Learning to Listen and Listening to Learn: Spoofed Audio Detection through Linguistic Data Augmentation

***Abstract*—Spoofed audio, both human or machine generated, causes deception and disinformation and as such is a societal challenge. This study advances the detection of spoofed audio through a novel approach that augments knowledge about the audio data using linguistics.**

Using perceptual methods, for English audio samples, experts in sociolinguistics listened for audio cues and used binary labels to indicate the perceived authenticity of the overall speech samples, based on phonetic and phonological features that occur frequently in spoken English. These Expert Defined Linguistic Features (EDLFs) were then used in supervised spoofed audio detection methods to augment AI models.

An ensemble method based on multi-domain features both from the audio data itself and the EDLFs was also created to evaluate the spoofed audio detection, and to show how EDLFs can improve the traditional ways of spoofed audio detection.

Our findings indicate that augmenting the audio data with expert-informed linguistic annotation increased the accuracy of spoofed audio detection significantly in both of the training and testing datasets across the evaluated single and ensemble models. Our findings show the promising avenue of augmenting audio data with perceptual linguistic techniques, as a method of human discernment, to enhance AI based approaches for spoofed audio detection. These features also create the foundation for direct linguistic annotations on new audio clips for robust spoofed audio detection.

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

There are several types of human and machine generated spoofed audio, including mimicry, replay attacks, and audio deepfakes, with more generation mechanisms emerging every day. Mimicry entails impersonating another person's voice; replay attacks entail replaying the manipulated recording of a speaker's voice, and can be generated using simply a mobile phone that does not require artificial intelligence (AI) methods. Deepfakes are synthetically generated or manipulated using AI methods and can easily pass as real.

Deepfakes are emerging as a particularly significant societal challenge as a mechanism for deception and disinformation. Both shallow and deep audio fakes [8], and resulting fraud and deception, are increasing as technology improves and as society relies further on remote environments.

Motivation: *In September 2021, the New York Times reported on a story in which Ozy media [19] arranged a call with Goldman Sachs and a YouTube partner to drive a $40 million investment in the media company. In attempting to close the deal, a phone call was arranged between key executives. During the call, the New York Times reports, the voice of the man representing Ozy media and YouTube "began to sound strange to the Goldman Sachs team, as though it might have been digitally altered." The situation turned out to be an attempt to impersonate the executive to close the deal. Ozy media folded after this incident. In another incident that occurred in early 2020, a bank manager in Hong Kong [3] authorized a $35 million transfer based on what turned out to be an AI generated voice, with stolen funds apparently traveling through the U.A.E. into bank accounts in the U.S. These situations indicate the necessity of further research on spoofed audio content.*

As we can see from this motivating scenario that there are some nuances in the speech of individuals that can help discern audio spoofs. We utilize this intuition to look at linguistics features to augment spoofed audio detection.

There are several approaches that use pure machine AI techniques to detect deepfakes and spoofed audio. For spoofed audio detection two main categories exist: Machine Learning (ML) and Deep Learning (DL) [1]. Machine learning models, such as Support Vector Machine, Random Forest, and K-NearestNeighbors, have not achieved very high metrics, with 0.67 accuracy at best [9] when applied on FoR dataset [16]. For Deep Learning methods, different neural network structures are used. ResNet is utilized for audio deepfake detection [5] and then is improved to achieve better metrics and meet the generalization challenge [4]. However, the most important drawback of ResNet is that it is computationally heavy, and is not always able to adapt to new data. Some researchers have used Temporal Convolutional Networks (TCN) [9] on audio deepfake samples, but not all of the spoofed audio types. The latest Automatic Speaker Verification and Spoofing Countermeasures Challenge was (ASVSpoof 2021 [26]). This challenge proposed four baselines for the three different types of attacks as three tasks. The three different tasks are Physical Access (PA- replay attack samples), Logical Access (LA- mostly text-to-speech and voice conversion samples), and Deepfake (DF- actual audio deepfake samples). Another method uses layer-wise neuron activation patterns of voices [20]. A new study [2] utilized a novel approach (creating a mathematical model of the human vocal tract). This study is mainly looking at vowels, however we include more features such as pause, consonant bursts, and intake/outtake of breath, that they are not accounted for through vocal tract measurements. Also,

This work is funded by NSF award # 2210011, Enabling Interdisciplinary Collaboration: Improving Human Discernment of Audio Deepfakes via Multi-level Information Augmentation

this study uses the samples generated by a single algorithm (tacotron 2 [18]) and a single type of attack (text-to-speech).

These current methods rely on traditional features and are frequently disrupted by adversarial models which constantly improve on deepfakes.

Our proposed approach addresses these issue to introduce the elements of human discernment to spoofed audio detection.

We propose an approach innovative route for mitigating audio-based misinformation, specifically by for improving spoofed audio detection, through a combination of AI techniques and human discernment. By incorporating linguistic annotations from sociolinguistic experts on the audio signals, our method demonstrates the benefits of augmenting AI models with knowledge about human language.

Contributions: Our key contributions are as follows:

- We propose a novel mechanism to more accurately detect spoofed audio using distinguishing linguistic features selected based on the knowledge of sociolinguistics experts; we call these Expert Defined Linguistic Features (EDLFs) and we use them to augment traditional acoustic features (such as Linear-frequency Cepstral Coefficients).
- Utilizing the augmented multi-domain features from the audio data itself and from the EDLFs as inputs for the machine learning and deep learning models, we propose an ensemble method showing marked improvements in spoofed audio detection.
- Our proposed models are effective even when applied to small datasets, a benefit compared to the large datasets required for the use of traditional AI-based detectors. These promising findings demonstrate the need for expert based data augmentation strategies for spoofed audio detection. This also creates a foundation for direct automated linguistic annotation of new audio clips for robust spoofed audio detection models.

The rest of the paper is organized as follows. Section II discusses our overall methodology. In section III we discuss the experimental results and findings, before concluding and sharing future directions in section IV.

II. Methodology

Figure 1 presents our overall methodology for spoofed audio detection through linguistic data augmentation.

A. Spoofed Audio Dataset

Although there are some valuable datasets for the purpose of spoofed audio detection [6; 16; 22; 23; 24; 25; 26], none of them have used state-of-the-art Voice Conversion (VC) algorithms such as ASSEM-VC [10], which can generate very realistic samples [1]. Additionally, some datasets focus on the Text-to-Speech (TTS) type of audio deepfake and ignore VC samples [6; 16].

The hybrid dataset used in this study is a combination of samples from some existing datasets and new generated

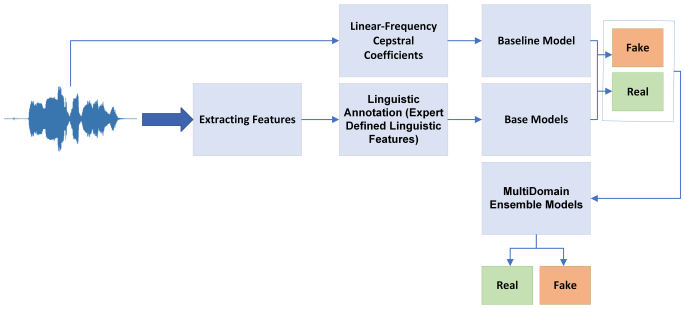

Fig. 1: Methodology for Spoofed Audio Detection through Linguistic Data Augmentation

samples. The details of the spoofed samples in our dataset are: 25 percent replay attack samples from [24], 30 percent TTS (generated by the algorithms MelGAN [11], Google WaveNet [14], and Descript samples [2]), 30 percent VC [3], and only 15 percent of mimicry samples [4]. Average of duration of the samples from each type is 3.5, 4.2, 4.1 and 2 seconds for replay attack, TTS, VC and mimicry respectively. The genuine samples are from FoR dataset [16], LJ Speech dataset [7], and Obama's public speeches with the average duration of 3.8 seconds.

The details of the recording environments of the replay attack samples are 52.4 percent "office", 29.6 percent "home" and other samples "studio", "anechoic room" and " analog wire". All of the samples are mono-channel except the mimicry ones plus one of the MelGAN samples which are two channel audios. Our dataset includes **344 audio samples, balanced by number of genuine versus spoofed samples**.

B. Extracting Features

Using the hybrid dataset described we extracted and utilized attributes from audio samples as follows: 1) Linguistic annotation of the Expert Defined Linguistic Features (EDLFs), 2) Linear-Frequency Cepstral Coefficients (LFCCs).

Definition: given a set of n audio clips $A = a_1, a_2, ..., a_i, ..., a_n$ each a_i has linguistic features ${a_i}^{EDLF}$ and audio features ${a_i}^{LFCC}$.

In the remainder of this section, we explain each type of feature sets and our methodology for the feature extraction and annotation phase.

1) Linguistic Annotation: The linguistic annotation was aimed to provide augmented features to incorporate human discernment perspective to the datasets. The linguistic annotation process was completed in multiple phases.

Figure 2 shows the different phases of the linguistic annotation methodology.

First, in a discovery phase, two co-authors who are sociolinguists reviewed a subset of 344 audio samples from the spoofed audio dataset in order to perceptually evaluate areas in which the human voice audio files were likely to diverge from

[1]https://mindslab-ai.github.io/assem-vc/

[2]https://www.descript.com/lyrebird

[3]Public VC github respiratory- https://mindslab-ai.github.io/assem-vc/

[4]https://www.youtube.com/watch?v=cQ54GDm1eL0

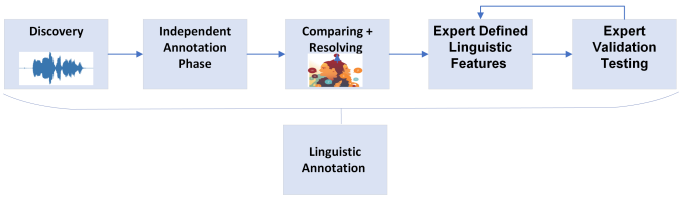

Fig. 2: Linguistic Annotation to Extract EDLFs

the spoofed files and/or were likely to be altered or absent in deepfakes, based on their knowledge of human language and of English language variation [13; 15].

After independently noting potential areas of divergence (Independent annotation phase), the sociolinguistics experts met to collaboratively to pick the most usable features in the Comparing + Resolving phase in terms of (a) frequency (features must be frequent enough in the dataset), (b) discernibility (features must be relatively easy to identify by ear), and (c) definability (features must be able to be clearly and reliably identified). Upon evaluating this sample dataset, the experts decided to focus on five features that commonly occur in spoken English, thereby providing pathways for future similar work using other English language audio clips. These features included: (1-4) a suite of phonetic and phonological features related to the production of spoken English and (5) an overall assessment of the audio sample's sound quality and voice quality. We call these Expert Defined Linguistic Features (EDLFs), as they are based on the sociolinguistics experts' perceptions of elements of an audio clip's sound and of the types of anomalies that may exist in it; they are discussed in detail in the following section. The sociolinguistics experts then mutually created criteria for annotating each EDLF, which are defined and described in the next section.

The sociolinguistics experts began by independently annotating the linguistic features for a subset of the audio samples, using the binary indicators of 1 and 0 for each of the linguistic features. In order to ascertain interrater reliability, the annotations from each expert were compared and found to have a convergence rate of approximately 90%. The few divergences in annotations were discussed between the experts and resolved. Having determined that they both consistently applied the linguistic annotations to the selected features, they then split the remaining audio samples from the total dataset and annotated the remaining samples accordingly.

Expert Defined Linguistic Features (EDLFs):
We next explain the five EDLFs: The experts focused on five features – (1) pitch, (2) pause, (3) word-initial or word-final consonant stops, (4) audible intake or outtake of breath, and (5) audio quality and voice quality – as those were features that met the criteria of frequency, discernibility, and definability as discussed above.

Pitch: Pitch was defined for this study as the perceived relative high or low tone of the speech sample. For a given sample, if any occurrences of pitch were perceived by the sociolinguistics experts to be anomalous – for example, pitch that is unusually higher or lower than expected or pitch that is unusually fluctuating or inconsistent – the sample received an annotation of 1. Samples in which pitch was perceived throughout as being usual or within a normal range of variation were annotated with a 0.

Pause: Pause was defined for this study as a break in speech production within a sample. For a given sample, if any occurrences of pausing were perceived by the experts to be anomalous – for example, 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), or an overly long or short pause – the sample received an annotation of 1. Samples in which pausing was perceived throughout as being usual or within a normal range of variation were annotated with a 0. *Word-initial or word-final consonant stops:* The sociolinguists considered the production of word-initial or word-final consonant stops – specifically the sounds of p, b, t, d, k, and g. For a given sample, if the production of any of the stops were perceived by the experts to be anomalous – for example, lack of a burst of air where one would be expected, or the addition of a burst of air where one would not be expected, or an unusually exaggerated or truncated burst (taking into account expected variation in global varieties of English) – the sample received an annotation of 1. Samples in which the production of consonant sounds was perceived throughout as being usual or within a normal range of variation were annotated with a 0. *Audible intake or outtake of breath:* For a given sample, if any intake or outtake of breath was heard by the experts, the sample received a label of Y for yes, indicating the presence of audible intake or outtake of breath. Samples in which no audible intake or outtake of breath were discerned by the experts were annotated with N, indicating absence of audible breath. In the preprocessing Y was encoded to 1 and N to 0. *Audio quality and voice quality:*The sociolinguists considered an overall qualitative estimation of the audio sample's audio quality and voice quality. For a given sample, if the sound or voice quality was perceived by the experts to be anomalous – for example, any disturbance or distortion to the sound, including sound that was perceived to be unusually tinny, robotic, or compressed – the sample received a label of 1. Samples in which sound and voice quality were perceived throughout as being usual or within a normal range of variation were annotated with a 0.

Definition: Given an audio clip a_i , with the linguistic feature set (EDLFs) where:

$$a_i^{EDLF} = [a_i^{breath}, a_i^{pitch-anomaly}, a_i^{audio-quality-anomaly}, a_i^{burst-anomaly}, a_i^{pause-anomaly}] \quad (1)$$

Quasi experimental results (Expert Validation Testing):
A quasi-experiment was applied to make sure the experts' judgments were unbiased and unaffected by the actual class labels of spoofed or genuine. In this phase, the sociolinguists

judged a completely new dataset including all types of the aforementioned spoofed audio without knowing the actual class of each clip. The EDLFs from this experiment gave us 0.75 accuracy (correctly classified instances) on unseen data using Multi Layer Perceptron. Both of the experts also yielded 0.75 accuracy in their final judgments for classifying the clips.This experiment thus validated the previous one, and demonstrated that the experts' judgments are aligned.

To establish the value that was added overall to the spoofed audio detection through the linguistic annotation, we also wanted to evaluate the methods based on the traditional audio features. We next discuss the Audio data features including LFCCs.

2) Linear-Frequency Cepstral Coefficients: a Linear cepstral representation of sound is called LFCCs.
Definition: When f_j is the jth frame of a_i, then:

$$LFCC_{f_j} = [{C_j}^1, {C_j}^2, ..., {C_j}^p, E_j] \tag{2}$$

Where ${C_j}^k$ is mandatory cepstral coefficient which represent the temporal properties of audio, and it can also be optional cepstral coefficients (such as delta or delta-delta cepstral coefficients) which represent the rate of change of the mandatory cepstral coefficients over time. The number of coefficients is p, and E_j is the energy coefficient which is also optional. Therefore, for an audio clip a_i which contains m frames, we have:

$${a_i}^{LFCC} = [{LFCC_i}^{f_1}, ..., {LFCC_i}^{f_j}, ..., {LFCC_i}^{f_m}] \tag{3}$$

where ${LFCC_i}^{f_j}$ represent the coefficients for the j_th frame of a_i. We provide raw audio data to the LFCC-LCNN detection model that captures these features. This method is one of the best baselines from ASVspoof 2021 competition, especially for logical access and deepfake tasks [26]. We use this model as a baseline to compare how it performs against EDLFs. The setup of the baseline based on the definition 2 is as follows: p=19, ${C_j}^k = {c_j}^k, {\Delta c_j}^k, {\Delta\Delta c_j}^k$ and it also contains E_j, for which ${\Delta c_j}^k = {c_j}^k - {c_j}^{k-1}$, and ${\Delta\Delta c_j}^k = {\Delta c_j}^k - {\Delta c_j}^{k-1}$.

C. Base and Baseline Models

We consider two types of features for evaluation, EDLFs vs the LFCC features described above. We evaluated the performance of EDLFs using several base AI models to quantify the accuracy of spoofed audio detection with each type of feature set. For all of the experiments using EDLF features, a test set (30 percent of the whole dataset, balanced in terms of different attack types and also spoofed vs genuine samples) as the unseen dataset, is kept out of the training steps. Anytime that we tuned hyper-parameters, the k-fold cross validation method with k=10 is empirically chosen and used. In particular we evaluated how the use of EDLFs impacted the base model performance against the baseline methods using LFCC-LCNN. Given a_i^{EDLF} as described in the definition 1 and the true class of each audio clip as $y = y_1, y_2, ..., y_n$, a function (f) maps a_i^{EDLF} to $y_i^{pred-EDLF}$ in a base model (B), such that:

$$B => f(a_i^{EDLF}|y_i) = y_i^{pred-EDLF} \tag{4}$$

We tried specific AI algorithms as our base models as follows: B= [MultiLayer Perceptron , Support Vector Machine, Random Forest Classifier, Euclidean distance, Logistic regression].

We also evaluated one of the top performing deep learning-based baseline methods from "ASVspoof 2021 [26]", (baseline 03) which is the Linear Frequency Cepstral Coefficients (LFCC)- Light Convolutional Neural Netrworks (LCNN). The purpose of this experiment is to show how EDLF-based models compare against the previous spoofed audio detection methods. Given ${a_i}^{LFCC}$ as described in the definition 3 and the y set, a convolutional neural network-based function (CNN) maps ${a_i}^{LFCC}$ to $y_i^{pred-LFCC}$ in the baseline model (BL), such that:

$$BL => CNN({a_i}^{LFCC}|y_i) = y_i^{pred-LFCC} \tag{5}$$

The LFCC-LCNN model [21] uses LFCCs features with a light convolutional neural network (LCNN). The back end of this model is the LCNN [12], but also includes Long short-term memory (LSTM) layers and average pooling. We used this model as a pre-trained model applied to the audio clips on our dataset. In the Asvspoof 2021 challenge, this baseline has a good performance on the Deepfake [5] and Logical Access [6] tasks. The output of this model is a score for each audio clip with a lower value as more likely spoofed.

D. MultiDomain Ensemble Models (MDEMs)

We also wanted to show if traditional baseline models can be improved by adding EDLFs. Thus we propose creating an ensemble model with the EDLF best performing model and LFCC-LCNN model. typically the ensemble output class is based on majority vote across individual base models.
Definition: Given a_i^{EDLF}, ${a_i}^{LFCC}$, y_i, B, BL and the predicted class labels achieved from them, such that:
If $y_i^{pred-LFCC} = spoofed \vee y_i^{pred-EDLFs} = spoofed$ then $y_i^{pred-ensemble} = spoofed$.
If $y_i^{pred-LFCC} = genuine \wedge y_i^{pred-EDLFs} = genuine$ then $y_i^{pred-ensemble} = genuine$.
When we have only two base models, the emphasis is on finding spoofed samples. Thereby, if at least one of the two models predict the case as spoofed, the final label was considered spoofed.

III. Experimental Results

In this section the experiments and associated results are described. We discuss the results with (a) EDLF base models, (b) baseline modelcomparison with LFCC-LCNN and (c) ensemble model combining the EDLF with the LFCC-LCNN model.

[5]https://competitions.codalab.org/competitions/32345
[6]https://competitions.codalab.org/competitions/32343

A. EDLF Base Models

We tested the EDLF features on several models including MultiLayer Perceptron , Support Vector Machine, Random Forest Classifier, Euclidean distance and Logistic regression. Logistic regression outperformed the others as depicted by the AUC score shown in figure 4. The accuracy (percentage of correctly classified instances) is 86.6 and 83.4 for the EDLFs train and test/unseen datasets respectively, using the logistic regression model. We use this as the best performing base model in the ensemble model.

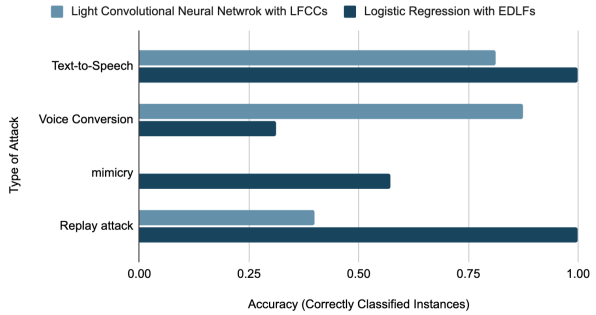

Fig. 3: Accuracy for Different Types of the Attacks for the test set

Given the distribution of attacks in our datasets we also wanted to see if EDLFs are better for some types of attacks vs others using accuracy of spoofed audio detection. As Figure 3 indicates, the accuracy for some types of spoofed audio samples are 1 (Replay Attacks and Text-to-Speech (Speech Synthesis)) indicating that EDLFs are best performing for these types of attacks. This also corroborates with the domain experts understanding since in the Voice Conversion linguistic features may be more closely preserved as part of the audio; similarly, in mimicry, in which a real human (not AI) is spoofing the audio by copying another person's voice, genuine linguistic features are also more likely to remain. The baseline model is better in detecting VC samples as the results in the Figure 3 present. Since the mimicry samples were two channel audios, and the baseline was designed for mono-channel audio processing, the baseline did not process them.

B. Linear Frequency Cepstral Coefficients (LFCC)- Light Convolutional Neural Netrworks (LCNN) Baseline

Linear Frequency Cepstral Coefficient (LFCCs) [26] is the input of the Light Convolutional Neural Networks [17] as our baseline model. The model gave us 0.698 and 0.595 as the TPR and accuracy for the training data respectively. For the test data TPR is 0.71 and accuracy is 0.6. To define a threshold based on the output score of the model, we empirically evaluated different thresholds, and checked the final metrics. Finally, the threshold = -3 is chosen. Figure 4 shows the comparison between the AUC scores obtained from the baseline and different base models on the unseen/test dataset. We can see that the LFCC-LCNN did not perform well as compared to the base models using EDLFs. We evaluate using EDLFs with LFCC-LCNN to gauge any improvements that EDLFs can provide.

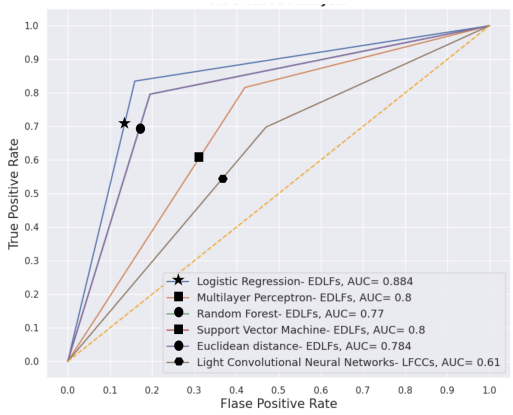

Fig. 4: The AUC score comparison between the baseline and different base models for the test set

C. MultiDomain Ensemble Models (MDEMs)

Each ensemble model included a combination of results from the base models using EDLF and baseline comparison model for LFCC-LCNN.

The results of the ensemble model on the unseen data are depicted in Figure 5.

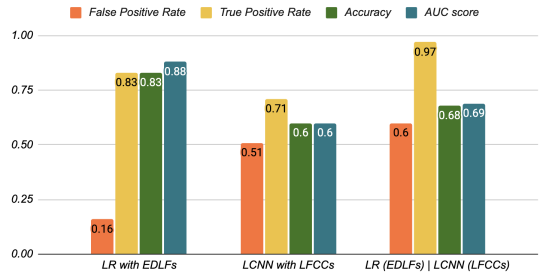

Fig. 5: The performance metrics for the base and ensemble models. LR: Logistic Rgression, EDLF: Expert Defined Linguistic Features, LFCC: Linear Frequency Cepstral Coefficient, LCNN: Light Convolutional Neural Network

As Figure 5 indicates, The EDLFs outperform the LFCC-LCNN model. However, when we combined LFCC-LCNN with EDLF in an ensemble model the performance of LFCC-LCNN also improved. This result, which is based on EDLFs, as compared to the baseline, can be considered a breakthrough in the field of spoofed audio detection, as it may work better than complex neural networks when facing small datasets.

D. Overall Findings

We summarize the overall findings of this study as follows:

- 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. These

EDLFs may help spoofed audio detection when facing lack of training data in an more effective way than complex neural networks.

- The improved performance of EDLFs indicates the value of using these 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 all other ensemble models.
- EDLFs also helped with performance improvement of the baseline in the ensemble model.

IV. Conclusion and Future Work

In this paper we have established that using features to augment the audio signals, based on sociolinguistic informed human discernment, improves the detection of audio deepfakes. In our approach experts have identified key linguistic features in our hybrid dataset that facilitated the AI algorithms' detection of audio fakes. In the future we plan to annotate these features on top of spectrograms for training of the AI algorithms automatically and on a much larger scale. Automating EDLFs extraction phase is being examined by the authors to be able to apply the method on large datasets.

The study also illustrates the importance of discernment as humans listen to audio signals. In our future work, we aim to develop tools and trainings, grounded in sociolinguistic understanding, so that the general public can learn about linguistic features and use this knowledge to more accurately identify spoofed audio.

We are also exploring causal discovery and inference among the feature sets to create the optimal feature set (causal feature selection). We are going to submit the EDLFs and the corresponding audio clips to a good venue to be available for other researchers.

V. Acknowledgment:

Authors would like to acknowledge support from NSF AWARD 2210011. The codes and audio samples will be available through one of the authors' GitHub repositories.

References

[1] Almutairi, Z., and Elgibreen, H. A review of modern audio deepfake detection methods: Challenges and future directions. *Algorithms 15*, 5 (2022), 155.

[2] 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.

[3] Brewster, T. Fraudsters cloned company director's voice in $35 million bank heist, police find. *Forbes, Editor's Pick 14* (2021).

[4] Chen, T., Kumar, A., Nagarsheth, P., Sivaraman, G., and Khoury, E. Generalization of audio deepfake detection. In *Odyssey 2020 The Speaker and Language Recognition Workshop* (2020), ISCA, pp. 132–137.

[5] Chen, Z., Xie, Z., Zhang, W., and Xu, X. ResNet and model fusion for automatic spoofing detection. In *Interspeech 2017* (2017), ISCA, pp. 102–106.

[6] Frank, J., and Schönherr, L. Wavefake: A data set to facilitate audio deepfake detection. *arXiv preprint arXiv:2111.02813* (2021).

[7] Ito, K., and Johnson, L. The lj speech dataset, 2017.

[8] Johnson, B. Deepfakes are solvable—but don't forget that "shallow-fakes" are already pervasive. *MIT Technology Review, Mar 25* (2019), 2019.

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

[10] Kim, K.-W., Park, S.-W., Lee, J., and Joe, M.-C. 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)* (2022), pp. 6997–7001.

[11] Kumar, K., Kumar, R., De Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., de Brébisson, A., Bengio, Y., and Courville, A. C. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems 32* (2019).

[12] Lavrentyeva, G., Novoselov, S., Tseren, A., Volkova, M., Gorlanov, A., and Kozlov, A. Stc antispoofing systems for the asvspoof2019 challenge. *arXiv preprint arXiv:1904.05576* (2019).

[13] Nycz, J. English phonetics. *The Handbook of English Linguistics* (2020), 323–343.

[14] Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).

[15] Reed, M., and Levis, J. M. *The handbook of English pronunciation*. John Wiley & Sons, 2019.

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

[17] Sahidullah, M., Kinnunen, T., and Hanilçi, C. A comparison of features for synthetic speech detection.

[18] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., and Wu, Y. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2018), pp. 4779–4783.

[19] Smith, B. Goldman sachs, ozy media and a $40 million conference call gone wrong. *The New York Times* (2021).

[20] Wang, R., Juefei-Xu, F., Huang, Y., Guo, Q., Xie, X., Ma, L., and Liu, Y. DeepSonar: Towards effective and robust detection of AI-synthesized fake voices. In *Proceedings of the 28th ACM International Conference on Multimedia* (2020), MM '20, Association for Computing Machinery, pp. 1207–1216.

[21] Wang, X., and Yamagishi, J. A comparative study on recent neural spoofing countermeasures for synthetic speech detection. *arXiv preprint arXiv:2103.11326* (2021).

[22] Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., Sahidullah, M., Vestman, V., Kinnunen, T., Lee, K. A., et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language 64* (2020), 101114.

[23] Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilçi, C., Sahidullah, M., and Sizov, A. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Interspeech 2015* (2015), ISCA, pp. 2037–2041.

[24] Wu, Z., Yamagishi, J., Kinnunen, T., Hanilçi, C., Sahidullah, M., Sizov, A., Evans, N., Todisco, M., and Delgado, H. Asvspoof: the automatic speaker verification spoofing and countermeasures challenge. *IEEE Journal of Selected Topics in Signal Processing 11*, 4 (2017), 588–604.

[25] Yamagishi, J., Todisco, M., Sahidullah, M., Delgado, H., Wang, X., Evans, N., Kinnunen, T., Lee, K. A., Vestman, V., and Nautsch, A. Asvspoof 2019: The 3rd automatic speaker verification spoofing and countermeasures challenge database.

[26] Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., Liu, X., Lee, K. A., Kinnunen, T., Evans, N., et al. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537* (2021).