USING SELF ATTENTION DNNS TO DISCOVER PHONEMIC FEATURES FOR AUDIO DEEP FAKE DETECTION

Hira Dhamyal, Ayesha Ali, Ihsan Ayyub Qazi, Agha Ali Raza

Lahore University of Management Sciences

ABSTRACT

With the advancement in natural-sounding speech production models, it is becoming important to develop models that can detect spoofed audios. Synthesized speech models do not explicitly account for all factors affecting speech production, such as the shape, size and structure of a speaker's vocal tract. In this paper, we hypothesize that due to practical limitations of audio corpora (including size, distribution, and balance of variables like gender, age, and accents), there exist certain phonemes that synthesized models are not able to replicate as well as the human articulation system and such phonemes differ in their spectral characteristics from bonafide speech. To discover such phonemes and quantify their effectiveness in distinguishing between spoofed and bonafide speech, we use a deep learning model with self-attention, and analyze the attention weights of the trained model. We use the ASVSpoof2019 dataset for our analysis and find that the attention mechanism picks most on fricatives: /S/,/SH/, nasals: /M/,/N/, vowels: /Y/, and stops: /D/. Furthermore, we obtain 7.54% EER on train and 11.98% on dev data when using only the top-16 most attended phonemes from input audio, better than when any other phoneme classes are used.

Index Terms— spoof, bonafide, countermeasure, attention, phonemes, deep neural network, senet, explainable, fair, small datasets, forensics, deepfake.

1. INTRODUCTION

Deepfake audios are increasingly being used to spread misinformation, and have become a threat to numerous voice-based systems like automatic speaker verification (ASV) and voice biometric systems. For example, audio deepfakes have been used to spread false political narratives which were regarded as a significant threat to the 2020 US presidential election and have been used to attack voice authentication systems in banks; on one occasion leading to a loss of USD 243,000 [1]. These examples highlight the extent of the misuse and harm that audio deepfakes can cause. There are several ways that deepfakes can be detected and in this paper we want to explore the task with respect to the human voice production system and its intricacies when producing speech sound units, which are absent in deepfake speech.

The human voice production system is a complex one; it is influenced by multiple bio-parameters of the individual speaker, which include the human vocal tract, its shape and size, thickness of the vocal folds, facial structure, state of the physical and mental health of the individual and lung capacity among others. These parameters affect the articulatory-phonetic units of speech. They vary across speakers and even within multiple utterances from the same speaker.

The various parts of the voice production system come into play when the air expelled from lungs is modulated by the mechanical vibrations of the vocal folds. The physical movements of the vocal tract change its shape and the dimensions of the various resonant chambers, introducing resonant patterns in the acoustic signal. Each distinct pattern characterized by the articulatory configuration of the vocal tract results in intelligible sound units known as phonemes. As phonemes result from a complex physical movements of the articulators and the vocal tract, they also carry speaker-specific characteristics in addition to linguistic cues [2].

Different measurements in the context of stop phonemes such as the voicing onset time (VOT) have been correlated with numerous parameters of a speaker like age [3], speaking rate [4], diseases like Parkinson's [5], depression [6]. Furthermore, the first three formant frequencies and their movements have been correlated with characteristics of the speaker such as speaker identity and speaker height [7].

While deep learning based audio synthesizers have been highly successful in generating an individual's voice, we think that the intricacies of the human voice production system may not be fully captured in such synthesis models. This incapacity is highlighted in scenarios where the data is scarce, especially when producing non-celebrity individual's voice. Synthesis models require huge amounts of data for model training and in low resource settings, constraint by the training data's distribution, do not perform well. Motivated by this, we hypothesize that there exist some phonemes or class of phonemes that the spoof generation models are not able to produce as well as others. Thus, studying the differences at the phoneme level will provide further insights into the nature of the deep learning based synthesizers or voice conversion models. Furthermore, it will also help in improving the deepfake detection models by focusing on the more distinguishing aspects of speech between bonafide and spoofed audios, by

adding explainability and interpretability into the DNN models.

In this paper, we quantify the phonetic differences between spoofed and bonafide speech by first modelling the problem as a classification task. We develop a deep learning based model for the task with self attention mechanism in place. The attention mechanism provides a way to quantify the phonemes that are highly attended to while doing the classification task. We later analyze the most attended phonemes. We find that fricatives including /S/, /SH/, nasals /M/, /N/ and vowel /Y/, stops including /D/ are among the most-attended phonemes. To test the efficacy of these selected phonemes, we compare the performance on the spoof detection task by using only parts of the input audio corresponding to specific categories of phonemes like vowels, stops, fricatives, and the top most attended phonemes. We find that when using the top 16 most attended phonemes, we obtain the best EER, lower than any other categorization, of 7.54% on the train set and 11.98% on the dev set. Our work would directly help in improving the performance of spoof detection, especially in low computational settings whereby the most distinguishing phonemes could be focused on to flag suspicious audios. Secondly, in ASV systems, the speaker could be asked to utter words including the distinguishing phonemes, making it easier to detect deceitful speakers. To the best of our knowledge, this is the first study that explores phonetic differences in spoof and bonafide speech using an explainable DNN approach.

2. BACKGROUND WORK

We discuss two different categories of related work. One focuses on the models and features that have been explored for the task of audio spoof detection. Since in this work we have used attention mechanism in order to isolate and identify the phonemes better at discerning spoofed and real speech, the second category we explore is related to studies using attention mechanism in order to add explainability into the deep learning model on speech based tasks.

Automatic detection of the audio spoof detection has been tackled using various deep learning frameworks, for example [8] uses recurrent convolutional structure for spoof detection, [9] uses resnet block architecture for the task, [10] uses time delay neural network architecture, [11] uses senet [12] model architecture for the task. Numerous papers have explored the distinguishing features between spoofed and bonafide, e.g [13] uses bispectral analysis and shows the efficacy of the features when used with simple machine learning models, [14] uses the voicing onset time and coarticulation, [15] studies the prosodic differences, [16] studies the specific frequencies which are more effective and found that lower frequencies ($<$ 1 kHz) and high frequencies ($>$ 7 khz) are the most useful, [17] uses mean pitch stability, its range and jitter as features.

Attention mechanism has been very effective at improving performance on a number of tasks (largely in NLP and speech domain), a survey of which can be found here [18]. Recently attention has been used solely as a way to add interpretability and explainability to deep neural models, which allows us to examine the internal workings of largely black box neural architectures. Chan et al., [19] showed that in speech recognition, acoustically similar units of speech have similar attention weights and the alignment between character and audio signal are identified correctly by the attention weights. Palaskar et al., [20] analyses attention in speech recognition task and find that highest attention weights are for the word end boundaries. Dhamyal et al., [21] uses attention weights to analyse the phonetic differences between acted and spontaneous emotional speech and found significant differences between phonemes in the two classes.

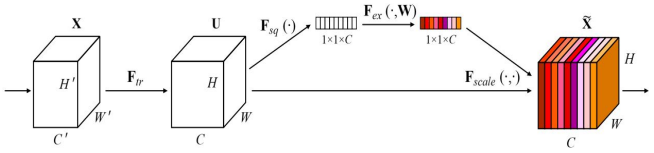

Fig. 1. Squeeze and excitation block used in the deep neural model

Specifically for audio spoof detection, [22] has focused on adding explainability by using GradCAM Binary maps which finds patches of the audio input spectrogram that the model most focuses on. However no further study has been done to find a pattern in those patches, i.e. whether the patches are focused on certain frequencies, or on specific sounds etc.

In contrast to earlier work on audio spoof detection, we analyze the vocal expression of phonemes in the spoofed and bonafide speech in order to identify and quantify the importance of phonemes that the spoof generation models are not able to synthesize well enough. Identifying such phonemes could help in better detection of spoofed speech and in text-dependent speaker verification/identification system where the text could include those specific phonemes, making it harder for adversaries to fool the system.

3. DEEP NEURAL MODEL

3.1. SENET

We use senet model architecture; specifically senet-50 which uses the squeeze and excitation and residual layers. Figure 1 shows a depiction of the squeeze and excitation layer that are used in the model. Such layers have been shown to outperform other DNN architectures for the audio deepfake detection task [23, 11]. To be able to capture which parts of the input the model most focuses on to predict the output label, we use self attention mechanism. Details of this are in the next section.

3.2. Self-Attention mechanism

In the senet architecture, given a input signal of length T, the model produces a sequence of T output vectors of dimension d, let this output be $\mathbf{H} = \{h_1, h_2, \ldots, h_T\}$, where h_t is the hidden representation of the input at time t. The self attention mechanism takes this hidden representation $\mathbf{H}$ as input and outputs the attention matrix A as follows:

$$A = \text{softmax}(g(\mathbf{H}^T W_1)W_2)$$

where W_1 is of dimension $d \times a$ and W_2 is of dimension $a \times r$. r represents the number of attention heads; 32 in our case and $g(.)$ is any activation function, we use ReLu. Softmax function is performed over the time dimension. As a result, A is of dimension $T \times r$, where the vector at time t is a weight vector that represents the weight of h_t. To get the weighted output $\mathbf{E}$, we do: $\mathbf{E} = \mathbf{H}A$, where $\mathbf{E}$ is of dimension $d \times r$. We do average pooling on $\mathbf{E}$ and then pass this through a linear layer to get the class probabilities for spoof and bonafide.

4. EXPERIMENT

4.1. Dataset

To analyze the performance of our countermeasure model, we use the ASVSpoof2019 Logical Access (LA) Dataset [24]. The synthetic audio samples in this dataset are created using a set of 6 different text to speech (TTS) systems and voice conversation (VC) algorithms, consisting of neural waveform model, vocoder model, waveform concatenation and spectral filtering, detailed in Table 1.

Since the dataset does not contain transcripts for the audio, we transcribe it using IBM Watson Speech-to-text API [25] and then use an HMM-based phoneme segmentor [26] to get word and phone boundaries. Figure 2 shows the frequency of all the phoneme in the training and the dev subset of the data. The total number of phonemes in the inventory are 40 (including silence).

4.2. Features

We use the Constant Q cepstral coefficient (CQCC) features based on the constant-Q transform. It provides a variable-resolution, time-frequency representation of the spectrum. Many deep learning countermeasure models are based on the CQCC features. We use a matlab implementation of CQCC to compute the features [23], where CQT is applied with maximum frequency of fmax $=$ fs$/2$ (where fs is the sampling rate of the given audio) and minimum frequency of fmin $=$ fmax$/2^9$ where 9 is the number of octaves. The number of bins per octave is set to 96. These parameters result in a time shift or hop of 8ms. Delta and delta-delta features are also appended. These parameters are empirically optimised for the audio spoof detection task [23].

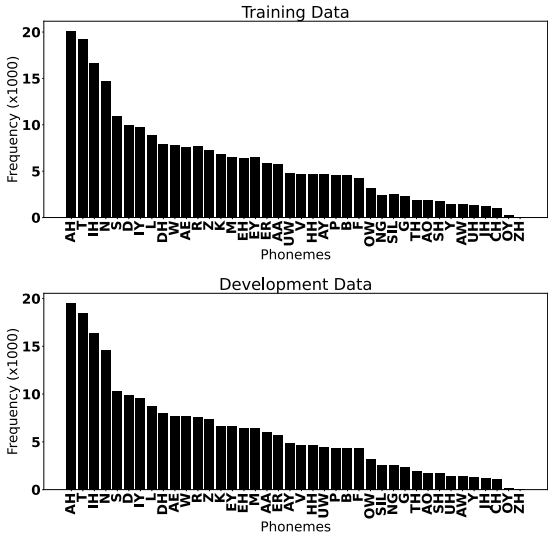

Fig. 2. Total number of phonemes in the train (on the top) and development data (on the bottom).

4.3. Training

Since we are using multi-head attention, where r (the number of attention heads) $= 32$, we need to encourage the attention heads to learn different aspects of the input audio. In order to do this, in addition to the cross entropy loss L_{ce}, we add a loss term L_p into the final loss function, where $L_p = \|(A^T A - I)\|_F^2$, where A is the attention matrix, I is the identity matrix and F is the Frobenius norm. This is similar to the loss terms used in [28]. The final loss term becomes

$$L = \lambda(L_{ce}) + (1 - \lambda)L_p$$

where λ is a hyper-parameter.

We train several models with the same senet-50 architecture and with different values of λ to extract the attention weights and analyze the similarities between each pair of attention weights. We train each model for 50 epochs, while using batch SGD with a learning rate of 0.001, decreasing at epochs 10 and 20 with decay rate of 0.0001, and use a batch size of 16.

5. RESULTS AND ANALYSIS

5.1. Spoofed speech detection model performance

To evaluate the counter measure model, we use Equal Error Rate (EER) and t-DCF cost. t-DCF evaluates the performance of a spoofed audio detection model in conjunction with a given ASV method. The ASV scores are provided as part of the ASVSpoof2019 challenge.

Table 2 shows the results of three different deep neural models with the same architecture as explained in Section 4.3, and trained in similar manner except for the λ used in the loss

Table 1. Spoof audio generation systems in ASVSpoof2019 LA train and dev subsets. [27]

System	Description
A01	NN based TTS system using VAE-LSTM as the acoustic model and Wavenet vocoder for waveform generator
A02	NN based TTS system using VAE-LSTM as the acoustic model and WORLD vocoder for waveform generator
A03	Feedforward NN as the acoustic model and WORLD vocoder for waveform generator
A04	A waveform concatenation TTS
A05	NN VC system using VAE as VC model and WORLD vocoder for waveform generator
A06	Transfer function based VC system
Total Train	No. of speakers: 20, bonafide: 2580, spoof: 22800
Total Dev	No. of speakers: 20, bonafide: 2548, spoof: 22296

function. The purpose for training several different models is to extract the attention weights for each of the audio files in the training and dev subsets and compare them. We hope that the attention weights are reproducible and no matter what loss functions are used for the deep neural network model, the interpretation that we assign to the attention weights is valid and reliable. The results achieved are comparable to the numerous studies on the data. Among the three models, we obtain better results when λ is decreased and the L_p loss term is given more weightage. However, when the weight of L_p increases more than 0.5, the result worsens.

Table 2. Several models are trained with the same architecture and same training procedure as described in Section 4.3. The λ for the loss function is different in each training.

Model	Train EER(%)	Dev EER(%)/t-DCF	Eval EER(%)/t-DCF
A(λ=1)	1.82	5.38 / 0.17	20.99 / 0.59
B(λ=0.9)	1.79	4.12 / 0.13	14.10 / 0.34
C(λ=0.6)	1.51	3.65 / 0.11	14.03 / 0.37

5.2. Attention weight analysis

Figure 3 shows examples of attention weights for individual files from two different models, and the highlighted peaks. The weights that are greater than the mean plus one standard deviation of the weight vector are chosen as the peaks, which represent the time steps where the attention is higher than the rest. To analyse the differences in the weights from each of

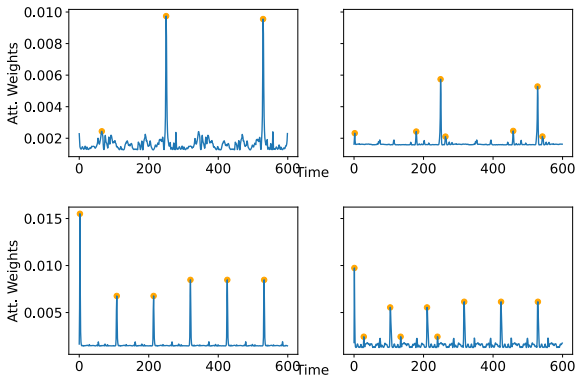

Fig. 3. Visualization of the attention weights obtained for two audios (rows) from two different models (columns). The similarities between the weights obtained from two different models can be observed from this example.

trained model, we measure the cosine similarity between all pairs of the weight vectors from the three trained models for all audios. Figure 5 shows a histogram of the cosine similarity scores. The mean cosine similarity for the train set is 0.74 and that of the dev set is 0.73. This indicates that the attention weights from multiple models are similar. Furthermore, by treating the weight vector as a signal, we perform dynamic time warping (DTW) Sakoe-Chiba's algorithm on each pair of the weight vector from the three models and calculate the DTW with an allowance of a window, to allow the weights to differ up to some specific number of frames. The idea is to allow the weights to be shifted a little bit given each phoneme spans more than one frame. Figure 6 shows the average DTW distance for all the audio file weights for three models with different window sizes. The differences in the distance are too small to be noticed in the figure, however as expected the higher the window size, the lower the distance.

5.3. Most Attended Phonemes

To identify the most attended phonemes, we only consider the peaks that are consistent in all the three models. Figure 4 shows the attended phonemes in train and dev set of the data. It can be observed that in both subsets of the data, fricatives like /S/, /SH/, nasals like /M/, /N/, stops lie /D/ and vowels like /Y/, /IH/, are predominant in the top most attended phonemes.

5.4. Phonemes subset

In order to access the phoneme subsets that are chosen by the self-attention model, we evaluate the trained model A (see Table 2 above) on the train and dev subsets by only using certain phoneme sounds in the input audio. For example, given a vowel /AA/ and an input audio, we only select parts of the audio where /AA/ is spoken, and patch them together (Note:

Table 3. Performance of model A (see table 2 above) on train and dev subsets when using only specific phoneme classes from the input audios.

Phoneme class	vowel			fricative			stop					nasal	voiced	unvoiced	top-16
	high	mid	low	voiced	unvoiced	sibilant	voiced	unvoiced	bilabial	velar	alveolar				
train EER (%)	17.76	**16.46**	23.14	22.45	**19.06**	19.13	24.96	**20.80**	27.21	26.54	21.50				
		9.36			15.32				17.44			23.07	18.46	**13.50**	**7.54**
dev EER (%)	21.84	**20.35**	27.39	27.81	24.20	**23.93**	31.07	**26.63**	32.00	30.57	28.22				
		12.15			20.27				20.95			28.31	24.71	**19.01**	**11.98**

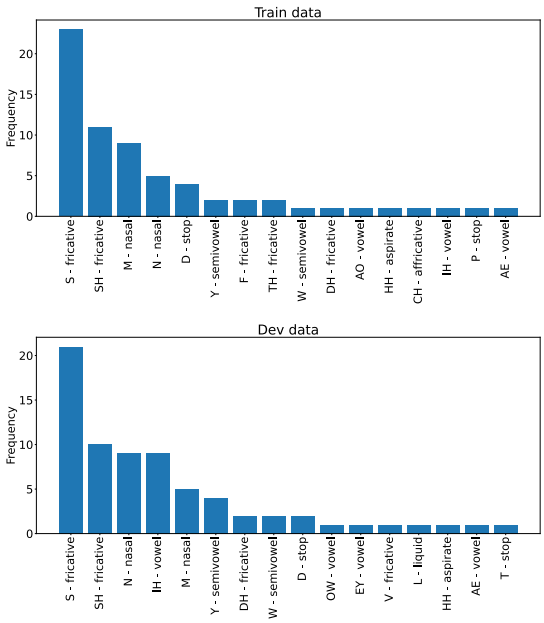

Fig. 4. Frequency of the attended phonemes in the train (up) and dev (bottom) subsets of the data.

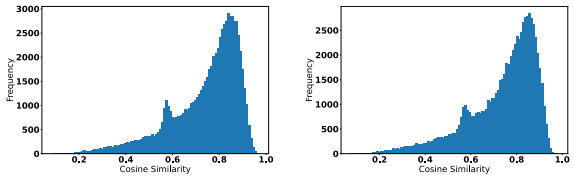

Fig. 5. Histogram of cosine similarity scores between weights obtained from all the models.

for all the analyses, we also include the silence regions of the audio). We do this for all the audios and pass them through the trained model. Table 3 shows the EER obtained when doing this over multiple subsets of the phonemes. *top-16* class includes the top 16 phonemes that are picked on by the self-attention model as the most attented phonemes, namely, 'S', 'SH', 'M', 'N', 'IH', 'Y', 'DH', 'W', 'D', 'OW', 'EY', 'V', 'L', 'HH', 'AE', 'T'; thus this class is a mixture of fricatives, nasals, stops and vowels.

From table 3, we observe that vowels perform best among

Fig. 6. Average DTW distance using different window sizes for each pair of audio file weights obtained from 3 models in the train set.

all other identified phoneme groupings achieving EER = 9.36% on train, 12.15% on dev. Among vowels, the vowel with height mid performs better than vowels with height high and low. Among fricatives, unvoiced fricatives perform better and also among stops, unvoiced stops perform better. Furthermore we also consider only unvoiced and voiced consonants, including fricatives and stops. Unvoiced consonants include 'K', 'P', 'T', 'F', 'TH', 'S', 'SH', 'HH' and voiced include 'B', 'D', 'G', 'V', 'DH', 'Z', 'ZH'. We find that unvoiced consonants perform better than voiced and better than most other classes, third to top-16 and vowels.

However when we use the top-16 attended phonemes exclusively for the classification task, we find that the attended combination of vowels, fricatives, and stops performs better than any single class of phonemes and gives the lowest EER, 7.54% on train, 11.9% on dev.

Finally, we also analyse the top attended phonemes for each individual TTS/VC systems. We find that the top-16 most attended phonemes in each system individually has 61% overlap on average with the top-16 phonemes reported earlier. The most frequent phonemes common across all systems include /S/,/IH/,/DH/,/EY/,/AE/,/T/,/N/, which are all present in the overall top-16 class as well.

6. LIMITATIONS AND FUTURE WORK

Even though attention has been used popularly for adding explainability into the neural models, some studies [29, 30] argue against the usefulness of exploring attention weights to add explainability to the task at hand. Being mindful of this,

we compare the attentions across multiple trained model in order to access the consistency across weights. We find that the attention weights obtained over multiple runs are very similar and thus gives us confidence in our results. Secondly we have used pronunciation dictionary to represent each word into a series of phones, called phonemic transcription, which is a particular form of broad transcription which disregards all allophonic differences. This creates a difference in the expected phoneme occurrence in the audio sample and the actual phonemes present in the pronunciation. This in turn may affect our analysis, however with the amount of data used we believe such discrepancies to be rare.

In future, we hope to extend a similar analysis to languages other than English, especially low resource languages. We would also like to explore the extend to which light weight ML models can perform when using only certain phonemes, in order to enhance our ability in low computational settings.

7. CONCLUSION

In this paper, we analysed the phonetic differences between spoofed and bonafide speech using ASVSpoof2019 corpus which contains spoofed speech from 6 different synthesizers. We employed a deep neural network based classifier composed of senet and residual layers, with self attention mechanism. The self attention mechanism plays a vital role by attending to the parts of input audio that are most useful for the classification of the audio into spoof or bonafide class. We later analysed these attention weights from three different trained models.

From our results, we conclude that fricatives like /S/, /SH/ and nasals, vowels like /Y/ and stops like /D/ are the distinguishing phonemes in spoofed and bonafide speech. When the 6 synthesizers used in the dataset are analyzed individually, we find that the top-16 class phonemes are consistently present in all the results. This validates our original hypothesis that machines are not able to model the workings of voice production very well. Furthermore, we also only use parts of the input audio which correspond to specific classes of phonemes. We find that when using the top-16 most attended phonemes, we get the lowest EER on both train and dev subsets, being 7.54% on train and 11.9% on the dev data. Next to this is the performance of vowels, where EER is 9.36% on the train set and 12.15% on the dev set. Identifying the phoneme groupings which are more distinguishing than others for audio spoof detection helps in better spoof detection models where the focus can be specifically on such phonemes and also helps in better protection of ASV systems against spoof attacks. It is also beneficial in building spoof detection models for low data and computational resource settings, especially for low resource languages or when building models for specific individuals where training data is harder to find or deploying models to devices where computational capacity is low. In such cases, we hope that simpler models are built to work off of the most distinguishing phonemes and obtain satisfactory performance.

8. REFERENCES

[1] Kim Hartmann and Keir Giles, "The next generation of cyber-enabled information warfare," in *2020 12th International Conference on Cyber Conflict (CyCon)*. IEEE, 2020, vol. 1300, pp. 233–250.

[2] Rita Singh, Bhiksha Raj, and Deniz Gencaga, "Forensic anthropometry from voice: an articulatory-phonetic approach," in *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2016, pp. 1375–1380.

[3] Richard J Morris and WS Brown Jr, "Age-related differences in speech variability among women," *Journal of Communication Disorders*, vol. 27, no. 1, pp. 49–64, 1994.

[4] Katrin Stölten, Niclas Abrahamsson, and Kenneth Hyltenstam, "Effects of age and speaking rate on voice onset time," *Studies in Second Language Acquisition*, vol. 37, no. 1, pp. 71, 2015.

[5] Emily Fischer and Alexander M Goberman, "Voice onset time in parkinson disease," *Journal of Communication Disorders*, vol. 43, no. 1, pp. 21–34, 2010.

[6] Alastair J Flint, Sandra E Black, Irene Campbell-Taylor, Gillian F Gailey, and Carey Levinton, "Acoustic analysis in the differentiation of parkinson's disease and major depression," *Journal of Psycholinguistic Research*, vol. 21, no. 5, pp. 383–399, 1992.

[7] Reinhold Greisbach, "Estimation of speaker height from formant frequencies," *International Journal of Speech Language and the Law*, vol. 6, no. 2, pp. 265–277, 2007.

[8] Akash Chintha, Bao Thai, Saniat Javid Sohrawardi, Kartavya Bhatt, Andrea Hickerson, Matthew Wright, and Raymond Ptucha, "Recurrent convolutional structures for audio spoof and video deepfake detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 1024–1037, 2020.

[9] Tianxiang Chen, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman, and Elie Khoury, "Generalization of audio deepfake detection," in *Proceedings of the Odyssey Speaker and Language Recognition Workshop, Tokyo, Japan*, 2020, pp. 1–5.

[10] Mari Ganesh Kumar, Suvidha Rupesh Kumar, MS Saranya, B Bharathi, and Hema A Murthy,

"Spoof detection using time-delay shallow neural network and feature switching," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019, pp. 1011–1017.

[11] Cheng-I Lai, Nanxin Chen, Jesús Villalba, and Najim Dehak, "Assert: Anti-spoofing with squeeze-excitation and residual networks," *Proc. Interspeech 2019*, pp. 1013–1017, 2019.

[12] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[13] Ehab A AlBadawy, Siwei Lyu, and Hany Farid, "Detecting ai-synthesized speech using bispectral analysis.," in *CVPR Workshops*, 2019, pp. 104–109.

[14] Hira Dhamyal, Ayesha Ali, Ihsan Ayyub Qazi, and Agha Ali Raza, "Fake Audio Detection in Resource-constrained Settings using Microfeatures," in *Proc. Interspeech 2021*, 2021.

[15] Yang Gao, Jiachen Lian, Bhiksha Raj, and Rita Singh, "Detection and evaluation of human and machine generated speech in spoofing attacks on automatic speaker verification systems," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 544–551.

[16] Dipjyoti Paul, Monisankha Pal, and Goutam Saha, "Spectral features for synthetic speech detection," *IEEE journal of selected topics in signal processing*, vol. 11, no. 4, pp. 605–617, 2017.

[17] Phillip L De Leon, Bryan Stewart, and Junichi Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[18] Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath, "An attentive survey of attention models," *arXiv preprint arXiv:1904.02874*, 2019.

[19] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[20] Shruti Palaskar and Florian Metze, "Acoustic-to-word recognition with sequence-to-sequence models," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 397–404.

[21] Hira Dhamyal, Shahan Ali Memon, Bhiksha Raj, and Rita Singh, "The Phonetic Bases of Vocal Expressed Emotion: Natural versus Acted," in *Proc. Interspeech 2020*, 2020, pp. 3451–3455.

[22] Bence Mark Halpern123, Finnian Kelly, Rob van Son12, and Anil Alexander, "Residual networks for resisting noise: analysis of an embeddings-based spoofing countermeasure," .

[23] Massimiliano Todisco, Héctor Delgado, and Nicholas WD Evans, "A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients.," in *Odyssey*, 2016, vol. 2016, pp. 283–290.

[24] "Asvspoof 2019: the automatic speaker verification spoofing and countermeasures challenge evaluation plan.," http://www.asvspoof.org/asvspoof2019/asvspoof2019 evaluation plan.pdf, 2019, [Online].

[25] IBM, "Watson speech to text - overview," `https://www.ibm.com/cloud/watson-speech-to-text/`, [Online; accessed 1-July-2021].

[26] Paul Lamere, Philip Kwok, Evandro Gouvea, Bhiksha Raj, Rita Singh, William Walker, Manfred Warmuth, and Peter Wolf, "The cmu sphinx-4 speech recognition system," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong*, 2003, vol. 1, pp. 2–5.

[27] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al., "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," *Computer Speech & Language*, vol. 64, pp. 101114, 2020.

[28] Yingke Zhu, Tom Ko, David Snyder, Brian Mak, and Daniel Povey, "Self-attentive speaker embeddings for text-independent speaker verification.," in *Interspeech*, 2018, vol. 2018, pp. 3573–3577.

[29] Sarthak Jain and Byron C Wallace, "Attention is not explanation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 3543–3556.

[30] Sofia Serrano and Noah A Smith, "Is attention interpretable?," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2931–2951.