Training Students to Listen Better: Advancing Deepfake Audio Discernment

Lavon Davis, PhD Student, Language, Literacy & Culture Program
Sara Khanjani, Noshaba Bhalli,PhD Students, Department of Information Systems
Dr. Vandana Janeja, Professor, Department of Information Systems
Dr. Christine Mallinson, Professor, Language, Literacy & Culture Program

Abstract

Audio deepfakes—spoofed audio, generated by or manipulated using Artificial Intelligence—can lead to deception and mis/disinformation. Our research aims to improve identification of audio deepfakes by incorporating insights from sociolinguistics and machine learning. This poster shares preliminary results from a pilot training program during Fall 2022, in which graduate research assistants led trainings with three undergraduate UMBC students to increase their sociolinguistic knowledge for discerning spoofed audio. Informal pre- and post-evaluations demonstrated improvement in students' ability to discern the authenticity of real and spoofed audio. The four students then shared their insights about the linguistic characteristics of audio deepfakes in an undergraduate UMBC Data Science course. Findings indicate the potential for further research into methods for increasing listeners' discernment of spoofed audio.

Objectives

- To train students to understand what deepfakes and spoofed audio are
- To train students to understand five expert-defined linguistic features (EDLFs)
- To train students to apply their understanding of EDLFs when listening to audio clips
- To determine whether training improves student accuracy in discerning spoofed audio

Research Question

Can training students to understand linguistic features improve their discernment of fake audio compared to legitimate audio?

Acknowledgments

This work was supported in part by Award#2210011. This work also included Data Science scholars supported through additional grants # 1915714 and # 2118285.
We also thank the three students who participated in our study, as well as students in IS 296, Foundations of Data Science, who listened to presentations by the three student participants.

Methods and Data

Two sociolinguists experts, Dr. Mallinson and Davis, listened to over 200 real and fake audio files to find human discernable audio cues helpful in identifying fake audios. Based on phonetic and phonological features in audio they defined 5 features. We refer them as Expert Defined Linguistic Features (EDLFs) as shown in figure 1. Our sociolinguistics experts found that an anomaly in these 5 EDLF is good indicator of fake audio.

Pitch	High or low tone of the speech sample.
Pause	Break in speech production within a sample.
Initial/Final Bursts	Lack or addition of burst of air at unexpected place in speech
Intake/ Outtake of breath	Presence or absence of any audible intake or outtake of breath
Sound Quality	Disturbance or distortion to the sound, including sound that was perceived to be unusually tinny, robotic, or compressed

Figure 1: Definition of 5 EDFLS

in Fall 2022, Davis and Khanjani led a four-week training with four undergraduate students, two from data science backgrounds and one from a social science background, with the goal of replicating the process of understanding five expert-defined linguistic features (EDLFs) and applying this knowledge to evaluate audio clips
As a pre-assessment, students were asked about their understanding of the 5 EDLFs. As expected, these students (who had no background in linguistics) had little understanding of the features and were unable to clearly define any of the 5 EDLFs. Subsequently, Davis presented and discussed the EDLFs, using examples from the sound clips in this project's datasets to illustrate each feature to increase students' knowledge and understanding. At the conclusion of the training, the graduate student trainer asked students to recall and discuss what was learned.

Results

Qualitative data obtained from students indicated that, having learned about the EDLFs, students were able to listen with a deeper intention in ways that improved their listening of and perceptions of genuine versus spoofed audio.

Our findings indicate the need to expand the training to improve other listeners' awareness of and ability to discern audio deepfakes—especially for college students, who are constantly exposed to social media and the potential for deception and misinformation/disinformation online.

As one student noted, "I learned about some of the formal [linguistic] indicators for a deepfake…, as well as training myself when to and when not to form a conclusion based on an anomaly on the authenticity of an audio file."

Conclusion

Preliminary findings indicate that the training process appears to effectively develop, enhance, and deepen students' linguistic understanding and their ability to discern audio deepfakes. Our ongoing work aims to further develop tools and trainings to guide listeners to more accurately discern audio deepfakes. As of March 2023, we have begun expanding our training by launching a pilot study with 40 undergraduate students and will complete the analysis of the pilot data by Summer 2023.

References

Nycz, J. (2020). English Phonetics. The Handbook of English Linguistics, 323-343.

Khanjani, Z., & Janeja, V. P. (2022). Audio Deepfake: a Survey, Frontiers in Big Data, doi: 10.3389/fdata.2022.1001063 Watson, G., Khanjani, Z., & Janeja, V. P. (2021). Audio Deepfake Perceptions in College Going Populations. arXiv preprint arXiv:2112.03351.

Kim, K. W., Park, S. W., Lee, J., & Joe, M. C. (2022, May). Assem-vc: Realistic voice conversion by assembling modern speech synthesis techniques. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6997-7001). IEEE.

FoR: A Dataset for Synthetic Speech Detection - Ricardo Reimao, Vassilios Tzerpos 2019 IEEE International Conference on Speech Technology and Human-Computer Dialogue (SpeD), pp. 1-10.