Evaluating Perceptions of Naturalness in AI-Generated Speech

Ogechi Anyamele, Saurav K. Aryal, PhD, Gloria Washington, PhD
AI4PC Lab, Institute for Human-Centered Artificial Intelligence
Howard University

INTRODUCTION

- Advances in neural text-to-speech have enabled realistic voice cloning, yet measuring how "natural" synthetic voices sound remains a challenge.
- Voice cloning is increasingly integrated into:
  1. Accessibility systems
  2. Entertainment platforms
  3. Virtual assistants and more
- Subtle differences in synthetic speech can reduce trust, affect usability, and shape overall user experience.
- Identifying factors that contribute to human-like speech and establishing reliable evaluation methods is essential for advancing systems.

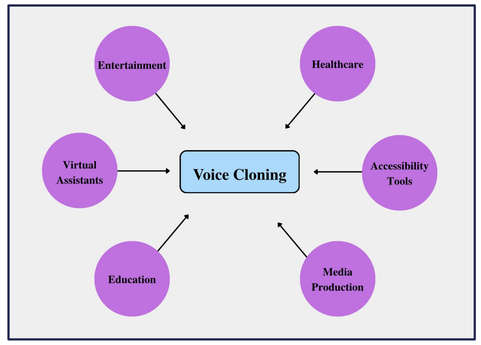

Figure 1. Applications for neural voice cloning technology

OBJECTIVES

- Investigate the relationship between training data quantity and perceived human-likeness in cloned voices using a neural voice cloning pipeline, Coqui XTTS.
- Explore and compare different approaches for evaluating speech naturalness to inform best practices in study design.

METHODS

Models:

Voice models were trained using the **Coqui XTTS** neural voice cloning framework, with speech samples ranging from short to extended durations.

Data:

- Varying lengths of training audio were used to simulate real-world scenarios.
- Each model generated synthetic speech outputs for evaluation.
- All outputs were evaluated at both the sample and holistic level.

Cloning Pipeline:

Models were trained under varying data conditions such as different time lengths of audio to create a system for analysis.

Figure 2. Diagram of end-to-end voice cloning pipeline

METHODS

Listening Study Design:

Synthetic speech outputs were evaluated through a qualitative listening study designed to capture differences in perceived naturalness across listener familiarity levels. Three distinct evaluator types were included to provide varied perspectives on the generated speech:

- **Self** — original speaker evaluating cloned voice samples
- **Familiar listener** — a listener who interacts with the speaker on a regular basis
- **Unfamiliar listener** — a listener with little to no prior exposure

This design allows for both subjective self-assessment and more objective evaluations from external listeners, highlighting how familiarity may influence perception. By including multiple listener types, the study can better identify patterns in perceived naturalness and potential biases in subjective ratings.

Each evaluator listened to a set of synthetic audio samples generated from models trained on varying amounts of speech data. The samples were presented in a consistent format to ensure comparability across conditions.

Collected Data:

- Naturalness ratings assigned to each synthetic audio sample
- Evaluator classifications (self, familiar, unfamiliar) associated with each rating
- A dataset combining factors like ratings and evaluator type for analysis

Evaluations:

Each evaluator rated samples on a five-point Likert scale for perceived naturalness. Ratings were compiled into a structured dataset for analysis. The dataset enabled comparison across evaluator types and training conditions to identify trends in perceived speech quality.

RESULTS

As components of the study were completed:

- Models trained under varying data conditions produced measurably different perceptual outputs across evaluator types.
- Self-evaluators and familiar listeners showed distinct rating patterns compared to unfamiliar listeners, suggesting familiarity bias as a factor in naturalness perception.
- The structured dataset supports ongoing analysis of the relationship between training data quantity and perceived human-likeness.

Together, these components establish a framework for further refinement in voice cloning applications.

3
Evaluator Types
(Self, Familiar, Unfamiliar)

5-pt
Scale for Rating

- The end-to-end system was successfully implemented and models were trained under varying data conditions.
- Listening evaluations were conducted across all three evaluator types to support ongoing comparative analysis.
- Ratings were compiled into a structured dataset enabling comparison across training data quantities.

RESULTS

Preliminary trends indicate that models trained with larger datasets generally produced speech rated as more natural, while smaller datasets resulted in lower perceived human-likeness. Differences in ratings across evaluator types suggest that listener familiarity can influence perceptions, highlighting potential biases in subjective evaluations. This dataset provides a basis for identifying optimal training data thresholds and for exploring more objective or automated metrics in future studies.

Insights:

Overall, results establish a foundational framework for examining how training data quantity relates to perceived naturalness in synthetic speech.

CONCLUSION

The initial study demonstrates strong potential for using an end-to-end neural voice cloning pipeline to systematically evaluate perceived speech naturalness. This approach provides a structured and repeatable framework for analyzing how different training conditions influence listener perception.

Limits:

- Small number of human raters (three evaluators)
- Potential subjectivity in naturalness perception across listener types
- Results represent a preliminary phase pending further analysis

Future Work:

- Expand evaluator pool for greater statistical reliability
- Refine model training conditions and data length intervals
- Explore automated metrics alongside perceptual evaluation
- Investigate alternative methods for measuring naturalness in evaluation

REFERENCES

1. OpenAI. 2023. *Bring Your Voice to Life: Getting Started with Coqui XTTS v2*. OpenAI Blog. https://blog.gopenai.com/bring-your-voice-to-life-getting-started-with-coqui-xtts-v2-e17c9e0e5ba7
2. Ayushi Pandey, Sebastien Le Maguer, and Naomi Harte. 2025. What is naturalness? In *Proceedings of the Speech Synthesis Workshop 2025 (SSW '25)*. ISCA Archive. https://www.isca-archive.org/ssw_2025/pandey25_ssw.pdf
3. Sajad Shirali-Shahreza. 2023. How should we define voice naturalness. In *Proceedings of the Sixteenth International Conference on Advances in Computer-Human Interactions (ACHI '23)*. IARIA, 270–280. https://personales.upv.es/thinkmind/dl/conferences/achi/achi_2023/achi_2023_4_270_28008.pdf
4. Jialu Zhang, Shiwei Dong, and Ge Yu. 1998. Total quality evaluation of speech synthesis systems. In *Proceedings of the International Conference on Spoken Language Processing (ICSLP '98)*. ISCA Archive. https://www.isca-archive.org/icslp_1998/zhang98_icslp.pdf
5. Braintrust. 2024. *How to evaluate voice agents*. Braintrust. https://www.braintrust.dev/articles/how-to-evaluate-voice-agents

ACKNOWLEDGEMENTS

This research project was supported in part by the Office of Naval Research grant N00014-22-1-2714. The work is solely the responsibility of the authors and does not necessarily represent the official view of the Office of Naval Research.