UNIVERSITÄT DES SAARLANDES

„Just Because We Camp, Doesn‘t Mean We Should: The Ethics Of Modelling Queer Voices“

Sigurgeirsson, A., Ungless, E.L. (2024) Just Because We Camp, Doesn't Mean We Should: The Ethics of Modelling Queer Voices.. Proc. Interspeech 2024, 3050-3054, 04.12.2025.

Amira Palisch, Julian Heß, Salome Djaoui 07.05.2026

Idee / Motivation des Papers

- Es wird geprüft, ob Text-To-Speech (TTS) Modelle die darauf ausgelegt sind so nah wie möglich an der Stimmlage und Aussprache des jeweiligen menschlichen Sprechers zu sein, in der Lage sind, akkurat die Stimmqualitäten die mit Queeren Stimmen (in diesem Experiment den Stimmen schwuler Männer) assoziiert werden (*gay voice*), wiederzugeben.

- Es wurden Modelle untersucht die zum Klonen von Stimmen verwendet werden, da diese repräsentativ für den Charakter der Menschen stehen, die auf diese Modelle für Kommunikation angewiesen sind und daher darauf vertrauen müssen, von dieser synthetischen Stimme angemessen repräsentiert zu werden.

Queere Stimmqualitäten

- Es werden verschiedene Stimmqualitäten bzw. Sprechweisen genannt, die häufiger mit queeren (hier: schwulen) Stimmen assoziiert werden:

  - Größerer Tonhöhenbereich (pitch range) der für manche Hörer „femininer“ klingen kann
  - Die Dauer und Qualität von Sibilanten
  - Hyper- bzw. Fehlartikulation von /s/- Lauten (leichtes lispeln)

- Diese Qualitäten synthetisch herzustellen wäre dementsprechend ausschlaggebend für eine erfolgreiche Replikation einer queer klingenden Stimme *(gay voice*).

Hypothesen

- H1: Verglichen mit den Differenzen zwischen der menschlichen (nicht queeren) Stimme und der synthetisch modellierten (nicht queeren) Stimme, wird die synthetische queere Stimme als weniger queer wahrgenommen werden als die menschliche queere Stimme nach der sie modelliert wurde.

- H2: Dementsprechend wird die wahrgenommene Ähnlichkeit zwischen synthetischer und menschlicher queerer Stimme deutlich zurückgehen.

- Es wird auch überprüft, welche Veränderungen an den Modellen die größte Auswirkung auf die Präzision der künstlichen queeren Stimmen hat.

Sprecher

- Es wurden aus dem „Ted-Lium 3“ Korpus schwule, US-Amerikanische Sprecher ausgewählt, die für beide Autoren eine wahrnehmbare *gay voice* hatten. Es variiert jedoch, wie stark ausgeprägt dieser Effekt bei den Sprechern ist.
- Die Kontrollgruppe wurde auf die selbe Weise ausgewählt: von den Autoren anhand der nicht-vorhandenen queeren Stimmmerkmale. Diese Gruppe identifiziert sich nicht als schwul.
- Für jeden *gay voice* Sprecher, gibt es einen Kontrollgruppensprecher der ungefähr das selbe Alter und die selbe Herkunft hat.

Sprecher

- Insgesamt gibt es 24 Sprecher; 20 davon sind weiß, 4 werden als afro-amerikaner wahrgenommen.

- Die Sprecher sind zwischen 29 und 58 Jahre alt. Alle Sprecher sind öffentlich als Schwul bekannt und/oder in einer öffentlichen Beziehung mit Männern.

- Alle Versuchspersonen sind US-Amerikanische bzw. Britische L1 Sprecher des Englischen

- Alle VPs sind Teil der LGBTQ+ Community, da es von besonderem Interesse ist ob die synthetische *gay voice* spezifisch von anderen Mitgliedern der Community erkannt werden kann

Stimuli

- Es wird die Wahrnehmung der *gay voice* in verschiedenen Entwicklungsstufen von TTS untersucht.
- Dafür gibt es drei Konditionen von Äußerungen:
  - **Lange Äußerungen (segment):** Drei 30sek Segmente pro Sprecher, ohne jegliche Arten Queerer Themen bzw. LGBTQ+ zugehörigem Vokabular
  - **„Ground-truth“ Äußerungen (Raw-utt):** 15 Äußerungen pro Sprecher ebenfalls ohne jegliches Queeres Vokabular
  - **VE-utt:** Alle Raw-utt Äußerungen wurden von einem voice-enhancement model aufgewertet
- Um ein TTS Modell nur auf *gay voice* zu trainieren, gab es zu wenig Daten. Daher wurde ein Modell genutzt, dass Stimmklone herstellt und auf 16k Stunden öffentlich zugänglicher Stimmdaten trainiert wurde.

Stimuli

- Es werden zwei verschiedene Arten der synthetischen Sprache untersucht:
  - „Copy-synth“: Der Zieltext und der Referenztext der verwendet wurde um das Modell auf diese Stimme zu trainieren (VE-utt) stimmen überein
  - „Synth“: Der Zieltext stimmt nicht mit dem Referenztext überein.
- Es wird erwartet, dass „Copy-synth“ die für dieses Modell optimalen Konditionen um die Sprecheridentität akkurat zu modellieren bietet.

Auswertungspunkte

- Es soll geprüft werden, ob eine Stimme als *gay voice* identifizierbar ist, nicht welche Sexualität dem Sprecher zuzuschreiben wäre.
- Versuchspersonen wurden gebeten auf einer Skala zwischen 1 und 7 zu bewerten, wie *gay* diese Stimme wahrgenommen wurde: 1 = Hetero 7 = Definitiv gay
- Versuchspersonen wurden geben auf einer Skala zwischen 1 und 5 zu bewerten, wie natürlich die jeweilige Stimme klingt: 1 = Schlecht 5 = Perfekt
- Versuchspersonen wurden gebeten auf einer Skala zwischen 0 und 100 zu bewerten, wie ähnlich die menschlichen Segmente zu der Synthetischen waren: 0 = komplett verschieden 100 = komplett gleich

Durchführung 1 (H1)

- Um sicherzugehen, dass die Gay- und Kontrollgruppe verschiedene *gay voice* Bewertungen haben, wurden 13 VPs rekrutiert, die jeweils 30 Segmente bewerten sollten. Daraufhin wurden 5 Sprecher beider Gruppen ausgeschlossen, sodass nur diejenigen übrig waren, die eine Bewertung über 5 für gay und eine unter 3 für Kontrolle erhalten haben.
- Um zu testen, ob und wann die *gay voice* in einem TTS Modell verloren geht, wurde jede VP gebeten zwei 30sek Segmente und eine randomisierte Menge an 30 Äußerungen (Raw-utt, VE-utt, Copy-synth) zu bewerten, nach dem Konzept wie *gay* und natürlich es klang.
- Jede Äußerung wurde ca. 4 Mal bewertet

Ergebnisse

- Wie erwartet war, hat die Kontrollgruppe eine sehr viel niedrigere Bewertung bei *gay voice* als die Queeren Sprecher.
- Segments vs. Copy-synth für die queeren Sprecher hatten eine als geringer wahrgenommene *gay voice*. Entgegen der Erwartungen ist diese Bewertung bei der Kontrollgruppe nach oben gegangen.
- Hypothese 1 hat sich bestätigt.

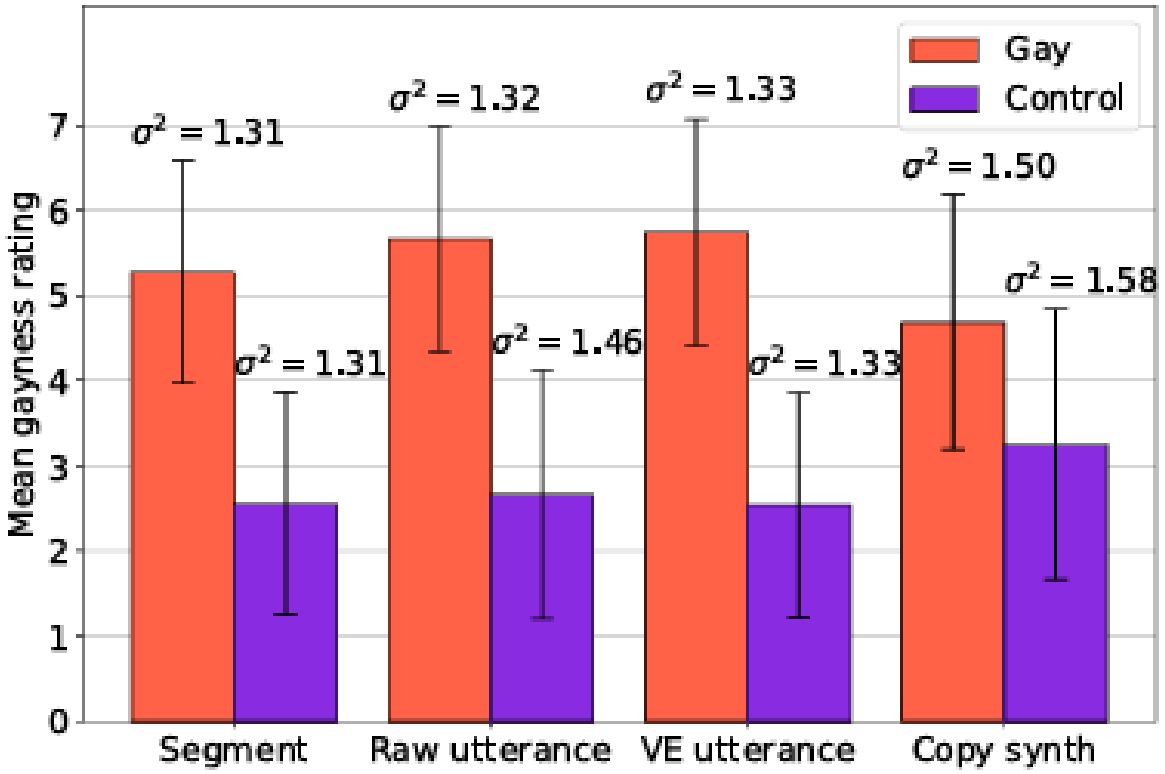

Figure 1: *Mean* gay voice *rating of both speaker groups in each clip type.*

Durchführung 2 (H2)

- Um zu testen ob die wahrgenommene Ähnlichkeit der Stimmen mit der wahrgenommenen *gayness* zurückgeht, wurde jeweils ein 30sek Segment und ein korrespondierender Synth der Sprecher (gay und Kontrolle) den VPs als Referenzen präsentiert.
- Diese sollten dann auf Ähnlichkeit bewertet werden.
- Die Präzision der Ähnlichkeit wurde anhand der Anzahl der korrekt assoziierten Paare gewertet.
- Jede VP hat 20 Äußerungen randomisiert bekommen und bewertet

Ergebnisse

- Nach einem Kruskal-Wallis Test wurde herausgefunden, dass es keine Signifikanten Zusammenhänge zwischen der wahrgenommenen Ähnlichkeit der synthetischen und menschlichen Sprecher vs. der queeren synthetischen und queeren menschlichen Sprecher gibt.
- Dementsprechend ist Hypothese 2 nicht bestätigt worden.

Ethische Konflikte

- Nachdem bestätigt wurde, dass TTS Modelle nicht ausreichend fähig sind *gay voice* nachzubilden, steht die Frage im Raum ob es nötig bzw. ethisch vertretbar ist diese Lücke auszubauen.

Pro	Con
Akkurate Repräsentation für Kommunikationshilfe bei Sprachbarrieren	Synthetische Queere Stimmen können einfach auf negative Weise zweckentfremdet werden um gegen diese Community zu hetzen.
Vorbeugen von Identitätsverlust für Menschen mit Einschränkungen, die auf eine synthetische Stimme angewiesen sind	Ein ausgebautes queer-voice-model kann verwendet werden um eine Art Radar für *gay voice* zu erstellen, was wiederum für Verfolgung queerer Menschen genutzt werden kann.

Unser Experiment

Unser Experiment: Idee

- Fragestellung: Gibt es Unterschiede in der Wahrnehmung der *gay voice* von L1 Sprechern des Deutschen vs. L2 Sprechern des Deutschen? Spielt das Geschlecht der Hörer*innen eine Rolle?
- Sprecher:
  - L1 Sprecher des Deutschen mit *gay voice*
  - L1 Sprecher des Deutschen ohne *gay voice*
  - L2 Sprecher des Deutschen mit *gay voice*
  - L2 Sprecher des Deutschen ohne *gay voice*
- Stimuli: (in randomisierter Reihenfolge)
  - 5 Nonsense Sätze mit *gay voice* und 5 Nonsense Sätze ohne *gay voice* von L1 Sprechern
  - 5 Nonsense Sätze mit *gay voice* und 5 Nonsense Sätze ohne *gay voice* von L2 Sprechern
- Versuchspersonen: 5 Frauen, 5 Männer (L1 Sprecher*innen)
- Bewertungssystem: Skala

Unser Experiment: Zeitplan

- Bis ende Mai alle Stimuli aufnehmen und Versuchspersonen rekrutieren
- Bis Mitte Juni (ca. 18.06.) alle Versuche Durchführen
- Bis Ende Juni alle Ergebnisse auswerten
- Bis Deadline die Präsentation erstellen
- 09.07. Deadline
- 14./17.07. Mini Konferenz