Auto Annotation of Linguistic Features for Audio Deepfake Discernment

Kifekachukwu Nwosu[1,2], Chloe Evered[1,3], Zahra Khanjani[1], Noshaba Bhalli[1], Lavon Davis[1], Christine Mallinson[1], Vandana P. Janeja[1,*†]

[1]University of Maryland, Baltimore County,
[2]Rochester Institute of Technology,
[3]Georgetown University

Abstract

We present an innovative approach to auto-annotate Expert Defined Linguistic Features (EDLFs) as subsequences in audio time series to improve audio deepfake discernment. In our prior work, these linguistic features – namely pitch, pause, breath, consonant release bursts, and overall audio quality, labeled by experts on the entire audio signal – have been shown to improve detection of audio deepfakes with AI algorithms. We now expand our approach to pilot a way to auto annotate subsequences in the time series that correspond to each EDLF. We developed an ensemble of discords, i.e. anomalies in time series, detected using matrix profiles across multiple discord lengths to identify multiple types of EDLFs. Working closely with linguistic experts, we evaluated where discords overlapped with EDLFs in the audio signal data. Our ensemble method to detect discords across multiple discord lengths achieves much higher accuracy than using individual discord lengths to detect EDLFs. With this approach and domain validation we establish the feasibility of using time series subsequences to capture EDLFs to supplement annotation by domain experts, for improved audio deepfake detection.

Introduction

In this paper, we focus on auto annotating linguistic features in audio signals to facilitate better detection of audio deepfakes. The prevalence of deepfakes – AI generated video, text, image, and audio – is increasing rapidly, as is research on deepfakes (Khanjani, Watson, and Janeja 2023). Although they can be used for entertainment purposes, deepfakes are also a nefarious mechanism used for fraud (Smith 2021), deception, and impersonation (Brewster 2021).

Techniques for audio deepfake detection have included deep neural network architectures such as ResNet (Chen et al. 2017), or Temporal Convolutional Networks (TCN) (Khochare et al. 2021). However, current methods that rely on automatic detection are "brittle" (Mai et al. 2023), and are frequently disrupted by adversarial models. Other approaches that do not rely solely upon automatic AI-based detection techniques include (Blue et al. 2022), which used articulatory phonetic features to identify fake English audio. Similarly, (Li, Ahmadiadli, and Zhang 2022) incorporated various acoustic measures in automatic detection analyses of real and fake audio, which improved algorithmic performance. While effective in improving audio deepfake detection, these methods and analyses are specialized and also may require an authentic audio sample for comparison.

Our interdisciplinary team recently showed that incorporating linguistic insights about English language variation can improve the detection of audio deepfakes (Khanjani et al. 2023). Prior to carrying out these studies, two team members who are sociolinguists reviewed a subset of 344 English audio samples from existing datasets commonly used for machine learning such as (Reimao and Tzerpos 2019) and new generated ones (Khanjani et al. 2023), in order to determine features for which the human voice audio files demonstrated perceptual variation, divergence, alteration, or absence compared to fake audio files. They then selected five phonetic and phonological features – which we call "Expert Defined Linguistic Features" (EDLFs) – that are frequent and easily discernible in spoken English: pitch, pause, word-initial and word-final consonant release bursts, audible intake or outtake of breath, and overall audio quality.

Using these expert annotations we trained AI models for deepfake detection and established that incorporating EDLFs in AI models improves detection of audio deepfakes (Khanjani et al. 2023). However, these EDLFs are at a signal level, showing a presence or absence of the feature in the audio signal. We thus aim to establish whether these EDLFs can be annotated inside the signal individually, as in Figure 1(a), to facilitate deepfake detection and help identify with a high level of explainability how and at which points the audio signals appear to be fake. We utilize time series subsequences, namely discords – anomalies in time series audio signals detected using matrix profiles (Yeh et al. 2016). Since discords can be detected with different lengths, we further create an ensemble to detect multiple discords across various lengths to capture different types of EDLFs.

Contributions: 1) We propose a novel methodology for auto annotation of linguistic features using time series discords. 2) We explore different discord lengths and their ability to annotate EDLF types. 3) Our linguistically informed approach enhances the unique nature of audio signals, laying groundwork for better detection of audio deepfakes.

*Corresponding Author: vjaneja@umbc.edu

†Authors would like to acknowledge support from the National Science Foundation, Award #2210011.

Methodology

Figure 1(b) shows our method for auto annotation of EDLFs. **Expert Defined Linguistic Features (EDLFs)**: In our work, we focus on the five EDLFs (Khanjani et al. 2023). The first four features are commonly occurring, variable, and distinguishing phonetic and phonological characteristics of spoken English: pitch (relative high or low tone of a speech sample), pause (break in speech within a speech sample), word-initial and word-final release bursts of consonant stops (/p/, /b/, /t/, /d/, /k/, and /g/), and audible intake or outtake of breath (at any point within a speech sample). As a fifth feature, we also included an overall qualitative estimation of the audio quality of a speech sample. For each sample, the sociolinguist team members perceptually 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. **Temporal Subsequences:** We utilize matrix profiles (Yeh et al. 2016) based discord detection to auto-annotate the EDLFs. Matrix profile provides a representation of time series to capture similarities and differences with a given length of subsequences. We scanned for discords, i.e, subsequences that were the most odd or dissimilar. Discord length is the key variable that determines the duration of the discords in seconds. The discord length (DL) used was 1K, 10K, 15K and 25K samples. DL translates to length in seconds as *DL/sample rate of audio*. The EDLFs are also of different lengths and may not show up in just one discord length. Combining these models with multiple discord length, we created an across-all-discord-length ensemble to evaluate the effect on accuracy. **Expert Linguistic Validation:** Sociolinguist team members validated if any EDLF overlapped with discords, if so, they provided start and end times for each EDLF as shown in figure 1(a). For across-all-discord-length ensemble if discord of any length overlapping with expert annotated EDLFs was considered at positive annotation. This helps account for EDLFs that align better with different discord lengths.

Experimental Results

We used a sample set of 50 audio clips containing 20 Text to Speech, 20 Voice Conversion, and 10 genuine clips. For details about clips, see (Khanjani, Watson, and Janeja 2023). **Evaluation by Discord Length:** Results of EDLF annotation by discord length are shown in Figure 2. Discord length of 25K had the highest accuracy and precision in annotating EDLFs, but it did not account for many of the EDLFs in the data, which were only captured by other discord lengths. **Across-all-discord-length Ensemble:** Performance across-all-discord-length ensemble method yielded much higher accuracy, precision, and recall than individual discord models; see Figure 2. Thus, we were able to account for all discords that were found in our dataset and reduced the numbers of false positives in the across-all-discord-length model. **Types of EDLFs Annotated:** Across the discord lengths, breath (45%) and pause (34%) were the most commonly found EDLFs, followed by consonant release bursts (10%) and audio quality (8%). The algorithm was able to find these EDLFs as they appeared on their own except for consonant release bursts, which was captured along side another EDLF. **Analysis of False Positives:** The sociolinguist team members re-examined the 142 instances of false positives –in which a discord did not overlap with an expert-annotated EDLF–and found that 48 of those discord intervals captured frication, i.e turbulent airflow in the vocal tract characteristic of a class of consonant sounds called fricatives, which are common in English speech (Zsiga 2013). This hypersensitivity to frication may be what allows discords to pick up on expert-annotated anomalous bursts, as the release of stop consonants involves a puff of air that may appear similar to voiceless fricatives in particular on a waveform and spectrogram. Nearly all instances of discords detecting frication occurred at a discord length of 1000, which is too short to capture most anomalous bursts in their entirety. Bursts judged to be anomalous by expert annotators were found by longer discord lengths (1K, 15K, and 25K), which we conclude are too long to capture a burst in isolation–as mentioned above, each discord that captured a burst also captured an additional EDLF. Thus, our approach would benefit from the addition of a discord length between 1K and 10K to improve auto annotation of anomalous bursts marked by domain experts.

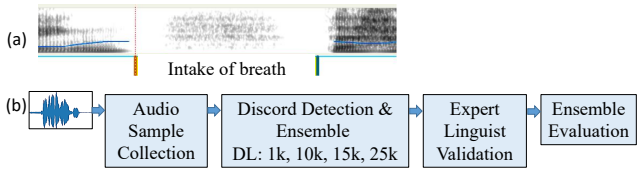

Figure 1: Auto Annotation of Linguistic Features

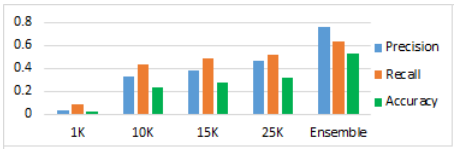

Figure 2: Discord Lengths and Ensemble Model evaluation

Conclusions and Future Work

In this paper, alongside expert validation we have established that time series mining provides an automated pathway to annotate EDLFs, which indicate distinguishing characteristics of real versus fake English audio for enhanced deepfake detection, and for better explainability as to when and where they occur in audio signals. In the future, we plan to test our approach on longer audio clips to validate and augment audio deepfake detection. Our model will be adjusted with more discord lengths to ascertain a possible correlation among discord lengths and types of EDLFs. With a linguistics-informed approach, we will also implement algorithms that check for repetitive patterns (namely motifs) in an audio to determine how those results correspond or differ from results produced with discords. This will also help us create other types of linguistics-informed ensembles for better annotation and audio deepfake detection.

References

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. In *31st USENIX Security Symposium (USENIX Security 22)*, 2691–2708.

Brewster, T. 2021. Fraudsters cloned company director's voice in $35 million bank heist, police find. https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/?sh=6037a7d37559. Accessed: 2023-08-05.

Chen, Z.; Xie, Z.; Zhang, W.; and Xu, X. 2017. ResNet and Model Fusion for Automatic Spoofing Detection. In *Interspeech*, 102–106.

Khanjani, Z.; Davis, L.; Tuz, A.; Nwosu, K.; Mallinson, C.; and Janeja, V. 2023. Learning to listen and listening to learn: Spoofed audio detection through linguistic data augmentation. In *Intelligence and Security Informatics: IEEE International Conference on Intelligence and Security Informatics, ISI*.

Khanjani, Z.; Watson, G.; and Janeja, V. 2023. Audio deepfakes: A survey. *Frontiers in Big Data*, 5: 1001063.

Khochare, J.; Joshi, C.; Yenarkar, B.; Suratkar, S.; and Kazi, F. 2021. A deep learning framework for audio deepfake detection. *Arabian Journal for Science and Engineering*, 1–12.

Li, M.; Ahmadiadli, Y.; and Zhang, X.-P. 2022. A Comparative Study on Physical and Perceptual Features for Deepfake Audio Detection. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 35–41.

Mai, K. T.; Bray, S.; Davies, T.; and Griffin, L. D. 2023. Warning: Humans cannot reliably detect speech deepfakes. *Plos one*, 18(8): e0285333.

Reimao, R.; and Tzerpos, V. 2019. For: A dataset for synthetic speech detection. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 1–10. IEEE.

Smith, B. 2021. Goldman Sachs, Ozy Media and a $40 million conference call gone wrong. The New York Times. https://www.nytimes.com/2021/09/26/business/media/ozy-media-goldman-sachs.html. Accessed: 2023-08-05.

Yeh, C.-C. M.; Zhu, Y.; Ulanova, L.; Begum, N.; Ding, Y.; Dau, H. A.; Silva, D. F.; Mueen, A.; and Keogh, E. 2016. Matrix profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In *2016 IEEE 16th international conference on data mining (ICDM)*, 1317–1322. Ieee.

Zsiga, E. C. 2013. *The sounds of language: An introduction to phonetics and phonology*, volume 7. John Wiley & Sons.