Thesis Project Portfolio

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

(Technical Report)

A Sociotechnical Analysis of Deepfake Audio Technology as Evaluated by Machine and Human Observers

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Vishnu Lakshmanan

Spring, 2025

Department of Systems and Information Engineering

Table of Contents

Executive Summary

This project aims to understand the differences in how demographic and technical factors impact how voices are viewed by humans and machines and which factors are important in making those determinations. My capstone project explores the technical dimensions of this threat by analyzing which demographic and technical factors contribute to the creation of realistic AI-generated voice clones. This research was motivated by the need to understand and quantify what makes synthetic voices sound real to human listeners and how that realism can be manipulated by malicious actors. In parallel, my STS research paper examines deepfake audio technology through the Social Construction of Technology (SCOT) framework, analyzing how social groups like financial institutions, shape the evolving perception and response to synthetic voice tools. Together, these projects form an analysis of voice cloning with the aim of understanding how it is viewed and what can be done to better understand these risks.

This project tackles the growing threat of AI-generated voice clones used in fraud and identity theft by identifying the factors that make cloned voices sound more realistic. A library of 336 cloned voices was created using four cloning tools, three training durations, and the addition of background noise. These were paired with 14 authentic voices across diverse demographics. Using optimization techniques, a representative subset of 67 cloned and 14 authentic voices was selected. These were evaluated by 449 human listeners through a Qualtrics survey using a visual analog scale, and by NISQA, a machine model for speech naturalness.
Human listeners rated authentic voices as more realistic overall, but certain cloned voices—those created with non-Lovo tools, shorter or longer training times, background noise, and based on young, male, non-Hispanic voices—were indistinguishable from real ones. This indicates that some clones can convincingly mimic human speech under specific conditions.Notably, NISQA scores did not align well with human perception, often penalizing background noise despite its positive effect on realism. These findings highlight the need to combine human and machine evaluations to strengthen voice authentication systems and better protect against deepfake threats.

This study investigates how deepfake audio is socially constructed as a cybersecurity threat by examining how human listeners and machine tools perceive voice authenticity. Grounded in the Social Construction of Technology (SCOT) framework, the research explores how various social groups interpret the meaning and implications of synthetic voice technologies. The significance lies in addressing a growing concern: fraudsters using cloned voices to exploit voice authentication systems. The study combines a large-scale human perception survey with machine analysis using the NISQA model. A total of 67 AI-generated voices, varied by cloning tool, training time, and background noise, were tested against 14 authentic samples. Survey participants evaluated voice realism on a visual analog scale, simulating real-world conditions like phone fraud scenarios, while machine ratings provided objective scores on naturalness.

The evidence reveals that human listeners generally perceive authentic voices as more realistic, but certain AI-generated clones. These findings align with SCOT by illustrating how technological meaning is shaped by the interpretive flexibility of different social groups. Machine evaluations, however, often diverged from human perception, especially penalizing background noise despite its realism-enhancing effect for listeners. The study concludes that voice authenticity is not only a technical challenge but a sociotechnical one, influenced by stakeholder values and institutional responses. This reinforces the need for interdisciplinary solutions that blend innovation, regulation, and public awareness.