Evince the artefacts of Spoof Speech by blending Vocal Tract and Voice Source Features

Tadipatri Uday Kiran Reddy , Sahukari Chaitanya Varun, Kota Pranav Kumar
Sreekanth Sankala, Kodukula Sri Rama Murty
Indian Institute of Technology Hyderabad, Telangana, India
{*ee19btech11038, ee19bech11040, ee19btech11051, ee20resch11011, ksrm*} *@iith.ac.in*

***Abstract*—With the rapid advancement in synthetic speech generation technologies, great interest in differentiating them from the natural speech is emerging. These advancements have produced results that can deceive some of the state-of-the-art spoof detection models. To prevent potential adverse effects, it becomes crucial to detect the spoof signals and predict the algorithm which generated them, which needs an understanding of the underlying attributes of spoof signals.**

In this paper, we propose a system that detects spoof signals and identifies the corresponding speech generation algorithm. We achieve 98% both in algorithm classification accuracy and spoof detection accuracy. The study emphasizes the parts of speech signals critical in identifying their authenticity by utilizing the Vocal Tract System(*VTS*) and Voice Source(*VS*) features. From experiments, we found that a VS feature-based system gives more attention to the transition of phonemes, indicating limitations of existing state-of-the-art TTS systems. In contrast, a VTS feature-based system gives more attention to stationary segments of speech signals. We proposed a couple of model fusion techniques to utilize the complementary information provided by these features to enhance classification performance. To validate the proposal, we analyzed the t-SNE plots for developed models which verified that feature fusion resulted in better clustering of the embeddings.

***Index Terms*—Synthetic Speech Attribution, X-vector, Bicoherence, LP Residual, Coarticulation**

I. Introduction

In the recent years, there has been a significant improvement in the performance of synthetic speech production models. This improvement has enabled the generation of synthetic signals, which can even fool some state-of-the-art spoof detection models. Failing to detect spoof signals has adverse consequences. For example, speech is used for user identification in many modern systems, and the ability to synthesise fake speech poses a huge security threat [1].

There are diverse methods of generating spoof speech [2], ranging from simple cut-and-paste methods to advanced Deep Neural Networks (DNNs) [3]–[6]. Utilizing high quality microphones and advanced TTS and VC systems makes problem of detecting spoof speech signals a complicated task.

There have been many studies which investigate the usage of different features to be used for differentiating spoof signals from natural signals. Some studies [7], [8] suggest that features consisting high frequency information have pronounced affects on spoof speech. Authors of [9] show that features conveying information of high frequency regions and detailed spectral characteristics are very useful for spoof detection. We utilise Mel Spectrograms and LP residuals, in order to exploit both magnitude and phase information of spoof speech.

Authors of [10] investigate attention given to phonemes in speech signals and shows that certain phonemes which are the highest attended, help make better predictions about spoofing when used for classification. We show similar explainability of our DNN models by analysing attention given to various graphemes.

We propose a spoof speech classifier using DNNs, with VTS or *Vocal Tract System* (Mel Spectrogram) and VS or *Voice Source* (Linear Prediction Residuals) features to identify spoof signals and classify them to their generators. Looking at LP Residuals is rather a seldom approach in contrast to the features used in the literature. After examining the proposed system's attention to the speech signals, we discovered that VS based-system emphasises phoneme transitions which is illustrated in V-C . At phoneme transitions, it is evident that source excitations change more significantly than vocal tract responses. This analysis helps understand the synthetic speech generator models better and make them more efficient by working on critical phoneme transitions, also referred to as co-articulations [11]. In contrast, VTS based-system emphasises the stationary regions of a speech signal, where the vocal tract response changes are significant. While studying these aspects, we also investigated the benefit of utilising the complementary information provided by the VTS and VS features for the attribution of speech samples [12].

The contributions of this paper are as follows,

- We propose DNN based spoof classifiers using VTS and VS features. Various fusion techniques were implemented to enhance the performance.
- We analysed the artefacts identified by the DNNs for classification

The paper is organised as follows. In Section II, we discuss about feature extraction methodologies employed in this work. In Section III, usage of higher-order correlations in this study is motivated. In Section IV, the proposed model architectures are described. In Section V, the results obtained with further study on the network behaviour are illustrated. Finally, In Section VI, we conclude the paper with future scope of our work.

II. Feature extraction

In this section, we briefly describe one of the speech production model known as Source-System model. Later we discuss

the features which could potentially capture the artefacts in spoof speech which are Log Mel Spectrogram capturing *VTS* features and Linear Predictive Residuals capturing *VS* features.

A. Speech Production Model: Source-System Model

Speech production is a very sophisticated task. The motor activities are both fast and accurate. The speech production mechanism can be briefly explained as follows: the lungs act as a source of energy moderating the air flow, while the vocal folds in the trachea operate the airflow from the lungs into quasi periodic puffs of excitation. The shape of the vocal tract determines the sound that is produced.

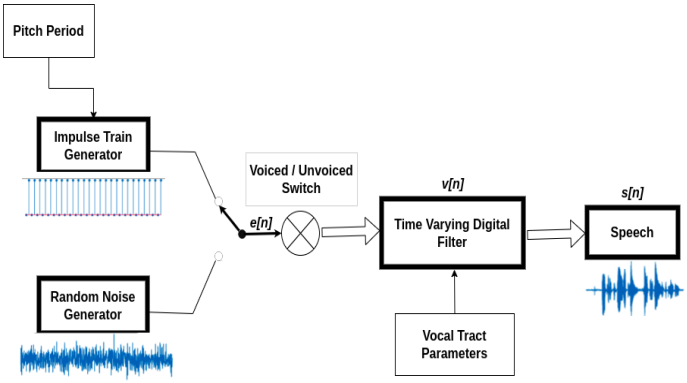


Fig. 1: Speech Production Model

In the literature many researches have proposed models [13], [14], [14]–[17] to describe the speech production system and model the non-linearities. However, in this study we restrict ourselves to Source-System Model as shown in Figure 1. This model assumes linearity and that the source and system work independently to produce speech. Here, the speech signal obtained is an outcome of a time-varying vocal-tract system driven by time-varying excitation [18], which can be modelled as 1.

$$s[n] = v[n] * e[n] \tag{1}$$

Let $s[n]$, $v[n]$ and $e[n]$ be the speech signal, impulse response of the vocal tract filter and excitation signal. The speech signal is the output of time varying digital filter ($v[n]$) with time varying excitation ($e[n]$).

Synthetic speech algorithms seek to mimic this natural system, where they may fail to incorporate subtle features of speech production such as

- *Co-articulation of phoneme sequences*,
  Phoneme transitions in synthetic speeches tend to have disfluencies because of abrupt phoneme transisitons in synthesis algorithms. [19] provides a good survey about phonetic variations in synthetic speeches.
- *Prosodic variations such as jitter, shimmer*,
  Natural speech have more variations in the excitation source making it highly unpredictable. In contrast, the excitation source of sythetic speech are more periodic and less noisy, as a result having less shimmer and jitter in the impulse train, as pointed out in [20].

B. Vocal Tract System Features : Log-Mel filter bank energies

The Vocal Tract System is a concatenation of acoustic tubes. The resonances of the vocal tract tubes manifest as formant frequencies in the spectorgram of the speech signal. Hence capturing the formant frequencies of the speech signal can convey the information about the vocal tract system being used in the speech production process. Many works [21]–[25] utilize the(*MFCC* (mel-frequency cepstral coefficients) and *LPCC* (linear prediction cepstral coefficients) features to capture the magnitude information.

Log Mel filter bank energies is one of the popular way to extract the information about the vocal tract system. These features are extracted from speech signals segmented into chunks of 25ms, which usually corresponds to 3-5 pitch periods. Short Time Fourier Transform (STFT) is then performed over these chunks, then followed by mel-scale warping and finally applying logarithm over the obtained features. These magnitude consists upto 2nd order statistics conveying about vocal tracts in speech production model.

C. Voice Source Features : Linear Prediction Residuals

The voice source features capture the vocal cords characteristics in speech production. The phase spectral information is complementary to information from VTS features [12], [26] for the fact that we are limiting to 2nd order statistics. Suppressing the magnitude envelope results in a relatively flat spectrum that characterizes the phase information. Linear Prediction (LP) analysis [27]–[29] is one such methodology where the magnitude spectral envelope is repressed by first finding the magnitude spectral envelope and then operated with an inverse formulation which results in residual rich in higher order correlations, with phase content. This feature is called the LP residual, which captures the voice source features. Analytically, the VTS response is modelled as a linear prediction filter.

$$s[n] = \sum_{k=1}^{K} a_i s[n-k] + Ge[n] \tag{2}$$

$$S(z) = V(z)E(z); \quad V(z) = \frac{G}{A(z)} = \frac{G}{1 - \sum_{k=1}^{K} a_k z^{-k}} \tag{3}$$

$$E(z) = V(z)^{-1} S(z) \tag{4}$$

LP Residual is obtained as result of inverse filtering the speech signal with estimated VTS response $H(z)$. For our study, we perform all pole modeling of LP Residual. The order is chosen such that we have one resonance for each kHz ($p = 4 + f_s/1000$) [30]. We ensure that the order is odd, to constraint the LP analysis with atleast one real pole [31].

According to Source-System model, excitation source of voiced sounds is impulse train in an ideal setting. However, natural signals are very off from the ideal scenario and excitation source does not have consistent period (jitter) nor constant amplitude (shimmer) making it less predictable as shown in Figure 3. On the other hand, LP Residuals of spoof signals are periodic and have relatively constant period making

them more predictable and this is consistent with all the spoof algorithms. This is a potential discriminative feature which can be exploited to classify the spoof signals.

III. Rationale for looking into Higher order correlations

Capturing the magnitude spectral features (2nd-order statistics) for speaker verification [32] tasks has been one of the standard approaches. But with evolution of technology, the synthetic speech production techniques have been able to fairly mimic these features indistinguishable from a normal speech signal. Therefore, using 2nd-order features is incomprehensible and as a result one should explore higher order statistical features to capture the discriminatory behaviour underlying within these synthetic speech generating algorithms. One such feature explored in the literature was *Bispectrum* [33] of a signal. This feature is a third-order statistic (or skewness), using this discrimination among different algorithms has enhanced as illustrated in Figure 2, we use *bicoherence* (Equation 5) which is normalised bispecturm making phase content more pronounced. We clearly see clear distinction among different algorithms, while GAN based TTS has cross white strips and are absent for the rest. However, there also exists high within the class variance.

$$\mathrm{B}_{coher}(\omega_1,\omega_2) = \frac{1}{W}\frac{\sum_{w=0}^{W-1} X(\omega_1)X(\omega_2)X^*(\omega_1+\omega_2)}{\sqrt{\sum_{w=0}^{W-1}|X(\omega_1)X(\omega_2)|^2\sum_{w=0}^{W-1}|X^*(\omega_1+\omega_2)|^2}} \quad (5)$$

This bolsters the utility of higher order features in the discrimination of spoof speech signals. Moreover, we can realize on an abstract level that the bispectrum captures the phase spectral information [34]. This motivates for the exploration of VS features as a discriminative evidence, recall VS features are obtained after suppressing VTS response. And one such VS feature exploited in this work is LP Residuals. By looking at the LP residuals of the synthesized algorithms, fine characterization in behaviours distinctive to certain synthesizer can be observed in Figure 3. The residuals of synthetic algorithms appear more periodic than that of the natural utterance. The jitter and shimmer in the natural residual have a lot more variation compared to the rest.

And as this information is not effectively captured by the VTS features, we can extract more exhaustive information from a speech sample by fusing both the features. There have been some established works [35]–[39] which stated that the source and the system features, complementary within themselves, could be utilized together for analysis of speech signals aiding in the task at hand.

In the further sections we propose a DNN based model to exploit VS and VTS features to classify them into synthetic generators. We fused the systems built on individual features using different paradigms and show a performance leap providing another evidence that VS and VTS are indeed complementary in nature.

IV. Model Architectures

We proposed two primary models, LP Residual based (*LPR-DNN*) and Log Mel Spectrogram based (*LMS-DNN*) , which are designed using DNN inspired by X-Vectors system [40]

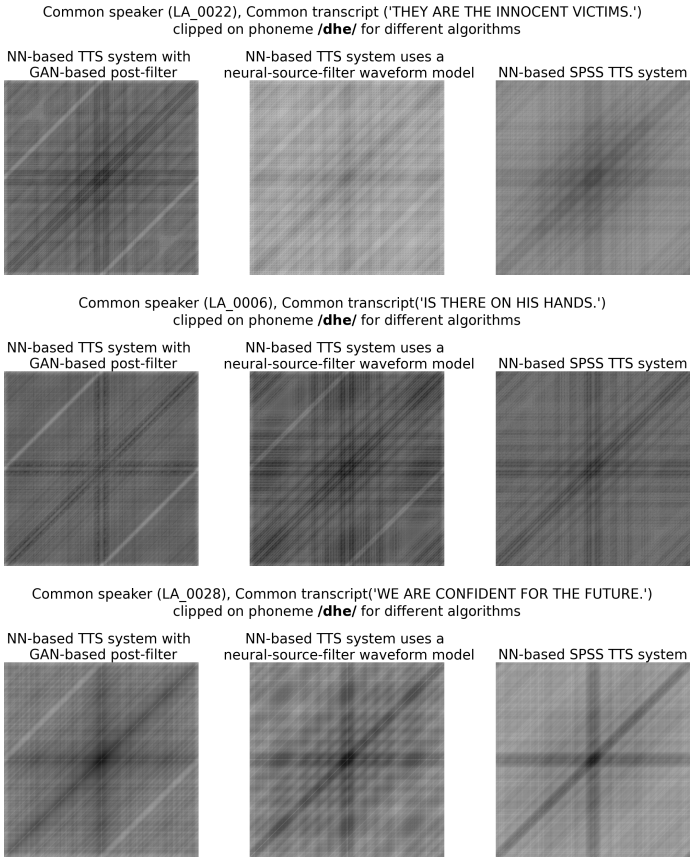


Fig. 2: Log-applied bicoherence magnitude plots for 3 different transcripts highlighting common phoneme for different speakers and different synthetic speech producing algorithms. These plots are displayed on a common intensity scale of [10,50].

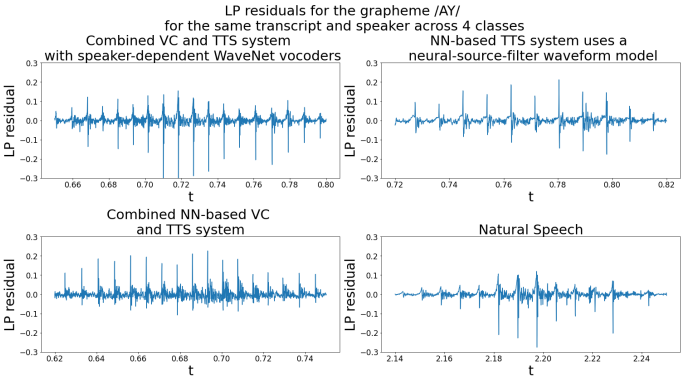
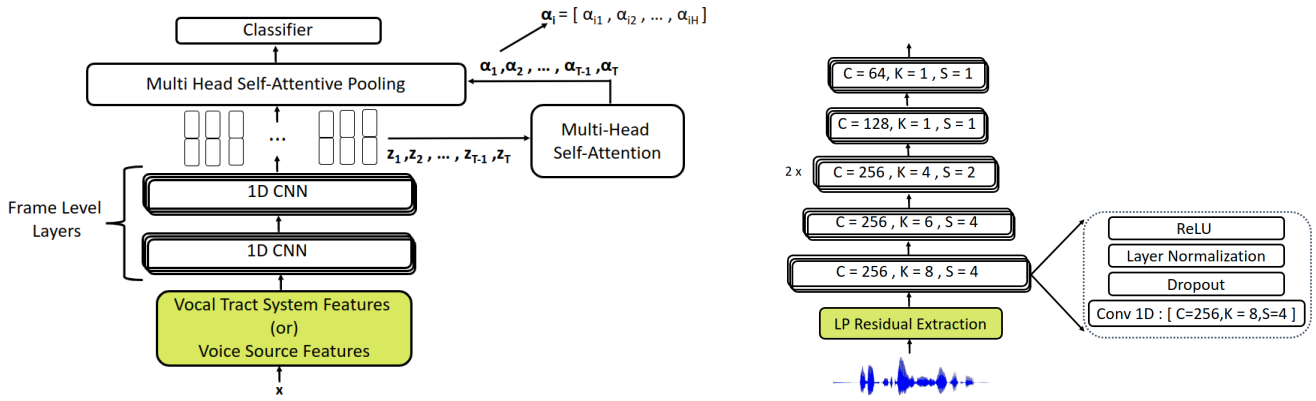


Fig. 3: LP residuals of grapheme $/AY/$ from different algorithms, for the sentence 'IS IT IN THE RIGHT PLACE?' of same speaker

with few additions like self-attention module [41], and custom feature extractor consisting of trainable filter banks as shown in Figure 4b. Furthermore, to exploit the complementary information present in VS and VTS features we fuse our systems in two ways, first at posterior scores and at architecture level. The complete diagram of proposed model architecture is shown in Figure 4a.



(a) Generic architecture (b) Feature extractor for LP Residuals

Fig. 4: Proposed Architecture

A. LPR-DNN: VS feature based DNN

Speech signal with length of N with Mth-order LP Coding we obtain (N-M-1) samples of LP Residuals, which is very sparse in time. Therefore, we extract low dimensional features using multiple filters and non-linear activation functions and another important motivation for us to introduce a feature extractor for LP Residuals is that, LP Residuals are obtained by inverse filtering the speech signal with an envelope, so called VTS Response, therefore the resultant spectrum of LP Residuals is relatively flat and traditional feature extraction methods will most likely be inefficient. Usage of LP Residuals for the task of Spoof detection/classification is studied very less in the literature, as a result, we propose a DNN based feature extractor consisting of 1d convolutional filter banks with non-linear activation functions as shown in Figure 4b. The receptive field [42] of the proposed feature extractor is 165 ($\approx$ 10ms), and we obtain 64 x T dimensional output, where T = (N-M-1)/165.

Later Frame Level Encoding is performed on these features using X-Vectors system, after obtain compact representation (128 x Time steps) we perform multi-headed self-attention [41] to mask out unimportant segments of extracted embeddings. Finally, we to obtain a fixed dimensional vector we apply statistical pooling [43] across time-axis producing a 1024-dimensional vector.

B. LMS-DNN: VTS feature based DNN

To exploit the vocal tract features, we build X-Vector network using Log-Mel Spectrogram features. Here, we use only single headed self attention module, therefore obtaining 256-diemensional vector.

C. Model Fusion

Fusing systems using different paradigms in speech community [44], [45] are very popular and are known to show a performance leap given complementary information available. As discussed earlier, VS and VTS features have complementary information, which we would like to exploit in this study by performing two different kinds of fusion.

1) LPR+LMS-DNN: Score level fusion:* One of such elementary fusion techniques are known as score level fusion, where we perform weighted average on logits obtained from different systems. We perform the same on using LPR-DNN and LMS-DNN with equal weightage.

2) LPR+LMS-DNN: Feature level fusion: Another sophisticated fusion paradigm is fusing information at architecture level, where we typically concatenate features in common latent space of different systems. As shown in Figure 5a, we concatenate features obtained after frame level encoding and then perform multi-headed self-attention, statistical pooling and logistic regression.

D. Motivation for Hierarchical Classifier

As witnessed in Figure 3, we observe a huge difference between spoof signals and natural signals, thus it is rational to assume that DNN can estimate the parameters faster and more accurately. Therefore, we employed two different classifier in our DNN architecture as shown in Figure 5b.

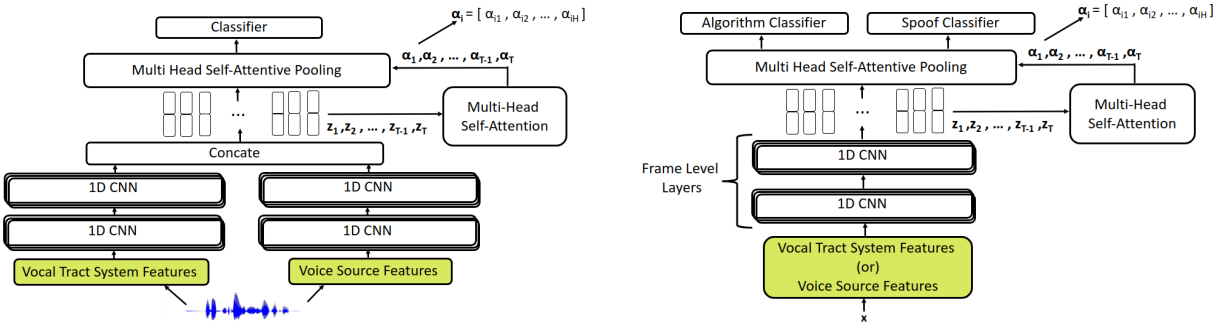
1) Vanilla Classifier (*VC*):
This is conventional logistic regression layer ($LR_1(.)$) to predict logits with k+1 classes with k is number of spoof algorithms and one natural class. The loss function is described in equation 6.

$$\mathcal{L}^{VC}_{(\boldsymbol{X},\boldsymbol{y})}(\boldsymbol{\theta}) = H(\boldsymbol{y}, LR_1\left(f\left(\boldsymbol{\theta}, \boldsymbol{X}\right)\right)) \tag{6}$$

Where *H(.)* is Cross entropy function, *f(.)* is function which returns embeddings from the DNN and $\boldsymbol{\theta}$ are parameters to DNNs.

2) Hierarchical Classifier (*HC*):
As the names suggests, we perform hierarchical classification with two levels, first we perform binary classification ($LR_2(.)$) to predict whether the signal is spoof or not. Later we use the VC to predict the spoof algorithm. The loss function is described in equation 6.

$$\mathcal{L}^{HC}_{(\boldsymbol{X},\boldsymbol{y})}(\boldsymbol{\theta}) = H(\boldsymbol{y}, LR_2\left(f\left(\boldsymbol{\theta}, \boldsymbol{X}\right)\right)) + \mathcal{L}^{VC}_{(\boldsymbol{X},\boldsymbol{y})}(\boldsymbol{\theta}) \tag{7}$$



(a) Block Diagram of Feature Level Fusion

(b) Block Diagram of Classifiers

Fig. 5: Architecture of LPR-LMS-DNN

Table I: Dataset composition (* represents common speakers)

Type	#Speakers	Proportion
Training	30+10*	33.25%
Validation	17+10*	16.11%
Evaluation	50	50.64%

V. Results and Discussion

In this section, performance evaluation and ablation analysis of the proposed architectures are studied. We train our models LPR+DNN, LMS+DNN, LPR+LMS-DNN, and LPR+LMS-DNN* and discuss about the obtained performances. Finally, we infer about the highlighted segments of spoof speech signal which are picked up by self-attention module in the model. Subsequently, we deduce the attributes of spoof speech signal which are exploited by the proposed model. A brief implementation details used for the studies are discussed below.

A. Data sets and Experimental details

The hardware specifications used for conducting the studies are Intel Xeon Silver 4114 CPU @ 2.20GHz, four NVIDIA GeForce RTX 2080 Ti GPUs. The software packages utilized are *PyTorch* library for training the neural networks and *scikit* for computing LP Residuals. This link [1] consists of source codes for replicating all the results reported in this paper.

For conducting the studies we utilize *Logical Access* portion of the data from the *ASVSpoof 2019* data corpus [46]. Custom data splits[2]were made since the work here is primarily concerned about synthetic speech classification but not speaker verification. We have created two versions of custom data splits; one has uniform algorithms among training, validation, and evaluation sets with the proportions of 40%, 10%, and 50% respectively, this set is referred as *CS1*; while the other has disjoint set of speakers in training and evaluation datasets, this set is referred as *CS2* The composition of custom data splits are shown in Table I.

We use speech signals synthesized at sampling frequency of 16KHz. The 23rd order of LP analysis [II-C] is done to compute LP residuals. Log Mel Spectrogram is computed with 80 mels, 512 bin FFT, window length of 400 samples (25ms), and hop length of 160 (10ms).

[1]https://github.com/TUdayKiranReddy/SPCUP2022

Table II: Results of various models (Algorithm Accuracy %/Spoof Accuracy %/Spoof EER %)

Classifier type	Backbone	CS1 Evaluation	CS2 Evaluation
VC	LPR-DNN	99/100/0.19	96/98/1.73
	LMS-DNN	96/99/1.47	96/99/1.31
	LPR+LMS-DNN*	**100/100/0.25**	99/99/0.7
	LMS-LPR-DNN	99/100/0.3	**98/99/0.1**
HC	LPR-DNN	97/99/0.52	96/98/2.368
	LMS-DNN	96/89/1.48	94/92/1.32
	LPR+LMS-DNN*	**99/99/0.19**	**98/99/0.853**
	LPR+LMS-DNN	99/99/0.636	98/98/0.704
HC with silence removal	LPR-DNN	95/92/1.73	94/95/0.75
	LMS-DNN	87/74/3.5	88/75/2.89
	LPR+LMS-DNN*	88/88/1.0	93/94/0.326
	LPR+LMS-DNN	**98/98/0.63**	**94/99/1.2**

We report both Classification accuracy (higher the better) and Equal Error Rate (*EER*) (lower the better) to evaluate the performance of our models.

B. Comparisons among variants of proposed models

The proposed models are hyper-tuned to achieve the best performance, and Table II consists of the performances achieved, to avoid over-fitting of DNNs due to distinct silent regions present in spoof algorithms we also experimented with silence removal using [47]. The below are the four major points which we would like to accentuate.

1) VS features are a cut above than VTS features: As we eye-balled the discriminatory features of VS, this is evident from the performance on LPR-DNN, which is atleast 3% superior than LMS-DNN. We can observe that EER also follows the same trend. This depicts that VS features captures the artefacts in the speech signal more prominently.

2) Performance leap with fusion techniques: As anticipated, fusion techniques showed better performance than LPR-DNN and LMS-DNN, which is a popular paradigm in DNN community to achieve better results. This performance leap is observed in all the studies performed such as, with/without silence removal, hierarchical classification and inference with

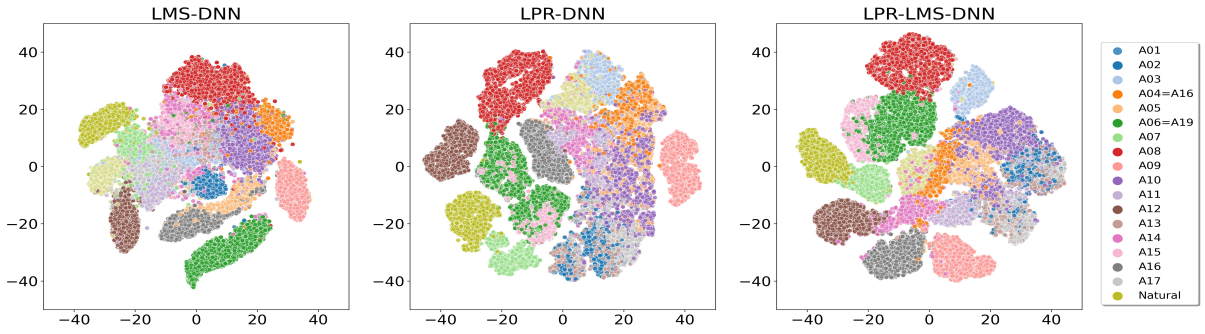


Fig. 6: t-SNE embeddings for processed vector from X-vectors

Table III: Effective receptive field and stride of proposed models

Feature used	Receptive Field	Effective Stride
LP Residual	2160 (135 ms)	64 (4 ms)
Mel Spectrogram	4960 (310 ms)	12 (0.75 ms)

distinct speaker. We plotted t-SNE [48] embeddings onto 2D space to evaluate the separability which are shown in Figure **??**, and it is observed that clusters in LMS-DNN and LPR-DNN embeddings are merged into others, whereas after performing feature fusion the cluster are unalloyed comparatively. This is yet another evidence for the theory that VS and VTS features consist of complementary information.

3) Over-fitting of VTS features: On the *CS1* data set, the performance dip when silence regions are removed in LMS-DNN is significant, implying that LMS-DNN is over-fitting the data by exploiting the amount of silence region present in synthetic speeches. In contrast, LPR-DNN shows robustness to amount of silence regions.

4) Robustness with different speakers: Speaker over-fitting is generic issue in speech recognition/verification, to test this we have created *CS2* dataset and train on it appropriately. The results obtained have very slight variations compared to *CS1*, implying that proposed DNNs are robust to new speakers and can identify the underlying algorithm characteristics.

C. Inference from the highlighted segments

In this section, we analyze the self-attentive modules and visualize the portion of the speech signal highlighted. We use models LPR-DNN and LMS-DNN and we retrained LPR-DNN with a single channel and single head self-attentive module, for the sake of simpler analysis and to have common ground. We pass a speech signal with an utterance of "*THERE WAS A SUBSTANTIAL EXPLOSION*" to LP residual and Mel Spectrogram-based Models. We look at the attention values, which is a vector of size T; where T is the number of time steps, which is decided by the effective receptive field and stride of the Time Delayed Neural Network(*TDNN*) as mentioned in III. For the visualisation, the attention maps are binarised, with their mean over time as the threshold.

From Figure 7, we find that LPR-DNN gives more attention to the transition of phonemes. In literature, it is known as

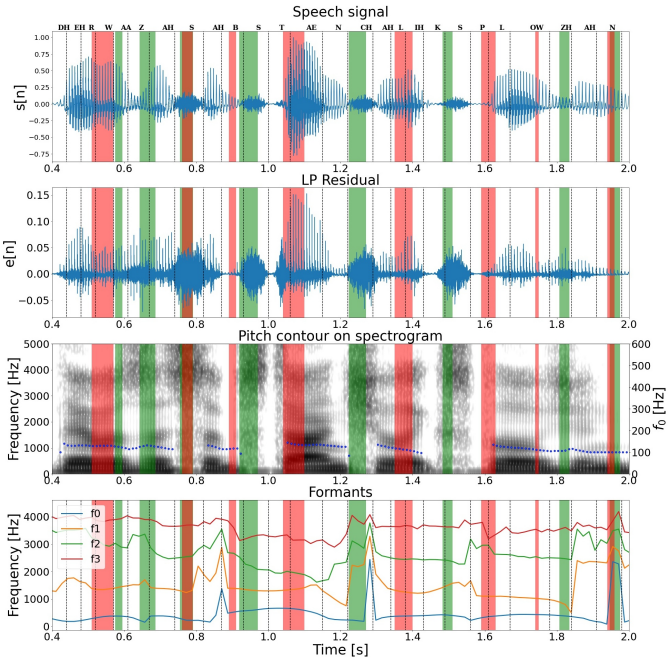


Fig. 7: Attention maps on a synthetic speech signal with an utterance of "*THERE WAS A SUBSTANTIAL EXPLOSION.*", NOTE:- GREEN bands indicate attention by LPR-DNN whereas RED represents LMS-DNN

coarticulation [49]. This is evident from formants plots where the green bands(LPR-DNN) are highlighted at the portions where formants are changed significantly, indicating change in phoneme. In contrast, LMS-DNN focuses on stationary parts of phonemes where the formants are relatively constant.

Typically, when there is a change from voiced-to-unvoiced sound or vice-versa, the source signal changes from random noise to impulse trains. This change is captured by LPR-DNN and it is different for different algorithms, therefore providing stronger discrimination compared to LMS-DNN.

VI. Conclusion

We propose a spoof speech algorithm classifier using DNN by using voice source (LP Residuals) and vocal tract (Mel Spectrograms) features, by combining evidences from these

two features we achieved a performance leap. Further investigation into the attention maps of DNN over the spoof speech, we observed that LPR-DNN is focusing on phoneme transitions and in contrast LMS-DNN is focused more on stationary portion of phonemes. We conclude that artefacts in spoof speech are deductive using higher order correlations and the research in spoof speech generation algorithms have to consider on improving natural phonetic transitions. This study could be utilized for enhancing the synthetic generation techniques by defending the defects identified. In particular, an exhaustive study on phonemes could be performed to identify which phoneme utterance could be more critical in spoof detection.

VII. Acknowledgement

We sincerely thank SIP Lab, IIT Hyderabad, for providing the computation power needed for the studies and the IEEE SP Cup 2022 organisers for the problem statement and dataset.

References

[1] E. Chandra and C. Sunitha, "A review on speech and speaker authentication system using voice signal feature selection and extraction," in *2009 IEEE International Advance Computing Conference*, pp. 1341–1346, 2009.

[2] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Classifiers for synthetic speech detection: A comparison," 09 2015.

[3] M. Schröder, M. Charfuelan, S. Pammi, and I. Steiner, "Open source voice creation toolkit for the mary tts platform," pp. 3253–3256, 08 2011.

[4] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *Arxiv*, 2016.

[5] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," 2021.

[6] Y. Tabet and M. Boughazi, "Speech synthesis techniques. a survey," in *International Workshop on Systems, Signal Processing and their Applications, WOSSPA*, pp. 67–70, 2011.

[7] J. Yang and R. K. Das, "Long-term high frequency features for synthetic speech detection," *Digital Signal Processing*, vol. 97, p. 102622, 2020.

[8] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," 09 2015.

[9] M. Sahidullah, T. Kinnunen, and C. Hanilçi, "A comparison of features for synthetic speech detection," 09 2015.

[10] H. Dhamyal, A. Ali, I. A. Qazi, and A. A. Raza, "Using self attention dnns to discover phonemic features for audio deep fake detection," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1178–1184, 2021.

[11] P. Menzerath and A. de Carvalho Lacerda, "Koartikulation, steuerung und lautabgrenzung : eine experimentelle untersuchung," 1933.

[12] K. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," vol. 13, pp. 52–55, 2006.

[13] A. Lacroix, "Speech production-physics, models and prospective applications," in *ISPA 2001. Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis. In conjunction with 23rd International Conference on Information Technology Interfaces (IEEE Cat.*, pp. 3–, 2001.

[14] L. Mitiche and A. B. H. Adamou-Mitiche, "Second order speech model based on ga's," in *2016 4th International Conference on Control Engineering Information Technology (CEIT)*, pp. 1–4, 2016.

[15] W. Zhu and H. Kasuya, "A new speech synthesis system based on the arx speech production model," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, vol. 3, pp. 1413–1416 vol.3, 1996.

[16] K. Iso, "Speech recognition using dynamical model of speech production," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 283–286 vol.2, 1993.

[17] C.-W. Kim, "Models of speech production," in *Formal Aspects of Cognitive Processes* (T. Storer and D. Winter, eds.), (Berlin, Heidelberg), pp. 142–158, Springer Berlin Heidelberg, 1975.

[18] R. Singh, B. Raj, and D. Gencaga, "Forensic anthropometry from voice: An articulatory-phonetic approach," in *2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pp. 1375–1380, 2016.

[19] R. Qader, G. Lecorvé, D. Lolive, and P. Sébillot, "Phonology modelling for expressive speech synthesis: a review," 07 2014.

[20] A. J. Rozsypal and B. F. Millar, "Perception of jitter and shimmer in synthetic vowels," *Journal of Phonetics*, vol. 7, no. 4, pp. 343–355, 1979.

[21] A. Meghanani, A. C. S., and A. G. Ramakrishnan, "An exploration of log-mel spectrogram and mfcc features for alzheimer's dementia recognition from spontaneous speech," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 670–677, 2021.

[22] F. Beritelli and R. Grasso, "A pattern recognition system for environmental sound classification based on mfccs and neural networks," in *2008 2nd International Conference on Signal Processing and Communication Systems*, pp. 1–4, 2008.

[23] W. J. Poser, "Douglas o'shaughnessy, speech communication: Human and machine. reading, massachusetts: Addison-wesley publishing company, 1987. pp. xviii 568. isbn 0-201-16520-1.," vol. 20, p. 52–54, Cambridge University Press, 1990.

[24] C. V. Cotton and D. P. W. Ellis, "Spectral vs. spectro-temporal features for acoustic event detection," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 69–72, 2011.

[25] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, and T. Sorsa, "Computational auditory scene recognition," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. II–1941–II–1944, 2002.

[26] R. Gonzalez, *Better Than MFCC Audio Classification Features*, pp. 291–301. 10 2013.

[27] G. Pop and D. Burileanu, "Towards detection of synthetic utterances in romanian language speech forensics," in *2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 80–84, 2021.

[28] K. S. R. Murty, V. Boominathan, and K. Vijayan, "Allpass modeling of lp residual for speaker recognition," in *2012 International Conference on Signal Processing and Communications (SPCOM)*, pp. 1–5, 2012.

[29] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[30] L. R. Rabiner, B. Gold, and C. K. Yuen, "Theory and application of digital signal processing," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 8, no. 2, pp. 146–146, 1978.

[31] T. Quatieri, *Discrete-time Speech Signal Processing: Principles and Practice*. Pearson Education, 2002.

[32] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 4, pp. 1–22, 2004.

[33] A. K. Singh and P. Singh, "Detection of ai-synthesized speech using cepstral bispectral statistics," in *2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 412–417, 2021.

[34] C. K. Kovach, H. Oya, and H. Kawasaki, "The bispectrum and its relationship to phase-amplitude coupling," *NeuroImage*, vol. 173, pp. 518–539, 2018.

[35]

[36] M. Faundez-Zanuy and D. Rodriguez-Porcheron, "Speaker recognition using residual signal of linear and nonlinear prediction models," 01 1998.

[37] W.-C. Hsu and J.-N. Sun, "A study of the usefulness of linear prediction residual for speaker recognition," *Advanced Science Letters*, vol. 9, pp. 754–761, 04 2012.

[38] W.-C. Hsu, W.-H. Lai, and W.-P. Hong, "Usefulness of residual-based features in speaker verification and their combination way with linear prediction coefficients," in *Ninth IEEE International Symposium on Multimedia Workshops (ISMW 2007)*, pp. 246–251, 2007.

[39] S. Mahadeva Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, no. 10, pp. 1243–1261, 2006.

[40] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5329–5333, 2018.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[42] A. Araujo, W. Norris, and J. Sim, "Computing receptive fields of convolutional neural networks," *Distill*, 2019. https://distill.pub/2019/computing-receptive-fields.
[43] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," in *Interspeech 2018*, ISCA, sep 2018.
[44] N. Chauhan, T. Isshiki, and D. Li, "Speaker recognition using fusion of features with feedforward artificial neural network and support vector machine," in *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, pp. 170–176, 2020.
[45] S. Toshniwal, A. Kannan, C.-C. Chiu, Y. Wu, T. N. Sainath, and K. Livescu, "A comparison of techniques for language model integration in encoder-decoder speech recognition," 2018.
[46] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-T. Hwang, Y. Tsao, H.-M. Wang, S. L. Maguer, M. Becker, F. Henderson, R. Clark, Y. Zhang, Q. Wang, Y. Jia, K. Onuma, K. Mushika, T. Kaneda, Y. Jiang, L.-J. Liu, Y.-C. Wu, W.-C. Huang, T. Toda, K. Tanaka, H. Kameoka, I. Steiner, D. Matrouf, J.-F. Bonastre, A. Govender, S. Ronanki, J.-X. Zhang, and Z.-H. Ling, "Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech," 2019.
[47] Y.-Y. Yang, M. Hira, Z. Ni, A. Chourdia, A. Astafurov, C. Chen, C.-F. Yeh, C. Puhrsch, D. Pollack, D. Genzel, D. Greenberg, E. Z. Yang, J. Lian, J. Mahadeokar, J. Hwang, J. Chen, P. Goldsborough, P. Roy, S. Narenthiran, S. Watanabe, S. Chintala, V. Quenneville-Bélair, and Y. Shi, "Torchaudio: Building blocks for audio and speech processing," *arXiv preprint arXiv:2110.15018*, 2021.
[48] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.
[49] P. Menzerath and A. de Carvalho Lacerda, "Koartikulation, steuerung und lautabgrenzung : eine experimentelle untersuchung," 1933.