Voice *"Cloning"* is Style Transfer

Kaitlyn Zhou[1,2] **Federico Bianchi**[2] **Martijn Bartelds**[2]
Anna Pot[3] **Yongchan Kwon**[2] **James Zou**[2,3]
[1]Cornell University
[2]TogetherAI
[3]Stanford University
kaitlynz@cornell.edu

Abstract

Artificially generated speech is increasingly embedded in everyday life. Voice cloning in particular enables applications where identity preservation is important, such as completing a recording, dubbing in a new language, or preserving the voices of individuals with speech loss. However, in our work, we find that despite the term, voice cloning does not faithfully *"clone"* an individual's voice. Instead, we find that widely-used voice cloning models systematically apply style transfer to source voices. As rated by human annotators, cloned voices are perceived as more authoritative, warm, customer-service-like, and human-like compared to their sources. Human annotators also report greater trust in cloned voices than source voices, and a greater willingness to disclose sensitive personal information to them. Our work furthermore shows that voice cloning leads to homogenization of speaker characteristics, as measured by reduced variance in accent, speaking rate, and the audio embedding space. Together, our results highlight a new set of limitations and risks of voice cloning technology and their potential impact on human behavior.

1 Introduction

A person's voice is a deeply personal marker of identity, conveying accent, affect, and individual character. Recent advances in text-to-speech (TTS) have enabled increasingly humanlike speech generation, with zero-shot voice cloning emerging as a prominent application. Most discourse around voice cloning has focused on harms in misuse: a more faithful clone can make impersonation, fraud, and other unauthorized uses more convincing. However, in settings where users intentionally provide their voices — such as enhanced presentations [Zheng et al., 2025] and pronunciations [Park et al., 2026], assistive technologies for individuals with speech impairments [Wairagkar et al., 2025, R Chavan et al., 2025], generative tools for personal well-being and self-reflection [Pataranutaporn et al., 2021], and multilingual media dubbing [Li et al., 2024] — vocal fidelity is precisely what gives the technology its value. In these cases, unfaithful cloning can also be harmful: it may distort how a person sounds, erase identity-linked vocal traits, or replace their voice with a more standardized version.

Although these systems are often described as *"cloning"* a voice, it remains unclear whether they actually preserve the speaker identity they claim to reproduce. We examine this assumption empirically and find that modern voice cloning systems fall short of faithful reproduction. Rather than preserving source voices, they systematically transform them, diminishing distinctive vocal traits and replacing them with more homogenized, socially preferred speaking characteristics.

Our study presents two key findings regarding modern voice cloning systems. First, we show how voice *"cloning"* systematically applies style transformations to source voices. These style transfers arise without explicit authorization from users and reflect limitations in voice cloning fidelity. Through annotations of paired source and cloned recordings, we observe that cloned voices sound significantly warmer, more authoritative, and more customer-service-oriented. Cloned voices in our study are often

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

perceived as more *"human-like,"* contributing to a phenomenon akin to hyperrealism — previously documented in visual generative models but not yet widely observed in speech [Miller et al., 2023, Lavan et al., 2025]. These vocal transformations influence more than just human perceptions: our behavioral evaluations indicate that listeners report greater trust in cloned voices and express higher willingness to engage in intimate conversations with them, with direct consequences for downstream human behaviors. This finding underscores the need for continued efforts to assess the downstream effects of synthetic voice.

Second, voice cloning homogenizes identity. It systematically imposes a *"native"* English accent, collapsing the variation present in non-native English speech and erasing markers of cultural origin and individual distinctiveness. Using speaker identity probe methods, we find that cloned voices lose the features that make source voices distinct from one another, with more identity confusion in cloned outputs relative to source recordings. This stylistic transformation compounds over time. Repeated cloning produces directional drift in audio embedding space, increased pitch in recordings, and altered emotional expression – all of which further distance outputs from the source speaker.

Our findings contribute to a deeper technical and interactional understanding of voice cloning, with direct implications for how these systems are deployed and governed. Our results highlight the need to transparently reveal systematic transformations introduced by these models—such as stylistic shifts and homogenization—that shape how voices are perceived and interpreted. This adds to a long line of work that discusses risks and limitations of voice cloning, such as impersonation, fraud, bypassing voice-based authentication [Hutiri et al., 2024, Du et al., 2025], labor displacement of workers whose voices and livelihoods these systems increasingly replace [Almeda et al., 2025], unauthorized use of voices [Agnew et al., 2024], and loss of personal identity [Leuenberger, 2025]. Taken together, this work suggests that deviations from faithful voice reproduction are not merely technical shortcomings but a new form of homogenization with implications for consumer protection, cultural diversity, and personal identity.

The wide-scale deployment of voice cloning technology is not inevitable but reflects choices made by companies, policymakers, and researchers [Bijker et al., 1987, Winner, 2017]. Our work aims to provide a clear empirical understanding of what these systems actually do to inform grounded technology design and policy responses.

2 Related Work

Research on voice cloning is part of a broader literature that includes speech synthesizers and text-to-speech systems [Arik et al., 2018, Le et al., 2023, Wang et al., 2023]. TTS is commonly used in speech technologies for synthetic voice generation, and voice cloning extends the TTS architecture. Voice cloning extracts a speaker-specific embedding from reference audio and incorporates it into the generation pipeline, thereby enabling the system to mimic a particular individual's voice.

Prior work primarily emphasized a model's ability to produce "natural" and humanlike voice [Shen et al., 2023, Ju et al., 2024, Anastassiou et al., 2024]. Meanwhile, other works have tried to add more control to the way zero-shot voice cloning is done by controlling aspects like the timbre, content, and style of a voice [Ji et al., 2025], the emotions a voice conveys [Chen et al., 2024], and even controlling the accent, rhythm, and intonation [Qin et al., 2024]. At times, this involves using natural language prompts to direct the style of the generation [Guo et al., 2023, Yang et al., 2023].

Despite rapid technical progress, voice cloning remains contested due to concerns about misuse, which include impersonation, fraud, and bypassing voice-based authentication [Hutiri et al., 2024, Du et al., 2025], the unauthorized use of voices in datasets [Agnew et al., 2024], and exaggeration of accents [Michel et al., 2025]. Studies from as early as 2023 show that audio deepfakes are indistinguishable from real voices [Mai et al., 2023]. Work in Human-Computer Interaction (HCI) has shown the economic impact on voice actors [Almeda et al., 2025]. Philosophical work has argued that cloned voices function as simulations detached from their embodied origins [Berkowitz and Sweeney, 2026] and risk the loss of personal and cultural identity [Leuenberger, 2025]. Legal challenges in voice cloning are complex [Lee and Sun, 2026, Berkowitz and Sweeney, 2025], and some scholars have noted the implications of false light[1] and defamation that could be applied to cloned voices [Wells-Edwards, 2022]. To this end, advances in natural language processing have

[1] https://www.law.cornell.edu/wex/false_light

made efforts to preserve the privacy of users while still leveraging the advances of voice cloning. For example, Platnick et al. [2024] discusses using voice cloning technology by closely matching the input of a voice to a similar preset consenting voice for cloning. Other design practices, such as those in South Korea, aim to explicitly mark generated AI voices for transparency to older adult customers Sang-Hun [2026].

3 Methods

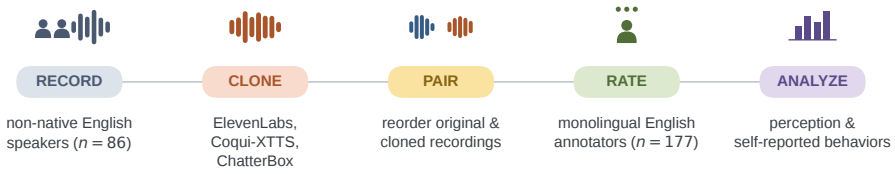


Figure 1: **Study pipeline.** Audio from n=86 non-native English speakers serves as reference for cloning on three models (ELEVENLABS, COQUI-XTTS, CHATTERBOX). Each source recording is paired with its clone and shown in randomized order to n=177 annotators.

3.1 Audio Data Collection

The experimental setup begins by recruiting $n = 86$ participants (over 18 years old, based in the U.S.) via Prolific. Our research goal is to understand how these voice cloning systems might impact a broad user population, thus we wanted our speakers to reflect this diversity. We recruit non-Native English speakers who have a wide range of foreign accents, some self-reporting none, while others report strong accents. Participants also vary in age and are sex balanced, Table 1. Participants started by providing information about their English language background, and then were asked to read aloud the Grandfather Passage [Van Riper, 1972], a nine-sentence standard text widely used in speech and language assessment.

Table 1: Speaker demographics ($N = 86$).

Variable	Value
Sex	
Male	43 (50.0%)
Female	43 (50.0%)
Age (years)	
Mean ± SD	38.3 ± 10.7
Range	19–64
Self-reported accent	
Mean ± SD	3.8 ± 3.2
Range	0–10

Each recording was automatically split into sentence-level audio clips using Whisper-based forced alignment [Radford et al., 2023], yielding up to nine clips per participant. We conducted manual quality control on each sentence clip, reviewing alignment boundaries and ensuring every clip was read correctly. Clips containing incorrect utterances were excluded. Because speakers varied in reading accuracy and speaking rate, and because recordings were capped at 90 seconds, some participants contributed fewer valid clips than others. To preserve speaker diversity, we retained all valid utterances from speakers rather than requiring a complete set of nine sentences. The final dataset contains 86 speakers and 699 valid sentence-level clips across the nine target sentences.

Audio clips were preprocessed by trimming silent segments and normalizing amplitude to a consistent level, common in speech processing pipelines [Labied et al., 2022, Keerio et al., 2009]. Upon publication, we will make this dataset available via Hugging Face for research purposes only. This study was IRB-approved, and all participants were paid $18 USD per hour. Details of the consent form, questionnaire, and full passage can be found in §A, along with how we protect participant privacy in §A.2 and task screenshots (Figure 14).

3.2 Voice Cloning

We evaluate three widely used TTS models — two open-source (`ChatterBox`, `Coqui-XTTS`) and one state-of-the-art proprietary model [`ElevenLabs V3`, Resemble AI, 2025, Eren and The Coqui TTS Team, 2021, ElevenLabs, 2026]. Open-source models were selected to reduce privacy risks by enabling greater control over speaker data, while `ElevenLabs` was included as a leading proprietary

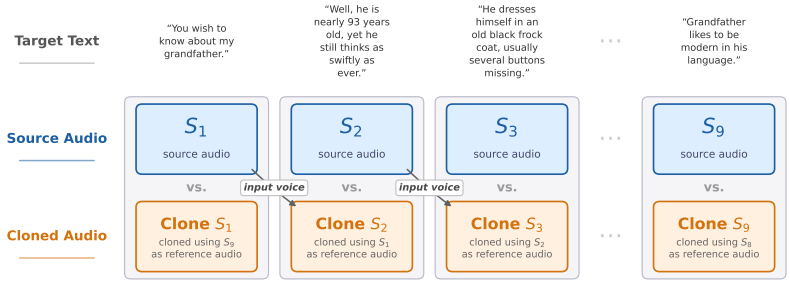


Figure 2: Illustration of *cross-sentence* voice cloning.

system that provides mechanisms for data removal and opt-out from model training. The dataset and code are available at the following links.[2][3]

Our unit of analysis is a *tuple* (s, ℓ), where s denotes the speaker and $\ell \in \{1, \ldots, 9\}$ indexes the sentence of the passage. Each tuple contains two audio samples: a human-produced source and a model-generated clone of the same utterance.

We generate voice clones using a *cross-sentence* cloning paradigm. In each cross-sentence pair, the model takes in S_ℓ^{source} as the reference audio and generates $S_{\ell+1}^{\text{cloned}}$, using the target text of sentence $\ell + 1$. The pairs wrap cyclically ($S_9^{\text{source}} \rightarrow S_1^{\text{cloned}}$). This offset of voice cloning ensures that the speech model isn't given speaker characteristics for the target cloned audio, but instead needs to extract generalizable speaker features from the reference audio. In human annotation evaluation, we realign source audio with cloned audio so that we are always comparing $S_{\ell+1}^{\text{cloned}}$ with $S_{\ell+1}^{\text{source}}$.

3.3 Voice Annotation

We present the source and cloned recordings to online annotators to examine how human listeners characterize both source and cloned speech recordings. Each annotation session contains clips from 10 distinct (s, ℓ) tuples, 10 source recordings and their 10 cloned counterparts. The 20 clips are then *globally shuffled*, so annotators are unaware of whether any given clip is human-produced or generated. For every clip, annotators provided ratings on five-point Likert-scale dimensions, including human-likeness, customer-service voice quality, authoritativeness, warmth, trustworthiness, and suitability for intimate conversation, see rationales in Table 2.

To control for potential confounding factors, each session is limited to one TTS model and one speaker-sex category. The rationale is that one model may sound substantially more authoritative than another, so combining clips from different models could introduce interaction effects in annotator judgments. In that case, it would be difficult to determine whether observed differences reflect the distinction between source and cloned audio or differences between the models themselves. The same concern applies to speaker sex: female voices could be rated differently from male voices, mixing female and male voices within a session could therefore similarly influence annotation outcomes.

Annotators were recruited via Prolific, with eligibility restricted to U.S.-based participants who reported English as their only language, to minimize potential ingroup–outgroup confounds between listeners and annotators. Each annotator could complete up to ten independent sessions. In total, we collected 4,000 annotations from 177 unique participants across four conditions: the three systems above, plus an `ElevenLabs` low-expressiveness ablation (§4). The analyses that compare the three systems therefore use the 3,000 annotations from 137 of those participants.

4 Voice Cloning as Style Transfer

Our first major finding is that voice cloning does not faithfully preserve speaker identity. We show that widely used voice-cloning systems often behave less like identity-preserved copying, and more like style-normalized transformation.

[2] https://huggingface.co/datasets/kzhou/voice_cloning_style_transfer
[3] https://github.com/kzhou-cloud/voice-cloning-public

Dimension	Main Question	Rationale
`humanlike`	*Does the clip sound humanlike?*	Measuring if cloned voices were as humanlike as source voices (intentionally ambiguous)
`customer service`	*Does the clip have a "customer service voice"?*	Measuring professional/assistant-like register (e.g., polished, deferential, scripted service tone).
`authoritative`	*Does the clip sound authoritative?*	Measuring perceptions of warmth and competence, adjusted for speech [Fiske et al., 2007].
`warm`	*Does the clip sound warm?*	Same as above
`trust`	*Would you trust instructions or answers from this voice?*	Measuring self-reported reliance and trust behaviors.
`intimate`	*Would you be comfortable having an intimate conversation with this voice?*	Measuring self-reported willingness to disclose personal information.
`native_english`	*Does the clip above sound like a native English speaker?*	Measuring perceived accent transfers. Stereotypical (exaggerated) accent generation known in TTS [Michel et al., 2025].

Table 2: Likert items rated per clip (scale 1–5). See figure 14 for details.

We find that cloned voices are seen as significantly more humanlike, more authoritative, more warm, and more customer-service-like. On a scale of 1 - 5, with an average score of 5 indicating that all annotators rated the audio files as "extremely warm", the source recordings had an average score of 2.4, while cloned recordings had a mean score of 2.8. In Figure 3, we show that the difference between the two distributions is statistically significant at the 95% confidence level based on the permutation test. We present the results in aggregate, but they are significant for each individual model as well. Because male and female voices differ substantially in speech acoustics [Titze, 1989, Smith et al., 2007], we further stratify our analysis by speaker sex,[4] and again the effects persist. See additional figures in the appendix, Figure 12 and 13. We analyze the shift toward perceived *"native"* English speech separately in Section 5.

For additional robustness, we also fit a linear mixed-effects model: `rating` $\sim$ `cloned` + `(1|speaker)` + `(1|annotator)` + `(1|sentence)`, with the TTS model as a fixed effect. This model accounts for our set-up where every speaker is rated by many annotators and every annotator rates many speakers. We also add per-clip audio-quality co-variates (z-scored duration, RMS energy, spectral centroid, and pitch).[5] Under both robustness checks our results hold as every trait remains significant with a 95% confidence interval excluding zero ($n = 3{,}000$ ratings, 86 speakers, 137 annotators; all $p < 0.001$), Appendix Tables 7 and 8.

Is the Perceptual Shift Just An Accent Effect? Since the perception of accent changed most dramatically, we wanted to understand if the other perception effects we saw were just a downstream result from the accent shift (§5). Here, we control for accent shift by restricting our dataset to speakers who self-report no foreign accent at all ($n = 16$), and we find that the cloned voices of these speakers are still rated significantly more humanlike, warmer, more customer-service-like, more trustworthy, more authoritative, and more native-sounding (all $p < 0.05$); only willingness to have an intimate conversation falls short of significance in this subgroup. If we increase to speakers who reported no minimal accent ($n = 30$), every trait is significant. Additionally, we conducted a mediation analysis using perceived English nativeness as the mediator. The direct effect of cloning holds for every trait we tested except warmth, whose gain is largely accounted for by the nativeness shift, Appendix A.6.

We also asked participants to indicate behavioral intentions upon listening to the source and cloned voices. Specifically, we asked how much they would trust answers and responses from the voices and how willing they would be to have intimate conversations with voices. In both cases, we see a significant increase in self-reported trust and in willingness to have an intimate conversation in

[4] We avoid collecting additional demographic information from participants and use speaker sex for demographic analysis, as provided by our annotation platform. We recognize the limitations and exclusionary nature of this classification and aim to have more inclusive measures of speaker identity for future work.

[5] `librosa.pyin` and `librosa.yin` were used to calculate pitch

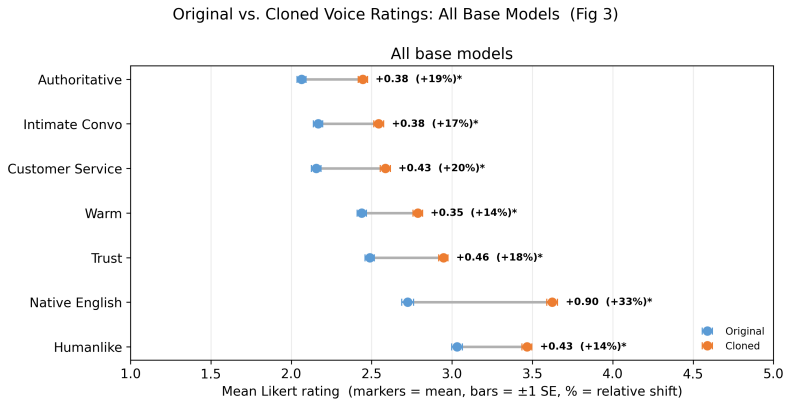


Figure 3: Rating differences between cloned and source voices across all three models (`ChatterBox`, `Coqui-XTTS`, `ElevenLabs V3`), standard error shown. All differences significant at 95% by permutation test ($p < 0.05$).

response to cloned voices compared to source voices. Again, these are stated perceptions and intentions elicited in a rating task rather than real self-incentivized behavior. Future work necessitates in-situ evaluations with self-incentivized users performing tasks that require reliance and information disclosure Zhou et al. [2025].

Our findings are in line with work from [Nightingale and Farid, 2022] that shows that generated faces are seen as more trustworthy than real human faces, but a departure from prior work which found text-to-speech technology to be less trustworthy, [Do et al., 2022]. Our findings also complement the findings of [Mogi et al., 2026], who found that engagement and disclosure are higher among one's own cloned voice rather than the cloned voices of family members or strangers. We discuss these findings and their implications in detail in §6.

4.1 Ablations

To better understand why the cloning process may fail to preserve the reference speaker's unique vocal characteristics, we examine two plausible mechanisms. One possible explanation is that **clip duration** affects fidelity, such that the source clips may be too short for the model to reliably capture speaker-specific characteristics. To test this, we run an ablation where we concatenate the first seven sentences (average of 37 seconds) and use these to generate the eighth. Figure 9 shows the same style transformations being reproduced.

A second possibility is that the observed transformations arise from the default **generation settings**, which can be tuned. `ElevenLabs` and `Chatterbox` provide similarity-to-speaker and expressiveness controls.[6] When comparing these settings to the defaults, we find that the defaults correspond to low expressiveness and high similarity in embedding space, suggesting that the generated voices are already highly similar to the source in terms of speaker fidelity, Figure 10. Evaluating `ElevenLabs` at low expressiveness also reproduces our results exactly, showing that these style transformations persist even at minimal expressiveness, see Figure 11.

5 Voice Cloning as Homogenization

In our results above, we illustrate that voice *"cloning"* is not actually faithfully capturing a speaker's speech characteristics, but that a number of stylistic transformations are being applied. Here, we show that these stylistic transformations are not arbitrary, but are actually converging towards a particular way of speaking, a form of voice homogenization.

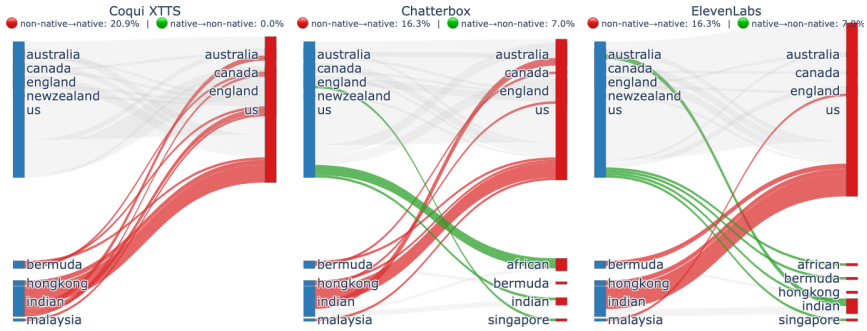


Figure 4: Shifts in classified accent after cloning. Sankey diagrams map source accent labels (left) to post-cloning labels (right). Red flows mark shifts toward dominant Anglophone varieties (U.S., U.K., Canadian, Australian, New Zealand), green the reverse, and gray labels that stay within the same broad category.

5.1 Homogenization in Accent and Cadence

One of the biggest differences between source audio and cloned audio is that cloned audio has a dramatic increase in the perceived English *"nativeness"* of the speech. Our non-native English speakers come from 22 unique language backgrounds, but their voice clones are judged to be significantly more similar to English *"native"* speech than their original voices. To examine this shift at a more granular level, we use `CommonAccent` [Zuluaga-Gomez et al., 2023], an open-source model trained to classify 16 English accents. We find that cloned voices are not only perceived as more *"native"*, but also shift toward a specific set of dominant Anglophone varieties, including U.S. and U.K. English as well as Commonwealth-associated varieties such as Canadian, Australian, and New Zealand English. This pattern is most pronounced for `Coqui-XTTS`, for which all cloned recordings are classified as Inner Circle varieties (Figure 4).

The cadence of the speakers is also transformed and homogenized, with average clip duration being slightly shorter (source vs cloned, 5.11 to 4.92 seconds). We visualize the entropy of source and cloned recording durations and find the entropy of cloned durations to be lower (source vs cloned, 2.06 to 1.86) than that of the entropy of source durations, highlighting a concentration of audio durations, Figure 15.

5.2 Cloned *"Speakers"* Are Less Distinguishable

In addition to audio recording analysis, we use classifier probes to assess how identifiable speakers are from their utterances. We represent every human and cloned clip as a 192-dimensional ECAPA-TDNN speaker embedding [Desplanques et al., 2020]. If voice cloning produces homogenization, then a classifier of fixed capacity should find it harder to distinguish between cloned speakers than between the source voices from which they were derived.

We build random forest and support vector machine (SVM) classifiers and train one set of classifiers on the source recordings and the other set of classifiers on the cloned recordings. Training data is sampled from speakers who have all nine valid utterances ($n = 43$), and we randomly select five as training examples and hold out the remaining four for evaluation. Probes are fit for each cloning system and we report means over 20 random splits.

Speaker identity is recoverable from source recordings but less so from their "cloned" counterparts (Table 3). A probe pooled over all three systems' clones drops from 94% to 81% (random forest) and from 99% to 87% (SVM); per system, the weakest is `ElevenLabs V3` at 61% and 72% (Table 3). Identification errors also scatter more widely: the mean number of distinct incorrect classifications per speaker rises from 0.23 on sources to 1.44 for `ElevenLabs V3`; suggesting that misidentification is spread across a broader set of identities. The margin between same-speaker and different-speaker cosine similarity also narrows from 0.378 on source embeddings to 0.232 on `ElevenLabs` cloned embeddings, Appendix Table 4.

[6]Temperature is also a tunable parameter, but leads more easily to degenerate behavior at non-optimal settings; to be investigated in future work.

Table 3: Speaker-identity probe on ECAPA-TDNN embeddings, per cloning system. *Spread* is the mean number of distinct incorrect speakers a true speaker's clips are assigned to, $n = 43$.

Metric	Source	ChatterBox	Coqui	ElevenLabs
Top-1 accuracy (Random Forest)	94%	79%	81%	61%
Top-1 accuracy (SVM)	99%	93%	93%	72%
Mean incorrect-speaker spread	0.23	0.79	0.71	1.44

5.3 Audio Embeddings Eventually Converge

Beyond the confusion between speaker identities observed in our classification probe, we find evidence that voice cloning applies a directional style transfer that will eventually converge across iterative rounds, further evidence of homogenization.

True voice cloning, in the strict sense, would produce a copy indistinguishable from the reference — and cloning that copy would yield the same result again. Some noise is inevitable in practice, since our reference and target texts differ, but if that noise were random, we would expect the mean of cloned embeddings to remain close to the source, even as individual clips drifted in embedding space. In our iterative cloning experiment, we take participants who have 9 valid sentences ($n = 43$) and repeatedly clone their voices using the cross-sentencing cloning method for 50 rounds. We then embed the cloned audio over time and track their movement, and we see that the transformation is systematic, directional, and convergent. After fifty rounds, embeddings cluster significantly closer together (with the radii of the approximate bounding sphere going from 366 to 336 in Euclidean distance, Figure 5).[7]

We further characterize this drift through cosine similarity, pitch analysis, and emotion classification using `NVIDIA's Audio2Emotion-v3.0` model [Chung et al., 2025]. We see a significant drop in cosine similarity from the source embedding across the rounds; we see a dramatic increase in pitches for both male and female speakers, Figure 5, and pronounced changes to the emotions classified with a significant increase in predictions of anger (potentially due to increasingly empathic reading of the passage, Figure 16).[8] While users are unlikely to clone the same voice fifty times in practice, this experiment serves as an illustrative example of the directional and convergent nature of the voice transformation.

6 Discussion of Potential Harms

Our findings show discernible differences in voice characteristics between human recordings and their cloned counterparts. Because we already mitigate key confounds through manual quality filtering, including human-read disfluencies and lower microphone quality, these differences are unlikely to be artifacts of recording quality or delivery. In addition, annotations are done in pairs — the same annotator rates the source and its clone — so cross-annotator differences do not drive these comparisons. What we report are therefore paired, within-annotator differences between source recordings and their cloned counterparts. Here, we discuss the potential limitations and harms of our findings.

Unfaithful Voice Cloning Voice cloning technology demonstrates limitations to faithfully clone voices. Instead, it applies systematic style transfers (desirable or not) to voices. This transformation has not been made explicit in these systems, and it is unclear if it is possible to perform voice cloning without these stylistic changes. The homogenizing effect of this style transfer over iterative cloning additionally puts the speech model at risk of modal collapse [Shumailov et al., 2024, Alemohammad et al., 2024], as synthetic audio data is often used to train and finetune models [Cornell et al., 2024, Zhao et al., 2024].

Behavioral and institutional risk at scale A potential risk of this systematic application of style transfer is the downstream impacts it poses to human behavior. As text-to-speech and voice cloning gain popularity, voices that sound more persuasive or trustworthy raise safety concerns. Widespread deployment could erode human agency and increase disclosure of sensitive information, and could

[7] Calculated using the distance between the centroid embedding and the farthest individual speaker embedding.

[8] Unlike §4, here, machine classification is used instead of humans to produce labels after 50 rounds of cloning.

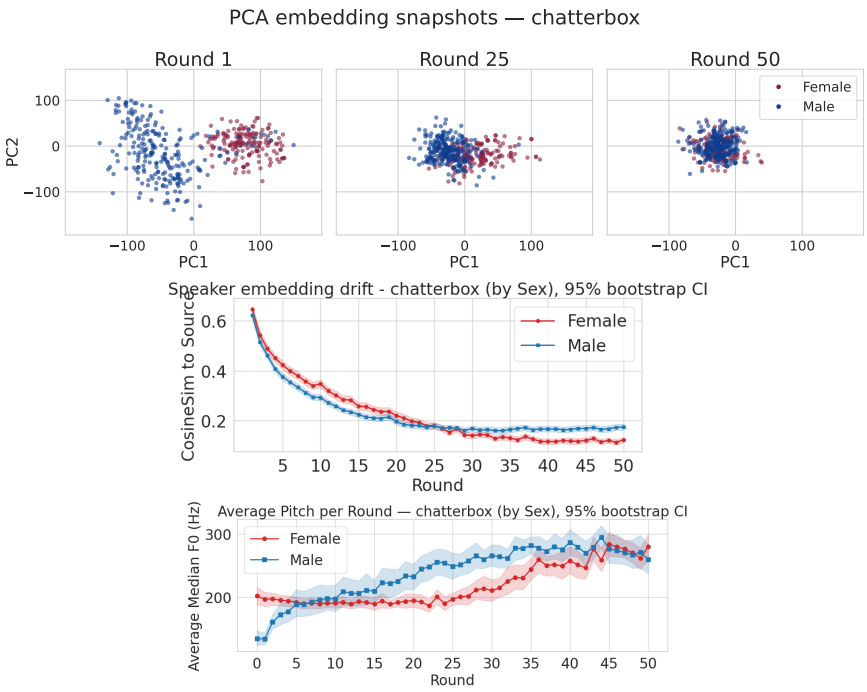


Figure 5: Changes to cloned audio across 50 rounds of repeated cloning with `Chatterbox`. (Top) Audio embeddings via PCA. (Middle) Cosine similarity to source. (Bottom) Pitch change from source; both show 95% bootstrap CIs of the mean.

make misuse even worse. For example, an insurance company whose synthetic voice sounds more persuasive to users disputing a claim, or a fraudulent caller who extracts more personal information from a target. These risks are heightened by the accessibility of current open-sourced or cheap zero-shot voice-cloning systems and by the small amount of reference data (as little as a few seconds) required to generate a convincing clone. Greater transparency about how generated voices are produced and labeled [El Ali et al., 2024] should be investigated in the scientific community and in the public policy sphere.

Normative risk: Whose voice counts as human? This failure of voice cloning fidelity matters not only for individual users who expect their voices to be preserved, but also for broader questions of labor and representation. When cloned voices are deployed in place of human speakers, they may not merely substitute for vocal labor, but reshape it according to narrow norms of what a voice should sound like. Voice cloning could contribute to reinforcing a narrow picture of what counts as a *"human-like"* voice. Cloned voices were rated as more human-like than their source recordings—a speech analogue of hyperrealism in generative media [Miller et al., 2023]. We hypothesize that listeners' prototype of a *"humanlike"* voice might value fluency and varieties of *"Standard"* English, making clones seem more humanlike than source recordings. Voice cloning may therefore encode and amplify existing language ideologies by making cloned voices sound more authoritative, warmer, more native-like, or even more "human" than the originals. Without clear watermarkings of generated speech, continued use of synthetic voices in downstream applications could further entrench the belief that to sound humanlike is to be fluent and to use standard English accents.

7 Conclusion

Voice cloning is often framed as a technology for faithfully preserving or replicating a person's voice. However, our findings show that voice cloning systems are actually applying style transformations on source voices in a homogenizing way. These transformations alter the way speakers are perceived, reduce vocal distinctiveness, and privilege standard accents. As speech technologies continue to grow in adoption, it is critical for researchers and policymakers to have a clear and empirical understanding of how these systems affect the voices they claim to preserve.

Scope and limitations. Our study is limited as we are analyzing zero-shot cloning of read English speech, recorded by U.S.-based speakers who are mostly non-native English speakers, across three systems and a single standard passage. We do not test spontaneous or conversational speech, non-English targets, or few-shot and finetuned cloning. As mentioned in the main text, the trust and disclosure outcomes are self-reported perceptions rather than observed behavior; establishing behavioral consequences requires in-situ study with incentivized users. Speaker sex is the binary category supplied by the recruitment platform and does not capture gender identity. Finally, our annotators are monolingual English speakers, so the perceptual shifts we report may not generalize to a broader, non-English and/or multilingual group.

Acknowledgments and Disclosure of Funding

Many thanks to all the online workers who participated in our studies! Funding provided by TogetherAI.

References

William Agnew, Julia Barnett, Annie Chu, Rachel Hong, Michael Feffer, Robin Netzorg, Harry H Jiang, Ezra Awumey, and Sauvik Das. Sound check: Auditing audio datasets. *AIES*, 2024.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go MAD. In *International Conference on Learning Representations*, volume 2024, pages 53581–53608, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/ebc042e767de551803ccfcc45e2454f5-Paper-Conference.pdf.

Shm Almeda, Robin Netzorg, Isabel Li, Ethan Tam, Skyla Ma, and Bob Tianqi Wei. Labor, power, and belonging: The work of voice in the age of ai reproduction. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 1238–1249, 2025.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. Seed-tts: A family of high-quality versatile speech generation models, 2024. URL https://arxiv.org/abs/2406.02430.

Sercan O. Arik, Jitong Chen, Kainan Peng, Wei Ping, and Yanqi Zhou. Neural voice cloning with a few samples. In *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/abs/1802.06006.

Adam Eric Berkowitz and Miriam E Sweeney. Look who's talking: Voice cloning as tension point between identity and data. *Philosophy & Technology*, 38(4):131, 2025.

Adam Eric Berkowitz and Miriam E Sweeney. Simulating voice and the simulacra of voice clones. *Philosophy & Technology*, 39(1):4, 2026.

Wiebe E. Bijker, Thomas P. Hughes, and Trevor J. Pinch, editors. *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*. MIT Press, Cambridge, MA, 1987.

Haozhe Chen, Run Chen, and Julia Hirschberg. Emoknob: Enhance voice cloning with fine-grained emotion control, 2024. URL https://arxiv.org/abs/2410.00316.

Chaeyeon Chung, Ilya Fedorov, Michael Huang, Aleksey Karmanov, Dmitry Korobchenko, Roger Ribera, and Yeongho Seol. Audio2face-3d: Audio-driven realistic facial animation for digital avatars, 2025. URL https://arxiv.org/abs/2508.16401. Authors listed in alphabetical order.

Samuele Cornell, Jordan Darefsky, Zhiyao Duan, and Shinji Watanabe. Generating data with text-to-speech and large-language models for conversational speech recognition. *Synthetic Data's Transformative Role in Foundational Speech Models*, 2024.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *Interspeech*, 2020.

Tiffany D. Do, Ryan P. McMahan, and Pamela J. Wisniewski. A new uncanny valley? the effects of speech fidelity and human listener gender on social perceptions of a virtual-human speaker. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391573. doi: 10.1145/3491102.3517564. URL https://doi.org/10.1145/3491102.3517564.

Jiachen Du, Hanyu Huang, Xinkai Zou, Shuzi Yin, Bingjie Gao, and Xinyi Fu. The social dynamics of voice cloning: Trust, privacy, and ethical tensions when sharing your ai voice replica. In *Companion Publication of the 2025 Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW Companion '25, page 307–311, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714801. doi: 10.1145/3715070.3749244. URL https://doi.org/10.1145/3715070.3749244.

Abdallah El Ali, Karthikeya Puttur Venkatraj, Sophie Morosoli, Laurens Naudts, Natali Helberger, and Pablo Cesar. Transparent ai disclosure obligations: Who, what, when, where, why, how. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400703317. doi: 10.1145/3613905.3650750. URL https://doi.org/10.1145/3613905.3650750.

ElevenLabs. Eleven v3. https://elevenlabs.io, 2026. General availability release. Accessed: 2026-04.

Gölge Eren and The Coqui TTS Team. Coqui TTS, January 2021. URL https://github.com/coqui-ai/TTS.

Susan T Fiske, Amy JC Cuddy, and Peter Glick. Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, 11(2):77–83, 2007.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096285.

Wiebke Hutiri, Orestis Papakyriakopoulos, and Alice Xiang. Not my voice! a taxonomy of ethical and safety harms of speech generators. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 359–376, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704505. doi: 10.1145/3630106.3658911. URL https://doi.org/10.1145/3630106.3658911.

Shengpeng Ji, Qian Chen, Wen Wang, Jialong Zuo, Minghui Fang, Ziyue Jiang, Hai Huang, Zehan Wang, Xize Cheng, Siqi Zheng, and Zhou Zhao. ControlSpeech: Towards simultaneous and independent zero-shot speaker cloning and zero-shot language style control. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6966–6981, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.346. URL https://aclanthology.org/2025.acl-long.346/.

Zeqian Ju, Yuancheng Wang, Kai Shen, Xu Tan, Detai Xin, Dongchao Yang, Yanqing Liu, Yichong Leng, Kaitao Song, Siliang Tang, Zhizheng Wu, Tao Qin, Xiang-Yang Li, Wei Ye, Shikun Zhang, Jiang Bian, Lei He, Jinyu Li, and Sheng Zhao. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models, 2024. URL https://arxiv.org/abs/2403.03100.

Ayaz Keerio, Bhargav Kumar Mitra, Philip Birch, Rupert Young, and Chris Chatwin. On preprocessing of speech signals. *International Journal of Signal Processing*, 5(3):216–222, 2009.

Maria Labied, Abdessamad Belangour, Mouad Banane, and Allae Erraissi. An overview of automatic speech recognition preprocessing techniques. In *2022 international conference on decision aid sciences and applications (DASA)*, pages 804–809. IEEE, 2022.

Nadine Lavan, Mairi Irvine, Victor Rosi, and Carolyn McGettigan. Voice clones sound realistic but not (yet) hyperrealistic. *PLoS One*, 20(9):e0332692, 2025.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale, 2023. URL https://arxiv.org/abs/2306.15687.

Jyh-An Lee and Xuan Sun. Vocal identity under siege by ai voice cloning technologies. *Singapore Journal of Legal Studies*, pages 46–76, 2026.

Muriel Leuenberger. The role of the voice for identity and implications for voice cloning technology. *Philosophy & Technology*, 38(4):175, 2025.

Yuang Li, Jiaxin Guo, Min Zhang, Ma Miaomiao, Zhiqiang Rao, Weidong Zhang, Xianghui He, Daimeng Wei, and Hao Yang. Pause-aware automatic dubbing using LLM and voice cloning. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors, *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 12–16, Bangkok, Thailand (in-person and online), August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.iwslt-1.2. URL https://aclanthology.org/2024.iwslt-1.2/.

Kimberly T Mai, Sergi Bray, Toby Davies, and Lewis D Griffin. Warning: Humans cannot reliably detect speech deepfakes. *Plos one*, 18(8):e0285333, 2023.

Shira Michel, Sufi Kaur, Sarah Elizabeth Gillespie, Jeffrey Gleason, Christo Wilson, and Avijit Ghosh. "it's not a representation of me": Examining accent bias and digital exclusion in synthetic ai voice services. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 228–245, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714825. doi: 10.1145/3715275.3732018. URL https://doi.org/10.1145/3715275.3732018.

Elizabeth J. Miller, Ben A. Steward, Zak Witkower, Clare A. M. Sutherland, Eva G. Krumhuber, and Amy Dawel. Ai hyperrealism: Why ai faces are perceived as more real than human ones. *Psychological Science*, 34(12):1390–1403, 2023. doi: 10.1177/09567976231207095. URL https://doi.org/10.1177/09567976231207095. PMID: 37955384.

Yamato Mogi, Wataru Akahori, and Naomi Yamashita. Exploring the effects of different chatbot voice identities on self-disclosure. CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3790546. URL https://doi.org/10.1145/3772318.3790546.

Sophie J. Nightingale and Hany Farid. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, 2022. doi: 10.1073/pnas.2120481119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2120481119.

Minju Park, Seunghyun Lee, Juhwan Ma, and Dongwook Yoon. Ai twin: Enhancing esl speaking practice through ai self-clones of a better me. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3790266. URL https://doi.org/10.1145/3772318.3790266.

Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022, 2021.

Daniel Platnick, Bishoy Abdelnour, Eamon Earl, Rahul Kumar, Zahra Rezaei, Thomas Tsangaris, and Faraj Lagum. Preset-voice matching for privacy regulated speech-to-speech translation systems. In Ivan Habernal, Sepideh Ghanavati, Abhilasha Ravichander, Vijayanta Jain, Patricia Thaine, Timour Igamberdiev, Niloofar Mireshghallah, and Oluwaseyi Feyisetan, editors, *Proceedings of*

the Fifth Workshop on Privacy in Natural Language Processing, pages 52–62, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.privatenlp-1.6/.

Zengyi Qin, Wenliang Zhao, Xumin Yu, and Xin Sun. Openvoice: Versatile instant voice cloning, 2024. URL https://arxiv.org/abs/2312.01479.

Durwa R Chavan, Prachi Moon, and Emma Dixon. Public reflections on the use of augmentative and alternative communication (aac) devices by people with i/dd in everyday life. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400706769. doi: 10.1145/3663547.3759720. URL https://doi.org/10.1145/3663547.3759720.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023.

Resemble AI. Chatterbox-TTS. https://github.com/resemble-ai/chatterbox, 2025. GitHub repository.

Choe Sang-Hun. South korea uses ai to help seniors with dementia. *The New York Times*, April 2026. URL https://www.nytimes.com/2026/04/28/world/asia/korea-ai-seniors-dementia.html.

Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers, 2023. URL https://arxiv.org/abs/2304.09116.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

David RR Smith, Thomas C Walters, and Roy D Patterson. Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled. *The Journal of the Acoustical Society of America*, 122(6):3628–3639, 2007.

Ingo R Titze. Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America*, 85(4):1699–1707, 1989.

Charles Van Riper. *Speech correction*. Prentice-Hall New York, 1972.

Maitreyee Wairagkar, Nicholas S Card, Tyler Singer-Clark, Xianda Hou, Carrina Iacobacci, Lee M Miller, Leigh R Hochberg, David M Brandman, and Sergey D Stavisky. An instantaneous voice-synthesis neuroprosthesis. *Nature*, 644(8075):145–152, 2025.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers, 2023. URL https://arxiv.org/abs/2301.02111.

Bryn Wells-Edwards. What's in a voice? the legal implications of voice cloning. *Ariz. L. Rev.*, 64: 1213, 2022.

Langdon Winner. Do artifacts have politics? In *Computer ethics*, pages 177–192. Routledge, 2017.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt, 2023. URL https://arxiv.org/abs/2301.13662.

Shuaijiang Zhao, Tingwei Guo, Bajian Xiang, Tongtang Wan, Qiang Niu, Wei Zou, and Xiangang Li. Advancing speech language models by scaling supervised fine-tuning with over 60,000 hours of synthetic speech dialogue data. *arXiv preprint arXiv:2412.01078*, 2024.

Qingxiao Zheng, Zhuoer Chen, and Yun Huang. Learning through ai-clones: Enhancing self-perception and presentation performance. *Computers in Human Behavior: Artificial Humans*, 3: 100117, 2025.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. REL-A.I.: An interaction-centered approach to measuring human-LM reliance. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11148–11167, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.556. URL https://aclanthology.org/2025.naacl-long.556/.

Juan Zuluaga-Gomez, Sara Ahmed, Danielius Visockas, and Cem Subakan. Commonaccent: Exploring large acoustic pretrained models for accent classification based on common voice. *Interspeech*, 2023.

A Methods Details

A.1 Grandfather Passage

You wish to know about my grandfather. Well, he is nearly 93 years old, yet he still thinks as swiftly as ever. He dresses himself in an old black frock coat, usually several buttons missing. A long beard clings to his chin, giving those who observe him a pronounced feeling of the utmost respect. When he speaks, his voice is just a bit cracked and quivers a bit. Twice each day he plays skillfully and with zest upon a small organ. Except in the winter when the snow or ice prevents, he slowly takes a short walk in the open air each day. We have often urged him to walk more and smoke less, but he always answers, "Banana oil!" Grandfather likes to be modern in his language.

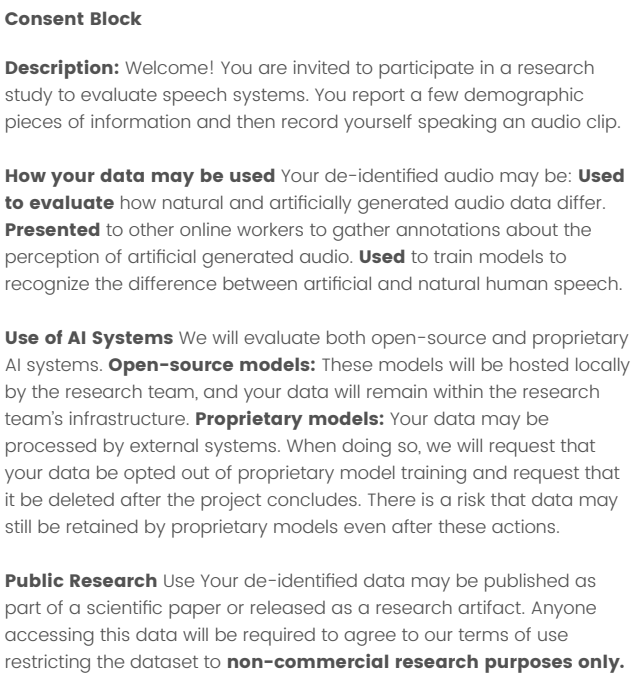


Figure 6: Screenshot of the consent block provided to the speaker.

A.2 Privacy Terms

The participants are asked to give their consent to use their audio data in voice cloning technology for research purposes only, and will be made aware that we will be: Opting out of proprietary model training. Asking for their data to be deleted from companies with proprietary models at the conclusion of this project.

For our research purposes, their data was also used in voice cloning technology on open-sourced models, which we will host locally (i.e., their data stays with our research team). For downstream use, their audio data may be used to perform analysis on how audio data and voice cloned data differ. presented to other online workers to gather annotations about the perception of artificial vs natural voice. Used to train models to recognize the difference between artificial and natural human speech. Shared anonymously online via a public research dataset that cannot be used for commercial purposes (explicit guidelines below).

Forbidden Uses of the public dataset include:

- Generating, enabling, or promoting hate speech, harassment, discrimination, misinformation, or culturally offensive or harmful content
- Beyond explicit research purposes, voice cloning, speaker impersonation, or the creation of synthetic voices intended to resemble or replicate any participant
- Attempting to identify, re-identify, or infer the identity of any participant, including attempts to extract personally identifiable information from the audio or associated metadata
- Any commercial, for-profit, or revenue-generating use, including product development, advertising, or monetized services
- Any use that misrepresents, stereotypes, or falsely attributes characteristics, language abilities, accents, or identities to the speakers
- Redistribution of the dataset under terms that conflict with or weaken these restrictions

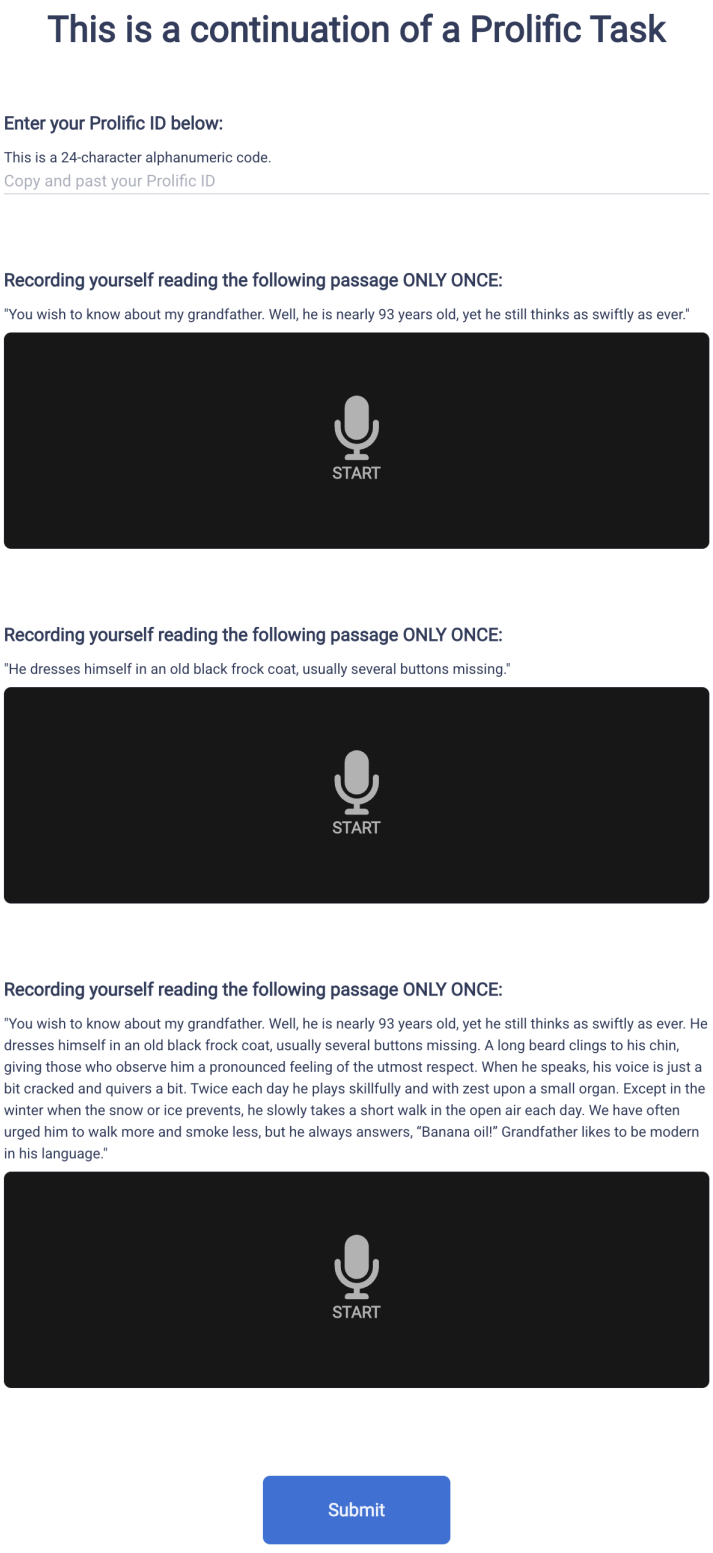


Figure 7: Screenshot of speaker task.

IIElevenLabs Sign up

(a) Data Subject Rights

Depending on where you are located or reside, you may have certain rights regarding the Personal Data we maintain about you and certain choices about what Personal Data we collect from you, how we use it, and how we communicate with you. Such rights may include the rights to:

- **Access Your Personal Data:** Receive confirmation of processing and request access the Personal Data we maintain about you. Most of your Personal Data can be accessed and exported by logging into your account. Where provided by applicable law, you may request that we include a description of the purpose of the processing of your personal data, third parties with which your data has been shared, and information regarding the safeguards for international data transfer.
- **Correct Your Personal Data:** Request to update and correct inaccuracies in your Personal Data. Most of your Personal Data can be accessed and corrected by logging into your account and accessing your account settings.
- **Delete Your Account:** Request to close your account by submitting your request using your account settings (see more information here).
- **Delete Your Personal Data:** Request that ElevenLabs delete all Personal Data we maintain about you including your account and the Voice Data associated with your account. This deletion also includes the personal data processed under the legal basis of consent. Where provided by applicable law, you may request that your Personal Data be anonymized rather than deleted.

Figure 8: ElevenLabs Privacy Terms

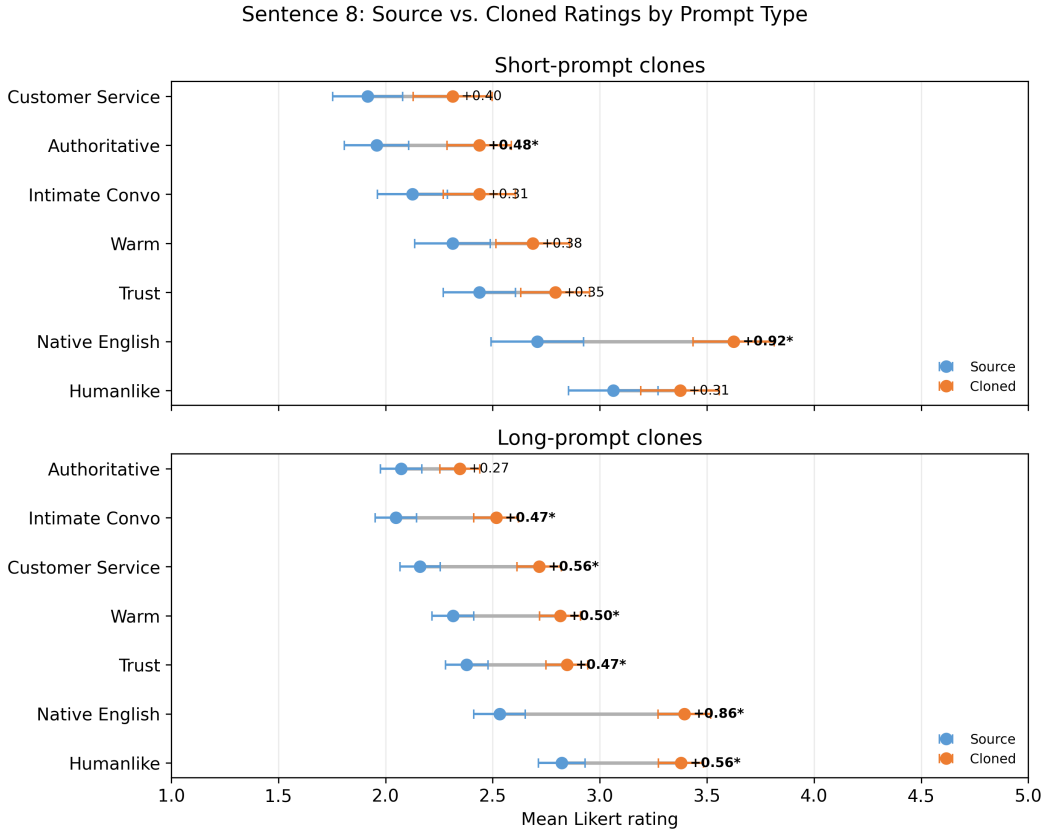


Figure 9: Comparison cloning with long vs short source clips (37 seconds versus 5 seconds). Long source clips are a concatenation of clips 1 through 7, whereas short clips are just clip 7.

A.3 Findings Ablations

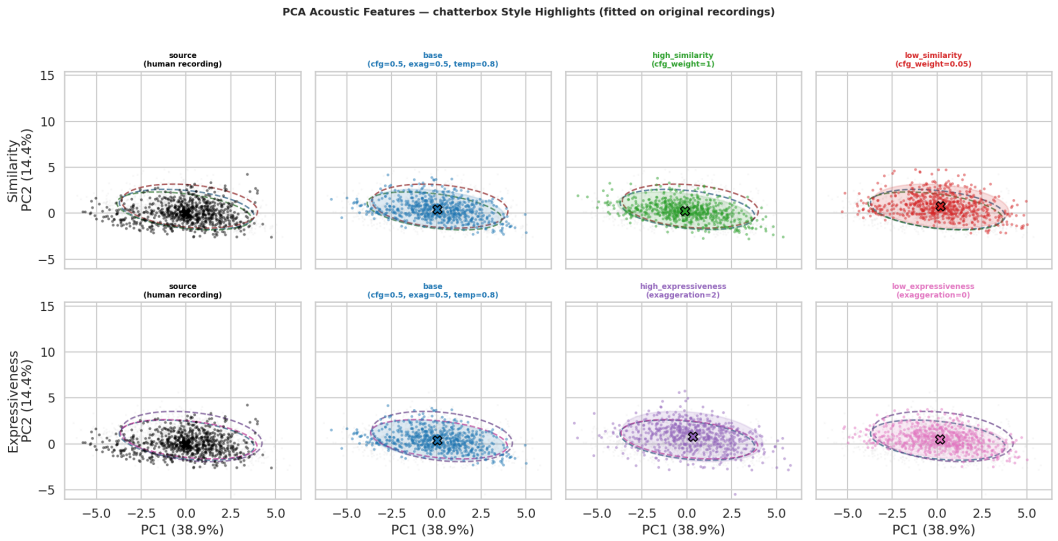


Figure 10: PCA projections of Chatterbox acoustic embeddings under different styles. Across settings, generated clips remain close to the source distribution in embedding space, suggesting that the observed style shifts are not primarily explained by reduced speaker similarity or increased expressiveness.

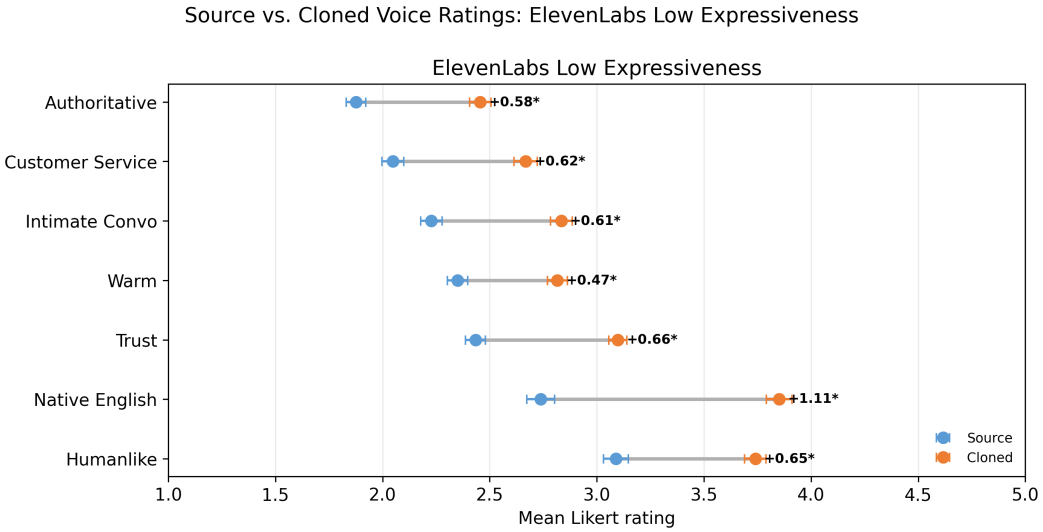


Figure 11: Human annotations on ElevenLabs clones under "low expressiveness"

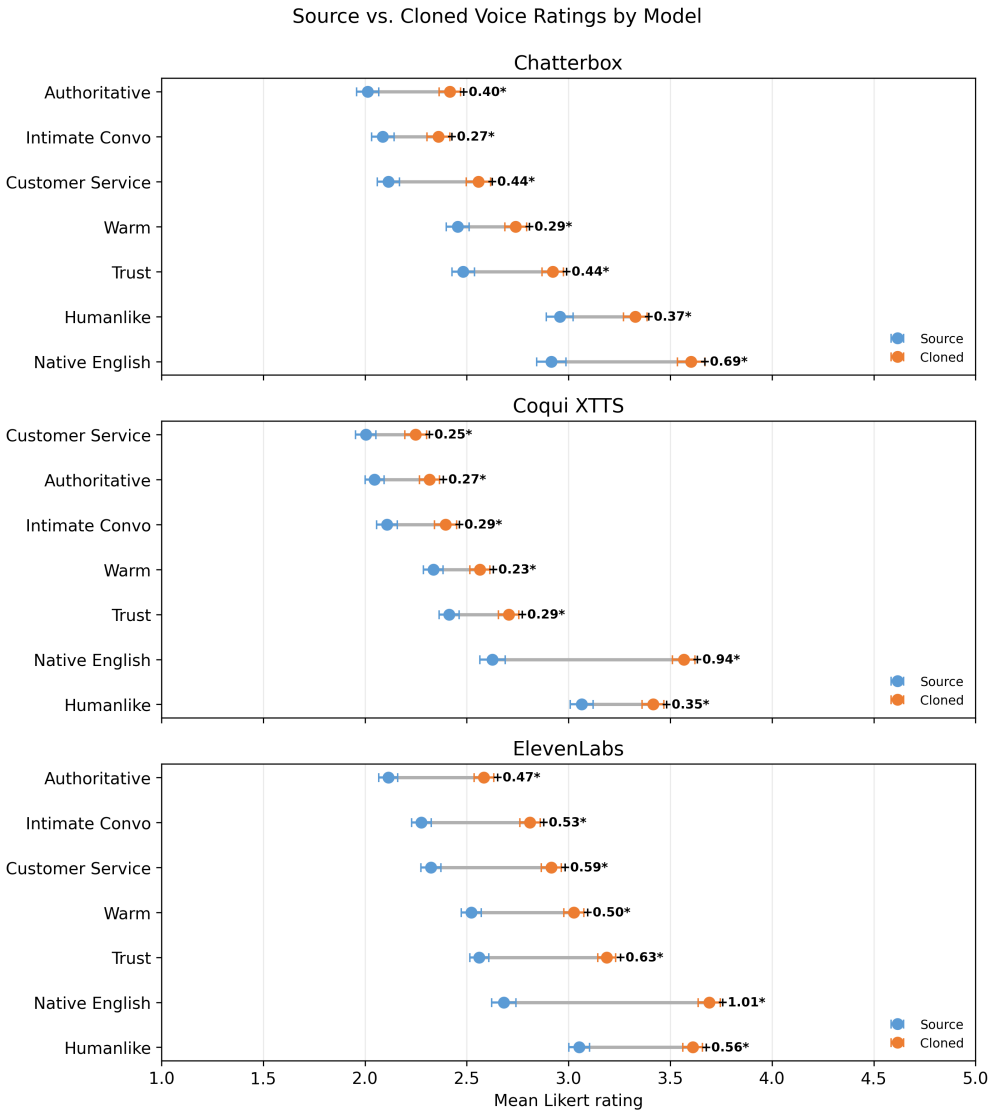


Figure 12: Rating differences between cloned and source voices by model.

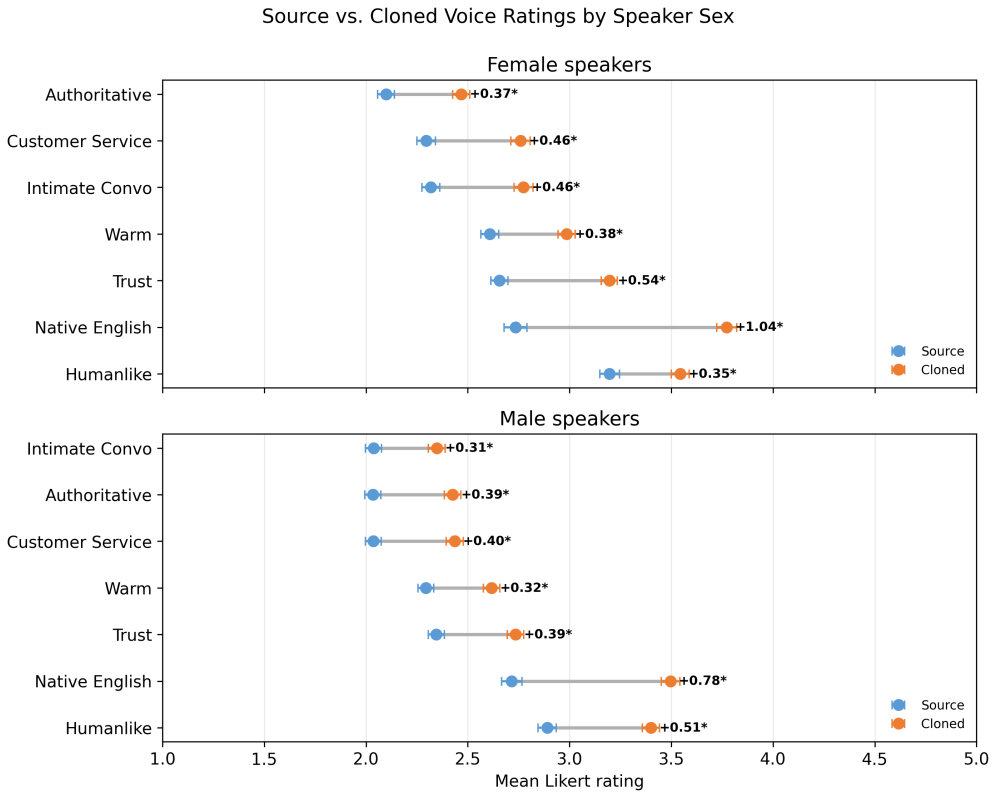


Figure 13: Rating differences between cloned and source voices by speaker sex.

A.4 Context Length

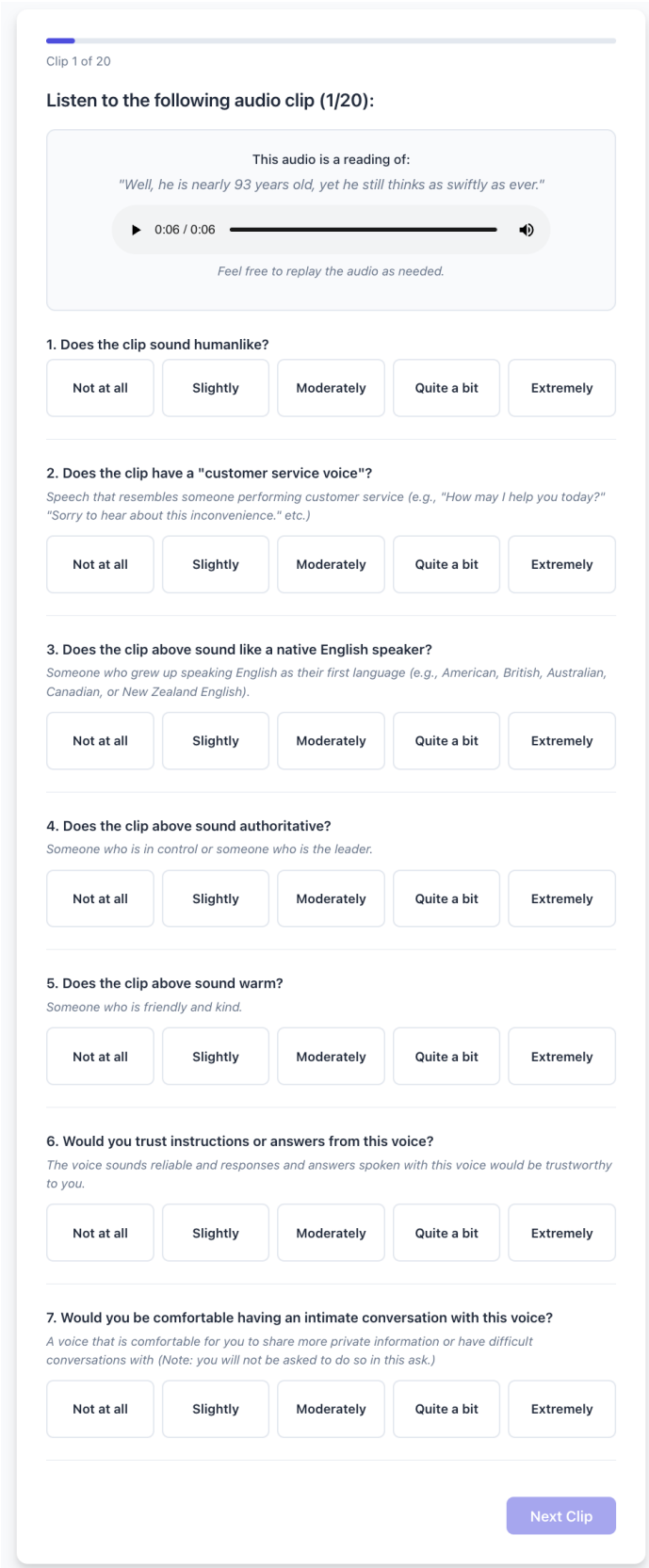


Figure 14: Screenshot of Annotation Task

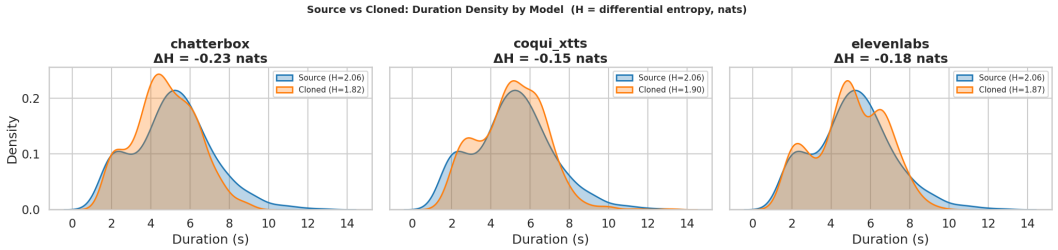


Figure 15: Change in entropy (denoted as nats) for duration distribution between source and cloned recordings

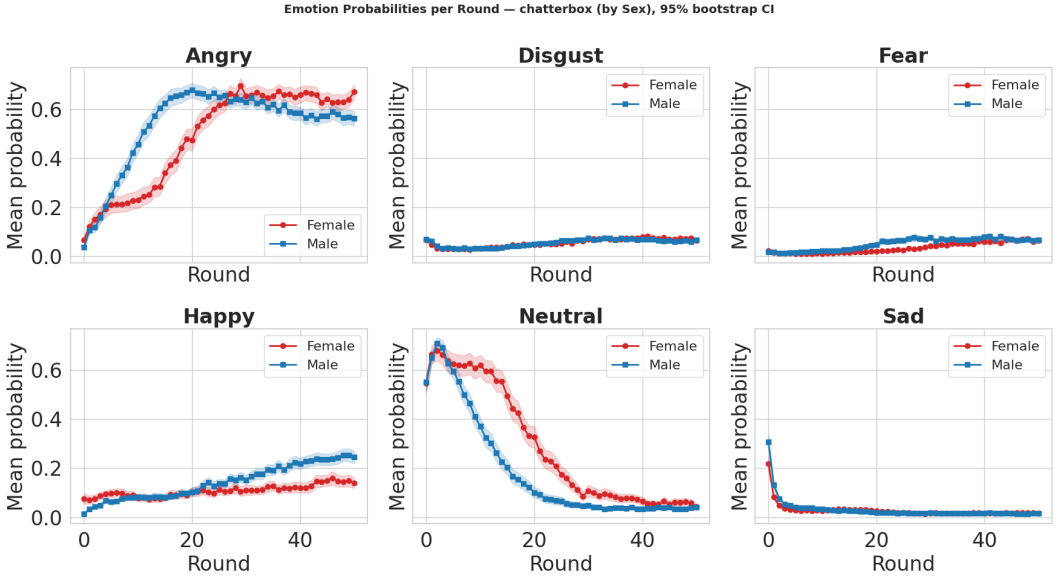


Figure 16: Change predicted emotion across the 50 iterative rounds of cloning, visualized with 95% confidence interval of the mean calculated via bootstrap resampling.

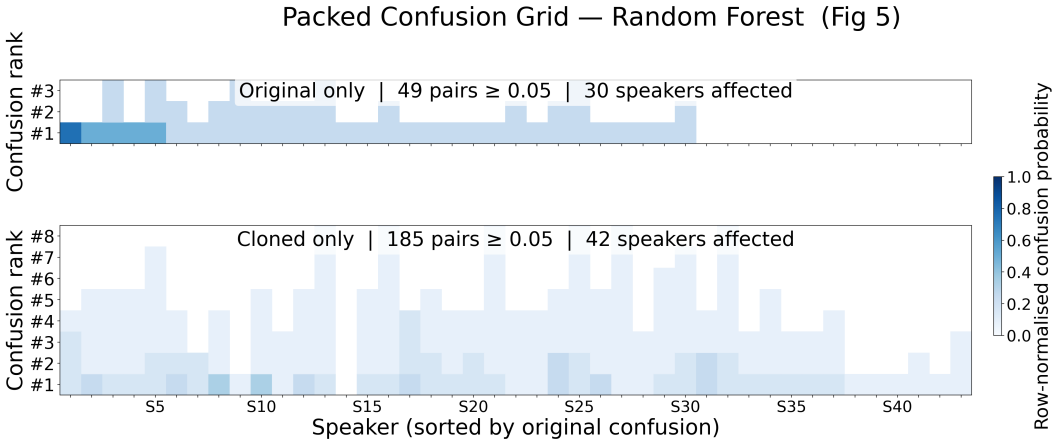


Figure 17: Probability distribution on incorrect speakers for source (top) and cloned (bottom) recordings. Each column represents a speaker, and each box in that column represents the probability placed on an incorrect speaker. The more boxes stacked in a column, the more confusion the classifier had in identifying the speaker's true identity. The top figure shows significantly less confusion than the bottom figure, indicating that the classifier confused cloned recordings more often and with more speakers.

A.5 Speaker-Verification Metrics on ECAPA Embeddings

Table 4: Speaker-verification metrics on ECAPA-TDNN embeddings. *Within* is the mean same-speaker cosine similarity and *between* is the mean different-speaker cosine similarity. Cloning closes the gap between mean same-speaker and different-speaker cosine similarity, where a larger gap suggest more separability between speakers. This is most noticeable for `ElevenLabs`.

Condition	**Within**	**Between**	**Gap**
Source	0.716	0.338	0.378
Clone `ChatterBox`	0.610	0.303	0.307
Clone `Coqui-XTTS`	0.607	0.281	0.326
Clone `ElevenLabs V3`	0.558	0.326	0.232

A.6 Accent Controls and Per-Model Effects

Speakers who report no accent. Table 5 limits our experiments to speakers based on their self-reported foreign-accent strength (LEAP-Q). Almost all the effects persist, even among speakers who rated themselves as having no accent.

Table 5: Clone−source rating differences restricted by self-reported accent strength. p-values from permutation tests; * denotes $p < 0.05$.

	No accent ($n = 16$)		**No/minimal** ($n = 30$)		**All** ($n = 86$)
Trait	Δ	p	Δ	p	Δ
Humanlike	$+0.28^*$	0.0050	$+0.39^*$	0.0002	$+0.43^*$
Customer Service	$+0.23^*$	0.0104	$+0.28^*$	0.0002	$+0.43^*$
Native English	$+0.24^*$	0.0138	$+0.38^*$	0.0002	$+0.90^*$
Authoritative	$+0.22^*$	0.0196	$+0.30^*$	0.0002	$+0.38^*$
Warm	$+0.20^*$	0.0380	$+0.28^*$	0.0002	$+0.35^*$
Trust	$+0.27^*$	0.0016	$+0.36^*$	0.0002	$+0.46^*$
Intimate Convo	$+0.13$	0.1930	$+0.22^*$	0.0022	$+0.38^*$

Mediation by perceived native-English. Table 6 decomposes the total cloning effect into the part transmitted through perceived native-English and the part that is not. The direct effect remains positive for all six traits, with a 95% CI excluding zero for five of them; warmth is the exception.

Table 6: Mediation analysis with cloning as treatment and perceived native-English as mediator, for every rated trait except the mediator itself; 1,000 speaker-cluster bootstraps. The direct effect is distinguishable from zero for all traits except warmth, listed last. c is the total effect, c' the direct effect controlling for the mediator.

Trait	**Total** c	**Direct** c'	c' **95% CI**
Authoritative	0.382	0.193	(0.124, 0.261)
Customer Service	0.429	0.215	(0.132, 0.287)
Trust	0.459	0.134	(0.068, 0.199)
Humanlike	0.435	0.102	(0.007, 0.196)
Intimate Convo	0.377	0.081	(0.013, 0.152)
Warm	0.348	0.071	(−0.005, 0.150)

Table 7: Clone−source effect per trait from crossed random-effects models (β, 1–5 Likert), with and without per-clip audio-quality covariates, and Cohen's d. Random intercepts: speaker, annotator, sentence; TTS model fixed. $n = 3{,}000$ ratings; all $p < 0.001$.

Trait	β **(95% CI)**	β **+ quality covariates**	**Cohen's** d
Warm	+0.35 [0.28, 0.41]	+0.35 [0.26, 0.43]	0.30
Authoritative	+0.38 [0.32, 0.44]	+0.37 [0.29, 0.45]	0.34
Trust	+0.46 [0.40, 0.52]	+0.44 [0.36, 0.52]	0.41
Humanlike	+0.43 [0.36, 0.51]	+0.40 [0.30, 0.50]	0.35
Customer Service	+0.43 [0.36, 0.50]	+0.44 [0.35, 0.53]	0.36
Native English	+0.90 [0.82, 0.98]	+0.78 [0.68, 0.88]	0.66
Intimate Convo	+0.38 [0.32, 0.44]	+0.36 [0.28, 0.44]	0.32

Per-model effects. Table 8 reports the crossed random-effects estimates separately for each system. The direction is consistent across all three, with `ElevenLabs V3` generally showing the largest shifts.

Table 8: Clone−source effect (β, Likert 1–5) by TTS system, from crossed random-effects models with random intercepts for speaker, annotator, and sentence. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Trait	`ChatterBox`	`Coqui-XTTS`	`ElevenLabs V3`
Warm	$+0.29$ [0.17, 0.40]***	$+0.23$ [0.12, 0.34]***	$+0.50$ [0.40, 0.61]***
Authoritative	$+0.40$ [0.29, 0.52]***	$+0.27$ [0.16, 0.38]***	$+0.47$ [0.37, 0.57]***
Trust	$+0.44$ [0.33, 0.55]***	$+0.29$ [0.19, 0.40]***	$+0.63$ [0.53, 0.72]***
Humanlike	$+0.37$ [0.22, 0.52]***	$+0.35$ [0.23, 0.48]***	$+0.56$ [0.44, 0.67]***
Customer Service	$+0.44$ [0.32, 0.57]***	$+0.25$ [0.13, 0.36]***	$+0.59$ [0.48, 0.70]***
Native English	$+0.69$ [0.54, 0.84]***	$+0.94$ [0.80, 1.07]***	$+1.01$ [0.88, 1.13]***
Intimate Convo	$+0.27$ [0.17, 0.38]***	$+0.29$ [0.18, 0.40]***	$+0.53$ [0.43, 0.64]***

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state both main findings: (1) voice cloning applies systematic style transfer (cloned voices rated more authoritative, warm, customer-service-like, human-like, trustworthy, with greater willingness for intimate disclosure), and (2) voice cloning homogenizes speaker identity (accent shifts toward dominant Anglophone varieties, reduced speaker distinguishability via classifier probes, convergence in embedding space over iterative cloning).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes],

   Justification: Section 6 discusses key limitations: behavioral measures are self-reported rather than in-situ (acknowledged with a call for future work in section 4);

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A].

   Justification: This paper presents empirical findings from a human subjects study and does not include theoretical results or mathematical proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 3 provides details of all experimental components: audio collection, voice cloning setup, annotation design, and analysis methods. Materials are screenshotted and included in the appendix. Audio examples are available via an anonymized OSF link. Code is available on anonymized GitHub.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   The source audio dataset will be released on HuggingFace upon publication for non-commercial research purposes only; audio examples are currently available via an anonymized OSF link. The analysis code is also linked.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Section 3 specifies experimental details. Other details include: data splits for classifier probes, default model settings for voice cloning with ablations on expressiveness and similarity controls, annotation session design, and the acoustic feature extraction for classifiers.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   All main rating comparisons (Figure 3) are tested with permutation tests ($p < 0.05$), with standard errors visualized. The iterative cloning analyses (Figures 5 and 16) report 95% confidence intervals calculated via bootstrap resampling, explicitly stated in figure captions.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   The paper does not report compute resources, as almost all computation could be run locally given the open-sourced models and API calls. Neglible amoutn of compute.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   The study is IRB-approved. All speakers provided informed consent, including disclosure of data uses and risks from proprietary model processing. Speakers were compensated at $18 USD/hour. Privacy protections include opt-out from proprietary model training, planned data deletion from ElevenLabs post-project, and dataset terms prohibiting re-identification, impersonation, and commercial use. The paper acknowledges limitations of binary sex classification.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 6 discusses negative impacts, including eroded human agency via increased trust in synthetic voices, enabling more convincing fraud and manipulation, homogenization of vocal identity erasing cultural and linguistic diversity, reinforcement of language ideologies privileging standard accents, labor displacement, and risk of modal collapse in models trained on synthetic audio.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The dataset will be released with explicit terms of use prohibiting impersonation beyond research purposes, re-identification of participants, commercial use, and redistribution that weakens protections. For proprietary model processing (ElevenLabs), opt-out from model training and data deletion were requested. Open-source models were run locally to minimize data exposure. Dataset access requires agreement to non-commercial research terms.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

13. Answer: [Yes]

14. Justification: All models and tools are cited: ChatterBox, Coqui-XTTS, ElevenLabs V3, CommonAccent, ECAPA-TDNN , librosa, Whisper, and NVIDIA Audio2Emotion. The Grandfather Passage is credited to Van Riper (1972).

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

15. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduces a dataset of 86 non-native English speakers reading the Grandfather Passage (699 sentence-level clips) plus voice clones from three models. Consent procedures and data use terms are documented in Figure 6. Forbidden uses are enumerated in The dataset will be released on HuggingFace for research purposes only upon publication.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

16. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

Justification: The paper includes: the full Grandfather Passage text, the consent block with data use disclosures, screenshots of the speaker recording task, screenshots of the annotation task, all Likert items with rationales, and compensation information ($18 USD/hour for speaker).

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

17. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer:[Yes]

Justification: The study is IRB-approved. Potential risks disclosed to participants include the possibility that data may be retained by proprietary models even after deletion requests. Participants consented to use of their audio in voice cloning for research, annotation by other crowd workers, and potential use in training detection models.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

18. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [No]

Justification: LLMs are not used as a core methodological component, just used for standard writing, editing, and code suggestions.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.