DR-SASV: A DEEP AND RELIABLE SPOOF AWARE SPEECH VERIFICATION SYSTEM

Amay Gada, Neel Kothari, Ruhina Karani, Chetashri Badane, Dhruv Gada, Tanish Patwa*

Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering, Mumbai
India

* Corresponding Author, e-mail: ruhina.karani@djsce.ac.in

Abstract: A spoof-aware speaker verification system is an integrated system that is capable of jointly identifying impostor speakers as well as spoofing attacks from target speakers. This type of system largely helps in protecting sensitive data, mitigating fraud, and reducing theft. Research has recently enhanced the effectiveness of countermeasure systems and automatic speaker verification systems separately to produce low Equal Error Rates (EER) for each system. However, work exploring a combination of both is still scarce. This paper proposes an end-to-end solution to address spoof-aware automatic speaker verification (ASV) by introducing a Deep Reliable Spoof-Aware-Speaker-Verification (DR-SASV) system. The proposed system allows the target audio to pass through a "spoof aware" speaker verification model sequentially after applying a convolutional neural network (CNN)-based spoof detection model. The suggested system produces encouraging results after being trained on the ASVSpoof 2019 LA dataset. The spoof detection model gives a validation accuracy of 96%, while the transformer-based speech verification model authenticates users with an error rate of 13.74%. The system surpasses other state-of-the-art models and produces an EER score of 10.32%.

Key words: automatic speaker verification, transformers, spoof detection, spoof-aware-speaker-verification.

1. INTRODUCTION

Voice-based authentication systems, also known as automatic speaker verification (ASV) systems, have become popular and effective substitutes for the present security techniques as a result of recent technological advancements. As there is no direct touch with the machine, unlike other systems, these do not provide discomfort or health dangers to the user.

While spoofing attacks attempt to use various technologies to alter the verification findings, leading to a severe performance decrease, ASV systems aim to verify the authenticity of target speakers. In a recent survey conducted by a leading Computer Security Company, McAfee, it was found that in India alone, 47% of the adults experienced voice related scams (which is twice the global average), of which 83% resulted in some kind of financial loss [1]. As technology reaches its zenith, the security of voice-based recognition and verification systems plummets, and it becomes imperative to prevent malicious attacks and maintain the system's integrity. In addition to target and impostor trials or utterances, reliability must be maintained in the presence of spoofed utterances, which are changed, synthesized, or specially produced utterances intended to mislead or manipulate the ASV system. This paper focuses on tackling spoofing attacks and strengthening existing ASV systems.

The Spoofing-Aware Speaker Verification (SASV) challenge 2022 aimed to encourage spoof aware ASV systems and motivated our proposal of a novel Deep Reliable - Spoof aware Automatic Speech Verification (DR-SASV) system.

ASV systems now in use are vulnerable to assaults like text-to-speech (TTS) and voice conversion (VC). To defend ASV systems from spoofing attacks, countermeasures (CM) systems that can differentiate between real, human speech and computer-generated speech are required. When integrated with existing ASV systems, performance can be significantly improved when dealing with imposters who use fake audio.

Conventional models such as gaussian mixture models (GMM) [2] and i-vector systems based on deep neural networks (DNNs) [3] have performed quite well in the past for spoof detection. The models were tested on the speech corpora: - National Institute of Standards and Technology (NIST) speaker recognition evaluation (SRE), VoxCelebs, and Speakers in the Wild (SITW). The Probabilistic Linear Discriminant Analysis (PLDA) model is used as a back-end model to deal with channel mismatches between enrolment speakers and evaluation speeches. Residual neural networks (ResNet) [4] with metric learning loss [5, 6] or discriminative training loss of softmax [7, 8] are popular alternatives to Time delay neural network (TDNN) [9] for extracting speaker embeddings.

Up until now, the majority of relevant research has solely examined the structure and effectiveness of stand-alone countermeasures. Nevertheless, there are several reasons why it is crucial to have spoofing detection and speaker identification techniques, like for the protection of sensitive data and integrity of a software, prevention of spreading of malware, and unauthorized network access. This paper proposes a two-part system that caters to the need for an integrated system for speech verification with support for spoof detection. We create the integrated spoof detection and speech verification system using state-of-the-art transformer models and deep convolutional neural networks. We go one step ahead in training the automatic speech verification model to be "spoof aware" and hence increase the reliability of the system.

2. RELATED WORK

Spoof detection and speaker verification have been a widely explored topic in the research community owing to its importance in the field of security. Massimiliano Todisco et al. [10] present a Gaussian back-end fusion approach to system combination

(ASV and Presentation attack detection (PAD)). Hye-jin Shim et al. [11] propose a novel E2E framework that jointly optimizes speaker Identification (SID), PAD, and the independent speaker verification (ISV) task. Jiakang Li et al. [12] focus on the joint text independent ASV and anti-spoofing system in a deep learning framework. Jiakang Li et al. [13] investigate the joint decision of anti-spoofing and ASV in a multi-task learning framework with contrastive loss. These works are directed towards and are limited to replay attacks.

Elaborate research has also been conducted in the field of speaker identification and verification [14, 15]. Joon Son Chung et al. [16] suggest an angular variation of the conventional networks that performs better than all current training functions. The goal of this study is 'open-set' speaker recognition of unseen speakers. Chao Li et al. [15] propose a deep residual CNN (ResCNN) inspired by residual networks (ResNets). Here, however, error cases are not fully explored; the model size is expensive and requires heavy CPU requirements and long training times.

Apart from this, a lot of work also exists where attempts have been made to incorporate anti-spoofing as well as speaker verification when creating a speech-based authentication system. Hemlata Tak et al. [16] utilize a bank of extremely basic classifiers, each with a front-end tuned to recognize various spoofing assaults and merge them at the score level via non-linear fusion. Zhongwei Teng et al. [17] propose a fully trainable end-to-end speaker-aware SASV system that integrates the total encoder from a light raw waveform encoder and a pre-trained ASV system. The method does not make use of an ensembler; moreover, the use of a larger dataset could have a considerable impact on the accuracy. Yuanjun Zhao et al. [18] propose an integrated spoofing-robust ASV system with an accommodation for both logical and physical condition attacks. The method proposed is jointly tuned to provide an adequate representation based on the integrated data from anti-spoofing and speaker verification tasks by utilizing multi-task learning. The suggested Speaker Recognition Automatic Speaker Verification (SR-ASV) system using Constant Q-transform (CQT) characteristics produces the best t-DCF results achieved by a single system on both partitions. Poor system performance is observed in high-quality replay attacks. Some potent spoofing methods based on generative models and neural networks still needs improvement. Haibin Wu et al. [19] propose a two level framework that first utilizes embeddings from CM models, propagating CM embeddings into a CM block to obtain a CM score. Petr Grinberg et al. [20] propose the usage of another Decision Tree ensemble method called boosting.

David Snyder et al. [21] focus at using embeddings taken from a feedforward deep neural network in place of i-vectors for text-independent speaker verification.

Many studies have focused primarily on replay and such individual attacks. There is a need for more comprehensive fusion strategies that consider a wider array of potential attacks and scenarios, such as voice conversion or deepfake attacks, which pose unique challenges and require innovative countermeasures, especially in the context of evolving spoofing techniques. Moreover, the proposed methodology offers the first of its kind two-factor authentication system, essential to tackle spoofed audios and speaker verification tasks.

This paper primarily aims to fulfil the following objectives:

- To propose a deep, reliable SASV system
- offer two-factor authentication whilst ensuring light-weightiness and computational viability.

The rest of the paper is structured as follows: Section 3 describes the proposed model followed by section 4 which showcases the results of the methodology. Section 5 talks about the future work and concluded the paper.

3. PROPOSED MODEL

Our work closely relates to creating an anti-spoof–speech verification and authentication system which incorporates replay, synthesized, and converted spoofing attacks. We incorporate state-of-the-art transformers and experiment with creating a Deep “spoof aware” ASV system: Deep Reliable - Spoof aware Automatic Speech Verification (DR-SASV) system, which essentially catches the spoofed audios that the spoof detection system misses detecting as a true positive. This, in a way, allows “2 factor authentication”. The usage of the lightweight MobileNetV2 further increases usability in mobile devices. The usage of the ResNet18 and MobileNetV2 models is a testament to the deepness of the proposed methodology.

While it is possible for humans to differentiate between lower ranges of frequencies of sound, the same does not apply to higher frequencies. Thus, we utilize the MEL Scale in our methodology as it best mimics a human's perception of sound by utilizing MEL Spectrograms. The conversion of frequency (in hertz) to MELscale is done using the given formula depicted in Eq. (1).

$$m = 1127.\ln(1 + f/100) \tag{1}$$

A spectrogram is a graphic representation that shows the frequency spectrum of an audio recording over time. A MEL spectrogram allows us to represent audio data in a more descriptive manner, in turn allowing us to use Deep learning techniques for audio classification. The MEL Spectrograms are leveraged in our speech embedder (CNN). The parameters used to encode the audio files into Mel Spectrogram are as follows: {Sr: None, n_fft: 512, hop_length: 128, n_mels: 165, fmin: 20, fmax: 8300, top_db: 80.}. Values are chosen after experimentation and finding the perfect balance between training efficiency, model generalization and GPU constraints.

The spectrogram of the target audio goes through a spoof detection model, and if it is a true negative, then only the audio pass through the speech verification model. During training, embeddings of three bonafide audios are fed to the transformer network in a bid to mimic a modern speech-based authentication system like Siri. The target audio is compared with a weighted (trainable) average of the 3 bonafide embeddings via cosine similarity. The block diagram of the system is given in the Figure 1.

Section 3.1 describes the dataset used which is initially converted to MEL spectrograms using the above-mentioned method, then used for training the proposed system and the ResNet18 model used for the spoof detection module. Section 3.2 describes the speech verification module and the inclusion of the MobileNetV2 model and a transformer in it.

3.1. Spoof Detection Model

The dataset used in this section is borrowed from the 2019 ASV Spoof Challenge [22]. Logical and physical access control are two alternative use case situations that are represented by the ASVspoof 2019 database. It is based on the Voice Cloning Toolkit (VCTK) corpus [23], a multi-speaker English speech database recorded at a sampling rate of 96 kHz in a hemi-anechoic chamber. It was produced using 107 speaker utterances, downsampled to 16 kHz at 16 bits per sample, from 46 male and 61 female speakers. Three speaker-disjoint sets made up of 107 speakers are divided between training, development, and assessment. There are 20 training speakers, 10 non-target and 10 target speakers, and 48 target and 19 non-target speakers in the training, development, and assessment sets, respectively. Target speakers employed in voice conversion or TTS adaption do not overlap among the three categories. Each record has reference to the audio file it corresponds to, the speakerID of one of the speakers, whether the sample is bonafide or spoofed, and what kind of spoof method was used to generate the spoofed sample. Additionally, the section mentions the preprocessing procedures for the spoof detection model which is followed by the section 3.2 describing the speech verification model proposed.

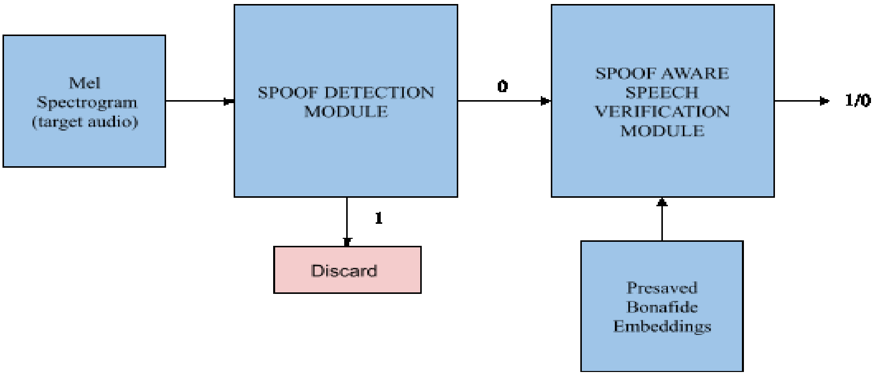

Mel Spectrogram (target audio)
SPOOF DETECTION MODULE
0
SPOOF AWARE SPEECH VERIFICATION MODULE
1/0
1
Discard
Presaved Bonafide Embeddings

Figure 1. System overview for the DR-SASV model

Baseline: The baseline contains an equal amount of “spoof (1)” and “bonafide (0)” audio samples giving a baseline accuracy of 50%. The essence of taking a baseline is to contextualize the results of trained models. The dataset that is mentioned above is used in this section of the overall model.

Preprocessing: First, the Audio to Spectrogram transform is created from a predefined config. that is the AudioConfig class. Then the silence is removed, and signals are resized to 2000 and mask the frequency and time to lessen background noise and enhance the clarity and quality of communication in noisy environments, with the number of masks equal to 2 in each case. The Silence Removal function reads the audio file, turns it into frames, and uses the Sliding Window Technique to check Voice Activity Detection (VAD) for each pair of frames. Voice-containing frames are gathered in a different list, and silence-containing frames are eliminated. The frames in this list are integrated to form an "Audio file". Our data is arranged in Data Blocks, which acts as a blueprint, later converting this into data loaders having a batch size of 256.

Training: The ResNet18 model is employed with batch size =256. Again, while larger batch sizes can accelerate training, they may also risk overfitting if not handled correctly. We chose a batch size of 256 after experimentation and validation to strike a balance between training efficiency, model generalization and GPU constraints.

The residual block consists of two 3x3 convolutional layers with the same total number of output channels. This ResNet18 model "knows" what spectrograms look like because it has already been trained by evaluating around 1.5 million images (ImageNet), including diverse images that are spectrograms. LabelSmoothingCrossEntropy is used as a loss function, shown by Eq. (2) to reduce overfitting and overconfidence. The system overview is given in Figure 2.

$$v_{ls} = (1 - a) \,.\, v_{hot} + a/k \tag{2}$$

where: k = number of label classes; a = smoothing hyperparameter

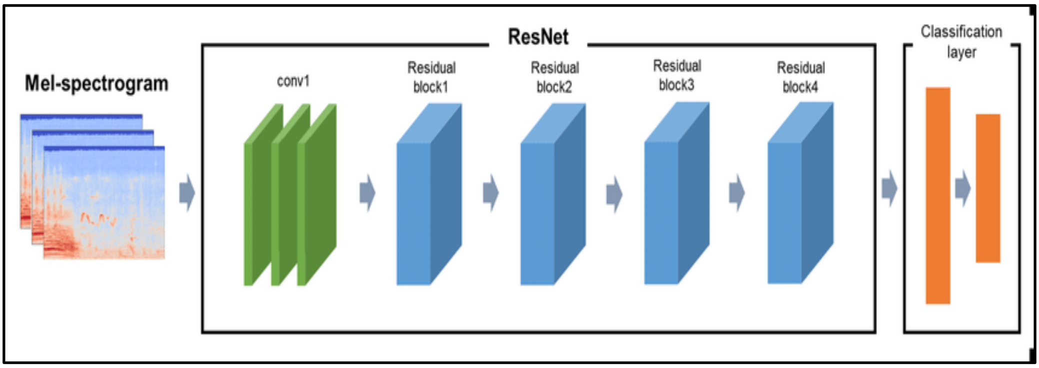


Figure 2. System overview for the spoof detection model

3.2. Speech Verification Network

Overview: The spoof detection model as discussed above, drastically reduces the spoofed data that is passed on the speech verification model. The aim of this model is to authenticate a speaker using upcoming deep-learning techniques. The model takes as input MEL-spectrograms corresponding to four audio files, three of which are the original (bonafide) audios of the speaker, and the remaining spectrogram maps to the target audio file that needs to be verified against the original sources. The model is divided into two parts, a speech embedder, and a verification network. The former is a Convolutional neural network, while the latter is a Transformer. Cosine similarity is computed for the original and the target embeddings to be used as input to a binary cross-entropy loss object. The speech embedder takes as input MEL-spectrograms of shape (165, 626, 3) and outputs one-dimensional speech embeddings of shape (256,1). The transformer takes as input 3 embeddings corresponding to the original sources and outputs an averaged embedding of shape (256,1), which is then compared with the speech embedding of the target audio file.

Data preparation: The first step in data preparation is creating MEL-spectrograms of the audio files. The parameters used to create the spectrograms effectively are previously mentioned (in section 3). We now prepare a dataset for the following task by carefully grouping 3 bonafide sources from the same speaker with a target source that can be bonafide or spoofed from the same or different speaker. The Table 1 shows the possible combinations and the verification label (1 - authorize and 0 - reject). The original dataset is split into train, test and val such that there are a different, unique set of speakers in each set. Using this, unique sets of 4 audio files which support the grouping principles are sampled, defined in Table 1.

Table 1. Define the grouping principles for creating the dataset for the verification model

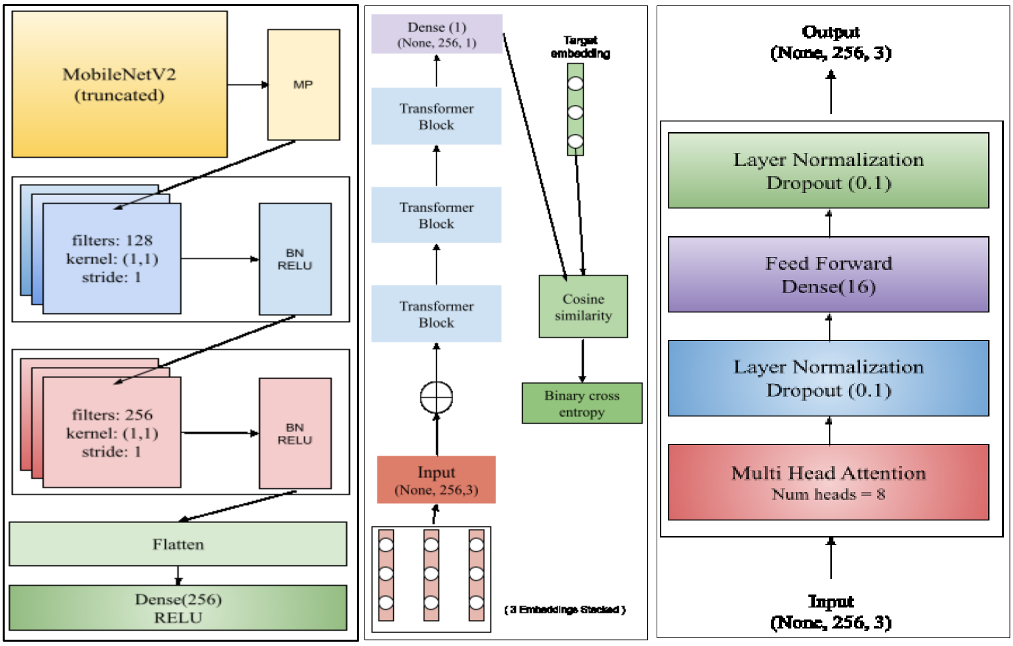
	Verification label
Speaker A x 3 + Speaker A bonafide	*1*
Speaker A x 3 + Speaker B bonafide	*0*
Speaker A x 3 + Speaker A spoof	*0*
Speaker A x 3 + Speaker B spoof	*0*

500000 quadruplets are generated for train, test and validation partitions.

Speech embedder: Speech embedder is an extension to the pretrained MobileNetV2 that embeds the MEL spectrograms into 256-dimension embeddings. The architecture for the same is defined in the Figure 3(a). Taking as input an audio spectrogram of shape (165, 626, 3), the model comprises residual blocks, depth-wise convolutions, and pointwise convolutions. Batch normalization is used to speed up the training and the RELU activation function introduces the necessary non-linearity. The main challenge, however, was to reduce the number of parameters to speed up training. Hence, using 1x1 convolutions on the truncated MobileNetV2 model proved extremely effective.

Verification network: The verification network is a transformer-based sequence model. The verification network used here, however, doesn't explicitly solve a sequence modelling problem. Instead, it acts as a machine that averages over multiple embeddings using dense layers and learns averaged features for the 3 embeddings that are passed as input. The architecture of the verification model is defined in the Figure 3(b). The combined input is passed through 3 transformer blocks, where it goes through multi-head attention (8 heads) and layer normalization followed by a 16-unit dense layer. In a transformer, computations are repeated multiple times in a parallel fashion by the attention module. The attention module has a query, key and value parameters which are split and passed through N separate heads. All the final attention calculations are combined together to produce the final attention score. The architecture of the transformer block is defined in Figure 3(c).

The output of the transformer network is a 256-dimension embedding, which is compared with the 256-dimension embedder output corresponding to the target audio using cosine similarity. The similarity score, however, is bounded between [-1, 1] and it is important for it to be normalized between 0 and 1 before passing it to the binary cross entropy loss object for training.



(a) (b) (c)

Figure 3. (a) Architecture for embedding audio Mel spectrograms for input to the verification model (b) Architecture for the Speech Verification model (c) Architecture for the transformer block used in the verification network

A summarized explanation of the entire workflow is given below:

1) Data Preparation:

Mel-Spectrogram Creation:

To prepare the data for the task, the first step is the creation of Mel-spectrograms from the audio files.

Dataset Formation:

The data is structured into quadruplets, each consisting of three bonafide (original) audio sources from the same speaker and one target audio source, which can be either bonafide or spoofed and may belong to the same or a different speaker.

Dataset Splitting:

The original dataset is divided into training, testing, and validation partitions, ensuring that each set contains a unique set of speakers.

2) Model Architecture:

i) Spoof Detection model: A ResNet18 model is applied to detect spoofed audios initially and discard them.

ii) Speech Embedder: The Speech Embedder plays a crucial role in converting the input Mel-spectrograms into meaningful embeddings for further processing. It is based on a modified MobileNetV2 architecture, adapted to suit the task of speaker verification.

Model Components:

- Input: The Speech Embedder takes Mel-spectrograms as input, with dimensions (165, 626, 3), representing the frequency, time, and channels.
- Residual Blocks: The model comprises residual blocks, which are a series of layers that allow the network to capture and learn complex patterns in the spectrogram data.
- Depth-wise Convolutions: Depth-wise convolutions are employed to capture spatial information effectively.
- Pointwise Convolutions: Pointwise convolutions are used to reduce the number of channels and control the model's capacity.
- Batch Normalization: Batch normalization is applied to accelerate training and ensure that the model generalizes well.
- ReLU Activation: The ReLU activation function introduces non-linearity into the model, enabling it to learn complex mappings.

Parameter Reduction:

1x1 convolutions were applied to the truncated MobileNetV2 model, proving to be highly effective in reducing the model's size while preserving its representational power.

iii) Verification Network:

It is composed of Transformer-based sequence modeling components, but it doesn't explicitly solve a sequence modeling problem as in natural language processing. Instead, it serves as a machine that processes multiple embeddings and computes an averaged feature representation.

Model Components:

- Input: The Verification Network takes as input three embeddings corresponding to the original sources and processes them to make an authentication decision.
- Transformer Blocks: The combined input passes through three transformer blocks, which are a fundamental component of the network architecture.
- Multi-Head Attention: Within each transformer block, multi-head attention with eight heads is applied to capture relationships and dependencies between different parts of the input embeddings.
- Layer Normalization: Layer normalization is used to stabilize and speed up training.
- Dense Layers: Following the attention mechanisms, the network employs 16-unit dense layers to further process the data and generate meaningful features.

Attention Mechanism:

In the transformer, computations are performed in parallel. The final attention calculations are combined to produce the ultimate attention score.

Cosine Similarity and Normalization: After the transformer processing, the network outputs a 256-dimensional embedding. This embedding is compared with the speech embedding of the target audio file using cosine similarity. The normalized score is then used as input to the binary cross-entropy loss object during training.

4. RESULTS AND DISCUSSIONS

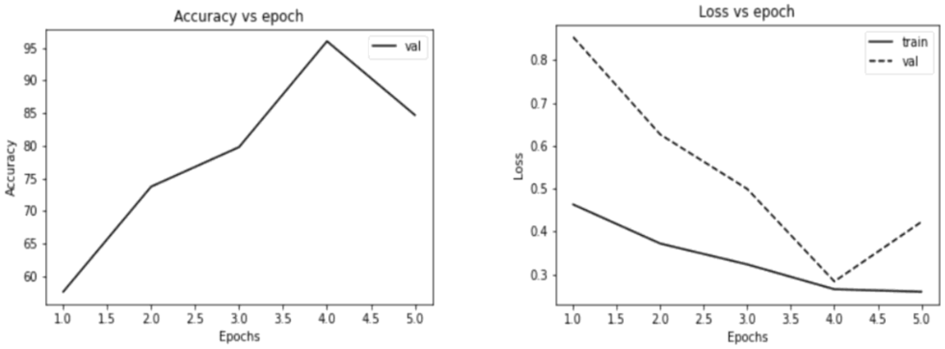
The system includes a spoof detection model based on ResNet18, which effectively filters out spoofed audios. This step significantly reduces the chances of false positives in the subsequent speaker verification process and makes the system deep in true sense. The DR-SASV system employs an end-to-end architecture that integrates spoof detection and speaker verification in a sequential manner making it truly deep and offering two-factor authentication. This design ensures that both aspects are considered in the verification process and along with the usage of the ASVSpoof 2019 LA dataset, makes it more reliable in identifying impostor speakers and spoofing attacks.

The proposed spoof detection model, which makes use of the XResNet-18 model on the 2019 ASV Challenge dataset gives a test accuracy of 95.34%, while the proposed speech verification model which makes use of a speech embedder and transformer, authenticates users with an accuracy of 86.26%. The combined architecture gives an EER score of 10.32%. Exponential learning decay is applied during training with a decay rate of 0.96 and decay steps being 100000. The formula for computing the learning rate at each epoch step is given by Eq. (3).

$$new_{lr} = initial_{lr} \; . decay\_rate^{\frac{step}{decay_{steps}}} \qquad (3)$$

where: new_{lr}= new learning rate; $initial_{lr}$=initial learning rate.

Early stopping at epoch 4 in the spoof detection model and at epoch 8 in the speech verification model prevent overfitting and allow generalization, which is important in a speaker-independent scenario. The learning curves for the spoof detection models are shown in Figure 4 (a) and Figure 4 (b) and the training curve for the speech verification model is given in Figure 5.



(a) *(b)*

Figure 4. (a)Accuracy curve for the spoof detection model (b) Loss vs epochs curve for the spoof detection model

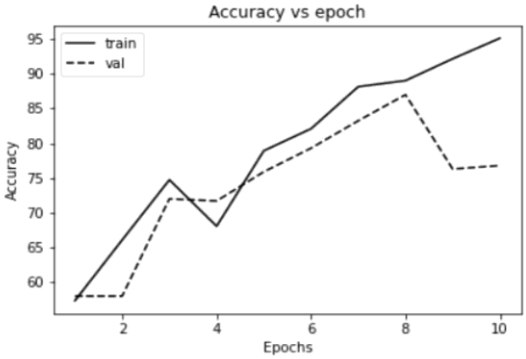


Figure 5. Training curve for the spoof verification model

Since the 3 audios corresponding to the input embeddings passed to the transformer are used only in the beginning to configure the averaged embedding, the verification model is of no use during test time. The target embeddings are directly compared with the stored average embedding.

The performance of both the models on the train, test and dev set is encapsulated in Table 2 while Table 3 and compares the EER scores of DR-SASV with other state-of-the-art SASV models with the purpose of offering a comprehensive view of the performance landscape and to position DR-SASV relative to other approaches. It is important to note that DR-SASV outperforms a majority of these models.

While there are systems which offer a lower EER score, DR-SASV promises faster computation, light-weightiness and a novel two-factor authentication concept.

Table 2. Performance of the spoof detection and speech verification model.

Accuracy	***Spoof Detection (%)***	***Speech verification (%)***
Train Set	*96.57*	*88.02*
Val set	*96*	*82.27*
Test set	*95.34*	*86.26*

Table 3. Comparing the EER score of the DR-SAV model with other state-of-the-art models.

Model/ Team	***EER %***
ECAPA-TDNN [25]	*23.81*
SASV 2022 Challenge Baseline-1 [26]	*19.31*
Integrated Replay [11]	*15.63*
MFCC and TDNN [9]	*11.89*
DR-SASV (our model)	***10.32***
SASV 2022 Challenge Baseline-2 [24]	*6.54*

Tandem [26]	*6.22*
HCCL [26]	*4.30*

5. CONCLUSION AND FUTURE WORK

This paper explores spoof detection on audio MEL-spectrograms using a pretrained Resnet18 model. The model is cascaded with a transformer-based speech verification model, which averages over three input audio embeddings. The averaged embedding then determines the fate of the target embedding.

The paper shows how a "spoof aware" verification system comprising of a ResNet18 model in conjunction with a spoof detection model comprising of a MobileNet V2 model produces commendable accuracy after being trained on the ASVSpoof LA dataset. Along with increasing reliability, the two-step verification saves a lot of time during runtime as the transformer module doesn't need to go through any computation. Furthermore, the system is speaker independent and is not biased toward any voice type (frequency and amplitude).

The system surpasses the baseline model described in the SASV Challenge 2022 and the significance of the DR-SASV system's performance lies in its effectiveness and reliability for the specific problem we aimed to address—spoof-aware speaker verification. With an Equal Error Rate (EER) of 10.32%, our system demonstrates its capability to distinguish between genuine speakers and impostors, a task of utmost importance in applications such as fraud detection, access control, and secure authentication.

Future work can include strengthening the verification system by retraining it on the spoofed inputs that the spoof detection model fails to detect as positives. False negatives from the spoof detection module passed into the verification system may cause false positives in the final output. Moreover, the accuracy, in terms of the EER score can be improved using a more powerful processor.

REFERENCES

[1] Bureau, B. B. 47% of Indians have experienced AI voice scams: Mcafee survey. *BusinessLine.* https://www.thehindubusinessline.com/info-tech/47-of-indians-have-experienced-ai-voice-scams-mcafee-survey/article66803142.ece. Accessed 2 May, 2023.

[2] Reynolds, D.A., Quatieri, T.F. and Dunn, R.B. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, vol. 10, no. 1–3, 2000, pp. 19–41. DOI:10.1006/dspr.1999.0361.

[3] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., Ouellet, P. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, 2011, pp. 788-798. DOI: 10.1109/TASL.2010.2064307.

[4] K. He, X. Zhang, S. Ren and J. Sun. Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778, DOI: 10.1109/CVPR.2016.90.

[5] Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., Cao, Y., Kannan, A., & Zhu, Z. Deep speaker: An end-to-end neural speaker embedding system. *arXiv.org*, 2017 URL: https://arxiv.org/abs/1705.02304.

[6] Chung, J. S., Huh, J., Mun, S., Lee, M., Heo, H. S., Choe, S., Ham, C., Jung, S., Lee, B.-J., & Han, I. In defence of Metric Learning for speaker recognition. *arXiv.org*, 2020. URL: https://arxiv.org/abs/2003.11982.

[7] Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. Sphereface: Deep Hypersphere embedding for face recognition. *arXiv.org*, 2018. URL: https://arxiv.org/abs/1704.08063.

[8] F. Wang, J. Cheng, W. Liu and H. Liu. Additive Margin Softmax for Face Verification. *IEEE Signal Processing Letters*, vol. 25, no. 7, 2018, pp. 926-930, DOI: 10.1109/LSP.2018.2822810.

[9] Todisco, M., Delgado, H., Lee, K.A., Sahidullah, M., Evans, N., Kinnunen, T., Yamagishi, J. Integrated Presentation Attack Detection and Automatic Speaker Verification: Common Features and Gaussian Back-end Fusion. Proc. *Interspeech 2018*, pp. 77-81, DOI: 10.21437/Interspeech.2018-2289.

[10] Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., Liu, X., Lee, K. A., Kinnunen, T., Evans, N., & Delgado, H. ASVSPOOF 2021: Accelerating progress in spoofed and Deepfake Speech detection. *arXiv.org*, 2021. URL: https://arxiv.org/abs/2109.00537.

[11] Shim, H., Jung, J., Kim, J., &amp; Yu, H. Integrated replay spoofing-aware text-independent speaker verification. *Applied Sciences*, vol. 10, no. 18, 2020. DOI: https://doi.org/10.3390/app10186292.

[12] J. Li, M. Sun and X. Zhang. Multi-task learning of deep neural networks for joint automatic speaker verification and spoofing detection. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC),* 2019, pp. 1517-1522, DOI: 10.1109/APSIPAASC47483.2019.9023289.

[13] Li, Jiakang & Sun, Meng & Zhang, Xiongwei & Wang, Yimin. Joint Decision of Anti-Spoofing and Automatic Speaker Verification by Multi-Task Learning With Contrastive Loss. *IEEE Access, 2020*. PP. 1-1. 10.1109/ACCESS.2020.2964048.

[14] Chung, J.S., Huh, J., Mun, S., Lee, M., Heo, H.-S., Choe, S., Ham, C., Jung, S., Lee, B.-J., Han, I. In Defence of Metric Learning for Speaker Recognition. Proc. *Interspeech*, 2020, pp.2977-2981, DOI: 10.21437/Interspeech.2020-1064.

[15] Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., Cao, Y., Kannan, A., & Zhu, Z. Deep speaker: An end-to-end neural speaker embedding system. *arXiv.org*, 2017. DOI: https://doi.org/10.48550/arXiv.1705.02304

[16] Tak, H., Patino, J., Nautsch, A., Evans, N., & Todisco, M. Spoofing attack detection using the non-linear fusion of sub-band classifiers. *Interspeech,* 2020. DOI: https://doi.org/10.21437/interspeech.2020-1844.

[17] Teng, Z., Fu, Q., White, J., Powell, M., & Schmidt, D. Sa-SASV: An end-to-end spoof-aggregated spoofing-aware speaker verification system. *Interspeech,* 2022. DOI: https://doi.org/10.21437/interspeech.2022-11029.

[18] Zhao, Yuanjun & Togneri, Roberto & Sreeram, Victor. Multi-task Learning-Based Spoofing-Robust Automatic Speaker Verification System. *Circuits, Systems, and Signal Processing,* vol. 41, no. 1-22, 2022. DOI: 10.1007/s00034-022-01974-z.

[19] Wu, H., Meng, L., Kang, J., Li, J., Li, X., Wu, X., Lee, H., & Meng, H. (2022). Spoofing-aware speaker verification by multi-level fusion. *Interspeech,* 2022. DOI: https://doi.org/10.21437/interspeech.2022-920.

[20] Grinberg, Petr & Shikhov, Vladislav. A Comparative Study of Fusion Methods for SASV Challenge 2022. *Interspeech,* 2022. DOI: https://doi.org/10.48550/arXiv.2203.16970

[21] Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. Deep neural network embeddings for text-independent speaker verification. *Interspeech,* 2017. DOI: https://doi.org/10.21437/interspeech.2017-620.

[22] Junichi, Y., Massimiliano, T., Md, S., Héctor, D., Xin, W., Nicolas, E., Tomi, K., Aik, L. K., Ville, V., & Andreas, N. ASVspoof. *3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database*, 2019. URL: http://urn.fi/urn:nbn:fi:att:4e69bd10-66b8-4b0a9854-c2b738ef721a. (Visited on 15.11.2022).

[23] Veaux, Christophe; Yamagishi, Junichi; MacDonald, Kirsten. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit, [sound]. *University of Edinburgh. The Centre for Speech Technology Research (CSTR*). 2017, DOI: https://doi.org/10.7488/ds/1994.

[24] J. Jung et al. SASV challenge 2022: A spoofing aware speaker verification challenge evaluation plan, *arXiv preprint* arXiv:2201.10283, 2022.

[25] Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized channel attention, propagation, and aggregation in TDNN based speaker verification. *Interspeech,* 2020. DOI: https://doi.org/10.21437/interspeech.2020-2650.

[26] SASV2022 Challenge Results. *SASV2022 Challenge Results SASV Challenge*, sasv-challenge.github.io/challenge_results/. Accessed 16 Sept. 2023.

Information about the authors:

Amay Gada– Received BTech in CS from DJ Sanghvi College of Engineering. Currently pursuing a Masters in CS from North Carolina State University. Areas of research: Scalable Systems, Deep learning and Security.

Neel Kothari – Pursuing BTech in CS from DJ Sanghvi College of Engineering. Areas of research Machine learning and Deep learning, Generative AI.

Ruhina Karani – Working as an Assistant Professor in Computer Engineering Department of D.J. Sanghvi College of Engineering. Her domain of expertise includes Machine Learning and Computer Vision.

Dr. Chetashree Bhadane – Currently working as an Assistant Professor in the Computer Engineering Department of D.J. Sanghvi College of Engineering.

Dhruv Gada – Pursuing BTech degree in CSE from DJ Sanghvi College of Engineering & BSc in Data science from IIT Madras. Areas of research: Language Models and Deep learning.

Tanish Patwa – Pursuing BTech degree in CSE from DJ Sanghvi College of Engineering. Areas of research: Machine learning and Deep learning.

Manuscript received on 24th August 2023