Deep Convolutional Networks with Attention for Identifying and Preventing End-to-End Audio Replay Attacks

MUSALI SUREKHA[1],KAMEPALLI UMA[2],
ASSOCIATE PROFESSOR[1], ASSISTANT PROFESSOR[2],
DEPARTMENT OF ECE
PBR VISVODAYA INSTITUTE OF TECHNOLOGY AND SCIENCE::KAVALI

Abstract

With automatic speaker verification (ASV) systems becoming increasingly popular, the development of robust countermensures against spoofing is needed. Replay attacks pose a sigindicant threat to the reliability of ASV systems because of the relative difficulty in detecting replayed speech and the ease with which such attacks can be mounted. In this paper, we propose an end-to-end deep learning framework for audio replay attack detection. Our proposed approach uses a novel visual attention mechanism on time-frequency representations of Uttrances based on group delay features, via deep residual learnIng (an adaptation of ResNet-18 architecture). Using a single model system, we achieve a perfect Equal Error Rate (EER) of 0% on both the development as well as the evaluation set of the AS spoof 2017 dataset, against a previous best of 0.12% on the development set and 2.76% on the evaluation set reported in the literature. This highlights the efficacy of our feature representton and attention-based architecture in tackling the challenging task of audio replay attack detection. Index Terms: Replay attack, group delay grams, end-to-end deep learning, visual attention, AS spoof 2017 dataset

Introduction

Automatic speaker verification (ASV) technology has several applications, including voice-based identification, pathologycall voice assessment [1] and forensic evidence evaluation [2]. These applications require the ASV systems to be robust against intentional circumvention using fake audio recordings, also known as ‘spoofing attacks. Spoofing attacks can be categoryrazed into four types: impersonation, replay, speech Synthessis, and voice conversion [3]. Due to the severity of these attacks, the Automatic Speaker Verification Spoofing and Counttreasures (AS spoof) Challenge [4] was launched in 2015, with the objective of enhancing the security of ASV systems against spoofing attacks. The AS spoof 2015 challenge ofcussed on speech synthesis and voice conversion attacks, while the AS spoof 2017 challenge [5] focused on replay attacks. The artifacts introduced by replay are very different from those introduced by voice conversion and speech synthesis. The AS spoof 2017 challenge task is to determine whether a given audio clip is a GENUINE human voice or a REPLAY recording. Replay attacks fool the ASV system by simply replaying a recording of a target speaker’s voice. Replay attacks are of key concern as they are relatively easy to perform and pose a significant threat to the reliability of an ASV system [6].

For instance, all smartphones provide high quality audio recording and playback, and hence can be used for replay attack. Detecting replay attacks using acoustic signal processing is considered hard, due to the unpredictable variation in the quality of a replay attack [5]. Artifacts introduced by naturally occurring factors such as reverberation may be confusable in some cases with those introduced by replay. Researchers also tried machine learning for detecting replay attacks, and found it to perform poorly, mainly due to the overfitting caused by the varyability in speech signals [7]. Such models do not generalize well to unseen acoustic environments that may be encountered in practice. Audio recordings using high-quality microphones in ideal acoustic environments can be indistinguishable from genuine speech signals. For the part of the AS spoof 2017 competition that required distinguishing between genuine human voice and replay recording, a total of 49 submissions were received. Only 20 of those 49 submissions outperformed the baseline spoof detection system, which was based on a Gaussian mixture model (GMM) back-end classifier with constant Q cepstral coefficient (CQCC) features [5]. This shows the difficulty of the challenge.

Deep convolutional networks on spectrograms performed the best, using an ensemble of three techniques – LCNNF F T, SV Mi−vest, and CNNF F T + RNN – to achieve an EER of 6.73% on the evaluation set and an EER of 3.95% on the development set of the AS spoof 2017 dataset [8]. Patil et al. [9] used VESA-IFCC (Variable length Teaser Energy Operator-Energy Separation Algorithm-Instantaneous Frequency Cosine Coefficients) and achieved an EER of 0.12% on the development set and an EER of 14.06% on the evaluation set. The organizers of the AS spoof 2017 competition achieved the best EER of 2.76% on the evaluation set, by creating a grand ensemble of the 21 best performing systems [8]. To further emphasize, the difficulty of reliably detecting replay attacks can be mainly attributed to the fact that the artifacts introduced by the recording and

playback get intertwined with other inessential sources of variability, such as recording and playback device-related artifacts, environmental noise, the speaker's voice identity, etc. Thus, it is important to propose models that can robustly identify the pertinent artifacts introdiced by the recording and playback process, while at the same time, ignore variability introduced by the 'other' factors, in order to generalize well to unknown scenarios. This necessitates the use of a feature representation with high spectral resolution to capture details present in spectral regions that contain discriminative information. Moreover, the model should also be able to selectively attend to these regions, so that it does not overfit on the other inessential variability factors. In this paper, we propose group delay (GD) grams ob.trained by concatenating a group delay function over Consecotime frames as a novel time-frequency representation of an Uttrance, for the end-to-end training of deep convolutional neureal networks for audio replay attack detection. The use of GDgrams provide a time-frequency representation with high spectrail resolution, which is required for robust replay attack detectton. Moreover, we propose a novel attention mechanism that

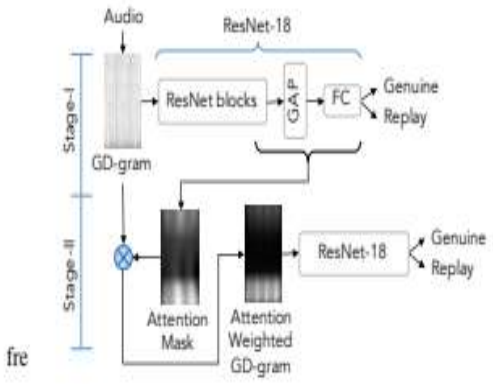

Fig. 1: Overview of the proposed framework for audio replay attack detection. (Note: GD: Group Delay, GAP: Global Average Pooling, FC: Fully Connected layer)

softly weights GD-grams, allowing the network to focus on the regions of the spectrum that contain high discriminative Informotion for replay detection. Our framework is based on adaptIng the ResNet-18 architecture [9] and using its Global Average Pooling (GAP) layer to provide attention maps for a second stage of discriminative training for improved performance. We achieve a perfect Equal Error Rate (EER) of 0% on both the development and evaluation sets of the AS spoof 2017 dataset.

Methodology

Our proposed framework (Figure 1) employs: (1) transfer learnIng of a pretrained convolutional neural network (CNN) for fast adaptation to the GD-grams extracted from utterances, (2) attensional weighting of the raw GD-grams from the first stage of training, and (3) another stage of transfer learning of a pretrained CNN on GD-grams weighted by soft attention for classossification. For both the stages, we used Deep Residual Network (ResNet) [9] as pretrained CNN and the weights were retrained after initialization. In the following subsections, we describe the three major components of our system – GD-gram, ResNet and Visual Attention – followed by a functional overview of our proposed framework.

Group Delay gram (GD-gram)

The short-time Fourier transform (STFT) of an input speech signil sequence x(n) can be expressed as:

$$X(\omega,t) = |X(\omega,t)|e^{j\theta(\omega,t)}, \qquad (1)$$

where |X (ω, t) | and θ (ω, t) are the magnitude spectrum and phase spectrum at frequency ω and time t, respectively. Group delay [10] is defined as the negative derivative of the phase spectrum of STFT:

$$\tau(\omega,t) = -\frac{d(\theta(\omega,t))}{d\omega}. \qquad (2)$$

As the implementation of Equation (2) requires the unwrapping of the phase spectrum, the group delay function can be alternatively calculated using only the amplitude values:

$$\tau(\omega,t) = \frac{X_R(\omega,t)Y_R(\omega,t) + Y_I(\omega,t)X_I(\omega,t)}{|X(\omega,t)|^2}, \qquad (3)$$

where R and I denote the real and imaginary parts. X (ω, t) and Y (ω, t) denote the STFT of x(n) and nix(n), respectively. We concatenate the group delay function (coefficients) of all frames of an utterance to form the GD-gram. This 2D matrix GD-gram is fed to the CNN as an input image (Figure 3).

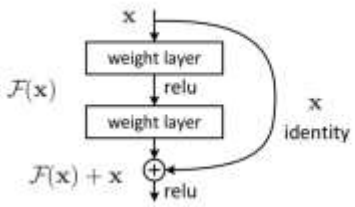

Fig. 2: Residual block: Basic building block of ResNet.

The group delay function has been previously applied in feature extraction tasks in speech processing [11, 10], where it has been proposed as an alternative to the magnitude spectrum. In our use case, a replayed speech signal passes through muletiple channels with the channel artifacts typically being introdiced in frequency bands with low signal to noise ratio. Thus, in order to robustly extract discriminative information from pertenant spectral regions, the time-frequency representation of the speech signal should provide high spectral resolution. Group delay functions have been shown to have higher spectral Resolotion in comparison with the magnitude spectrum [12]. Moreover, GD-gram contains both power and phase spectrum Informotion [11, 13], thus making it a good feature representation for end-to-end learning for spoof detection. The approach of using phase spectrum information for replay attack detection is novel. Previously, phase information has been used for detectIng speech synthesis and voice conversion attacks [14].

Deep Residual Network (ResNet)

Deep residual learning [9] enables the training of CNNs that are substantially deeper than the architectures preceding it. It alleviates the problem of vanishing gradients in deep CNNs by introducing skip connections that enable gradient flow across a large number of layers. The skip connections cause the outputs to learn a residual mapping. The residual block forms the basic building block of a ResNet (Figure 2). If the desired mapping to be learned is H(x), the stacked residual layers learn the residual mapping, F(x) = H(x) − x. Thus, the original mapping to be learned becomes F(x) + x. ResNet uses the Rectified linear unit (REL) activation function. In our work, we use the ResNet-18 model [9], which consists of layers in the following order: 7 × 7 convolution layer, eight residual blocks, Global Average Pooling (GAP) layer, foollowed by a fully connected layer with SoftMax. Along with ResNet-18, we also use dropout [15] to regvulgarize our model. Dropout combats the issue of overfitting by preventing activations from becoming strongly correlated. CNNs effectively utilise local Spectro-temporal correlations in time frequency representations of speech, such as GD-grams. However, using dropout in convolutional layers results in the scaling of the learning rate by the dropout probability, in case there is a strong correlation between adjacent pixels. Hence, we use spatial dropouts [16] in which entire feature maps are dropped out to regularize the network.

Visual attention Li et al.

showed through the F-ratio metric that high frequency bands have great discriminative capability for audio replay attack detection [7]. They used inverted Mel warping to emphasissize the high frequency bands and demonstrated that it improves performance of spoof detection on the AS spoof 2017 developmint set (EER improved from 12.37% to 7.50%). To capture the discriminative information contained within specific regions, we propose a visual attention mechanism based on class activation mapping [17] for replay attack detection. Class activation maps (CAM) using global average pooling (GAP) utilizes the implicit attention present in CNNs. The GAP layer was introduced to act as a structural regularize and prevent overfitting [18]. It has since been shown that remarkable results in localizing the discriminative regions of an image can be achieved using CAMs after being trained just on image level labels [17]. For instance, CAM can detect an object without usprevision of the object's location in an image. CAM leverages the ability of GAP to retain the localization capability of the last layers of a convolutional neural network. The GAP layer in ResNet-18 outputs the spatial average of all the activation maps after the last convolution layer. At the end, a fully-connected layer is used to predict the class, using the weights attached to each unit in the GAP layer. For classicflying c classes, Sc denotes the output of the SoftMax layer,

$$S_c = \sum_k w_k^c \sum_{x,y} f_k(x,y) = \sum_{x,y} \sum_k w_k^c f_k(x,y), \quad (4)$$

where fh (x, y) denotes the activation value of unit k in the last convolutional layer at location (x, y). The weights of the fullyconnected layers are denoted as w. The CAM of a class Mc is obtained by

$$M_c(x,y) = \sum_k w_k^c f_k(x,y). \quad (5)$$

The CAM is computed as a weighted sum of the feature maps of the last convolution layer. Mc (x, y) represents the reelevince of the activation of the grid (x, y) for classifying the miage as belonging to class c (Genuine or Replay). As replay detaction is a binary classification problem, we use a single class activation map Mpret (x, y), which is the activation map corespending to the class predicted by the Stage-I network, to softly weight the GD-gram. The class activation map Mpret (x, y) is then up sampled to the size of the input GD-gram to generate the attention mask A (x, y). Unlike images where the horizontal and vertical axes have the same meaning, GD-grams are timefrequency representations with x representing the time axis and y representing the frequency axis. The soft attention weighted (AW) output GDAW (x, y) is given by

$$GD_{AW}(x, y) = GD(x, y) * A(x, y). \quad (6)$$

As the attention mask A (x, y) highlights the spectral regions that are relevant in differentiating between genuine and replayed speech, AW GD-grams act as time-frequency representations of speech in which the regions of the spectrum that are important for spoof detection gets emphasized, for further discriminative training. Such representations are essential for the model to generalize to spoof attacks 'in the wild'. To summarize, the combination of the GD-gram timefrequency representation and the attention-based convolutional neural network architecture are significant departures from exsiting approaches in the literature, and contributes to the high performance of our proposed framework.

3. Experiments

Dataset

In this work, we focus on the AS spoof 2017 replay attack dataset. The dataset consists of (a) training and development sets of genuine/replay labelled audio examples, along with metadata about the speech content, devices and replay environment, and (b) an evaluation set of both known and unknown Conditons (Table 1). The evaluation set is comprised of a combineton of replay environments, playback devices and speakers that are not part of the development data to evaluate model performmince in unforeseen conditions. The AS spoof 2017 dataset is based on the red dots data collection project, processed through various replay conditions. The dataset consists of speech data collected from 177 replay sessions in 123 unique replay configurations, with 42 differEnt speakers (Table 1). A replay configuration means a unique combination of room, replay device and recording device, while a session refers to a set of source files sharing the same replay configuration. The speech signals were collected in highly varying acoustic-conditions. Different quality of playback and recording devices were used. In order to simulate spoofing attacks 'in the wild', the training set has only 3 replay configorations with speech from 10 speakers, whereas the evaluation set has 110 highly heterogeneous acoustic replay configurations with speech from 24 speakers.

4. Results

The results of our proposed approach, 'GD-ResNet-18 with attension' model on the development and evaluation set of the AS spoof 2017 dataset is summarized in Table 2, along with the best results reported in the literature. The organizers of the AS spoof 2017 challenge created a grand ensemble of the 21 best performing systems submitted to the challenge to obtain an EER of 2.76% on the evaluation set. However, using ResNet-18 with attention on GD-grams yields an EER of 0% and HTER of 0% on the evaluation set. Previously, Pal et al. [23] achieved a nearly 0 overall average EER (0.05%) on the AS spoof 2015 dataset comprising of voice conversion and speech synthesis attacks. An experiment without using the attention mechanism and just the stage-I network resulted in an EER of 12.77%. GDgrams provide a time-frequency representation with high spectrail resolution. In the absence of the second stage of discrimennative training, there is an increased tendency to overfit on the spectral artifacts caused by inessential audio factors. Thus, the attention mechanism is crucial for leveraging the discriminatetime information present in GD-grams. Experimentation with the magnitude spectrogram as input representation (instead of GD-grams) resulted in an EER of 13.14% on the evaluation set for stage-I. However, its performance degraded to 16.29% with the addition of an attention mechanism. This further highlights the importance of learning the attention masks from the group delay domain, which offers higher resolution, for the second stage of discriminative training.

The maximum Area Under the curve of the Receiver Operaacting Characteristic (AUROC=1) is obtained on the evaluation set of the corpus showing that the model is perfect in its divefermentation between replay and genuine

utterances. The evilaction set was designed to assess the limits of replay attack detaction and provides spoofing attacks 'in the wild' with replay attacks from 110 replay configurations whereas the training set was composed of only 3 replay configurations. The remarkable improvement of our model over the previaoust state-of-the-art can be primarily attributed to two factors: (1) the higher spectral resolution offered by GD-grams along with inclusion of phase information, and (2) the ability of the visual attention mechanism to attend to spectral regions containIng discriminative information that allows the model to generalsize well to unseen replay configurations. Specifically, the GAP layer of the stage-I ResNet is used to identify regions of interEst in the raw GD-gram representation of the speech signal and to weight the GD-gram to generate an Attention Weighted GDgram before passing it to the stage-II ResNet for classification. In Figure 3, it is visible that the Attention Weighted varyants are more discriminative than the raw GD-grams. The raw GD-gram after being softly weighted by the attention mask resalts in a representation where certain regions in the spectrum are emphasized relative to other regions. Using this intermentdate representation for another stage of discriminative training allows our framework to tune itself to place emphasis on the discriminative information present in raw GD-grams. This further validates the hypothesis that it is important to emphasize the discriminative frequencies, and deemphasize frequencies that are more impacted by inessential factors of variability in speech, in order to achieve good results for replay attack detection.

5. Conclusion

In this paper, we propose an end-to-end deep learning framework for audio replay attack detection based on raw GD-grams. We highlight the importance of utilizing discriminative Informotion contained within specific regions of the spectrum by proposing a visual attention mechanism to allow our model to focus on the regions pertinent to replay attack detection, along with a time-frequency representation of speech with high spectrail resolution, to tackle the challenges associated with audio replay attack detection. We achieved an EER of 0% on both the development and evaluation sets of the AS spoof 2017 dataset.

6. References

[1] C. Grenouille, G. Patchouli, J.-F. Bonastre, M. Azzarello, A. Giovanni, and A. Ghio, "Application of automatic speaker recognition techniques to pathological voice assessment (dysphonia)," in Proceedings of European Conference on Speech Communication and Technology (Eurospeech 2005). ISCA, 2005, pp. 149–152.

[2] A. M. T. S. B. Adikari, S. Devadithya, A. R. S. T. Bandara, K. C. J. Dharmawardane, and K. C. B. Wavegedara, "Application of automatic speaker verification techniques for forensic evidence evaluation," in 2014 19th International Conference on Digital Signal Processing, Aug 2014, pp. 444–448.

[3] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: A survey," Speech Communication, vol. 66, no. Supplement C, pp. 130 – 153, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639314000788

[4] Z. Wu, J. Yamagishi, T. Kinnunen, C. Hanili, M. Sahidullah, A. Sizov, N. Evans, and M. Todisco, "Asvspoof: The automatic speaker verification spoofing and countermeasures challenge," IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 4, pp. 588–604, June 2017.

[5] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in INTERSPEECH 2017, Annual Conference of the International Speech Communication Association, Stockholm, Sweden, 2017. [Online]. Available: http://www.eurecom.fr/publication/5235

[6] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in 2014 International Conference of the Biometrics Special Interest Group (BIOSIG), Sept 2014, pp. 1–6.

[7] L. Li, Y. Chen, D. Wang, and T. F. Zheng, "A Study on Replay Attack and Anti-Spoofing for Automatic Speaker Verification," ArXiv e-prints, Jun. 2017.

[8] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection slides," http://www.asvspoof.org/slides-ASVspoof2017- Interspeech.pdf, 2017.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[10] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," IEEE Transactions on Audio, Speech, and Language Processing , vol. 15, no. 1, pp. 190–202, 2007.

[11] R. M. Hegde, H. A. Murthy, and G. R. Rao, "Application of the modified group delay function to speaker identification and discrimination," in ICASSP, vol. 1. IEEE, 2004, pp. I–517.

[12] J. Sebastian, M. Kumar, and H. A. Murthy, "An analysis of the high resolution property of group delay function with applications to audio signal processing," Speech Communication, vol. 81, pp. 42–53, 2016.

[13] J. M. K. Kua, J. Epps, E. Ambikairajah, and E. Choi, "Ls regularization of group delay features for speaker

recognition," in Tenth Annual Conference of the International Speech Communication Association, 2009.

[14] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, D. Erro, and T. Raitio, "Toward a universal synthetic speech spoofing detection using phase information," IEEE Transactions on Information Forensics and Security, vol. 10, no. 4, pp. 810–820, 2015.

[15] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." Journal of machine learning research, vol. 15, no. 1, pp. 1929–1958, 2014.

[16] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 648–656.

[17] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.

[18] M. Lin, Q. Chen, and S. Yan, "Network in network," arXiv preprint arXiv:1312.4400, 2013.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in CVPR09, 2009.

[20] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[21] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudashev, and V. Shchemelinin, "Audio replay attack detection with deep learning frameworks," Proc. Interspeech 2017, pp.82–86, 2017.