Evaluating the Importance of Demographic and Technical Factors in Creating Authentic-Sounding AI-Generated Human Voice Clones

V. Lakshmanan, D. Ferri, R. Agarwal, B. Kaur, F. Mysha, P. Lim, and G. J. Gerling
University of Virginia, Systems & Information Engineering, Charlottesville, VA, USA, gg7h@virginia.edu

Abstract— **Business and governmental institutions face growing threats from synthetic audio deepfakes due to advances in voice cloning and artificial intelligence. By accessing a short recording of a person's voice, malicious actors can clone it to say anything they like. This poses serious risks of fraud, identity theft, and loss of trust. While much prior research has explored defensive postures, limited works have considered the factors that make a cloned voice sound authentic. This effort investigates factors leading to more authentic sounding AI-generated clones of the human voice. A voice library of about 350 short samples was created, spanning a range of demographic (age, gender, ethnicity) and technical factors (cloning tool, training time, background noise). Using optimization techniques, a subset of 81 voices (67 cloned and 14 authentic) were selected for an online survey with human listeners (n=449). Each voice was also assessed by the NISQA speech quality and naturalness model. Overall, human listeners perceived authentic voices as more realistic than cloned voices. However, subsets of cloned voices of certain technical and demographic factors were indistinguishable from authentic voices. Finally, human and machine generated ratings did not correlate, indicating that NISQA may evaluate voice authenticity in ways distinct from human listeners.**

I. Introduction

Rapid advancements in artificial intelligence (AI) are poised to increase human productivity. Like any new capability, however, such technology can also be used toward malicious ends. An emerging concern is the impact of deepfakes: hyper-realistic, synthetic media that learn patterns from public datasets and produce human-like responses [1]. At the core of many deepfake technologies lies a powerful machine learning framework, such as Generative Adversarial Networks (GANs), designed specifically to generate convincing synthetic media. GANs use two neural networks: a generator that creates the content and a discriminator that evaluates its authenticity. With continuous feedback and refinement, this dynamic drives both networks toward progressively greater accuracy and realism [2]. Therefore, traditional methods of fraud detection are struggling to keep pace with AI-generated deepfakes, which are increasingly used in scams and identity theft [3]. Realistic voice clones may enable unauthorized access to an institution's authentication system and put an individual at risk of a personal security breach.

Unlike traditional alterations using Computer-Generated Imagery or manual modifications to graphics, AI-generated deepfake tools allow anyone, without a need for significant prior experience or expertise, to create lifelike imitations [3]. Indeed, commercial and open-source cloning tools, such as Eleven Labs, Lovo, and FineVoice, are low cost and easy to use. With just a few audio samples scraped from the public domain and/or social media sources, a victim's voice can be cloned with alarming accuracy [4].

Red team tactics are helping uncover vulnerabilities in legacy systems, exposing weaknesses fraudsters may exploit. These tactics often use live, but non-customer, accounts to execute fraud attacks and evaluate customer protection measures [5]. Other efforts have conducted machine assessments of voice authenticity, for example, the NISQA speech quality and naturalness model. Yet others have used human listeners and found they can detect deepfakes only 73% of the time [1]. For the time being and probably well into the future, we will need to leverage both human and machine assessments in the evaluation of voice authenticity. Toward that end, we need to obtain a better understanding of key traits that make some voice clones more authentic and thereby more dangerous.

The work described herein adopts the perspective of a red team tasked with developing AI-generated voice clones capable of deceiving both human listeners and machine detection algorithms. We explore factors that drive how human listeners and machine assessments assess the relative authenticity of cloned voices. Factors underlying the creation of voice clones include those of nature both demographic (age, gender, and ethnicity) and technical (cloning tool, training time, and background noise). We seek to identify those factors most vulnerable to misuse, with the ultimate purpose of strengthening deepfake detection tools, which may thereby better protect individuals from future threats.

II. Methods

To investigate the factors that influence how authentic a cloned voice sounds, a comprehensive methodology was designed and centered around 6 key variables: 3 technical factors consisting of cloning tools, training time, and presence of background noise and 3 demographic factors consisting of speaker age, gender, and ethnicity. A total of 4 voice cloning tools, including 3 commercially available and 1 open-source platform, were used to generate a large library of cloned voices. Training samples of 15, 30, and 60 seconds were uploaded into each tool, and background noise was added to the outputs to simulate real-world audio conditions. Voices were cloned from original recordings of persons with diverse

demographic characteristics, including having a Hispanic background with Spanish as a first language. To ensure the study remained feasible within survey response limits, a subset of 67 cloned voices were selected from the full set using an optimization approach that balanced all technical and demographic factors, in addition to the 14 authentic voices that were provided by speakers. These selected voices were then evaluated using both an automated voice quality assessment tool and a human survey, with 449 participants rating the perceived authenticity of each voice through a standardized listening interface.

A. Technical and demographic factors

To tease apart factors that lead to making a cloned voice sound more real, a set of 6 factors was identified, Fig. 1. These factors were derived from conversations with subject matter experts with experience in red team tactics. With respect to the technical factors, three commercially available (Eleven Labs, FineVoice, and Lovo) and one open-source (F5-TTS) tools were used to produce the cloned voices. These tools allow a user to upload a training sample consisting of 15 seconds to 30 minutes of speech. Then, to better study the impact of model training times on clone authenticity, versions of clones were made at 15, 30, and 60 seconds. A reasonable assumption is that these are durations of audio clips or videos that a fraudster would have access to when cloning a victim's voice. Next, background noise was added by merging a cafe or coffee shop type of noise with the cloned voice output by the tool to create a combined voice. Half of all the voices were free of background noise with noise inserted into the remaining half. Note that noise was selected and adjusted to reflect a realistic volume level that could be experienced during a phone conversation.

With respect to the demographic factors, clones were made of both males (5) and females (9) and from a range of different ages (21-78). Another important consideration was the whether the speaker was a native English speaker (11) or had a Hispanic background with Spanish as a first language (3).

B. Procedure for creating 336 cloned voices

We first recruited 14 individuals to create authentic voice samples, of demographics as listed in the paragraph above. Each person read and recorded their voice for a sample script that took about 60 seconds to read. Each voice was then cut into two portions of 15 and 30 second samples.

All three samples (15, 30, 60 seconds) were uploaded into each cloning tool for training, representing varying levels of voice data availability that might realistically be encountered in real-world scenarios. Therefore, across the four cloning tools, a total of 168 unique clones were generated (14 individual speakers * 3 time durations * 4 cloning tools). Note that each cloned voice was given one of five different scripts to output to create variety in the messages.

Each of the 168 cloned voices was then duplicated and combined with an audio file with background noise, as noted in *II. Methods, A*. This produced a set of 336 unique cloned voices. Therefore, the complete library of voices consisted of 14 authentic voices and 336 cloned voices.

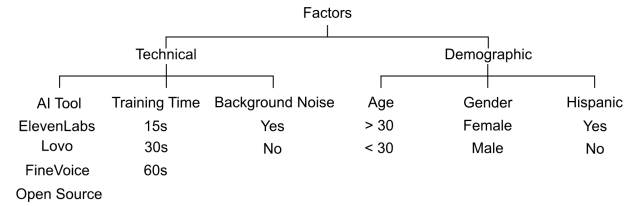

Fig. 1. Factors included in voice cloning and selection of subsets of cloned voices for the human survey experiments.

C. Optimization to downsample cloned voices

To afford sufficient statistical power analysis, responses from 30 human listeners were required per authentic and cloned voice. Moreover, we anticipated receiving feedback from 405 survey respondents and asking each respondent to evaluate 6 voices (5 cloned and 1 authentic). To attain 30 responses per voice, our library of 336 cloned voices needed to be filtered to 67. Therefore, we performed an optimization to choose an appropriate set of 67 cloned voices, with constraints set to ensure an even spread numerically across the technical and demographic factors.

In particular, the branch and bound method of nonlinear optimization was selected, implemented via OpenSolver in Excel [6]. The objective function was set to select exactly 67 cloned voices. For the technical factors of AI tool and training time, the voices were selected such that these factors were distributed evenly across their levels. Each of the 14 voice speakers must not have exceeded 6 total instantiations of their voice, while ensuring a similar number of background/non-background noise voices. For the demographic factors, to keep the proportion of Hispanic cloned voices consistent with 21.4% of voices, 15 (22.3% of 67) Hispanic derived voices were selected, while gender constraints were determined similarly (35.7% male voices resulted in 23 male cloned voices, for 34.3% of 67).

The results led to the selection of the following set of 67 cloned voices, with gender (23 male, 44 female); ethnicity (15 Hispanic, 52 non-Hispanic); training time (23 of a 15 second training time, 22 of a 30 second training time, 22 of a 60 second training time); training tool (17 using Eleven Labs, 17 Lovo, 17 FineVoice, and 16 F5-TTS); and background noise (34 with background noise, 33 without background noise).

D. Machine evaluation using NISQA

In addition to human participant survey methods, we sought to evaluate the voices using machine assessment. After considering several tools, including ASVTorch, the tool NISQA was selected. NISQA, which stands for Non-Intrusive Speech Quality Analysis, predicts speech quality of a given sample [7]. The tool's "NISQA - TTS" model can score the naturalness of a voice sample, by analyzing its acoustic properties through a deep neural network trained on human-annotated speech quality data. The model extracts key

features, including spectral characteristics, temporal variations, and distortions caused by noise, reverberation, or compression. By comparing these features to patterns observed in natural human speech, NISQA assigns a Mean Opinion Score (MOS) that reflects perceived quality. Voices generated by AI tools or modified through synthetic processes often exhibit subtle anomalies in pitch variation, background noise consistency, and articulation smoothness, which NISQA detects as deviations from natural speech. NISQA outputs a Mean Opinion Score (MOS) on a scale from 1 to 5, where 1 represents very unnatural or heavily distorted speech, and 5 indicates highly natural, human-like voice quality. A score closer to 5 suggests that a voice exhibits smooth articulation, natural pitch variation, and minimal distortions. A score closer to 1 may indicate robotic intonation, noticeable artifacts, or unnatural pauses. As a sanity check, over the 336 cloned voices, the NISQA software produced results covering a range from 1.35 (poor) to 5.0 (exceptional).

E. Human participant evaluation of voices

To understand how human listeners evaluate voice authenticity, several options were evaluated ranging from in-person observations to online surveys. First, a pilot study was conducted with a group, in person, with 4 participants over 30 minutes. During the first step, each participant was asked to rate confidence that a given voice was real on a scale of 1 (definitely not) to 5 (definitely). Before the second step, the participants listened to a 30 second clip of the speaker's authentic voice and then re-evaluated the same set of voices. From this pilot study, several key insights emerged: using the same script repeatedly was problematic, 30 second audio clips were too long for participants, and directly asking whether the audio sounded like the speaker did not significantly affect the ratings. These findings prompted the development of a shorter survey to be conducted online, enabling a larger pool of participants to complete a similar set of questions in less time.

Therefore, an online survey was created (approved by the local Institutional Review Board) using the Qualtrics software. In the survey, a participant was asked to first read and agree to the informed consent, to practice listening and rate one sample voice, to listen and rate six voices, and finally to answer two questions about their familiarity with English and non-native speakers. Each of the voices played to completion for between 7 and 14 seconds before the participant was asked to rate that voice, using a visual analog scale (VAS) with endpoints "Definitely Fake" and "Definitely Real." The choice of a VAS was made to obtain ratio scale data, as opposed to discrete and/or Likert scale data. Ratio scale data are readily analyzable via conventional statistical analysis, e.g., t-tests and ANOVA. The generation of ratio scale data is also useful for direct comparison to machine scoring (i.e., NISQA model). A slate of 6 random voices was presented to each human listener. To ensure that each of the 81 (67 cloned and 14 authentic) voices was evaluated by at least 30 unique listeners, the target number of participants was set to 405. To ensure each voice reached the required minimum number of responses, a quota was set per voice such that if a voice was evaluated 30 times it would not reappear, allowing for the randomization of all voices. The final two questions in the online survey asked participants to self-evaluate their familiarity with comprehending English and their exposure to non-native English speakers.

III. Results

A. Analysis of authentic versus cloned voices

To examine group-level differences, individual-level comparisons were plotted for each of the 81 unique voices. Across all responses, the authentic voice samples totaled 449 ratings, while the cloned voices accounted for 2245 ratings. For each voice, mean survey responses and standard deviations were computed, with most authentic voices receiving between 29 and 36 responses, and cloned voices receiving around 30 responses each. Authentic voices consistently showed higher average ratings than their cloned counterparts, with nearly all differences reaching statistical significance in two-sample t-tests, Fig. 2, upper. When plotting the coefficient of variation (CV) against each voice's mean, voices with higher mean ratings exhibited lower CVs, indicating stronger participant agreement and confidence in identifying those voices as real, Fig. 2, lower. In contrast, voices with lower means tended to show greater relative variability, suggesting increased uncertainty when participants rated a voice as potentially fake.

B. Analysis of technical and demographic factors

Survey responses revealed that certain technical factors had a noticeable effect on the perceived realism of cloned voices, Fig. 3 leftmost column. Among the cloning tools, Lovo exhibited significantly lower ratings, falling below the neutral midpoint of the scale. Training time also played a role as both shorter and longer (60 seconds) training durations led to more realistic-sounding clones compared to those trained on 30 seconds of data, which is not obviously explainable. Additionally, the presence of background noise improved realism of its clones. As well, demographic characteristics of the voice source also influenced its clone's realism, Fig. 3 rightmost column. Clones based on male voices were more convincing than female voices, and younger voices under the age of 30 tended to receive higher realism ratings. One of the clearest differences appeared for ethnicity, where clones based on non-Hispanic voices were consistently more realistic.

C. Analysis removing combinations of factors

To understand what combination of factors most effectively narrows the gap between cloned and authentic voices, survey responses were compared across two filtered subsets of the cloned voice data, Fig. 4. The first subset removed clones associated with technical factors that previously led to lower realism ratings, including voices generated with Lovo, 30-second training times, and no background noise. The second subset filtered clones based on demographic factors that had

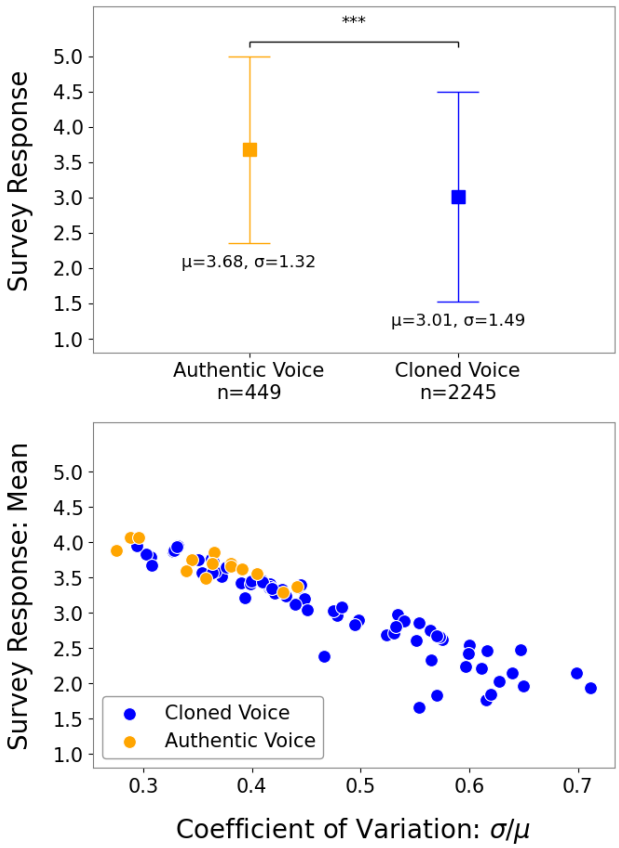

Fig. 2. Survey participants evaluation of aggregate set of authentic and cloned voices. (Upper) The authentic voices were perceived to be more realistic (1.0 = definitely fake, 5.0 = definitely real) than the cloned voices at a statistically significant level. The mean value for the set of authentic voices was 3.68, above the neutral value of 3.0, while the mean value for the set of cloned voices was 3.01, near the neutral level. (Lower) Survey responses per each of the 81 unique authentic and cloned voices are plotted with their mean value representing 29-36 survey responses for each of the authentic voices, and around 30 survey responses for each of the cloned voices. This mean is plotted against each unique voice's coefficient of variation, which is a way to normalize the variance. We see that for voices with a higher mean, the CV is lower, indicating the survey respondents indicate greater confidence in the voices they rank as closer to definitely real, and for voices with lower means, the CV is lower, indicating that survey respondents indicate lesser confidence in the voices they rank as closer to definitely fake.

been rated lower overall, specifically voices from sources over the age of 30, Hispanic, or female. Each subset showed improved performance when compared to the full set of cloned voices. Most notably, when both sets of filters were applied simultaneously, leaving only those clones created using stronger technical factors and based on younger, male, non-Hispanic voice sources, the resulting responses were no longer statistically different at a significant level from the authentic voice responses. This suggests specific combinations of factors can effectively blur the line of perceived realism between cloned and authentic voices.

In Fig. 5, survey responses were also analyzed to examine how authentic and cloned voices compared when controlling for speaker identity. For each of the 14 voice speakers, the survey ratings for the authentic voices were compared directly

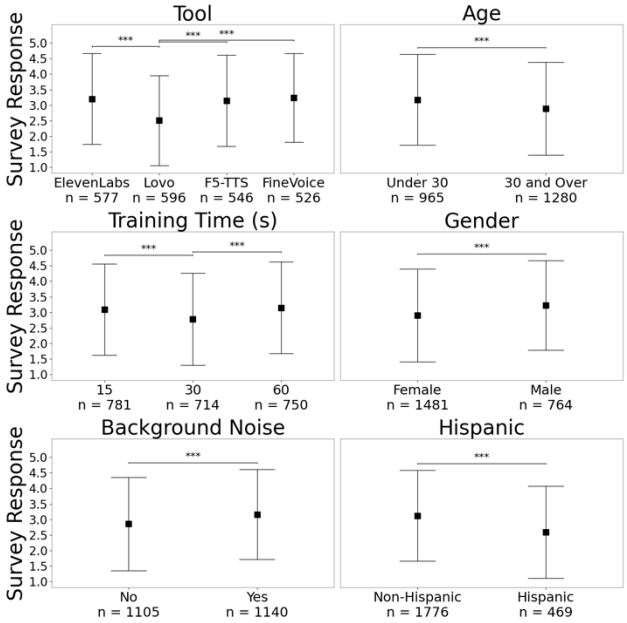

Fig. 3. Survey participants evaluation of technical and demographic factors, for the cloned voices alone. (left column) In terms of technical factors, the Lovo AI voice generator led to statistically lower survey responses (with a mean value below 3.0, neutral) as compared to the other tools. Voice clones trained on input data of 15 and 60 seconds were rated as more realistic than those based upon 30 seconds of data. Voice clones with background noise were rated higher than those without. (right column) Clones build from the voices of those under 30, non-Hispanic, and male were rated more realistic.

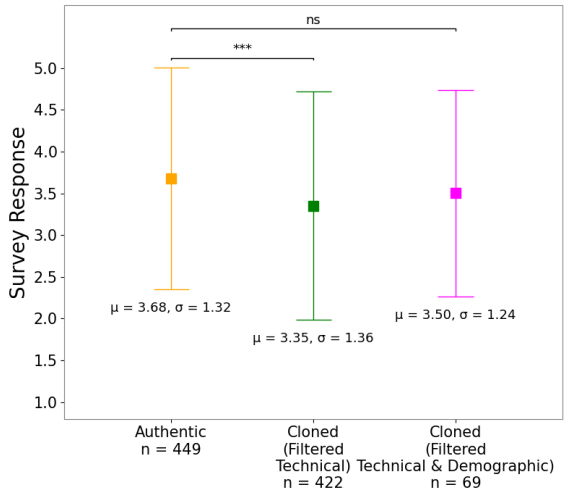

Fig. 4. Survey participants evaluation showing aggregate of responses for authentic voices (449), in comparison to two subsets of the cloned voices. In green (second bar) are the subset of 422 voices of the entire set of 2245 cloned voices where significantly lower technical factors from Fig. 3 were removed (Lovo, 30 s training time, and no background noise). In magenta (third bar) are the subset of 69 voices where the technical factors, as well as significantly lower demographic factors from Fig. 3 were removed (over 30 years old, Hispanic, and female). When both these technical and demographic factors were removed, the remaining 69 cloned voices (generated using the 3 non-Lovo training tools, 15 or 60 s training times, background noise; based on individuals under 30 years old, non-Hispanic, and male) yield no statistically significant difference from the authentic set of voices. These factors therefore make it easiest to train clones to perform at the level of authentic voices.

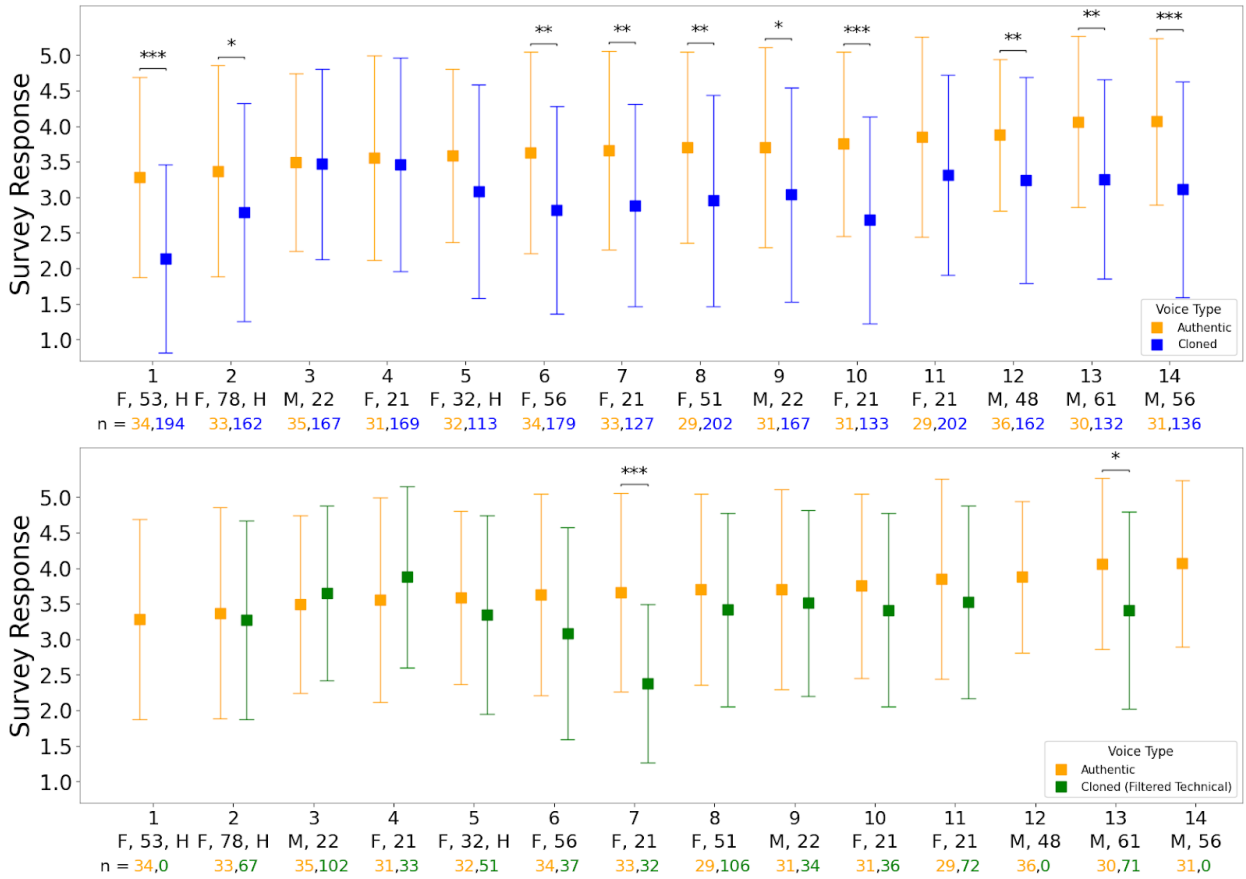

Fig. 5. Survey responses for authentic and cloned voices, per each of the 14 voice speakers around which the clones were built. (Upper) For 10 of the 14 voice speakers, their authentic voices (n=29-36 survey responses per voice speaker) were perceived as more realistic than their corresponding cloned voices (n=113-202 survey responses), at a statistically significant level. (Lower) To create the "Cloned (Filtered Technical)" set, those statistically significant technical factors from Fig. 3 were removed to create a subset of 641 cloned voice responses that did not include any clones with Lovo, 30s training time, and no background noise. When comparing this filtered subset with the authentic voice responses, only 2 of the 14 voice speakers are now differentiable between their cloned and authentic voice sets. Therefore, cloned voices that isolate particular technical factors (i.e., tool, training time, background noise) can be used to create cloned voices that perform at the level of authentic voices.

with those of their cloned counterparts. In nearly every case, the authentic voices were rated as more realistic. However, when the subset of clones using certain technical factors identified in Fig. 3 were isolated, excluding those created with Lovo, 30-second training, or no background noise, the performance of cloned voices improved notably. For this filtered group, only two speakers showed a meaningful gap in realism ratings between authentic and cloned responses. These results highlight that the selection of specific technical factors alone can elevate the cloned voices to match the perceived realness of authentic voices.

D. Comparison of survey and NISQA results

In Fig. 6, authentic voices cluster towards the upper-right, indicating high scores from both the human listener survey and NISQA machine evaluation. In contrast, the cloned voices exhibit no clear relationship between these two factors. This discrepancy highlights a potential misalignment between machine and human ratings. Notably, all cloned voices with background noise seem to be in the left 2 quadrants, indicating that NISQA scores are lower with the added background noise. Indeed, some clones without background noise were perceived by NISQA as equally natural as the authentic voices. Overall, the plot shows that while authentic voices cluster in the high scoring quadrant for both NISQA and survey scores, cloned voices with background noise tend to have lower NISQA scores but a wide range of survey response scores, and cloned voices without background noise have a wide range of scores across both axes.

E. Participant characterization

Survey participants were asked to report their level of English comprehension and general exposure to non-native English speakers. No meaningful differences were found, suggesting that neither aspect of listener background impact the evaluation of voice realism, Fig. 7.

IV. Discussion

This work sought to evaluate authentic and cloned voices by both human listeners and a machine algorithm. Data from an online survey with 449 respondents revealed that people readily perceive authentic voices as more realistic than clones

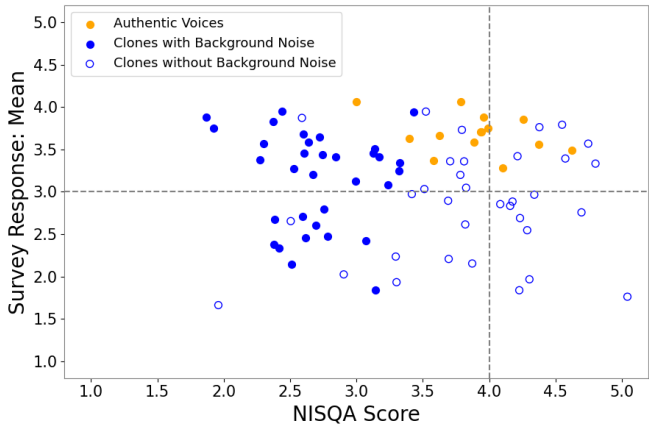

Fig. 6. Relationship between NISQA and human listener scores across three sets of voices. The x axis represents the NISQA score (1 = least natural, 5 = more natural). The y axis shows the mean survey response per voice (1 = definitely fake, 5 = definitely real). Authentic voices cluster towards the upper right, indicating agreement between methods. In contrast, cloned voices vary more widely, with several scoring high in NISQA, but with low human listener ratings, suggesting that perceived realness is not solely determined by background noise. Some cloned voices with background noise score comparably to authentic voices in both dimensions, indicating added background noise may sometimes improve perceived authenticity.

(Fig. 2, upper). Moreover, an analysis of demographic factors showed that clones built from source voices that are non-Hispanic, male, and under 30 years old are perceived as more realistic than those which are Hispanic, female, and over 30 years old (Fig. 3, rightmost column). Similarly, clones created using the Lovo tool, 30 seconds of training, and no added background noise were rated significantly less realistic than other corresponding categories (Fig. 3, leftmost column). When clones from the lowest-performing technical and demographic groups were removed, the remaining subset was indistinguishable from authentic voices (Fig. 4). These results suggest that while cloned voices generally lag in believability, certain configurations convincingly mimic authentic voices.

Interestingly, NISQA did not exhibit sensitivity to the same factors that influenced human perception (Fig. 6). While background noise impacted NISQA scores, its ratings did not align with human evaluations. This indicates that realism, as perceived by humans, is distinct from naturalness as measured by NISQA, likely because NISQA focuses on signal quality while people rely on additional acoustic and contextual cues. Human and machine evaluations each have limits and need to be used together. Human listeners notice social-contextual cues, such as background noise, that models like NISQA are not trained to interpret. However, machines offer a level of consistency and scalability that humans cannot. NISQA's misalignment with human ratings may stem from its training to assess speech quality rather than realism. It likely emphasizes features like clarity and pitch variation, which don't always correspond to perceived authenticity, and may ignore linguistic signals used by humans. Tools such as Microsoft's Deepfake Detection API or newer adversarial systems focused on believability may offer better alignment with human perception if trained on relevant criteria.

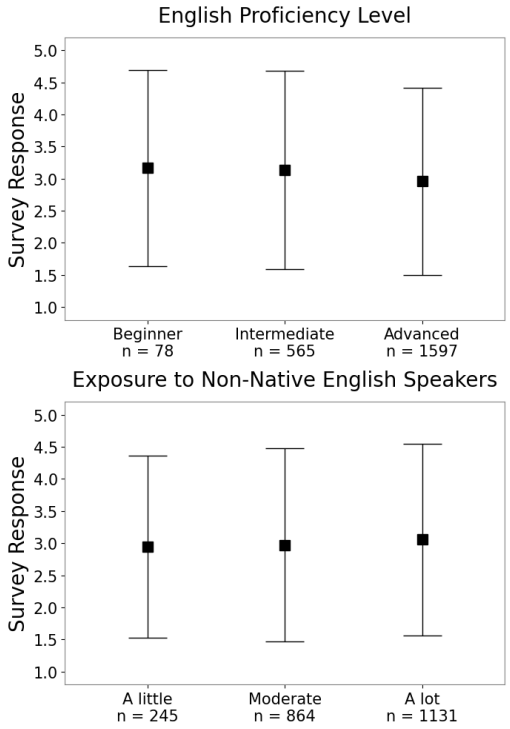

Fig. 7. Survey participants evaluation of their own abilities. Consistent mean values between the groups suggest that greater proficiency or exposure to English speakers does not influence the aggregate survey responses.

ACKNOWLEDGEMENTS

We would like to thank Greenway Solutions, especially Patrick Shaw, for their support, feedback, and assistance.

REFERENCES

[1] K. T. Mai, S. Bray, T. Davies, and L. D. Griffin, "Warning: Humans cannot reliably detect speech deepfakes," PLOS ONE, vol. 18, no. 8, p. e0285333, Aug. 2023, doi: 10.1371/journal.pone.0285333.

[2] "Signals | Emerging Tech Trends 2024 | Q1." Accessed: Apr. 13, 2025. [Online]. Available: http://innovationinsights.mastercard.com/signals-emerging-tech-trends-2024-q1

[3] E. M. Al-dahasi, R. K. Alsheikh, F. A. Khan, and G. Jeon, "Optimizing fraud detection in financial transactions with machine learning and imbalance mitigation," Expert Systems, vol. 42, no. 2, p. e13682, Feb. 2025, doi: 10.1111/exsy.13682.

[4] A. Kassis and U. Hengartner, "Breaking Security-Critical Voice Authentication," in 2023 IEEE Symposium on Security and Privacy (SP), May 2023, pp. 951–968. doi: 10.1109/SP46215.2023.10179374.

[5] F. M. Teichmann and S. R. Boticiu, "An overview of the benefits, challenges, and legal aspects of penetration testing and red teaming," Int. Cybersecur. Law Rev., vol. 4, no. 4, pp. 387–397, Dec. 2023, doi: 10.1365/s43439-023-00100-2.

[6] A. C. Pangia and M. M. Wiecek, "A branch-and-bound algorithm for parametric mixed-binary nonlinear programs," J Glob Optim, vol. 91, no. 3, pp. 466–468, Mar. 2025, doi: 10.1007/s10898-024-01447-4.

[7] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," 2021, doi: 10.48550/ARXIV.2104.09494.