Battling Misinformation through Interdisciplinary Collaboration

Zahra Khanjani, Vandana P. Janeja, Christine Mallinson
(Information Systems / Language, Literacy & Culture / UMBC)

- Climate change is a real phenomenon caused by human activities, and the vast majority of scientists agree on this fact. However, a surge of climate misinformation is obscuring the truth, leading to potentially harmful outcomes (Environmental Defense Fund). Social media is inundated with climate myths in various forms, including videos, images, audio, and text.
- Deepfakes, in any of these formats, pose a significant threat by amplifying misinformation, especially given the availability of open-source generative models online and social media's power to rapidly spread false information.
- In this study, we focus on audio as an example of social media content, given its presence in both audio and video formats.
- We illustrate how interdisciplinary collaboration can enhance AI tools for detecting deepfakes, as well as improve human abilities to discern truth, particularly at the audio level. This method can be adapted to other forms of deepfakes, such as video and text, because our study incorporates linguistic insights, which are relevant across these formats.
- Augmenting data representations with phonetic and phonological features of natural speech extracted by sociolinguistics, which we call **Expert Defined Linguistic Features (EDLFs)**, helped improve spoofed audio detection, both in the areas of (1) AI detectors' performance as well as in (2) human spoofed audio discernment.

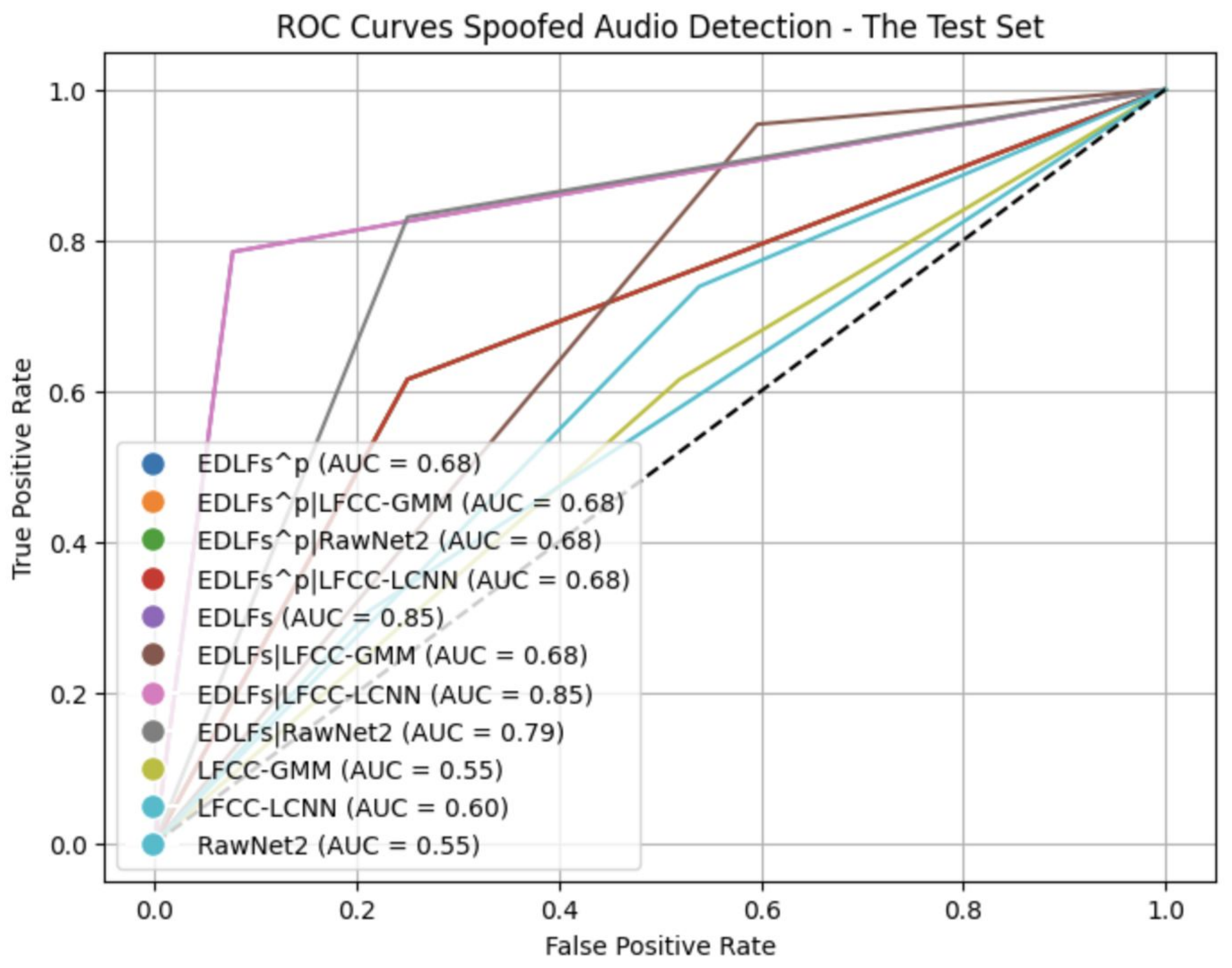
Linguistic Based AI Augmentation

840 audio files balanced in terms of spoofed vs genuine samples

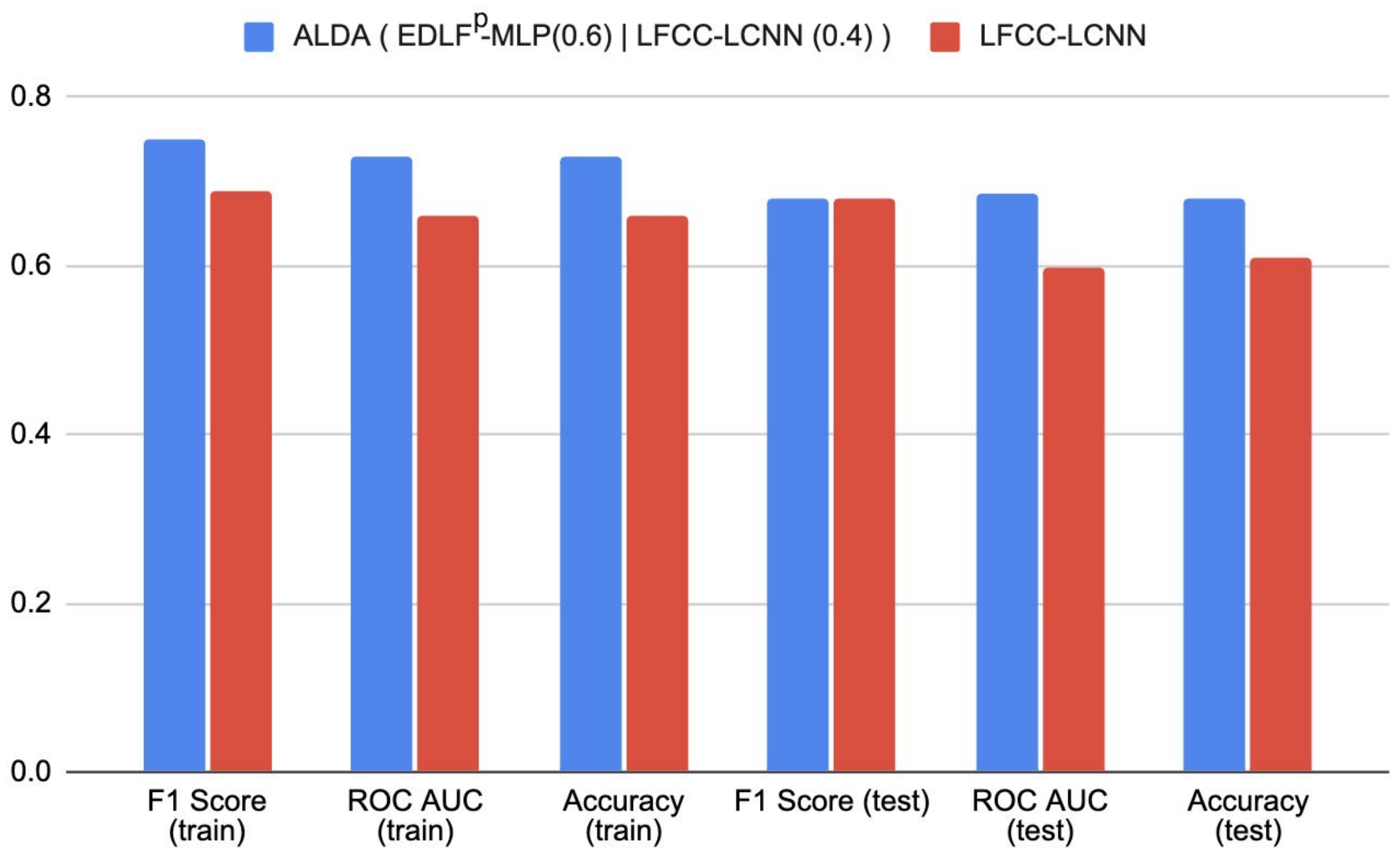
For spoofed samples, we include multiple types of attacks (Text-to-Speech, Voice Conversion, Replay and Mimicry)

15% of the whole data is held out - unseen/ test set - for which the types of attack distributions are preserved.

AI Models: Results



While using EDLFs as input features gives the highest AUC score in spoofed audio detection, predicted EDLFs outperform the common baselines as well.



We gave 0.6 weight to the labels from MLP fed by the EDLFs_p and 0.4 weight to the labels from the baseline, in the voting process

Five Expert Defined Linguistic Features

Pitch	Relative high or low tone of a speech sample
Pause	Break in speech production within a speech sample
Word Initial/Final Consonant Bursts	Release bursts of consonant stops /p/, /b/, /t/, /d/, /k/, and /g/
Intake/Outtake of Breath	Presence or absence of any audible intake or outtake of breath
Audio Quality	Overall qualitative estimation of the audio quality of a speech sample

EDLFs include commonly occurring, variable, and distinguishing phonetic and phonological characteristics of spoken English. *For each sample, the sociolinguist team members perceptually identified and identified the* ***presence or absence of these features and annotated any anomalies in their production****. As such, the labels indicate potential linguistic characteristics of real versus fake audio.*

Conclusions and Future Work

- We utilized EDLFs both to train AI based detectors and to train students to better discern spoofed audio.
- We saw improvements by using EDLFs in addition to traditional features in AI models. For AI models, we introduced ALDAS, which establishes a mechanism of auto labeling linguistic features and their viability for spoofed audio detection. These auto labeled features significantly improve common baselines of ASVspoof 2021. We demonstrate the importance of developing models in cooperation with and under the supervision of linguists and domain experts who are able to inform auto-annotation models and validate results.
- Detecting fake audio speech is difficult for humans, and equipping listeners with tools to help them spot it proves to be an ongoing challenge. Our findings indicate that the reading passages educating students about audio deepfakes, which served as the control, had a positive, though limited, impact on students' discernment accuracy points to the need for digital media literacy education in tandem with efforts to improve listener accuracy based on audible cues. Efforts to improve the public's ability to spot misleading content should take a holistic approach that incorporates insight from across disciplines.
- While our work flags content as potential deepfakes, highlighting possible misinformation, we recognize the need for an additional layer of scrutiny: If content is identified as fake, is the information itself false? Conversely, if the content is genuine and from a verified source, is the information it conveys accurate? This aspect is beyond the scope of our current work.

User Training

Over the course of two semesters and two phases of pilot studies, we developed a short training module that introduces students to five EDLFs as potential indicators of real vs. fake speech.

In Fall 2023, a doctoral student trained 264 students across nine undergraduate classes. Six classes received sociolinguistic training, while three served as a control and received only a reading passage about audio deepfakes.

In a pre-test, students listened to 20 short real and fake English clips and labeled each as real, fake, or unsure. After receiving either the training or the reading passage (control), students completed the same assessment eight weeks later as a post-test.

User Training: Results

- The training **increased confidence** for some students, yet a decrease in their unsurety did not always come with an similar increase in deepfake discernment accuracy.
- Training led students to be **more skeptical** of genuine speech samples, leading them to label real clips as fake
- The control group, who received a short reading about audio deepfakes, **also showed improvement, but very slightly.**

References

Khanjani, Z., et al., 2023, Learning to listen and listening to learn: Spoofed audio detection through linguistic data augmentation. IEEE ISI
Mallinson, C., et al., (2024). A place for (socio) linguistics in audio deepfake detection and discernment: Opportunities for convergence and interdisciplinary collaboration. *Language and Linguistics Compass*, *18*(5), e12527.
Nwosu, K., Evered, et al.,(2023). Auto Annotation of Linguistic Features for Audio Deepfake Discernment. In *Proceedings of the AAAI Symposium Series* (Vol. 2, No. 1, pp. 242-244)
Khanjani, Z., Watson, G., & Janeja, V. P. (2023b). Audio deepfakes: A survey. *Frontiers in Big Data*, 5, 1001063.

This work was supported by the National Science Foundation, Award #2210011.

HDR Conference 2024