DO YOU REMEMBER? OVERCOMING CATASTROPHIC FORGETTING FOR FAKE AUDIO DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

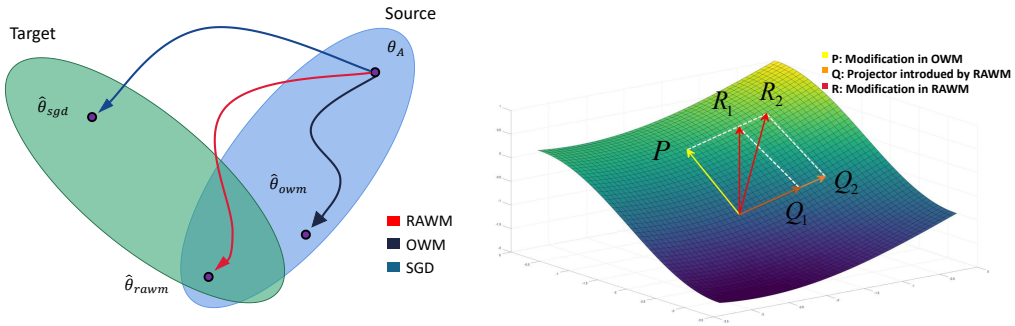
Current fake audio detection algorithms achieve promising performances on most datasets. However, their performance may be significantly degraded when dealing with audio of a different dataset. The orthogonal weight modification to overcome catastrophic forgetting does not consider the similarity of some audio, including fake audio obtained by the same algorithm and genuine audio, on different datasets. To overcome this limitation, we propose a continual learning algorithm for fake audio detection to overcome catastrophic forgetting, called Regularized Adaptive Weight Modification (RAWM). Specifically, when fine-tuning a detection network, our approach adaptively computes the direction of weight modification according to the ratio of genuine utterances and fake utterances. The adaptive modification direction ensures the network can detect fake audio on the new dataset while preserving its knowledge of previous model, thus mitigating catastrophic forgetting. In addition, orthogonal weight modification of fake audios in the new dataset will skew the distribution of inferences on audio in the previous dataset with similar acoustic characteristics, so we introduce a regularization constraint to force the network to remember this distribution. We evaluate our approach across multiple datasets and obtain a significant performance improvement on cross-dataset experiments.

1 INTRODUCTION

Currently, fae audio detection has attracted increasing attention since the organization of a series of challenges, such as the ASVspoof Challenge (Wu et al., 2015; Kinnunen et al., 2017; Todisco et al., 2019; Yamagishi et al., 2021) and the Audio Deep Synthesis Detection Challenge (ADD) (Yi et al., 2022). In these competitions, deep neural networks have achieved great success. Currently, large-scale pre-trained models have gradually been applied to fake audio detection and achieved state-of-the-art results on several public fake audio datasets (Tak et al., 2022; Martín-Doñas & Álvarez, 2022; Lv et al., 2022; Wang & Yamagishi, 2021). Although fake audio detection achieves promising performance, it may be significantly degraded when dealing with audio of another dataset. The diversity of audio proposes a significant challenge to fake audio detection across datasets (Zhang et al., 2021b;a).

Some approaches have been proposed to improve detection performance across datasets. Monteiro et al. (2020) proposed an ensemble learning method to improve the detection ability of the model for unseen audio. Wang et al. (2020) designed a dual-adversarial domain adaptive network to learn more generalized features for different datasets. Both methods require some audio from the old dataset, but in some practical situations, it is almost impossible to obtain the old data. Zhang et al. (2021b) proposed a data augmentation method to extract more robust features for detection across datasets, which is only suitable for the datasets with similar feature distribution. Ma et al. (2021) proposed the first continual learning method for fake audio detection, called Detecting Fake Without Forgetting (DFWF) inspired by Learning without Forgetting (LwF) (Li & Hoiem, 2017). The DFWF improves the detection performance by fine-tuning on the new dataset and overcomes catastrophic forgetting by introducing regularization. However, the regularization degrades the performance of the DFWF on the new dataset.

To address these issues, we propose a continual learning approach, called Regularized Adaptive Weight Modification (RAWM). Because the genuine audio has more similar feature distribution



(a) (b)

Figure 1: Schematic of SGD, OWM, and RAWM. **(a)**, With RAWM, the optimization process searches for configurations that lead to great performance on bath dataset 1 (blue area) and dataset 2 (green area). A successful optimized configuration $\hat{\theta}_{rawm}$ stops inside the overlapping subspace. However, the configuration $\hat{\theta}_{sgd}$ obtained by SGD is optimized without considering forgetting, and the configuration $\hat{\theta}_{owm}$ obtained by orthogonal weight modification (OWM) can not reach to the region of dataset 2. **(b)**, the RAWM adaptively modify weight direction by introducing a projector that is orthogonal to the projector $\boldsymbol{P}$ proposed by OWM.

than the fake audio on different datasets, we hope that the parameter updating direction of genuine audio in the new dataset is similar to that of genuine audio in the old dataset, and minimize the interference of the fake audio to the previously learned knowledge. Our approach can adaptively modify the direction of weights according to the ratio of genuine audio and fake audio of each batch in the process of fine-tuning. Specifically, if the proportion of fake audio is larger, the modified direction is closer to the orthogonal projector of the subspace spanned by all previous input; if the proportion of genuine audio is larger, the modification is closer to the previous input subspace. For the fake audio of the new dataset, we modify the model weights like the orthogonal weight modification (OWM) (Zeng et al., 2019), which makes the new fake audio not interfere with the learned knowledge and overcomes catastrophic forgetting. However, when the genuine audio of the new and old datasets is quite different, the effect of the above method is not obvious. By introducing regularization constraints, our approach forces the model to automatically remember the inference distribution without requiring prior knowledge of datasets. The optimization process of RAWM is compared with that of the Stochastic Gradient Descent search (SGD) and OWM in Figure 1a. In addition, compared with the experience-replay-based method in continuous learning, RAWM does not require previous data, which makes this method suitable in most situations.

Contributions: We propose a regularized adaptive weight modification algorithm to overcome catastrophic forgetting for fake audio detection. We demonstrate the process of computing the modified direction of weights adaptively (Sec. 4.1). Second, we introduce regularization to the case that genuine audio with different feature distribution between new and old datasets. (Sec. 4.2). The experimental results show that our proposed method outperforms several continual learning methods in learning new knowledge and overcoming forgetting, including Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017), LwF, OWM, and DFWF.

2 Related Work

In continual learning, overcoming catastrophic forgetting methods can be divided into the following categories. The regularization methods perform a regularization on the objection function or enforce the important weights of previous tasks (Kinnunen et al., 2017; Zenke et al., 2017b; Aljundi et al., 2018; 2019; Mallya & Lazebnik, 2018; Serra et al., 2018). The dynamic architecture methods reserve their previous knowledge by introducing additional layers or nodes and growing model architecture (Rusu et al., 2016; Schwarz et al., 2018); (Yoon et al., 2017). The memory-based methods remember their previous data to be used later to prevent gradient updates from deviating from their previous tasks (Lopez-Paz & Ranzato, 2017; Castro et al., 2018; Wu et al., 2019; Lee et al., 2019). The Bayesian inference methods compute posterior distribution by multiplying the previous posterior by the likelihood of the dataset belonging to the new task (Nguyen et al., 2017). The natural

gradient descent methods approximate the Fisher information matrix in EWC using the generalized Gauss-Newton method to fast gradient descent (Tseran et al., 2018; Chen et al., 2019). The continual learning methods, such as the EWC measures the importance of weights by Fisher matrix, the LwF introduces regularization constraints to the model weights and the OWM modifies the direction of weights orthogonal to the previous input subspace, have achieved great success in image classification (Zeng et al., 2019; Kirkpatrick et al., 2017), object detection (Perez-Rua et al., 2020), semantic segmentation (Cermelli et al., 2020), lifelong language learning (de Masson D'Autume et al., 2019) and sentence representation (Liu et al., 2019).

However, the approximation of the regularization methods, such as EWC and LwF, will produce error accumulation (Zenke et al., 2017a; Huszár, 2017; Ma et al., 2021). In contrast, our proposed method only needs the current inputs and projector for the last task, thus reducing the memory decline in long task sequences. Compared with the OWM algorithm, our method can adaptively modify the direction according to the type of audio, thus improving the performance on new datasets and reducing forgetting. For DFWF, we relax the regularization constraint and introduce a direction modification, thus achieving better performance.

3 BACKGROUND

3.1 ORTHOGONAL WEIGHT MODIFICATION

The OWM algorithm overcomes catastrophic forgetting by modifying the direction of weights on the new task. The modified direction $\boldsymbol{P}$ in the OWM is orthogonal to the subspace spanned by all inputs of the previous task. The standard algorithm for orthogonal projector requires the matrix consists of all previously trained input, which is inaccessible in most continual learning situations. To address this issue, the OWM constructs the orthogonal projector by an iterative method similar to the RLS algorithm (Shah et al., 1992) as Eq 1.

$$\begin{aligned} \boldsymbol{P}_l(i,j) &= \boldsymbol{P}_l(i-1,j) - \mathbf{k}_l(i,j)\overline{\mathbf{x}}_{l-1}(i,j)^T \boldsymbol{P}_l(i-1,j) \\ \mathbf{k}_l(i,j) &= \frac{\boldsymbol{P}_l(i-1,j)\overline{\mathbf{x}}_{l-1}}{\alpha + \overline{\mathbf{x}}_{l-1}(i,j)^T \boldsymbol{P}_l(i-1,j)\overline{\mathbf{x}}_{l-1}(i,j)} \end{aligned} \tag{1}$$

The $\overline{\mathbf{x}}_l(i,j) \in \mathbb{R}^s$ represents the output of the lth layer in response to the mean of the ith batch inputs on jth dataset, α is a hyperparameter decaying with the number of tasks. This iterative algorithm only needs the current inputs and orthogonal projector for the last task, thus avoiding loading data from the previous task.

3.2 LEARNING WITHOUT FORGETTING

The LwF algorithm is proposed for continual learning based on the idea of model distillation, where the previous knowledge is viewed as a soft label to constitute a penalty term. In learning a new task, the training of the model consists of two stages. In the first stage, the shared parameters and the output part parameters of the previous task are frozen, and the parameters of the output part of the new task are randomly initialized. The model is trained on new data until loss convergence. In the second stage, the parameters of the previous task are reserved. All parameters in the network are trained jointly with the distillation loss function. The loss function in the second stage can be written as:

$$\boldsymbol{L}_{lwf} = \lambda_0 \boldsymbol{L}_{old}(\mathrm{y}_o, \hat{\mathrm{y}}_o) + \boldsymbol{L}_{new}(\mathrm{y}_n, \hat{\mathrm{y}}_n) + \boldsymbol{R}_{lwf}(\hat{\theta}_s, \hat{\theta}_o, \hat{\theta}_n) \tag{2}$$

where λ_0 is a ratio coefficient; $\mathrm{y_o}$ and $\mathrm{y_o}$ are the old and new ground truth, respectively; $\hat{\mathrm{y}}_o$ and $\hat{\mathrm{y}}_n$ are the predicted output of the old and new model, respectively. $\hat{\theta}_s$, $\hat{\theta}_o$ and $\hat{\theta}_n$ represent the train parameters of the shared part, the output part of the old task, and the output part of the new task, respectively. All parameters are constrained by the regularization term $\boldsymbol{R}_{lwf}$.

4 PROPOSED METHOD

We propose a continual learning method, called Regularized Adaptive Weight Modification (RAWM), to overcome forgetting for fake audio detection. Our method is inspired by OWM, which

was proposed to overcome catastrophic forgetting and has achieved promising performance in image classification.

On most fake audio datasets, genuine audio is concentrated in one region, and its feature distribution is relatively concentrated and unified. Under the same acoustic conditions, the feature distribution of genuine audio has a smaller variance than that of fake audio (Ma et al., 2021; Yan et al., 2022). This situation is similar to the sample replay in the experience-replay-based method, a continual learning method to overcome catastrophic forgetting by replaying previous training data. Those utterances with similar feature distribution between datasets can be regarded as the samples reserved from the old dataset and replayed on the new dataset. In the training process on the new dataset, they should be trained without any modification. In this view, modifying the direction of weights orthogonally is inappropriate for utterances in the new dataset that could potentially conform to the old dataset's feature distribution. To address this issue, we soften the modification direction of OWM to be closer to the previous inputs subspace. Specifically, we introduce a projector $\boldsymbol{Q}$ orthogonal to the modification direction $\boldsymbol{P}$ proposed by OWM. The norm of $\boldsymbol{Q}$ is variable on the ratio β of genuine audio and fake audio in a batch of data. The adaptive modification direction $\boldsymbol{R}$ is shown in Figure 1, which is variable to the norm of $\boldsymbol{Q}$. The norm of projector $\boldsymbol{Q}$ increases with the number of genuine audio increasing, resulting in the modification direction $\boldsymbol{R}$ being closer to the subspace spanned by previous inputs. On the contrary, the modification direction is closer to the direction of the projector $\boldsymbol{P}$, thus reducing the impact on learned knowledge and overcoming catastrophic forgetting.

However, there are also some datasets whose feature distribution of genuine audio is quite different from that of other datasets (Müller et al., 2022). When they are selected as new datasets, modifying the direction according to the method mentioned above will skew the inference distribution of genuine audio in the old dataset. To address this issue, we introduce a regularization term to force the new distribution of inference to be similar to the old one. This method does not require prior knowledge of the dataset and reduces forgetting of the learned knowledge.

Forward propagate the input of the i batch in the j dataset, the regularized adaptive weight modification can be written as:

$$\begin{aligned} \boldsymbol{W}_l(i,j) &= \boldsymbol{W}_l(i-1,j) + \gamma(i,j)\Delta\boldsymbol{W}_l^{BP}(i,j) && when\ j=1 \\ \boldsymbol{W}_l(i,j) &= \boldsymbol{W}_l(i-1,j) + \gamma(i,j)((1-\eta)\boldsymbol{R}_l(j-1)\Delta\boldsymbol{W}_l^{BP}(i,j) + \eta\Delta\boldsymbol{W}_{l_{reg}}^{BP}(i,j)) && when\ j>1 \end{aligned} \tag{3}$$

where $\boldsymbol{W}_l(i,j) \in \mathbb{R}^{s\times v}$ represents the connection weights between the lth layer and the $(l+1)$th layer; γ represents the learning rate of this network; $\Delta\boldsymbol{W}_l^{BP}(i,j)$ and $\Delta\boldsymbol{W}_{l_{reg}}^{BP}(i,j)$ represents the standard backpropagate weight modification introduced by cross-entropy loss and regularization, respectively; $\boldsymbol{R}$ represents the modification direction in our method. The η is a coefficient measuring the attention degree of the old task.

The adaptive direction modification (AWM) is described in Section 4.1 and the regularization (REG) is described in Section 4.2. We show the process of our algorithm in Algorithm 1.

4.1 Adaptive Weight Modification

We consider a feed-forward network consisting of $L+1$ layers, indexed by $l = 0, 1, \cdots, L$ with the same activation function $g(\cdot)$. We first introduce a projector $\boldsymbol{Q}$ that is orthogonal to the weights modification direction $\boldsymbol{P}$ proposed by the OWM algorithm. The orthogonal projector can be written as Eq 4:

$$\boldsymbol{Q} = \beta[\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}] \tag{4}$$

where the projector $\boldsymbol{P}$, which is orthogonal to the subspace spanned by all previous inputs, can be calculated as Eq 1 The construction of the orthogonal projector $\boldsymbol{Q}$ is mathematically sound (Haykin, 2002; Ben-Israel & Greville, 2003; Bengio & LeCun, 2007). To verify the modification direction according to the ratio of genuine audio and fake audio, we introduce the β defined as:

$$\beta = \frac{\mathrm{N_g}+1}{\mathrm{N_f}+1} \tag{5}$$

in which $\mathrm{N_g}$ and $\mathrm{N_f}$ represent the number of genuine and fake audios in a batch, respectively. By adding one to both the numerator and denominator, β can be calculated when all the batch audios are genuine. As illustrated in Eq 4, the norm of projector $\boldsymbol{Q}$ is proportional to the ratio β. Our approach defines the modified direction $\boldsymbol{R}$ of weights as:

$$\boldsymbol{R} = \boldsymbol{P}_N + \mathrm{m}\boldsymbol{Q}_N \tag{6}$$

Algorithm 1 Regularized Adaptive Weight Modification

Require: Training data from different datasets, γ (learning rate), m (constant hyperparameter), T_{reg} (constant hyperparameter).

for every dataset *j* **do**
  repeat
  for every batch *i* **do**
    repeat
    if $j = 1$ **then**
      $\boldsymbol{W}_l(i,j) = \boldsymbol{W}_l(i-1,j) + \gamma(i,j)\Delta\boldsymbol{W}_l^{BP}(i,j)$ ▷ $\Delta\boldsymbol{W}_l^{BP}$ gradient by standard BP method
    else
      $\mathbf{k}(i,j) = \boldsymbol{P}_l(i-1)\overline{\mathbf{x}}_{l-1}(i,j)/[\alpha + \overline{\mathbf{x}}_{l-1}(i,j)^T\boldsymbol{P}_l(i-1,j)\overline{\mathbf{x}}_{l-1}(i,j)]$
      $\boldsymbol{P}_l(i,j) = \boldsymbol{P}_l(i-1,j) - \mathbf{k}(i,j)\overline{\mathbf{x}}_{l-1}(i,j)^T\boldsymbol{P}_l(i-1,j)$ ▷ $\overline{\mathbf{x}}_l$ output of the mean of inputs
      $\beta = \dfrac{\mathrm{N_g}+1}{\mathrm{N_f}+1}$ ▷ $\mathrm{N_g}$, $\mathrm{N_f}$: genuine, fake audio number
      $\boldsymbol{Q} = \beta[\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}]$
      $\boldsymbol{P}_N = \dfrac{\boldsymbol{P}}{||\boldsymbol{P}||}$
      $\boldsymbol{Q}_N = \dfrac{\boldsymbol{Q}}{||\boldsymbol{I}-\boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}||}$
      $\boldsymbol{R} = \boldsymbol{P}_N + \mathrm{m}\boldsymbol{Q}_N$
      $\hat{y}_o(i) = \frac{y_o(i)^{1/T_{reg}}}{\sum \mathbf{y}_o(i)^{1/T_{reg}}}$
      $\hat{y}_n(i) = \frac{y_n(i)^{1/T_{reg}}}{\sum \mathbf{y}_n(i)^{1/T_{reg}}}$
      $\Delta\boldsymbol{W}_{l_{reg}}^{BP} = -\nabla(\hat{\mathbf{y}}_o(i)\cdot\log\hat{\mathbf{y}}_n(i))$
      $\boldsymbol{W}_l(i,j) = \boldsymbol{W}_l(i-1,j) + \gamma(i,j)((1-\eta)\boldsymbol{R}_l(j-1)\Delta\boldsymbol{W}_l^{BP}(i,j) + \eta\Delta\boldsymbol{W}_{l_{reg}}^{BP}(i,j))$
      end if
    until loss plateaus
  end for
end for

$$\boldsymbol{P}_N = \frac{\boldsymbol{P}}{||\boldsymbol{P}||}, \quad \boldsymbol{Q}_N = \frac{\boldsymbol{Q}}{||\boldsymbol{I}-\boldsymbol{P}(\boldsymbol{P}^T\boldsymbol{P})^{-1}\boldsymbol{P}||} \tag{7}$$

where m is a constant to constrain the norm of projector $\boldsymbol{Q}$ to prevent gradient explosion or gradient vanishing in the backward process; $\boldsymbol{P}_N$ and $\boldsymbol{Q}_N$ are identity projectors normalized by $\boldsymbol{P}$ and $\boldsymbol{Q}$, respectively. Normalization is to prevent the case that the change of β has little effect on the modified direction because of the large norm gap between $\boldsymbol{P}$ and $\boldsymbol{Q}$. In the back-propagate process, the direction of network weights is modified as Eq 8.

$$\begin{aligned} \boldsymbol{W}_l(i,j) &= \boldsymbol{W}_l(i-1,j) + \gamma(i,j)\Delta\boldsymbol{W}_l^{BP}(i,j) \qquad when\ j=1 \\ \boldsymbol{W}_l(i,j) &= \boldsymbol{W}_l(i-1,j) + \gamma(i,j)\boldsymbol{R}_l(j-1)\Delta\boldsymbol{W}_l^{BP}(i,j) \qquad when\ j>1 \end{aligned} \tag{8}$$

4.2 Regularization

The method mentioned above considers modification of genuine audio for it has similar feature distribution on different datasets. In this section, we introduce a regularization item to force the model to remember the previous inference distribution, thus reducing the forgetting when the feature distributions of genuine audio between old and new datasets have great differences. We consider two feed-forward networks. Both the teacher and student models are trained on the old dataset with the same parameters $\boldsymbol{\theta}_t$, $\boldsymbol{\theta}_s$. The parameter of the teacher model $\boldsymbol{\theta}_t$ is frozen in the process of training on the new dataset and the parameter of the student model $\boldsymbol{\theta}_s$ is fine-tuned. We normalize the forward output $\mathbf{y}_t(i)$, $\mathbf{y}_s(i)$ on the ith batch and force the student model to remember the previous knowledge by optimizing the loss function as:

$$\boldsymbol{L}_{reg}(\hat{\mathbf{y}}_o(i), \hat{\mathbf{y}}_n(i)) = -\hat{\mathbf{y}}_o(i)\cdot\log\hat{\mathbf{y}}_n(i) \tag{9}$$

$$\hat{y}_o(i) = \frac{y_o(i)^{1/T_{reg}}}{\sum \mathbf{y}_o(i)^{1/T_{reg}}}, \quad \hat{y}_n(i) = \frac{y_n(i)^{1/T_{reg}}}{\sum \mathbf{y}_n(i)^{1/T_{reg}}} \tag{10}$$

where T_{reg} is a constant hyperparameter. The weight modification of this regularization can be written as Eq 11.

$$\Delta\boldsymbol{W}_{l_{reg}}^{BP} = \nabla\boldsymbol{L}_{reg} \tag{11}$$

This modification $\Delta \boldsymbol{W}_{l_{reg}}^{BP}$ is fused with the adaptive modification $\Delta \boldsymbol{W}_{l}^{BP}$ as Eq 3 in RAWM.

5 EXPERIMENTS

5.1 DATASETS

We conduct our experiments on four fake audio datasets. The models are trained using the training set of the ASVspoof2019LA dataset and are evaluated using some of the evaluation sets, including ASVspoof2019LA, ASVspoof2015, VCC2020, and In-the-Wild.

ASVspoof 2019 LA Dataset (Todisco et al., 2019) is the sub-challenge dataset (30 males and 37 females) containing three subsets: training, development, and evaluation. The training set and development share the same attack including four TTS and two VC algorithms. The bonafide audio is collected from the VCTK corpus (Veaux et al., 2017). The evaluation set contains totally different attacks.

ASVspoof2015 dataset (Wu et al., 2015) is an open-source standard dataset of genuine and synthetic speech in the ASVspoof2015 challenge. The genuine speech was recorded from 106 speakers (45 males and 61 females) with no significant channel or background noise effects. The spoofing speech is generated using a variety of speech synthesis and voice conversion algorithms.

In-the-Wild dataset (Müller et al., 2022) contains a set of deep fake audio (and corresponding real audio) of 58 politicians and other public figures collected from publicly available sources, such as social networks and video streaming platforms. In total, 20.8 hours of genuine audio and 17.2 hours of fake audio were collected. On average, each speaker had 23 minutes of genuine audio and 18 minutes of fake audio.

VCC2020 dataset (Zhao et al., 2020) is collected from Voice Conversion Challenge 2020. This dataset contains two subsets: a set of genuine audio provided by organizers and a set of fake audio provided by participating teams. Different from the previous three datasets, VCC2020 is a multilingual fake audio dataset, including English, Finnish, German and Mandarin.

We divide the genuine and fake audios of the VCC2020 dataset into four subsets. A quarter is used to build the evaluation set, a quarter to build the development set, and the rest to be used as the training set. The In-the-Wild dataset is divided in the same way as the VCC2020. The ASVspoof2015 is the most similar to the ASVspoof2019LA for their audios are collected from the same datasets or conversion algorithms. The audio of the In-the-Wild dataset is collected from the real world and the audio of the VCC2020 dataset is multilingual. The detailed statistics of the datasets are presented in Table 1. The Equal Error Rate (EER), which is widely used for fake audio detection and speaker verification, is applied to evaluate the experimental performance.

5.2 EXPERIMENTAL SETUP

Fake audio detection Model: We use the pre-trained model Wav2vec 2.0 (Baevski et al., 2020) as the feature extractor and the self-attention convolutional neural network (S-CNN) as the classifier. The parameters of Wav2vec 2.0 is loaded from the pre-train model XLSR-53 (Conneau et al., 2020), which is trained by a large-scale multilingual database. The classifier S-CNN contains three 1D-Convolution layers, one self-attention layer, and two full connection layers, according to the forward process. The input dimension of the first convolution layer is 256 and the hidden dimension of all convolution layers is 80. The kernel size and stride are set to 5 and 1, respectively. The hidden dimension of all full connection layers is 80 and the output dimension of the last is 2.

Training Details: We fine-tune the model weights including the pre-trained model XLSR-53 and the classifier S-CNN. All of the parameters are trained by the Adam optimizer with a batch size of 2 and a learning rate γ of 0.0001. The constant m and T_{reg} in RAWM are set to 0.1 and 2, respectively. The α is initialized to 0.00001 for convolution layers, 0.0001 for the self-attention layer, and 0.1 for full connection layers. The norm in normalization of projector $\boldsymbol{P}$ and $\boldsymbol{Q}$ is the $\boldsymbol{L}^2$ norm. All results are (re)produced by us and averaged over 7 runs with standard deviations.

Table 1: Statistics of experimental datasets.

Dataset	ASVSpoof2019 ($\mathbf{S}$)		ASVSpoof2015 ($\mathbf{T}_1$)		VCC2020 ($\mathbf{T}_2$)		In-the-Wild ($\mathbf{T}_3$)	
	#Real	#Fake	#Real	#Fake	#Real	#Fake	#Real	#Fake
Train	2,580	22,800	3,750	12,625	1,330	3,060	9,431	5,908
Dev	2,548	22,296	3,497	49,875	665	1,530	4,715	2,954
Eval	7,355	63,882	9,404	184,000	665	1,530	4,717	2,954

Table 2: The EER(%) of our baseline on multiple evaluation sets.

Model	$\mathbf{S}$	$\mathbf{T_1}$	$\mathbf{T_2}$	$\mathbf{T_3}$
Baseline	0.258	24.532	46.503	91.473

5.3 BASELINE

In this section, we train our model on the training set of the ASVspoof2019LA dataset with increasing training steps. Table 2 shows the detection performance on multiple evaluation sets. The detection performance of Model-4 is very close to the state-of-the-art result for fake audio detection in the same dataset (Nautsch et al., 2021). For this reason, we select Model-4 as our baseline. Although the model achieves promising detection performance on the evaluation set of the ASVspoof2019LA, its detection accuracy degrades significantly on other datasets. The models have achieved the lowest cross-dataset detection EER on the ASVspoof2015 dataset among three unseen datasets, which verifies our view that the fake audio generated by the same algorithms has similar feature distribution, while the feature distribution of fake audio generated by different algorithms is quite different. This difference damages the cross-dataset detection performance. For more results, see Table 7 in appendix.

5.4 THE EFFECTIVENESS OF THE η FOR OUR METHOD

Sequence training between two datasets: We perform some experiments to evaluate the effectiveness of η in RAWM, which represents the attention degree to learned knowledge, in Table 3. We can easily observe that the RAWM achieves great performance on new datasets and overcomes catastrophic forgetting on previous datasets, especially in the experiment on $\mathbf{S} \rightarrow \mathbf{T}_1$. By comparing the experimental results of three cross-datasets, we observe that when the new and old datasets have similar feature distribution (Table 3a), the performance of learning new knowledge and overcoming forgetting improves with the increasing of η ($\eta < 1$); When the feature distribution of the new and previous datasets is greatly different (Table 3b, Table 3c), the effectiveness of the RAWM achieves the best when $\eta = 0.50$. In addition, the EER is increased when $\eta < 0.50$, which shows that the regularization is conducive to performance on new dataset and overcoming forgetting.

Sequence training on four datasets: We also present the EER of our method with different η on multiple evaluation sets, in Table 4a. We observe that our method significantly reduces the degradation of the detection performance across datasets. The RAWM achieves the lowest EER on these results when $\eta = 0.50$, which demonstrates that in the scenario of continual learning on multiple datasets, the same attention degree to the new and old datasets is more conducive to learning and overcoming forgetting. In addition, the result on dataset $\mathbf{T_3}$ shows that the RAWM with smaller η has better performance on the new dataset; The results on $\mathbf{S}$, $\mathbf{T_1}$ and $\mathbf{T_2}$ show that the model with larger η is more effective in overcoming forgetting.

5.5 ABLATION STUDIES FOR OUR METHOD

Sequence training between two datasets: In this section, we compare our proposed method with adaptive weight modification without regularization ($-$REG) and orthogonal weight modification without regularization ($-$AWM). Table 5 presents their EER on three evaluation sets. We observe that RAWM achieves similar EER to $-$REG on the new dataset, both of them are superior significantly to $-$AWM, which shows that the adaptive weight modification has a significant positive impact on learning new knowledge, while regularization impacts little. For overcoming forgetting,

Table 3: The EER(%) on evaluation sets of our method with different η. All experiments are trained using the training set in order to $\mathbf{S} \rightarrow \mathbf{T_k}$ and are evaluated using the evaluation set on $\mathbf{S}$ and $\mathbf{T_k}$

(a)

η	$\mathbf{S}$	$\mathbf{T_1}$
Baseline	0.258	24.532
0.00	1.643	0.256
0.20	1.424	0.431
0.25	1.175	0.311
0.50	0.878	0.257
0.75	**0.666**	**0.247**
1.00	3.123	0.343

(b)

η	$\mathbf{S}$	$\mathbf{T_2}$
Baseline	0.258	46.503
0.00	1.413	3.845
0.20	1.334	4.288
0.25	1.275	3.994
0.50	**1.237**	**3.721**
0.75	1.262	4.571
1.00	4.234	4.566

(c)

η	$\mathbf{S}$	$\mathbf{T_3}$
Baseline	0.258	91.473
0.00	4.126	1.457
0.20	3.490	1.848
0.25	2.975	1.593
0.50	**2.038**	**1.425**
0.75	2.482	2.271
1.00	2.453	2.598

Table 4: The EER(%) on four evaluation sets. All experiments are trained using training set in order to $\mathbf{S} \rightarrow \mathbf{T_1} \rightarrow \mathbf{T_2} \rightarrow \mathbf{T_3}$ and are evaluated using evaluation sets.

(a) The EER(%) of the RAWM with different η.

η	$\mathbf{S}$	$\mathbf{T_1}$	$\mathbf{T_2}$	$\mathbf{T_3}$
Baseline	0.258	24.532	46.503	91.473
0.00	1.845	1.127	3.916	1.410
0.20	1.724	1.003	4.120	1.367
0.25	1.699	0.945	4.017	1.529
0.50	**1.508**	**0.641**	**3.850**	**1.163**
0.75	1.636	0.873	3.975	2.454
1.00	2.714	1.621	3.875	2.325

(b) The EER(%) of our method compared with others.

Method	$\mathbf{S}$	$\mathbf{T_1}$	$\mathbf{T_2}$	$\mathbf{T_3}$
Baseline	0.258	24.532	46.503	91.473
Fine-tune	7.068	2.841	5.674	2.543
EWC	5.569	3.444	4.510	2.129
OWM	4.083	2.167	4.480	2.472
LwF	2.714	1.621	3.875	2.325
DFWF	3.476	3.735	7.345	3.114
RAWM(Ours)	**1.508**	**0.641**	**3.850**	**1.163**

when the feature distribution of the new and old datasets is similar (Table 5a), the EER of the $-$REG on the old datasets is much lower than that of the $-$AWM and higher than that of the RAWM, which shows that the adaptive weight modification and regularization can significantly reduce the forgetting of previous knowledge in this case; When the languages of the new and old datasets are different (Table 5b), the EER of RAWM in the previous datasets is similar to that of the $-$REG and much lower than that of the $-$AWM, which also proves that the adaptive weight modification has a significant positive impact on overcoming forgetting, while regularization has little impact in this scenario; When the feature distribution of the new and old datasets is quite different (Table 5c), the EER of the $-$REG is similar to that of the $-$AWM and much higher than that of the RAWM, which shows that in this case, regularization has a great benefit in overcoming forgetting, while the effect of adaptive weight modification is not obvious.

Sequence training on four datasets:

In this section, we present the results of the ablation study on four evaluation sets in Table 5d. We observe that the EER of $-$REG to $-$AWM degrades more obviously than that of RAWM to $-$REG on all evaluation sets, indicating that adaptive weight modification has a more obvious benefit in learning new knowledge and overcoming forgetting than regularization for sequence training on multiple datasets. Compared with sequence training between two data sets (Table 5), the effect of RAWM on overcoming forgetting is not obviously degraded, which shows that our method is effective in overcoming forgetting for long sequence learning.

5.6 Comparison of our method with other methods

Sequence training between two datasets: We compare our method with several methods in Table 6. The EWC, LwF, and OWM as three mainstream continual learning methods achieve great success in image classification, object detection, and other fields. The DFWF is the first continual learning method to overcome forgetting for fake audio detection. The results demonstrate that fine-tuning without modification (Fine-tune) forgets previous knowledge obviously. The forgetting of RAWM is one-tenth that of Fine-tune on Table 6a and the EER on the new dataset of RAWM is also half that of Fine-tune. We observe that EWC and OWM achieve similar performance in three experiments. The performance of LwF outperforms EWC and OWM on the new dataset; The overcoming forgetting of LwF is also superior to that of EWC and OWM when the new and previous datasets have a greatly different feature distribution (Table 6c). The DFWF is more effective in overcoming forgetting than

Table 5: The EER(%) on evaluation sets of the ablation studies. (a), (b) and (c) are trained using the training set in order to $\mathbf{S} \rightarrow \mathbf{T_k}$ and are evaluated using the evaluation set on $\mathbf{S}$ and $\mathbf{T_k}$; (d) is trained in order to $\mathbf{S} \rightarrow \mathbf{T_1} \rightarrow \mathbf{T_2} \rightarrow \mathbf{T_3}$ and are evaluated using evaluation sets.

(a)

Method	**S**	$\mathbf{T_1}$
RAWM	**0.666**	**0.247**
−REG	1.643	0.256
−AWM	2.448	0.500

(b)

Method	**S**	$\mathbf{T_2}$
RAWM	**1.237**	**3.721**
−REG	1.413	3.845
−AWM	3.086	5.432

(c)

Method	**S**	$\mathbf{T_3}$
RAWM	**2.038**	**1.425**
−REG	4.126	1.457
−AWM	4.857	2.663

(d)

Method	**S**	$\mathbf{T_1}$	$\mathbf{T_2}$	$\mathbf{T_3}$
RAWM	**1.508**	**0.641**	**3.850**	**1.163**
−REG	1.845	1.127	3.916	1.410
−AWM	4.083	2.167	4.480	2.472

Table 6: The EER(%) of our method compared with various methods. All experiments are trained using the training set in order to $\mathbf{S} \rightarrow \mathbf{T_k}$ and are evaluated using the evaluation set on $\mathbf{S}$ and $\mathbf{T_k}$

(a)

Method	**S**	$\mathbf{T_1}$
Baseline	0.258	24.532
Fine-tune	7.324	0.510
EWC	2.832	0.500
OWM	2.448	0.500
LwF	3.123	0.343
DFWF	1.849	0.689
RAWM(Ours)	**0.666**	**0.247**

(b)

Method	**S**	$\mathbf{T_2}$
Baseline	0.258	46.503
Fine-tune	8.755	5.647
EWC	3.494	5.289
OWM	3.086	5.432
LwF	4.234	4.566
DFWF	1.874	7.355
RAWM(Ours)	**1.237**	**3.721**

(c)

Method	**S**	$\mathbf{T_3}$
Baseline	0.258	91.473
Fine-tune	20.976	2.679
EWC	5.039	2.615
OWM	4.857	2.663
LwF	2.453	2.598
DFWF	5.324	3.275
RAWM(Ours)	**2.038**	**1.425**

the EWC, OWM, and LwF, but its performance on the new dataset is inferior to the three continual learning methods. Compared with others, our method achieves lower EER on new and previous datasets of all experiments, which demonstrates that RAWM has promising learning and overcoming forgetting performance on sequence training between two datasets, regardless of whether the new and old datasets have similar feature distributions (Table 6a, Table 6b) or the same languages (Table 6c).

Sequence training on four datasets: Finally, We compare our method with several continual learning methods for sequence training on four datasets in Table 4b. We observe that the performance of DFWF is inferior to others on the last three datasets. This result shows that the learning effect of DFWF is not effective as other continual learning methods. Most methods achieve lower EERs than fine-tuning, and the best result of overcoming forgetting and continual learning is our proposed RAWM algorithm. This indicates that our method outperforms other methods not only in sequence training between two datasets but also in sequence training on multiple datasets.

6 Conclusion

In this work, we propose a continual learning algorithm for fake audio detection, called RAWM, that can adaptively modify the weight direction in the training process on the new dataset. Our method overcomes catastrophic forgetting by softening the orthogonal modification proposed by OWM and introducing regularization. The experimental results demonstrate that our method performs better than four continual learning methods in learning new knowledge and overcoming forgetting, including EWC, LwF, OWM, and DFWF. The results also show that our method is effective in overcoming catastrophic forgetting in both scenarios of sequence training between two datasets and multiple datasets. Our method does not require previous data; thus it can be applied to most classification networks. Despite our results, we have yet to study the impact of different acoustic conditions and noise interference on forgetting, and exploring these questions will form the focus of our future studies.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 139–154, 2018.

Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11254–11263, 2019.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

Adi Ben-Israel and Thomas NE Greville. *Generalized inverses: theory and applications*, volume 15. Springer Science & Business Media, 2003.

Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 233–248, 2018.

Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9233–9242, 2020.

Yu Chen, Tom Diethe, and Neil Lawrence. Facilitating bayesian continual learning by natural gradients and stein gradients. *arXiv preprint arXiv:1904.10644*, 2019.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

Cyprien de Masson D'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. Episodic memory in lifelong language learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Simon S Haykin. *Adaptive filter theory*. Pearson Education India, 2002.

Ferenc Huszár. On quadratic penalties in elastic weight consolidation. *CoRR*, abs/1712.03847, 2017. URL http://arxiv.org/abs/1712.03847.

Tomi Kinnunen, Md. Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas W. D. Evans, Junichi Yamagishi, and Kong-Aik Lee. The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In Francisco Lacerda (ed.), *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, pp. 2–6. ISCA, 2017. URL http://www.isca-speech.org/archive/Interspeech_2017/abstracts/1111.html.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Kibok Lee, Kimin Lee, Jinwoo Shin, and Honglak Lee. Overcoming catastrophic forgetting with unlabeled data in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 312–321, 2019.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Tianlin Liu, Lyle Ungar, and Joao Sedoc. Continual learning for sentence representations using conceptors. *arXiv preprint arXiv:1904.09187*, 2019.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Zhiqiang Lv, Shanshan Zhang, Kai Tang, and Pengfei Hu. Fake audio detection based on unsupervised pretraining models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9231–9235. IEEE, 2022.

Haoxin Ma, Jiangyan Yi, Jianhua Tao, Ye Bai, Zhengkun Tian, and Chenglong Wang. Continual learning for fake audio detection. *arXiv preprint arXiv:2104.07286*, 2021.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 7765–7773, 2018.

Juan M Martín-Doñas and Aitor Álvarez. The vicomtech audio deepfake detection system based on wav2vec2 for the 2022 add challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9241–9245. IEEE, 2022.

João Monteiro, Md. Jahangir Alam, and Tiago H. Falk. An ensemble based approach for generalized detection of spoofing attacks to automatic speaker recognizers. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pp. 6599–6603. IEEE, 2020. doi: 10.1109/ICASSP40776.2020.9054558. URL https://doi.org/10.1109/ICASSP40776.2020.9054558.

Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263*, 2022.

Andreas Nautsch, Xin Wang, Nicholas W. D. Evans, Tomi H. Kinnunen, Ville Vestman, Massimiliano Todisco, Héctor Delgado, Md. Sahidullah, Junichi Yamagishi, and Kong Aik Lee. Asvspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech. *IEEE Trans. Biom. Behav. Identity Sci.*, 3(2):252–265, 2021. doi: 10.1109/TBIOM.2021.3059479. URL https://doi.org/10.1109/TBIOM.2021.3059479.

Cuong V Nguyen, Yingzhen Li, Thang D Bui, and Richard E Turner. Variational continual learning. *arXiv preprint arXiv:1710.10628*, 2017.

Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M Hospedales, and Tao Xiang. Incremental few-shot object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13846–13855, 2020.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*, pp. 4528–4537. PMLR, 2018.

Joan Serra, Didac Suris, Marius Miron, and Alexandros Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pp. 4548–4557. PMLR, 2018.

Samir Shah, Francesco Palmieri, and Michael Datum. Optimal filtering algorithms for fast learning in feedforward neural networks. *Neural networks*, 5(5):779–787, 1992.

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *arXiv preprint arXiv:2202.12233*, 2022.

Massimiliano Todisco, Xin Wang, Ville Vestman, Md Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*, 2019.

Hanna Tseran, Mohammad Emtiyaz Khan, Tatsuya Harada, and Thang D Bui. Natural variational continual learning. In *Continual Learning Workshop@ NeurIPS*, volume 2, 2018.

Christophe Veaux, Junichi Yamagishi, Kirsten MacDonald, et al. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2017.

Hongji Wang, Heinrich Dinkel, Shuai Wang, Yanmin Qian, and Kai Yu. Dual-adversarial domain adaptation for generalized replay attack detection. In Helen Meng, Bo Xu, and Thomas Fang Zheng (eds.), *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pp. 1086–1090. ISCA, 2020. doi: 10.21437/Interspeech.2020-1255. URL https://doi.org/10.21437/Interspeech.2020-1255.

Xin Wang and Junichi Yamagishi. Investigating self-supervised front ends for speech spoofing countermeasures. *arXiv preprint arXiv:2111.07725*, 2021.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 374–382, 2019.

Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md Sahidullah, and Aleksandr Sizov. Asvspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Sixteenth annual conference of the international speech communication association*, 2015.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*, 2021.

Xinrui Yan, Jiangyan Yi, Jianhua Tao, Chenglong Wang, Haoxin Ma, Tao Wang, Shiming Wang, and Ruibo Fu. An initial investigation for detecting vocoder fingerprints of fake audio. *CoRR*, abs/2208.09646, 2022. doi: 10.48550/arXiv.2208.09646. URL https://doi.org/10.48550/arXiv.2208.09646.

Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al. Add 2022: the first audio deep synthesis detection challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9216–9220. IEEE, 2022.

Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017.

Guanxiong Zeng, Yang Chen, Bo Cui, and Shan Yu. Continual learning of context-dependent processing in neural networks. *Nature Machine Intelligence*, 1(8):364–372, 2019.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995. PMLR, 2017a. URL http://proceedings.mlr.press/v70/zenke17a.html.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pp. 3987–3995. PMLR, 2017b.

You Zhang, Fei Jiang, and Zhiyao Duan. One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters*, 28:937–941, 2021a.

You Zhang, Ge Zhu, Fei Jiang, and Zhiyao Duan. An empirical study on channel effects for synthetic voice spoofing countermeasure systems. *arXiv preprint arXiv:2104.01320*, 2021b.

Yi Zhao, Wen-Chin Huang, Xiaohai Tian, Junichi Yamagishi, Rohan Kumar Das, Tomi Kinnunen, Zhenhua Ling, and Tomoki Toda. Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. *arXiv preprint arXiv:2008.12527*, 2020.

A APPENDIX

A.1 RESULTS OF OUR BASELINE

Table 7: The EER(%) on multiple evaluation sets. Model-1 to Model-6 are the models trained using the ASVspoof2019LA training set with increasing training steps.

Model	**Evaluation Sets**			
	$\mathbf{S}$	$\mathbf{T_1}$	$\mathbf{T_2}$	$\mathbf{T_3}$
Model-1	3.751	**6.316**	**7.670**	**75.198**
Model-2	2.975	8.517	10.000	78.477
Model-3	1.794	9.988	26.165	85.436
Model-4	**0.258**	24.532	46.503	91.473
Model-5	0.259	25.698	44.741	91.824
Model-6	0.262	27.872	49.726	92.113