Transdisciplinarity in Audio Deepfake Discernment with Expert-in-the-loop AI Models

Vandana Janeja
University of Maryland, Baltimore County

Faculty Collaborator: Christine Mallinson (Language Literacy and Culture, UMBC)
UMBC PhD students: Zahra Khanjani (IS), Noshaba Bhalli(IS), Lavon Davis (LLC)
Undergraduate Students: Chloe Evered (Linguistics-Georgetown), Kifekachukwu Nwosu (Computer Science-RIT)

Talk at the Federal Information Integrity R&D Interagency Working Group (IIRD IWG), March 22, 2024

Funded by NSF award #2210011

Fake or Real

a **e** i

b **f**

c **g**

d **h**

- d
- e
- f

Fake or Real

a **e** i

b **f**

c **g**

d **h**

- d real
- e TTS
- f TTS

Jim Gray 2007, "after you have captured the data, you need to curate it before you can start doing any kind of data analysis".

*Hinton has raised an alarm over the flood of false content where the average person will not be able to know truth from fiction, 2023**

Natural language is not a synonym for English

Miriam Webster's 2023 word of the year is "Authenticity" while also identifying a crisis of authenticity. +

*https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html

+Teresa Nowakowski, Merriam-Webster's 2023 Word of the Year Is 'Authentic', Smithsonian Magazine, Nov. 29, 2023, https://www.smithsonianmag.com/smart-news/why-merriam-websters-2023-word-of-the-year-is-authentic-180983329/

Audio Deepfakes: what and why?

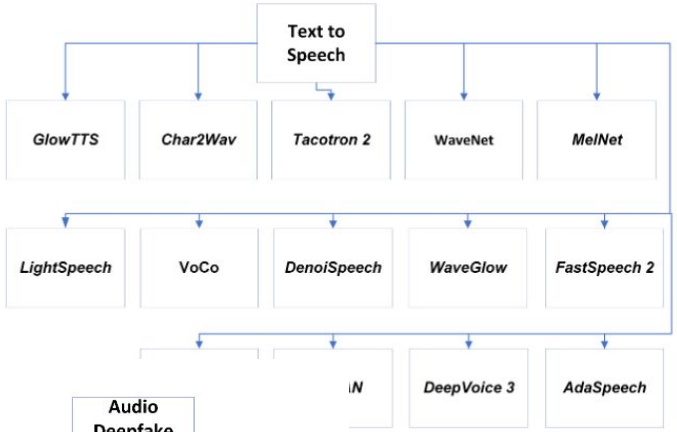
Why Audio?

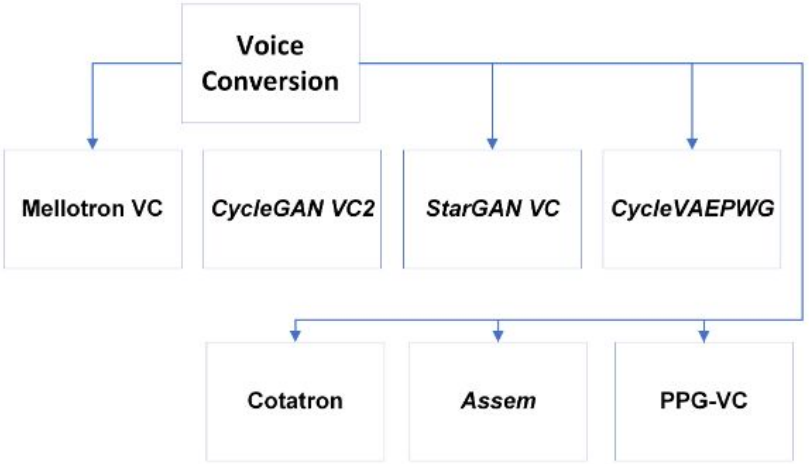
- Recent incidents of Fraud using spoofed audio

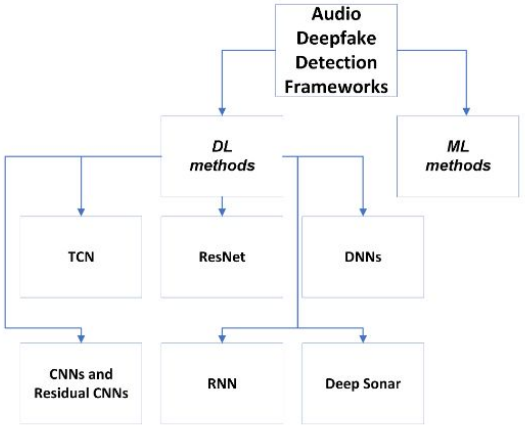
Image: Tero Vesalainen (Shutterstock)

Audio Deepfakes Landscape

Spoofed Audio Generation and Detection







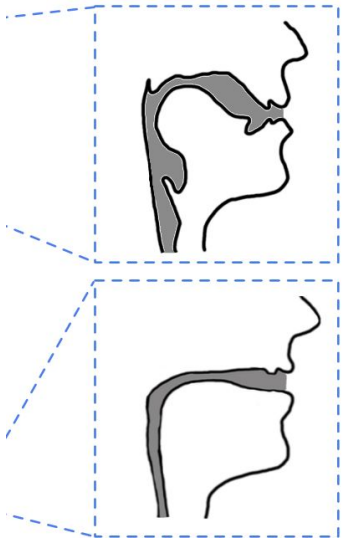
Spoofed audio countermeasures typically rely on improving algorithms to “catch” fakes, leading to a vicious cycle

Khanjani, Z., Watson, G., & Janeja, V. P. (2023). Audio deepfakes: A survey. Frontiers in Big Data, 5, 1001063.

Using Human Features

Articulatory phonetic techniques

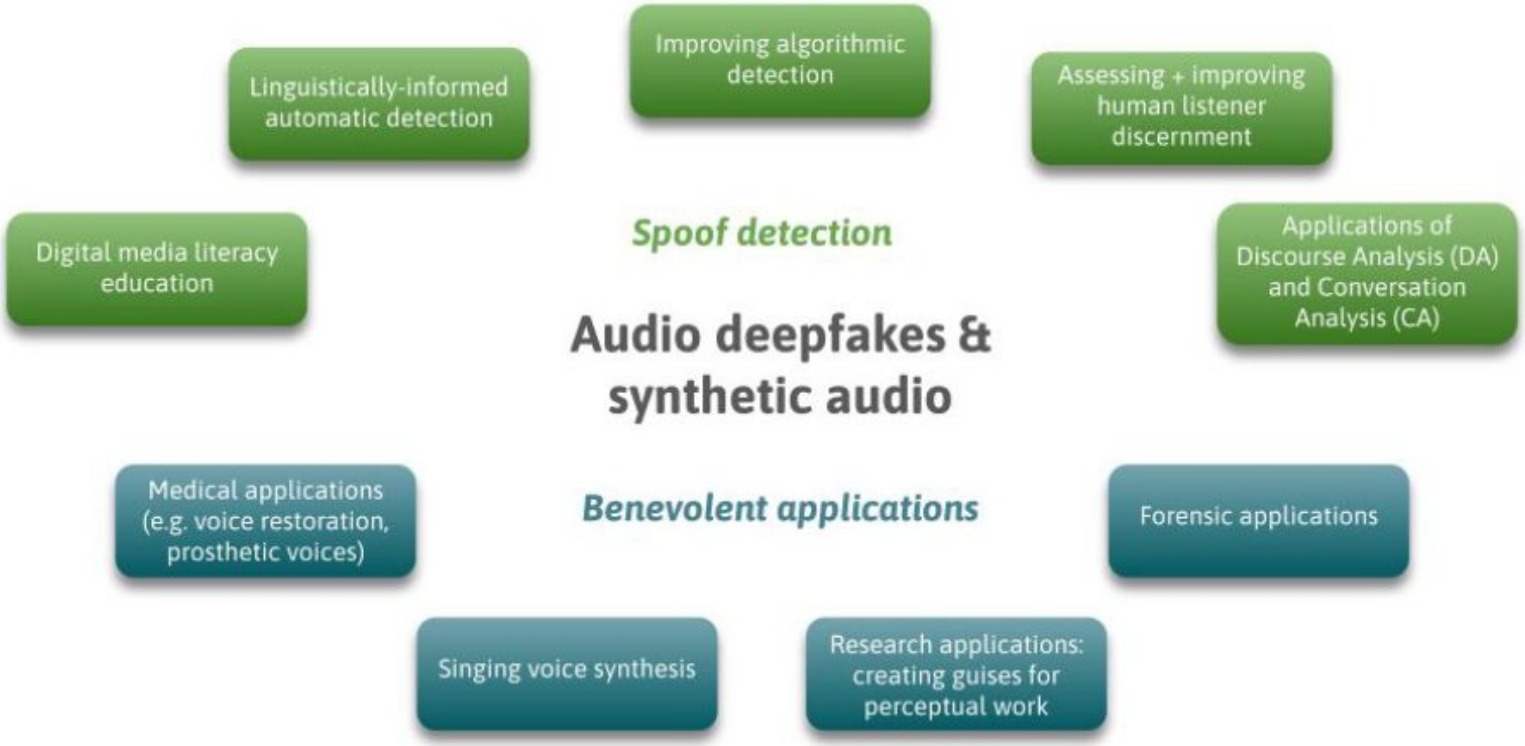
- To identify spoofed English audio by discerning that the clips in question were impossible or highly unlikely to have been produced in a human vocal tract.
- Type of attack (TTS), generative algorithm (Tacotron 2)
- The figure shows An anatomical approximation of a deepfaked model (bottom), which no longer represents a regular human vocal tract (top) and instead is approximately the dimensions of a drinking straw.



Resonator pre-processing, vowels, training, and methodology to do the analysis and an authentic audio sample for comparison.

BLUE, L., WARREN, K., ABDULLAH, H., GIBSON, C., VARGAS, L., O'DELL, J., BUTLER, K., AND TRAYNOR, P. Who are you (i really wanna know)? detecting audio {DeepFakes} through vocal tract reconstruction. In 31st USENIX Security Symposium (USENIX Security 22) (2022), pp. 2691–2708

Generative AI - Good, Bad and Ugly



What we did

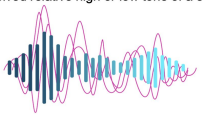
A place for Expert-in-the-loop

Linguistic Data Augmentation based on the knowledge of sociolinguistics experts--Strengthening AI with human knowledge and Strengthening human discernment with human knowledge

Expert Defined Linguistic Features (EDLFs)

Pitch

Pitch: the perceived relative high or low tone of a speech sample.



PITCH

- Defined for this study as the perceived relative high or low tone of the speech sample.
- Anomalous occurence of pitch-the sample received an annotation of 1
  - unusually higher or lower than expected, or
  - unusually fluctuating or inconsistent
- Normal occurrence usual or within a normal range of English language variation annotated with a 0

Pause

Pause: a break in speech production within a speech sample.

Pause

- A break in speech production within a sample.
- Anomalous Pause-the sample received an annotation of 1
  - lack of a pause where one would be expected,
  - addition of a pause where one would not be expected (such as between words of a phrase),
  - an overly long or short pause
- Pause as usual or within a normal range of English language variation annotated with a 0.

Lack of a burst of air where one would be expected, the addition of a burst of air where one would not be expected, or an unusually produced burst at the beginning or end of a word.

Bursts: Word-initial or word-final consonant stops

- The sounds /p/, /b/, /t/, /d/, /k/, and /g/
- Anomalous received an annotation of 1
  - lack of a burst of air where one would be expected,
  - The addition of a burst of air where one would not be expected,
  - An unusually exaggerated or truncated burst
- Production of consonant sounds perceived as usual or within a normal range of English language variation annotated with a 0

Expert Defined Linguistic Features (EDLFs)

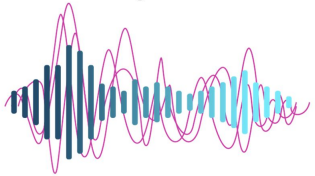
Audio Quality

Any disturbance or distortion to the speech signal

Tinny
Nasal
Echo
Compressed
Buzzing
Robotic

Pitch

Pitch: the perceived relative high or low tone of a speech sample.



Pause

Pause: a break in speech production within a speech sample.

Initial and Final Consonant Bursts

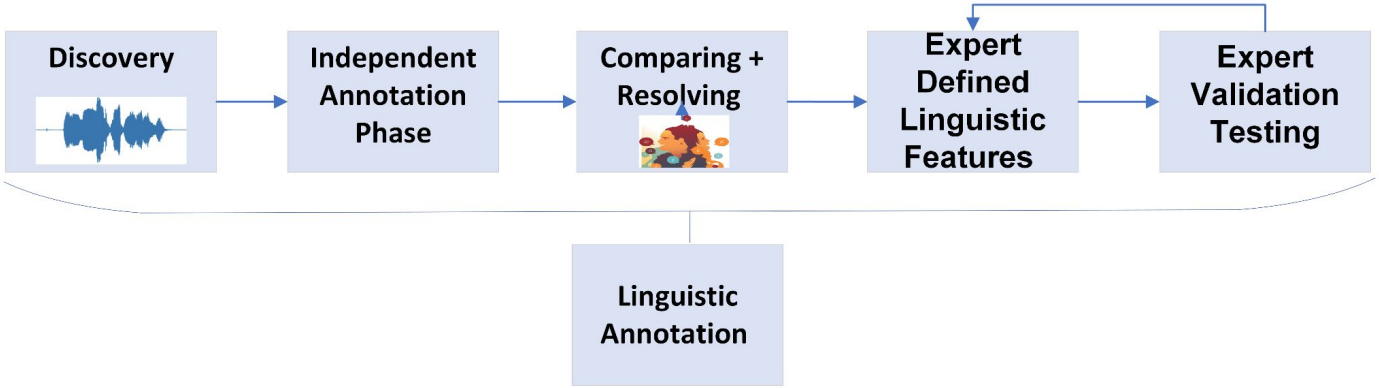
Lack of a burst of air where one would be expected, the addition of a burst of air where one would not be expected, or an unusually produced burst at the beginning or end of a word.

Breath

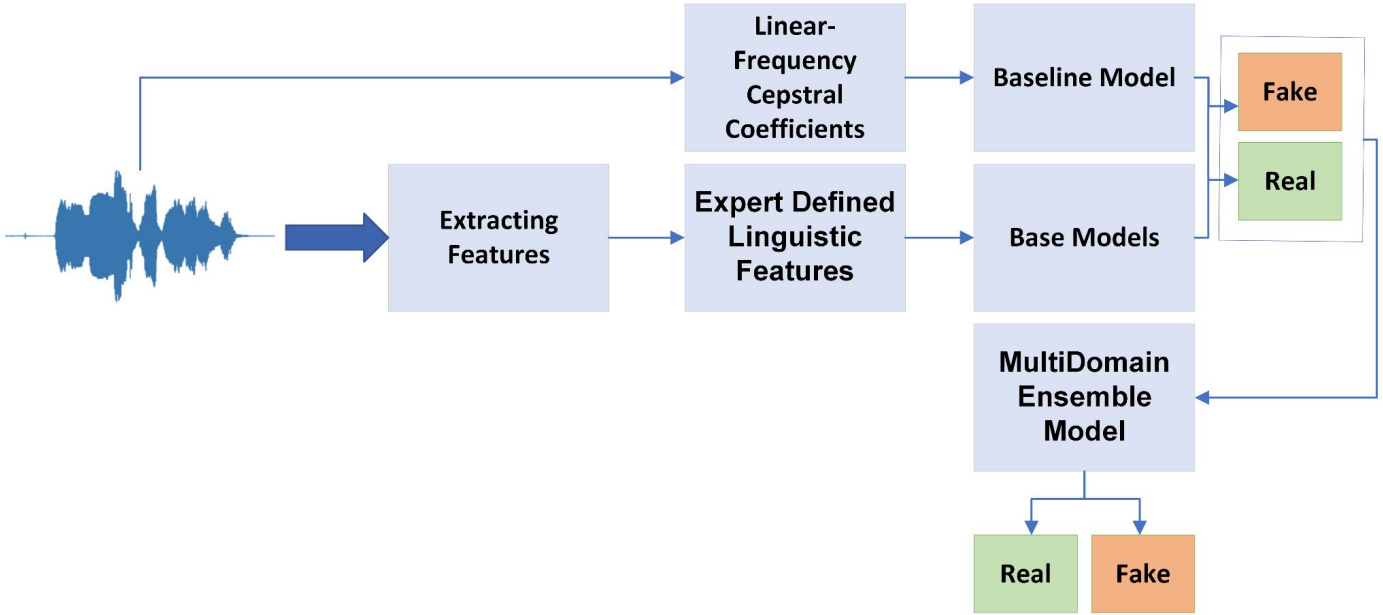
Any intake or outtake of breath

Augmenting AI Models

Linguistic Labelling

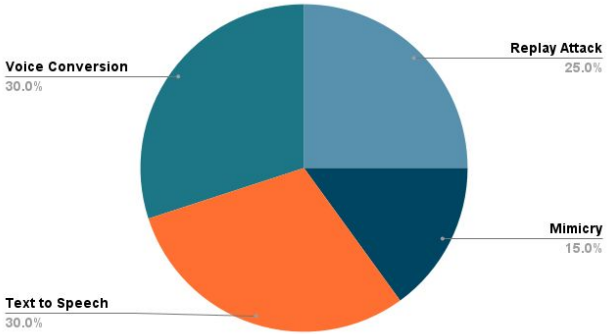


Augmenting AI models



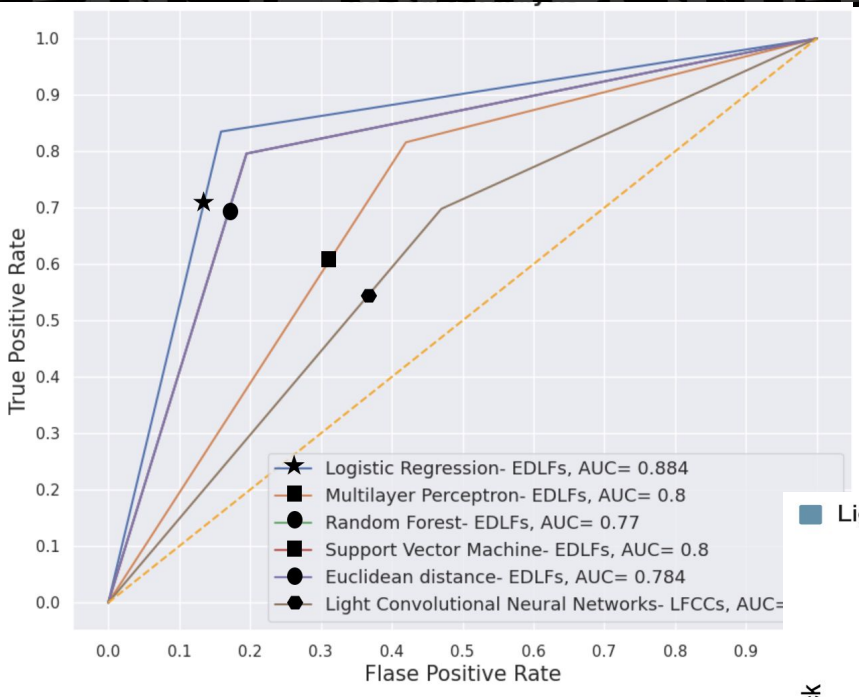
Khanjani, Z., Davis, L., Tuz, A., Nwosu, K., Mallinson, C., & Janeja, V. P. (2023, October). Learning to listen and listening to learn: Spoofed audio detection through linguistic data augmentation. In *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)* (pp. 01-06). IEEE.

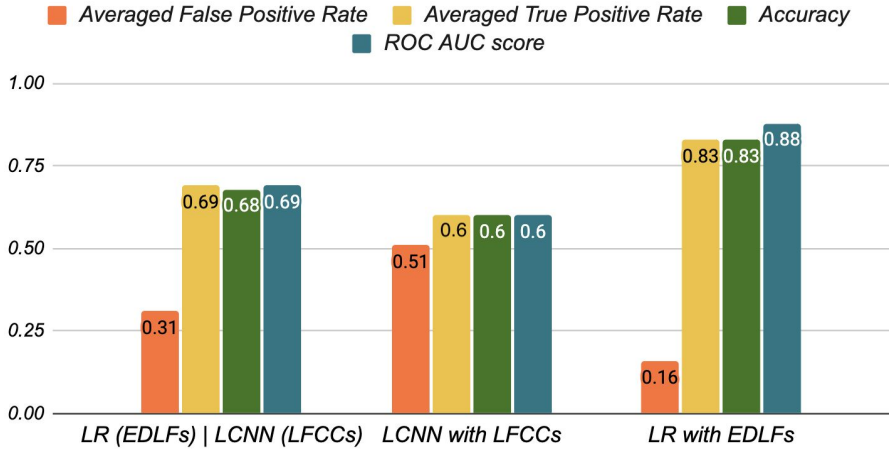
Spoofed Audio Dataset

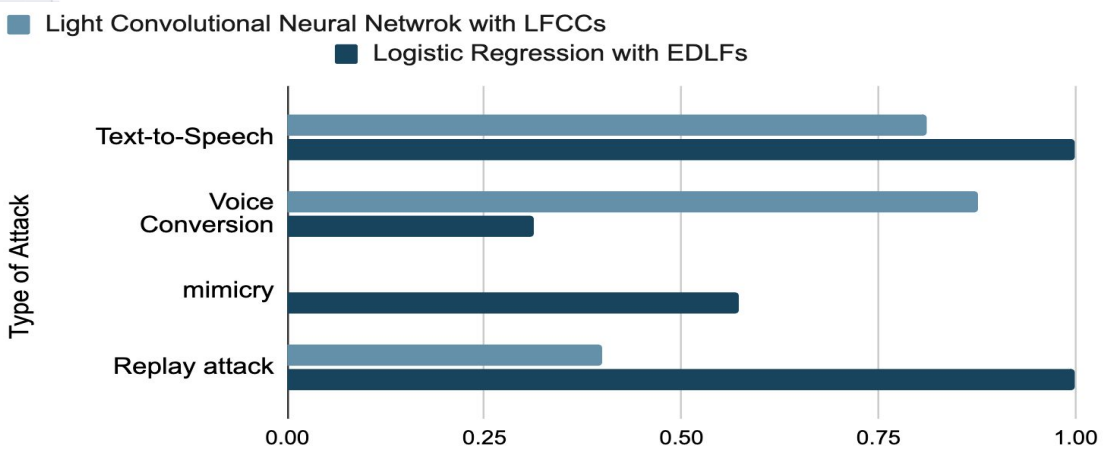


- Multiple types of spoofed audio
- State-of-the-art VC methods included
- Subset of available datasets with added samples

Results


True Positive Rate
Flase Positive Rate
Logistic Regression- EDLFs, AUC= 0.884
Multilayer Perceptron- EDLFs, AUC= 0.8
Random Forest- EDLFs, AUC= 0.77
Support Vector Machine- EDLFs, AUC= 0.8
Euclidean distance- EDLFs, AUC= 0.784
Light Convolutional Neural Networks- LFCCs, AUC=

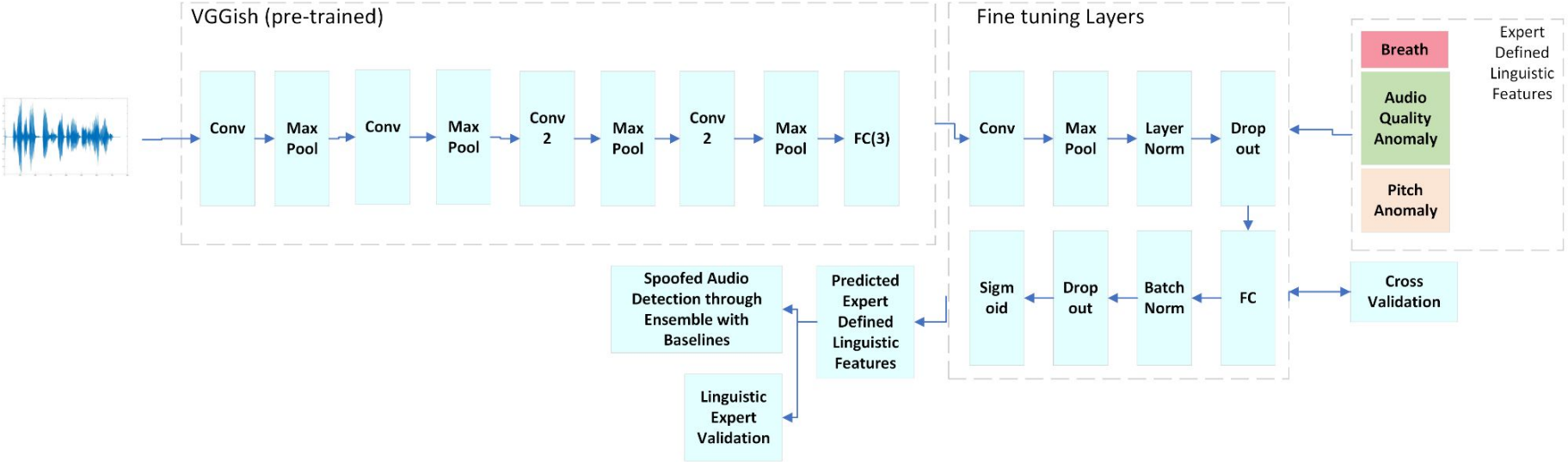

Averaged False Positive Rate
Averaged True Positive Rate
Accuracy
ROC AUC score
0.31
0.69
0.68
0.69
0.51
0.6
0.6
0.6
0.16
0.83
0.83
0.88
LR (EDLFs) | LCNN (LFCCs)
LCNN with LFCCs
LR with EDLFs


Light Convolutional Neural Netwrok with LFCCs
Logistic Regression with EDLFs
Type of Attack
Text-to-Speech
Voice Conversion
mimicry
Replay attack

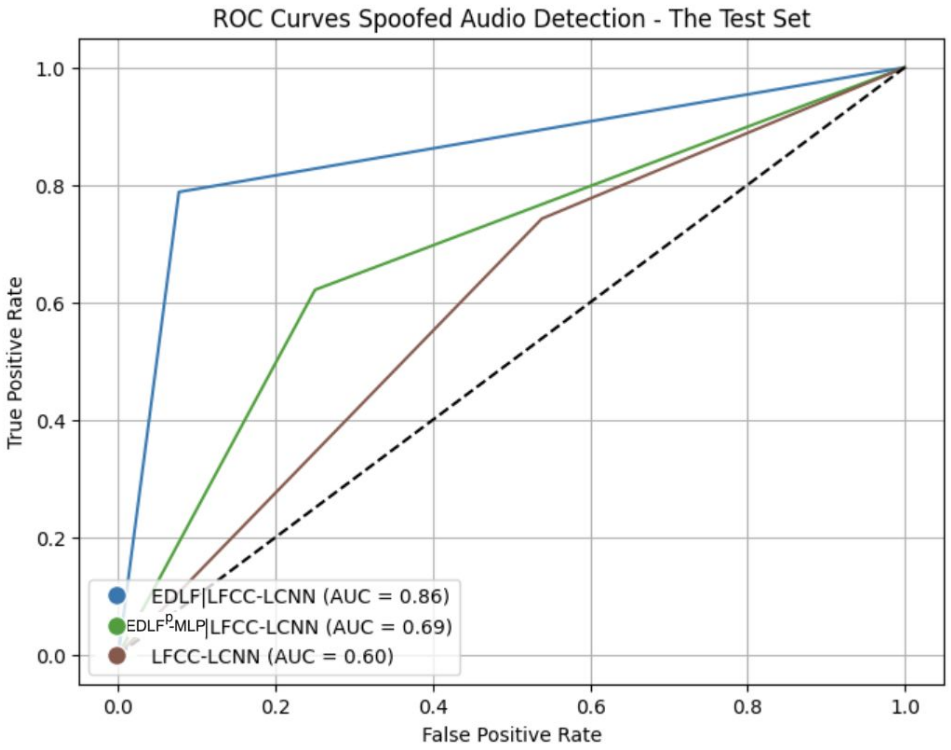
What we found?

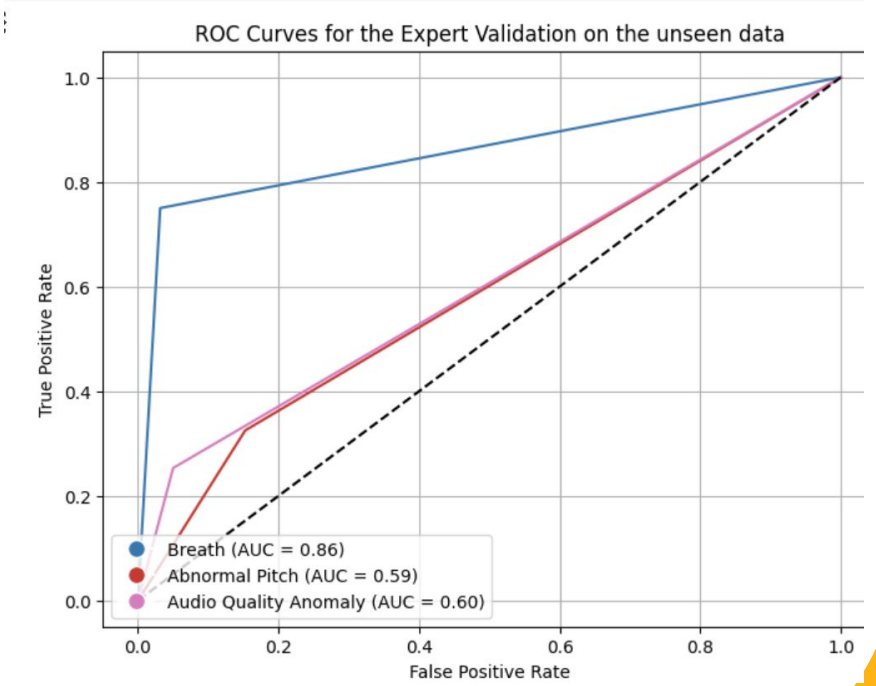
- Expert Defined Linguistic Features (EDLFs) as input feature set to a machine learning classifier such as Logistic Regression shows substantially improved performance in detecting speech synthesis and replay attacks.
- The improved performance of EDLFs indicates the value of using linguistic features as annotations on audio signals.
- This is especially useful if auto annotation techniques can work in conjunction with experts to train better spoofed audio detection models.
- EDLF-based models alone outperformed the ensemble model.
- EDLFs also helped with performance improvement of the baseline in the ensemble model
- Scaling of the labels is a challenge

Auto Labelling



How did AI auto labelling do as compared to Experts?





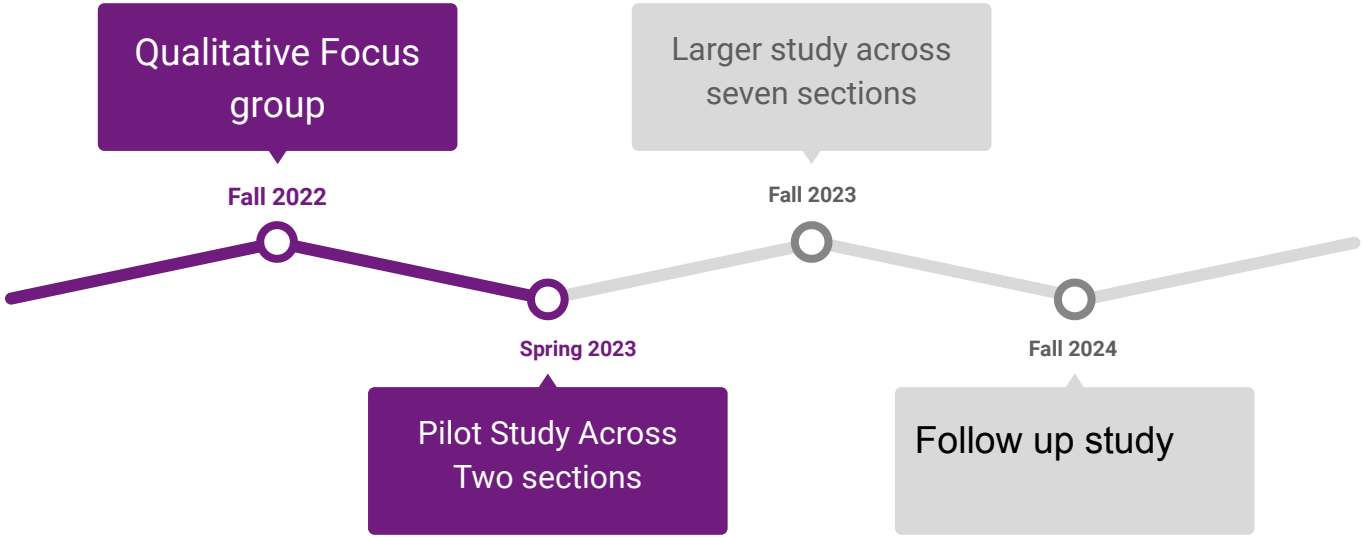
Augmenting AI: What else should we think about?

- Auto labelling
- Auto annotation of features as subsequences*
- Embedding other linguistic knowledge
- Multilingual variations
- Bias in deepfake audio
- Differently abled listeners
- Can we listen better as humans?

**Nwosu, Kifekachukwu, Chloe Evered*, Zahra Khanjani, Noshaba Bhalli, Lavon Davis*, Christine Mallinson, and Vandana Janeja. “Auto Annotation of Linguistic Features for Audio Deepfake Discernment.” Assured and Trustworthy Human-Centered AI (ATHAI). Workshop paper delivered at the fall symposium, Association for the Advancement of Artificial Intelligence. Arlington, VA: October.

Training: Augmenting Human Knowledge

Is there any benefit to providing sociolinguistic training to improve undergraduate students' knowledge of, familiarity with, and discernment ability regarding audio deepfakes?



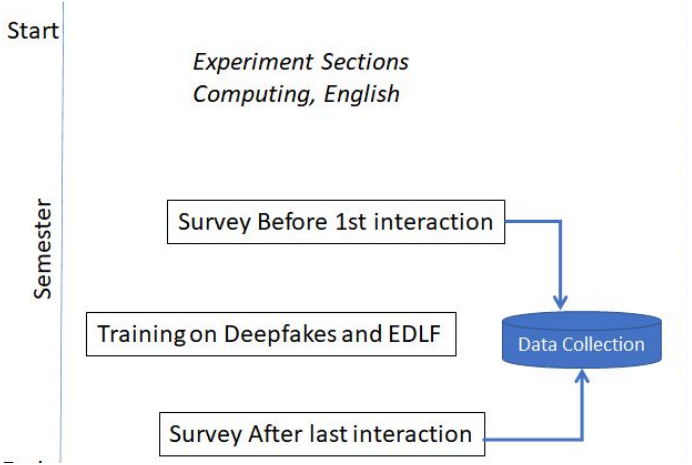
Part I: Qualitative Pilot (focus group)

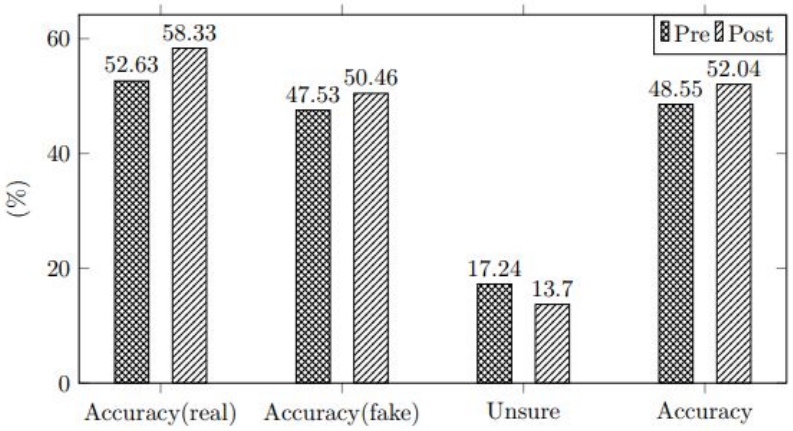
- Four one-hour training sessions with three undergraduate students with no background in linguistics
- Students were able to **listen with a deeper intention** and **explain concepts from the training** to peers with minimal understanding

"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** [about] the authenticity of an audio file."

"After the training I am **confident** to be able to distinguish [anomalous EDLFs] in an audio clip, **listen much more carefully**, considering the **context** of audio recordings, speaker background, additional noise etc., and **approach this task without jumping straight to assumptions**."

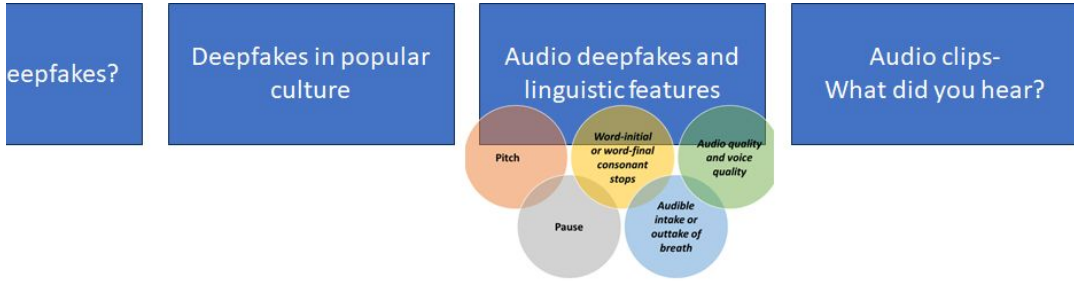
Part II: Pilot Study (Spring 2023)





- 27 students across two introductory undergraduate courses
- Pre-survey
  - 20 audio clips (half real, half fake)
  - Real, fake, or unsure?
  - Open-ended questions
- 20-minute training session
  - Based on longer training session from Fall 2022
- Post-survey
  - Administered one month after the pre-survey

Debrief



Ongoing

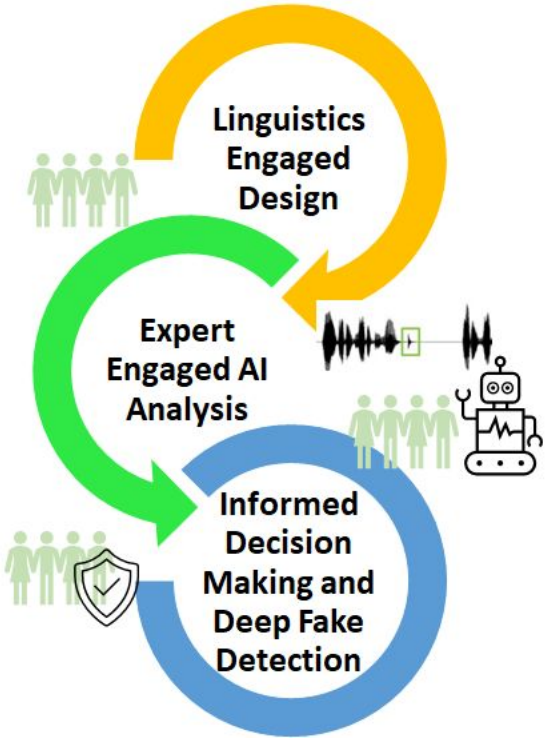
- Phase III: with over 7 sections (experiment and controls)
- Deep dive for understanding significance of findings
- Unsurity - Overconfidence or improvement or increased skepticism?
- Exploring - Other types of training (linguistics and readings), type of training, length of training, longitudinal follow through

What worked and What did not

Lessons

- Training humans for discernment is hard but possible
- Linguistic features led to improvements in AI model performance
- Complex audio scenarios should be tackled right away as a research agenda
- Need for infrastructure (people, software and data)
- Generative AI tools need guardrails and legislation, need for advocacy
- Transdisciplinarity is not easy but very fruitful when done right
- Need to develop trust and partnership
- Lot more is possible

Opportunities



Test Yourself and Learn More: Online and In person Exhibit

- https://mdata.umbc.edu/umbc-cadvc-deepfake-gallery-exhibit/

Hear from Our Students

- Retrieving the Social Sciences: https://socialscience.umbc.edu/episode-51/

THE CENTER FOR SOCIAL SCIENCE SCHOLARSHIP AT UMBC

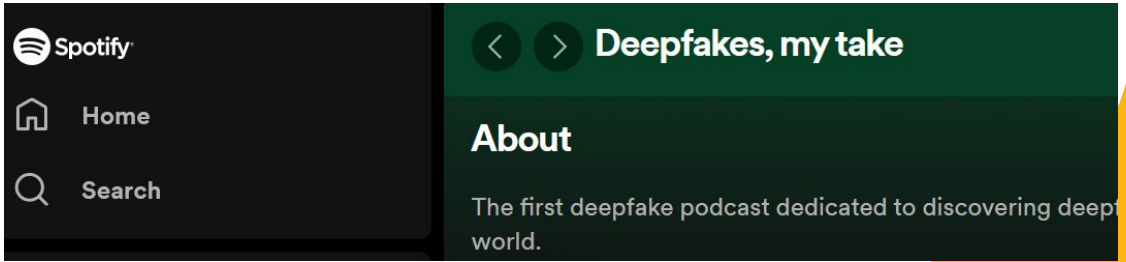
Ep.51: Using Interdisciplinarity to Tackle Audio Deepfakes

Monday Dec 18, 2023

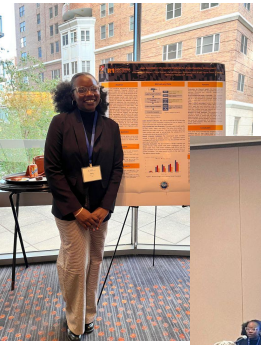
On today's episode I speak with two talented undergraduate researchers associated with the ongoing NSF-funded EAGER award led by Drs. Christine Mallinson and Vandana Janeja of UMBC.

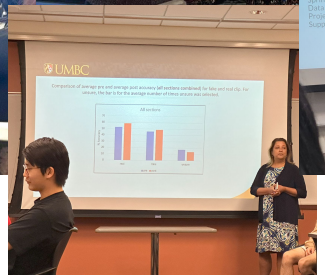
Kiffy Nwosu is an undergraduate computer science student from Maryland who has worked as a researcher at UMBC since high school, and is now a student at the Rochester Institute of Technology. Chloe Evered, originally of Houston, Texas, is a recent graduate of the Georgetown University department of linguistics with a minor in Chinese. Chloe is now pursuing a master's degree in linguistics, also at Georgetown.

- Deepfakes My Take By Gabrielle Watson



UMBC





MData lab

NSF

THE CENTER FOR SOCIAL SCIENCE SCHOLARSHIP AT UMBC

References

Bleaman, I. L., Webber, J. J., & Lo, S. K. (2023). Speech Synthesis in the “Mother Tongue”: Designing, Training, and Evaluating a Text-to-Speech System for Yiddish. *Journal of Jewish Languages, 11*(1), 15–43. https://doi.org/10.1163/22134638-bja10034

Blue, L., Warren, K., Abdullah, H., Gibson, C., Vargas, L., O’Dell, J., Butler, K., AND Traynor, P. (2022). Who are you (I really wanna know)?: Detecting audio deepfakes through vocal tract reconstruction. *Proceedings of 31st USENIX Security Symposium (USENIX Security 22)*, *USENIX Association,* 2691–2708. https://www.usenix.org/conference/usenixsecurity22/presentation/blue

Brewster, T. (2021). Fraudsters cloned company director’s voice in $35 million bank heist, police find. *Forbes*. https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions

J. Shen et al. (2018). Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Canada,* 4779–4783. doi: 10.1109/ICASSP.2018.8461368

Main, N. (2023, May 22). *Man scammed by deepfake video and audio imitating his friend*. Gizmodo. https://gizmodo.com/deepfake-ai-scammer-money-wiring-china-1850461160

Wang, R., Juefei-Xu, F., Huang, Y., Guo, Q., Xie, X., Ma, L., & Liu, Y. (2020). DeepSonar: Towards effective and robust detection of AI-synthesized fake voices. *Proceedings of the 28th ACM International Conference on Multimedia (MM ’20), Association for Computing Machinery,* 1207–1216. https://doi.org/10.48550/arXiv.2005.13770

Some of our publications for reference

* Indicates Undergraduate Students, Italics indicates Graduate students

- *Nwosu, Kifekachukwu*, Chloe Evered*, Zahra Khanjani, Noshaba Bhalli, Lavon Davis*, Christine Mallinson, and Vandana Janeja. “Auto Annotation of Linguistic Features for Audio Deepfake Discernment.” Assured and Trustworthy Human-Centered AI (ATHAI). Workshop paper delivered at the fall symposium, Association for the Advancement of Artificial Intelligence. Arlington, VA: October.
- Mallinson, Christine, *Lavon Davis, Chloe Evered*,* Vandana Janeja*, Noshaba Basir Bhalli, Zahra Khanjani, Nehal Naqvi*, and Kifekachukwu Nwosu**. “Learning to Listen: Training Undergraduate Students for Better Discernment and Detection of Audio Deepfakes.” American Association of Applied Linguistics: Houston, TX. March.
- Mallinson, Christine, Vandana Janeja, *Zahra Khanjani, Lavon Davis, Noshaba Basir Bhalli, Chloe Evered*, and Kifekachukwu Nwosu*.* “Incorporating Sociolinguistic Insights and Techniques to Enhance AI Based Methods for Audio Deepfake Detection: An Interdisciplinary Approach.” Linguistics Society of America: New York, NY. January.
- *Zahra Khanjani, Lavon Davis, Anna Tuz, Kiffy Nwosu*,* Christine Mallinson and Vandana Janeja, Learning to Listen and Listening to Learn: Spoofed Audio Detection through Linguistic Data Augmentation, 20th Annual IEEE International Conference on Intelligence and Security Informatics (ISI) , October 2023 Accepte