S**ai**FETY: An Integration of Audio Protection and Ethical Data Collection Comparisons Within Txt2Vid

Lucas Caldentey, Avrick Altmann, Yan Li Xiao, Fenet Geleta
Mentors: Arjun Barrett, Laura Gomezjurado, Pulkit Tandon

1. Abstract

As a result of globalization and massive technological advancements, multi-media communications have begun running excessively on internet traffic. This reliance on digital connections further increased due to the recent Covid-19 pandemic. From the daily dependence on News channels and social media live streaming to peer-to-peer online meetings, the world's primary form of transmitting information is now digital. With the decline of human-to-human interaction, it is critical to not only have a stable and reliable metric to converse but also an effective way to ensure the safety and ethicality of all users. We introduce an extended version of Txt2Vid with more clear and developed stances toward user safety. Specifically, we focused on comparing the data collection methods between Txt2Vid and other video communication platforms. Additionally, we developed an audio key authentication system using text-dependent voice verification (Novikova) that prevents users from falsely using the voices and information of others. With these implementations, we hope to smoothen the transition and comfort of the public as AI becomes more and more prevalent in our lives. And show people from all over the world that deep fakes can be used safely and positively to bring us closer together.

2. Introduction

In an era marked by the rapid expansion of video streaming and visual media sharing, video compression technologies have become indispensable for ensuring efficient data transmission and bridging the digital divide on a global scale. Among these transformative technologies stands Txt2Vid, a cutting-edge video compression pipeline that goes beyond conventional solutions by specializing in deepfake and Artificial Intelligence (AI). Txt2Vid opens new possibilities for video communication, promising innovative applications and immersive experiences. However, as AI and deepfake technologies continue to advance, the need to bolster user security and prioritize ethical data handling becomes increasingly pressing for platforms like Txt2Vid. By seamlessly integrating AI-generated content into videos, users can now personalize and enhance their visual narratives. This capability offers an exciting frontier for creative expression and interactive storytelling. However, for such technologies to gain widespread acceptance and adoption, addressing the critical issues of user security and data privacy is essential.

In the quest for secure and reliable video communication, this research paper sets forth a two-fold objective. Firstly, we propose the introduction of a text-dependent voice verification system designed to establish a robust metric for user authentication. This novel system aims to mitigate the risks of unauthorized access and protect user data with a new level of assurance. By confirming the identity of

users through voice verification, we enhance the platform's security measures and foster an environment of trust. Additionally, our research critically analyzes the data collection methods employed by Txt2Vid and other video-calling platforms to identify potential deficiencies and vulnerabilities. The safeguarding of user privacy is not only a legal and ethical imperative but also vital for building user trust. We endeavor to propose ethical measures that ensure data is handled responsibly, transparently, and with due regard for user consent. Through a rigorous examination of data practices, we seek to raise awareness of potential privacy concerns and provide concrete recommendations for improvements.

The significance of this research lies in its potential to elevate Txt2Vid and similar platforms into secure and trustworthy video communication solutions. By enhancing user confidence in the platform's security and privacy protections, we anticipate a ripple effect, leading to greater trust among users and making Txt2Vid an attractive option for diverse video communication needs.
To validate our hypothesis, we conducted surveys with participants, representing diverse demographics and usage scenarios, to gather valuable insights into user perceptions of security and trust in Txt2Vid. Incorporating their feedback into our analysis, we aim to uncover valuable insights that will help shape the platform's security enhancements and further improve its user experience.

3. Related research

Text-dependent verification and text-independent verification are two different approaches to voice authentication. In text-dependent verification, the user is required to provide a specific, predetermined phrase or text during the verification process. This predetermined text serves as a reference for comparison, ensuring that the user's voice is authenticated against a known and unique reference, making it difficult for unauthorized individuals to impersonate the user's voice. On the other hand, text-independent verification does not require any specific phrase; it allows voice authentication based on any spoken content without the need for a reference text. We choose to incorporate text-dependent verification in the Txt2Vid platform due to its numerous advantages. Text-dependent audio biometrics significantly reduces the risk of unauthorized access or voice manipulation, making it an ideal choice for protecting against deepfakes and potential voice-based attacks. Moreover, as a new platform with limited data, implementing text-dependent verification is more straightforward and practical, allowing us to establish a strong security foundation from the outset. As Txt2Vid evolves and accumulates more data, we may explore combining text-dependent and text-independent methods for even greater security. By doing so, Txt2Vid can generate voices that closely resemble the intended users while safeguarding against emerging threats, thereby enhancing the overall user experience and content integrity.

4. Methodology

The research design encompasses a combination of exploratory and descriptive methodologies. Exploratory research is undertaken to delve into the potential benefits of implementing text-dependent voice verification, while descriptive research aims to quantify user perceptions through survey results.

Furthermore, the study introduces an experimental component to thoroughly evaluate the performance of the Txt2Vid system in accurately identifying different voices. This comprehensive analysis of the text-dependent voice verification system aims to determine its overall effectiveness and user acceptance.

The target population comprises primarily individuals in the age range of 10-19 who are regular users of various platforms. Participants were recruited through an email list and directed to a dedicated website to engage in both the survey and voice verification testing. The collected qualitative data measured both user perceptions and preferences regarding the voice verification system. Simultaneously, the voice verification testing generated audio inputs to assess the system's accuracy and safety in verifying different voices. Prior to participation, individuals received email invitations with a request for informed consent. During voice verification testing, participants were prompted to recite diverse lines to gauge the system's performance effectively. The study's variables encompass user perceptions, preferences, and voice verification accuracy, which were measured through qualitative analysis of survey responses and comparison of voice inputs with the system's verification results. Thematic analysis was utilized to understand user perceptions derived from survey responses, while voice verification data was analyzed to evaluate the system's accuracy, considering the qualitative nature of the data. It is important to recognize several limitations of this research. The age range of 10 to 19 might not fully represent all potential users of video-calling platforms, limiting the generalizability of the findings to other age groups. Moreover, relying on self-reported survey responses may introduce social desirability bias, influencing participants to provide answers they perceive as more favorable. Additionally, voice verification testing conditions may not fully simulate real-world scenarios, potentially affecting the system's accuracy. Despite these limitations, the research endeavors to offer valuable insights into user perceptions and the feasibility of the text-dependent voice verification system for video-calling platforms. This outcome will contribute to future improvements and enhance security measures in Txt2Vid and AI-based media platforms.

5. Experimentation

Throughout this research project, we tested various methods and models for voice authentication. The voice biometric model we chose was built off of a pre-trained Convolutional Neural Network (CNN). This allows the network’s hidden neurons to have the same weights and bias values per layer. With each layer focusing on specific features (such as pitch, speed, or tonation for example), CNN networks can increase their complexity and more accurately distinguish individual audio. We found it important for voice biometrics to have this property as more complex layering could lead to more natural speech from users while testing. Our research is primarily focused on two facets of voice biometrics: an internal examination of the voice biometric feasibility, and an external analysis of the security’s integration amongst people. Firstly, we studied the reliability of voice biometrics from a day-to-day standpoint. Factors like sickness, sore throats, or even the time of day are all factors that can change people’s voices suddenly and are not studied enough when developing audio networks. To ensure the voice biometric was capable of recognizing the user's voice even when sick or sore, a recording set of 25 audio clips was created by each group member to create a 100-audio clip database alongside an additional 25 audio clips

of a notably sick and sore voice. Each different biometric model was trained on 5 clips from each group member's voice and tested for accuracy with the remaining 20 clips. The CNN could accurately verify 90% of the verification data generated by group members, and could accurately verify 100% of the 20 audio clips generated by 2 group members.

To further test our voice biometric and spread awareness for the safe use of deep fakes and AI, we created a flask app due to a website's relative ease of use for consumers. A link to our site can be found below within our bibliography. To create an account, users are prompted for a username and password. After the password is hashed using the SHA algorithm, both are stored in an SQL database. The user is also prompted to record 5 voice recordings reading the following sentences.

1. "I'm extremely excited for the SHTEM program this year."
2."The quick brown fox jumped over the lazy dog."
3."The hungry purple dinosaur ate the kind, zingy fox, the jabbering crab, and the mad whale."
4. "With tenure, Suzie'd have all the more leisure for yachting, but her publications are no good."
5. "The beige hue on the waters of the loch impressed all, including the French queen."

These sentences were chosen for their significance to the program or for their high phonetic coverage–meaning they encompass many different sounds in the English language. Using the Python backend flask provides, these files are passed into the AI, and their weights are stored as a ".npy" file generated by Numpy. On login, users are prompted to record a random sentence that the AI verifies and matches to their generated .npy file. The program also utilizes Open AI's Whisper STT, to confirm that users actually read the correct sentence. This prevents potential imitation through any audio clip of the user's voice by requiring the voice to read the specific sentence. If the authentication is successful and the password hash is correct, the user is logged in.

6. Results

6.1 The Demographic of Our Research Participants

For our research, we primarily focused on the feedback of participants from the age range of 10-19 to reflect the young population of video-calling platform users. There is a relatively equal number of males and females, people who identify as Asians make up the majority of our research group, and the amount of time participants spend on video-calling platforms per week ranges from under 1 hour to 6-10 hours. With half of people spending 1-5 hours per week on video-calling platforms, our research participants are regular users of various platforms, providing us with an understanding of how Txt2Vid compares to platforms like Zoom and Google Meet.

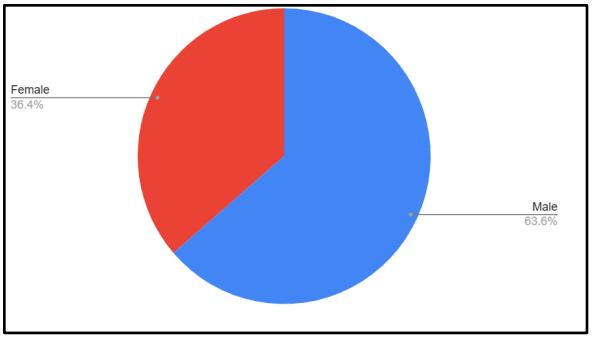


Figure 1: The gender identities of our research participants.

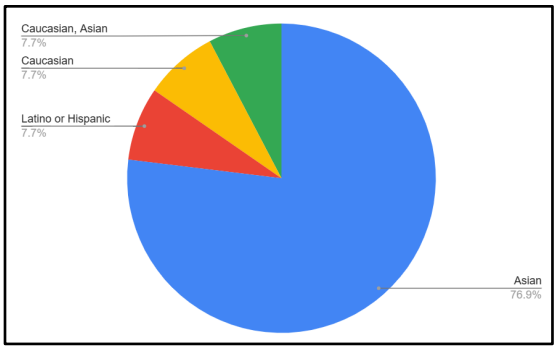


Figure 2: The race of our research participants.

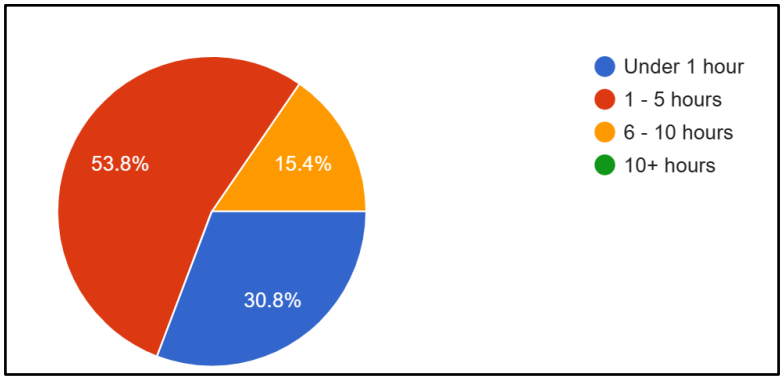
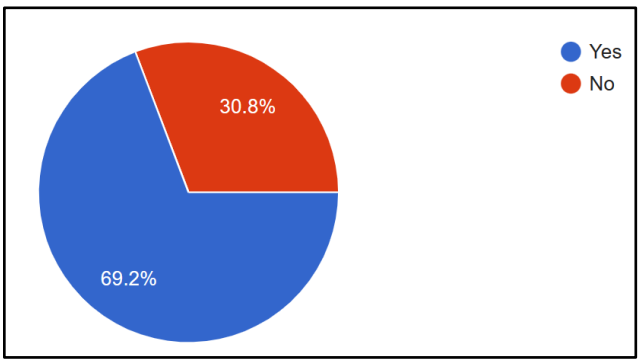


Figure 3: The amount of time participants spend on video-calling platforms per week.

6.2 Performance Analysis

To evaluate the performance of our voice-verification system, we gathered information on research participants' feedback on the likelihood of them using Txt2Vid with the voice-verification system versus without. Additionally, users were asked to rate the level of convenience of the voice-verification system, whether it successfully verified their voice, and feedback on how we can make the system better. Our participants report a 69.2% success rate, with one reporting that the system worked on her second attempt and one attempting 4 times with all failing. Although our current success rate is not viable for the industry, our results demonstrate the concept for future voice biometrics.



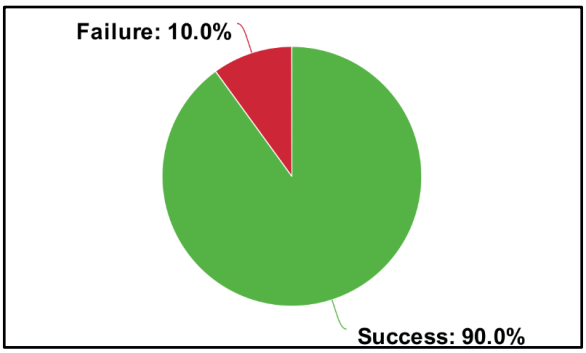


Figure 4: Participants reported the accuracy of our voice-verification system.

Figure 5: The success rate of our initial training data.

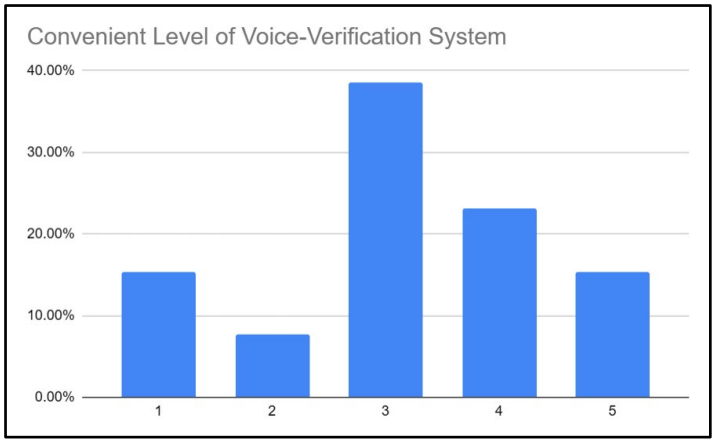


Figure 6 : Participants reported the convenience level of the voice-verification system with 1 being inconvenient and 5 being convenient.

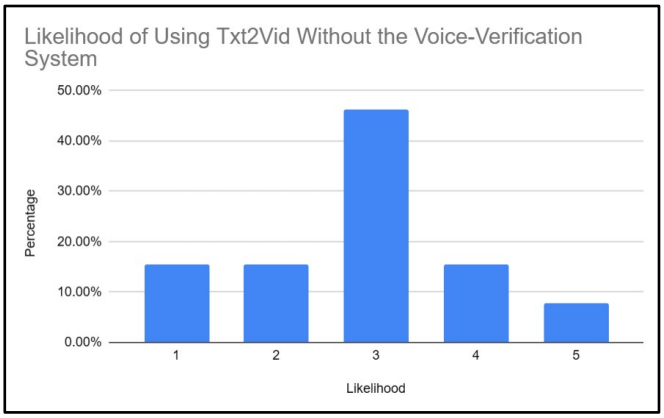


Figure 7: Likelihood of participants using Txt2Vid without the addition of the voice-verification system with 1 being unlikely and 5 being likely.

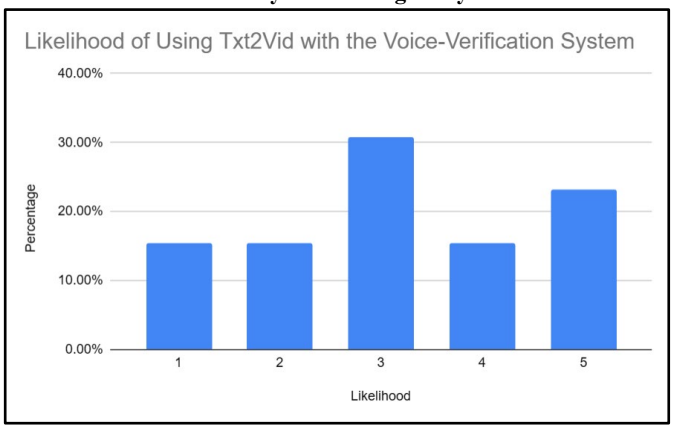


Figure 8: Likelihood of participants using Txt2Vid with the addition of the voice-verification system

Out of the 13 people that we surveyed, 53.85% of them reported no change in the likelihood of them using the Txt2Vid platform, 30.77% reported an increase with an average of 1.75 shift toward the “Likely” side of the scale, and 15.38% reported a decrease. As shown in the comparison of Figure 7 with Figure 8, there is a decrease in the number of participants feeling neutral about Txt2Vid before and after the addition of the voice-verification system, and a more left-skewed graph with a 15.4% increase in the number of participants reporting the likelihood of them using Txt2Vid being 5, which is highly likely. This serves as evidence that the addition of the voice-verification system to Txt2Vid decreased the reluctance to widely use this platform and increased user privacy protection. However, with 53.85% of participants reporting no change in the likelihood of them using the platform and 15.38% reporting a decrease, more must be done to increase the user base of Txt2Vid and make it effective in verifying the voices of our users.

7. Future Research

When evaluating the feedback we got from our research participants, 46.15% of them suggested implementing a voice recorder to the website, 7.7% suggested spreading the word involving the ethical concerns involving Txt2Vid, 23.1% suggested improving the interface, and 15.38% reported the system as inaccurate. Therefore, in the future, we hope to improve user experience by adding a voice recording device to the website, improving the user interface, increasing the accuracy of the voice-verification system, and implementing the system to the Txt2Vid platform. The addition of a voice recorder will significantly facilitate the process of voice enrollment and encourage users to opt-in to voice biometric data collection. For this project, we focused on developing a highly accurate voice-verification system outside of the Txt2Vid platform rather than integrating it into the system. We wish to further improve the accuracy of our voice-verification system with an increase in the CNN’s training and the quantity of data collected. Additionally, we wish to complete the implementation of this new security measure into the Txt2Vid platform.

8. Conclusion

In closing, this research underscores the critical significance of ensuring secure and ethical video communication platforms in our rapidly evolving digital age. The surge in visual content consumption, coupled with the transformative impact of the Covid pandemic, vividly underscores the need for dependable and morally guided user experiences. Through the introduction of an extended version of Txt2Vid, enriched with text-dependent voice verification, we've not only established a formidable user authentication mechanism but have also adroitly addressed concerns surrounding unauthorized access and potential voice impersonation. Furthermore, our in-depth scrutiny of data collection practices brings to light potential vulnerabilities, underscoring the importance of responsible data handling and the crucial aspect of user consent. By embracing these security enhancements, our research seeks to foster an environment of user trust, promoting the harmonious integration of AI-generated content within the realm of video communication. In a broader context, this endeavor not only augments the capabilities of Txt2Vid but also paves the way for a more secure and cohesive digital future, where AI-driven

technologies can be harnessed responsibly and meaningfully. Looking ahead, we anticipate that these efforts will serve as a stepping stone for continued advancements in secure video communication platforms, fostering innovation while ensuring user safety remains a guiding principle.

9. Acknowledgements

The authors of this paper would like to thank Arjun Barrett, Laura Gomezjurado, and Pulkit Tandon for their continuous support and guidance throughout our research. We would also like to thank Sylvia Chin and Pr.Weissman for providing us with this incredible opportunity and overseeing the SHTEM program.

Bibliography

Caldentey, L., Altmann, A., Xiao, Y. L., & Geleta, F. (2023, July 30). Saifety Shtem Project. SaiFETY. https://drago314.pythonanywhere.com

- This is the site created and tested for our research. Please feel free to look through it and test the voice biometric yourself.

Eric-Urban. (n.d.). *Speaker recognition quickstart - speech service - azure AI services*. Speaker Recognition quickstart - Speech service - Azure AI services | Microsoft Learn. https://learn.microsoft.com/en-us/azure/ai-services/speech-service/get-started-speaker-recognition?tabs=script&pivots=programming-language-csharp

Molla, R. (2017, June 8). *An explosion of online video could triple bandwidth consumption again in the next five years*. Vox. https://www.vox.com/2017/6/8/15757594/future-internet-traffic-watch-live-video-facebook-google-netflix

Novikova, Evgeniia (2023, January 8). *What is text-dependent and text-independent voice biometrics* - neuro.net blog. What Is Text Dependent And Text Independent Voice Biometrics https://neuro.net/en/blog/what-is-text-dependent-and-text-independent-biometrics

Omri Wallach. Graphics/Design. (2021, September 20). *The world's most used apps*, by downstream traffic. Visual Capitalist. https://www.visualcapitalist.com/the-worlds-most-used-apps-by-downstream-traffic/

Tandon, P., Chandak, S., Pataranutaporn, P., Liu, Y., Mapuranga, A. M., Maes, P., Weissman, T., & Sra, M. (2022, April 3). Txt2Vid: *Ultra-low bitrate compression of talking-head videos via text*. arXiv.org. https://arxiv.org/abs/2106.14014

Team, I. (n.d.). *What are Convolutional Neural Networks?* Accessed, 8/18/23 https://www.ibm.com/topics/convolutional-neural-networks

Team, O. (Ed.). (2020, May 4). K*eeping the internet up and running in times of crisis* - OECD. OECD: Better Policies for Better Lives. https://www.oecd.org/coronavirus/policy-responses/keeping-the-internet-up-and-running-in-times-of-crisis-4017c4c9/