$), und beide wurden als signifikant professioneller bewertet als Modell 2 (Real Voice > Modell 1 > Modell 2; alle Unterschiede signifikant), siehe Figure 10.
- **Natürlichkeit:** Die Real Voice wurde als signifikant natürlicher bewertet als Modell 1, und beide wurden als signifikant natürlicher bewertet als Modell 2 (Real Voice > Modell 1 > Modell 2; alle Unterschiede signifikant), wie in Figure 11 dargestellt.

- **Verständlichkeit:** Modell 1 wurde als signifikant verständlicher bewertet als Modell 2 ($p = 0.006$). Es gab *keine* signifikanten Unterschiede in der Verständlichkeit zwischen der Real Voice und den beiden Klonen. Details in Figure 12.
- **Gesamteindruck:** Die Real Voice erzielte einen signifikant besseren Gesamteindruck als Modell 1, und beide erzielten einen signifikant besseren Gesamteindruck als Modell 2 (Real Voice > Modell 1 > Modell 2; alle Unterschiede signifikant), ersichtlich in Figure 13.

Schlussfolgerung (Conclusion):

Die Studie zeigt, dass der **“Modell 1”** in der Wahrnehmung der getesteten Stichprobe (hauptsächlich audio-unerfahrene, gut gebildete Personen mittleren Alters) **erstaunlich gut abschneidet**. In direkten Präferenzvergleichen war er der Realen professionellen Sprecherin statistisch ebenbürtig und wurde dem “Modell 2” desselben Anbieters signifikant vorgezogen.

Obwohl die **Real Voice** bei den subjektiven Qualitätsbewertungen in den Kategorien **Professionalität, Natürlichkeit und Gesamteindruck signifikant am höchsten** bewertet wurde, schlug sich dies nicht in einer signifikant höheren Präferenz gegenüber Modell 1 nieder.

Der **“Modell 2”** schnitt in dieser Studie **konsistent am schlechtesten** ab. Er wurde sowohl Modell 1 als auch der Real Voice signifikant unterlegen präferiert und erhielt auch bei den Qualitätskriterien Professionalität, Natürlichkeit, Verständlichkeit (im Vergleich zu Modell 1) und Gesamteindruck die signifikant niedrigsten Bewertungen. Lediglich bei der Klangqualität gab es keine signifikanten Unterschiede zwischen den drei Stimmen.

Zusammenfassend lässt sich sagen, dass **Modell 1 eine überzeugende Alternative darstellt**, die in der Präferenz von Laien kaum vom menschlichen Original zu unterscheiden ist, auch wenn bei detaillierter Qualitätsbewertung der Mensch in wichtigen Aspekten wie Natürlichkeit und Professionalität noch überlegen ist. Das untersuchte Modell 2 konnte in diesem Vergleich nicht überzeugen.