Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

# Deep Learning based DFWF Model for Audio Spoofing Attack Detection

**Kottilingam Kottursamy**

Associate Professor, Department of CSE, SRM Institute of Science and Technology, Kattankaluthur, Chennai, India

**E-mail:** k.kottilingam@gmail.com

## Abstract

One of the biggest threats in the speaker verification system is that of fake audio attacks. Over the years several detection approaches have been introduced that were designed to provide efficient and spoof-proof data-specific scenarios. However, the speaker verification system is still exposed to fake audio threats. Hence to address this issue, several authors have proposed methodologies to retrain and finetune the input data. The drawback with retraining and fine-tuning is that retraining requires high computation resources and time while fine-tuning results in degradation of performance. Moreover, in certain situations, the previous data becomes unavailable and cannot be accessed immediately. In this paper, we have proposed a solution that detects fake without continual-learning based methods and fake detection without forgetting in order to develop a new model which is capable of detecting spoofing attacks in an incremental fashion. In order to retain original model memory, knowledge distillation loss is introduced. In several scenarios, the distribution of genuine voice is said to be very consistent. In several scenarios, there is consistency in distribution of genuine voice hence a similarity loss is embedded additionally to perform a positive sample alignment. The output of the proposed work indicates an error rate reduction of up to 80% as observed and recorded.

**Keywords:** Deep learning, fake without forgetting, data training, continuous learning, fake audio detection

## 1. Introduction

Ever since the interaction of automatic speaker verification spoofing challenge series fake audio detection has gained much attention. To prevent these voice spoofing attacks, several approaches have been proposed over the years [1]. There are two aspects in which


*DOI: https://doi.org/10.36548/jaicn.2022.3.004*
*Received: 23.07.2022, received in revised form: 20.08.2022, accepted: 05.09.2022, published: 15.09.2022*

fake audio detection is studied and analysed by users. The first methodology involves neural networks for effective classifiers. The second methodology involves signal processing methodology for robust acoustic features. Though these methodologies have provided impressive results it is observed that when unseen spoofing attacks occur there is a noticeable degradation in their performance [2]. According to the ASB spoof 2019 challenge the top 10 highest performing systems submitted a medium equal error rate of around 0.1% to 15.6% when exposed to A16 [3] and A17 [4] attack. This indicates that the currently prevalent system lacks efficiency and requires counter measures in order to improve the performance of the system against unseen attacks.

In recent years a number of methods have been proposed to detect fake audio on unseen data. There are three models strained jointly as an ensemble approach as indicated by authors in [5]. The output indicates that the model of strained is capable of outperforming previously existing models. In [6] the authors have solved out of the main data set problems using dual adversarial domain adaptation framework. But the drawback with these methodologies is that they require new as well as the original data. As the need for speech synthesis technologies and voice conversions increase there is way for new spoofing attacks and advanced replied devices emerging in a consistent manner. However this is a time consuming effort [7] that requires several data sources to formulate mixed data [8] and also requires heavy storage consumption [9]. Moreover privacy is concerned certain special cases will prevent access to old data.

## 2. Literature Review

This part of the paper examines the current trends and previously existing methodology that have been introduced over the years to detect voice-based spoofing attacks with advancement in voice synthesis systems and speaker recognition technology [4]. Despite the imperfect identification of human speakers, there is a high level of accuracy in determining the apt speaker recognition system [8]-[10]. Hence one of the popular alternatives is that of a speaker recognition system along with biometric authentication methodology. Similarly in recent years automated user verification by machines have become commonly used and are highly recommended. In this methodology, the voice of the speaker is captured and compared with that of the existing speaker profiles with the help of the AuSpeaker recognition system [11]. When there is a match on comparison, the speaker is provided access to the system with recognition. In the initial stages Gaussian Mixture models

[12] were used as the parametric methods. However, with the recent development, deep learning methodologies have been used to improve accuracy and reduce overhead. However, the existing human susceptibility to spoofing determines the classical synthesis attacks and impersonation. It was determined that it is possible to fool the humans based on a single impersonal attack [8] based on the age of the person. A traditional survey format is used in classical synthesis attack measurements that differentiates between the Festvox-synthesized voices [9] and real voices resulting in 50% of the time.

Training model which is fine tuned is a splendid way to push the performance as far as spoofing attacks are concerned. Hence it is crucial to build a novel technology that is able to continually learn new knowledge on continuous exposure with the help of knowledge learnt previously. In recent years continue learning has become an important aspect which uses peach recognition and computer vision [10]. The primary purpose is to face the issues in fine tuning such as forgetting information due to new information and training. In our proposed work a number of sequential training tasks were investigated based on ASV spoof data with several spoofing types. The output results thus obtained indicate that find tuning performs better than Detecting Fake Without Forgetting (DFWF) [13] approach wherein past knowledge is forgotten or decreased. Thus with the help of continual learning on the fake audio detection, it is possible to rectify the issues involving distinguishment of unseen spoofing in audio. Similarly other survey and data formats are used to include brain scanning methodology [14] which identifies the neural activity of the participants. Hence it is not possible to determine the differences in neural activity statistically if either the synthetic or real speakers are played.

## 3. Proposed Methodology

For continuous learning, three major approaches are commonly used. The neural network structure can be changed dynamically based on the new task using the dynamic architecture approach. While training a new task, the replay and buffer experiences can be recorded as examples in replay approaches. The learning process can be constrained by adding a loss function to hand-craft regularization in the regularization approaches. This work focuses on the regularization approach. Catastrophic forgetting may occur due to the fine-tuning and excessive time required for retaining mixed data. The tradeoff between performance and resources can be overcome using the deep learning based DFWF method. Data is not stored, however, the previous data is remembered by the system. During training,

the positive sample alignment (PSA) and learning without forgetting (LwF) constraints are added to $L_{original}$, which represents the original loss. Hence, $L_{total}$, the total loss is given by the following expression.

$$L_{total} = L_{original} + xL_{LwF} + yL_{PSA} \quad (1)$$

Here, the significance of the respective terms are controlled by the hyper-parameters $x$ and $y$. The model is generally trained with $L_{original}$, which is substituted with the cross entropy loss. The following sections explain the details of $L_{PSA}$ and $L_{LwF}$ losses. The deep learning based DFWF technique is introduced for learning without forgetting (LwF). In comparison with the original network, the output probabilities of the current network are close with respect to the new data. This is made possible with the knowledge distillation loss used by LwF while training new tasks. Speech recognition applications commonly use the student-trainer knowledge distillation model which is similar to the proposed technique. However, $L_{LwF}$, an additional knowledge distillation loss is achieved by including LwF.

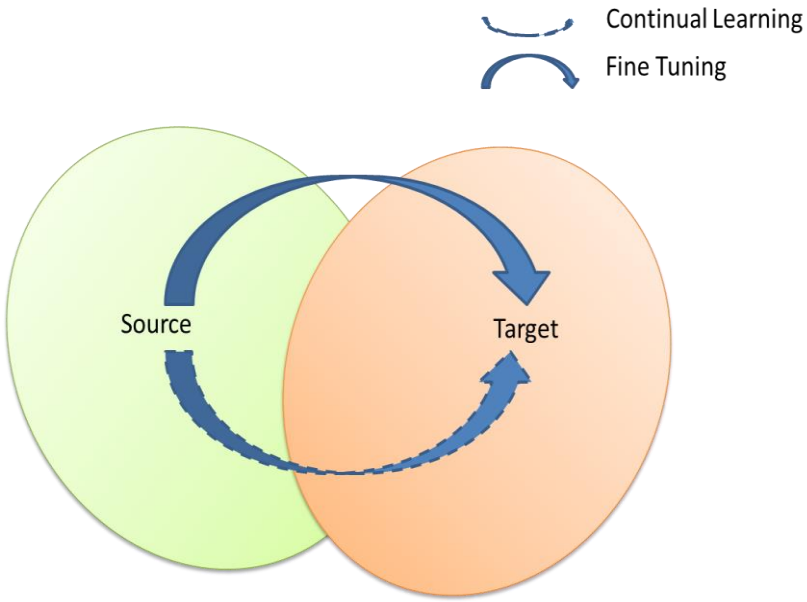



**Figure 1.** Difference between Continual Learning and Fine Tuning

When compared to different types of fake audio, the feature distribution is more consistent in genuine speech while working with the classification task for genuine or fake in positive sample alignment. The difference between continual learning and fine tuning is represented in Fig.1. The new spoofing data is inconsistent with the previous spoofing data in the new spoofing scenario. However, there is a similarity in various situations while comparing the features of genuine speech.

For genuine samples, from the original model, more knowledge can be gathered by focusing more on the genuine sample while being introduced to new data. Among various data resources, the feature distribution of genuine speech is constrained by the PSA while the fake audio detection characteristics are considered. The genuine/positive embeddings between the original and current model are evaluated using the cosine distance.

## 4. Experiments and Results

The physical access (PA) and logical access (LA) subsets of the ASV spoof 2019 dataset are used for the purpose of experimentation. Speech synthesis and voice conversion spoofing attacks of 19 different varieties are available in the LA subset. Nine unique replay and 27 unique acoustic configurations are available in the PA subset. In both PA and LA subsets, five spoofing attacks are selected separately such that they are completely different from each other.

This series of unseen spoofing attacks are used for evaluating the performance of the proposed model. The subsets chosen from the LA subset are named L1, L2, L3, L4 and L5. Further, these subsets are classified into evaluation, training and development sets. The subsets chosen from the PA subset are named P1, P2, P3, P4 and P5. The subsets are classified based on their divisions in PA for evaluation, training and development.

Average equal error rate (AEER), equal error rate (EER), False Accept Rate (FAR) and False Reject Rate (FRR) techniques are used for evaluating the performance of the proposed technique. The accuracy of the model is higher when the EER value is lesser. In sequential training, the AEER model is evaluated. Among various spoofing types, the arithmetic mean of EER is evaluated.

Figure 1 represents the comparison of the performance spoofing types when there is an incremental increase. The LA and PA attacks and their corresponding AEER% are represented over sequential training. FT represents the fine-tuning model, STC represents the sequential fine-tuning model, MC represents the multi-condition training model and PT represents the pre-trained model. During the ith training step, the average performance of the current data as well as the previous data under i attacks is given by AEER.

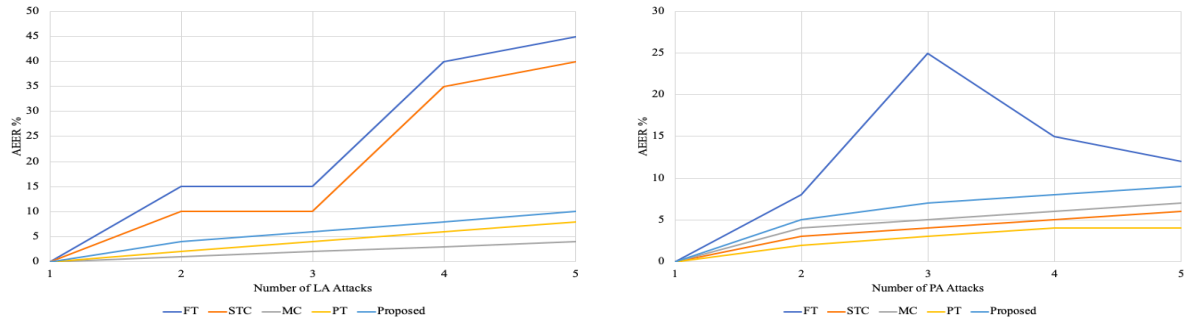
Several experiments are conducted to identify the catastrophic forgetting phenomenon and fine-tune the model with new spoofing attacks. When tested on PA, there is a significant increase in the EER of the base model. The performance of the model improves on the PA

after the fine-tuning process. However, there is a sharp increase in the EER under the LA subset from 5% to 20%.

This occurs due to the mismatch between the PA and LA attack distribution. When old data is used, a performance decrease is observed along with a deviation from the previous optimal parameters when new data is used for fine-tuning the model. This unseen spoofing data scenario is handled with the proposed deep-learning-based DFWF technique. When a gradual occurrence of the unseen spoofing attack is observed the benefits of the DFWF technique are more significant.

The PA and LA attack sequential training results show that with the increase in the number of spoofing attacks, there is a degradation in the performance of the fine-tuned model during the sequential training from L1 to L5. There is a dramatic increase in the AEER value of the fine-tuned model from 0% to 45%. However, for the deep learning-based DFWF model, the AEER value remains at 8%.When compared to the fine-tuning task, there is a relative reduction in the value of deep learning-based DFWF by a value of 80%.

An AEER value of around 7% is achieved by the deep learning-based DFWF model during sequential training from P1 to P5. This value is close to that of the MC training model around the last step. With the fine-tuning strategy, an obvious fluctuation is observed in the trend of AEER. Among various spoofing attacks, the similarity observed can be the reason for these fluctuations. Some amount of previous knowledge can be retained while training with current data if there is some similarity between the features of the previous spoofing types and new spoofing types.




(a) (b)

**Figure 1.** Comparison of spoofing types and their performance under (a) LA attacks, and (b) PA attacks

In order to analyse the effect of two components used in a methodology an ablation study is carried out. LwF is the first constraint factor that is incorporated on input data with output distribution of the model. PSA is the next constraint that works on speech input which is genuinely embedded. However both the factors are incorporated to retain key parameters that correlate with the original model. Thus training that involves PSA and LwF similar to Deep learning with DFWF while training that doesn't involve both PSA and LwF corresponds to fine tuning. In this work every setting takes into consideration optional hyperparameters. The result thus obtained indicates that on sequential training the two components can be used together to improve performance of the system. When they are used together optimal average EER is obtained with deep learning based DFWF. However we are yet to carry out research on multi condition training where the proposed methodology does not meet the expected level of outcome.

The primary reason behind this is that please data remains inaccessible with DFWF, while the multi condition model holds access to all data. According to the textbook the upper limit of performance can be determined with the results of the multi condition continual learning system. It is observed that it is not possible to differentiate between the PA spoofing speech because of the mixing in genuine speech and PA spoofing speech. On fine-tuning, it is possible to distinguish between the PA attacks. However, the LA attacks that previously existed will remain spread out over the genuine speech. On apply deep learning based DFWF, it is possible to differentiate between the PA and LA attacks in a more professional manner resulting in accuracy.

## 5. Conclusion

During fake audio detection, the unseen spoofing attacks are resolved with the deep learning based DFWF technique which works on the principle of continuous learning. Based on the previous data, the detection capability of the model is preserved while the embedding of genuine speech is prevented from deviating from the original distribution using PSA based on the LwF. The catastrophic forgetting problem can be mitigated while the computing resource and time can be saved using this technique. In certain scenarios, as previous data need not be accessed, the training is faster and similar to multi-condition training. With the increase in the spoofing data, the effectiveness of the deep learning based DFWF technique also increases in an incremental manner to learn new spoofing attacks. This is validated by the experimental results observed from the tests conducted on the different settings of

sequential training. Future work is directed towards exploring advanced techniques for continual learning in fake audio detection.

## References

[1] Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., & Narayanan, S. (2022). A review of speaker diarization: Recent advances with deep learning. Computer Speech & Language, 72, 101317.

[2] Adeel, A., Gogate, M., & Hussain, A. (2020). Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments. Information Fusion, 59, 163-170.

[3] Qawaqneh, Z., Mallouh, A. A., & Barkana, B. D. (2017). Deep neural network framework and transformed MFCCs for speaker's age and gender classification. Knowledge-Based Systems, 115, 5-14.

[4] Subramanian, A. S., Weng, C., Watanabe, S., Yu, M., & Yu, D. (2022). Deep learning based multi-source localization with source splitting and its effectiveness in multi-talker speech recognition. Computer Speech & Language, 75, 101360.

[5] Jacob, I. J., & Darney, P. E. (2021). Design of deep learning algorithm for IoT application by image based recognition. Journal of ISMAC, 3(03), 276-290.

[6] Middya, A. I., Nag, B., & Roy, S. (2022). Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. Knowledge-Based Systems, 244, 108580.

[7] Kumar, T. S., & Senthil, T. (2021). Construction of hybrid deep learning model for predicting children behavior based on their emotional reaction. Journal of Information Technology, 3(01), 29-43.

[8] Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. Neurocomputing, 234, 11-26.

[9] Fernandez-Lopez, A., & Sukno, F. M. (2018). Survey on automatic lip-reading in the era of deep learning. Image and Vision Computing, 78, 53-72.

[10] Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, 101894.

[11] Chen, J. I. Z., & Hengjinda, P. (2021). Early prediction of coronary artery disease (CAD) by machine learning method-a comparative study. Journal of Artificial Intelligence, 3(01), 17-33.

[12] Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, 101894.

[13] Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, 101894.

[14] Guo, J., Xu, N., Qian, K., Shi, Y., Xu, K., Wu, Y., & Alwan, A. (2018). Deep neural network based i-vector mapping for speaker verification using short utterances. Speech Communication, 105, 92-102.

## Author's biography

**Kottilingam Kottursamy** is currently working as an associate professor in the Department of CSE, SRM Institute of Science and Technology, Kattankaluthur, Chennai, India. His area of research includes bio statistics, machine learning algorithms, database tuning, high speed networks, information sciences & analytics.