https://doi.org/10.2352/EI.2023.35.4.MWSF-372

Synthetic Speech Attribution Using Self Supervised Audio Spectrogram Transformer

Amit Kumar Singh Yadav, Emily R. Bartusiak, Kratika Bhagtani, and Edward J. Delp
Video & Image Processing Laboratory, School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA

Abstract

The ability to synthesize convincing human speech has become easier due to the availability of speech generation tools. This necessitates the development of forensics methods that can authenticate and attribute speech signals. In this paper, we examine a speech attribution task, which identifies the origin of a speech signal. Our proposed method known as Synthetic Speech Attribution Transformer (SSAT) converts speech signals into mel spectrograms and uses a self-supervised pretrained transformer for attribution. This transformer is pretrained on two large publicly available audio datasets: Audio Set and LibriSpeech. We finetune the pretrained transformer on three speech attribution datasets: the DARPA SemaFor Audio Attribution dataset, the ASVspoof2019 dataset, and the 2022 IEEE SP Cup dataset. SSAT achieves high closed-set accuracy on all datasets (99.8% on ASVspoof2019 dataset, 96.3% on SP Cup dataset, and 93.4% on DARPA SemaFor Audio Attribution dataset). We also investigate the method's ability to generalize to unknown speech generation methods (open-set scenario). SSAT has high performance, achieving an open-set accuracy of 90.2% on the ASVspoof2019 dataset and 88.45% on DARPA SemaFor Audio Attribution dataset. Finally, we show that our approach is robust to typical compression rates used by YouTube for speech signals.

Introduction

With deep learning [1–4], it is possible to generate high quality, semantically consistent speech which is perceptually indistinguishable from speech recorded by human speaker. This development is advantageous for voice-based applications such as eLearning, virtual assistants, and commercials. However, it can also enable convincing spoofing attacks, such as voice conversion [5], impersonation [6], and cloning [7]. These synthetic speech attacks have been used to spread misinformation and target financial fraud. An impersonator using synthetic speech targeted a $40 million financial transaction with Goldman Sachs in 2021 [8]. In 2022, synthetic speech was used to spread misinformation during Russia-Ukraine war, where a deepfake video showed Ukrainian President Volodymyr Zelensky surrendering to Russia [9]. Therefore there is a need to develop methods to detect synthetic speech. In a large-scale financial fraud and misinformation campaigns, it is possible that the same speech synthesizer is used to create and spoof vast amounts of speech signals to target different people. Thus, attributing the speech synthesizer used to generate and spoof speech signals can provide more information about how the campaign is spreading and may even point us to its source.

Detecting and attributing synthetic speech to its source becomes challenging due to the diverse methods for synthetic speech generation. Common approaches for synthesizing speech include waveform concatenation [10] (a simple cut and paste method), source-filter modeling of speech signal using vocoders [11], and deep learning-based methods [1–4, 12]. Handcrafted features (*e.g.*, cepstral coefficients) [13–15], spectrograms [16–18], and time-domain speech analysis [19] have been used for synthetic speech detection. In the last few years, new deep learning-based synthetic speech generation method have been proposed [1–4, 12].

In this paper, we propose Synthetic Speech Attribution Transformer (SSAT) for synthetic speech attribution, which identifies the source of a speech signal. If the speech is spoken by a human, we classify it as authentic/bona fide. If the speech signal is synthetic, we identify the generation method used to create it. We examine both closed-set and open-set attribution scenarios. In a closed-set scenario, we evaluate our approach only on the speech generation methods present in the training set. In an open-set scenario, we also evaluate on methods which are not present in the training set (we refer to them as unknown methods). We investigate and compare several approaches for open-set attribution. Finally, we investigate robustness of SSAT against compression for data rates of 16kbps or above. Our attribution results show improvement in terms of balanced accuracy compared to other approaches [20, 21], especially in the open-set scenario.

The rest of the paper is organized as follows: in the Related Work section, we discuss common representations of speech signals, synthetic speech detection methods, and methods for synthetic speech attribution. In the Proposed Method section, we describe our method. The experimental setup, dataset used for our experiments, and results are mentioned in the Experiments and Results section. Finally, we conclude the paper with a discussion of results and directions for future research.

Related Work

Existing methods for detecting synthetic speech or manipulation in a speech signal [13, 14, 22–24] use approaches based on Gaussian Mixture Model (GMM), Support Vector Machine (SVM), and neural networks. These approaches are used to process temporal and spectral hand-crafted features such as Cochlear Filter Cepstral Coefficients (CFCCs) [23], Constant Q Transform (CQT) [22], Constant Q Cepstral Coefficients (CQCCs) [24], Mel Frequency Cepstral Coefficients (MFCCs) [13], and Linear Frequency Cepstral Coefficients (LFCCs) [22] for synthetic speech detection. In [22], features such as log power magnitude spectrogram LFCCs, and CQT are used to train a Recurrent Neural Network (RNN) [25] for synthetic speech detection. Other methods for detecting synthesized speech use a RNN [25] to capture artifacts in the time-domain speech signal [19]. Recent methods

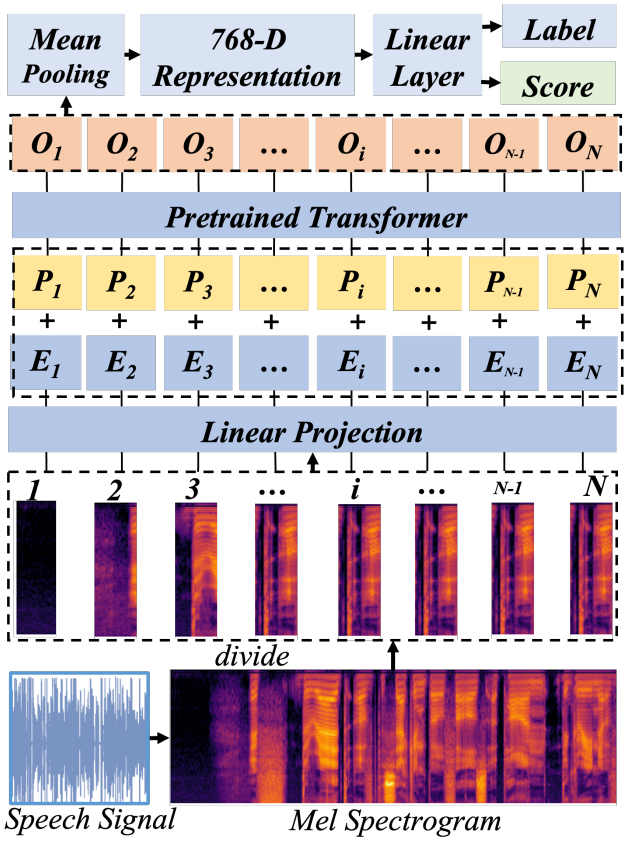


Figure 1: The block diagram of our proposed method Synthetic Speech Attribution Transformer (SSAT). $\boldsymbol{F}_i$ represents the i^{th} region, $\boldsymbol{P}_i$ represents its positional encoding, $\boldsymbol{E}_i$ is vector representation of $\boldsymbol{F}_i$, and $\boldsymbol{O}_i$ is the 768-dimensional representation corresponding to the i^{th} region.

use spectrogram, a 2D representation of the speech signal for synthetic speech detection and attribution [17, 21]. In a spectrogram, the vertical axis represents frequency bands while the horizontal axis represent time [26]. In a mel-spectrogram, the frequencies are represented in the mel scale [16]. In [27] and [17, 28], spectrogram is processed using an Efficient Convolutional Neural Network (EfficientCNN) [29] and transformer [30], respectively, for synthetic speech detection. Methods based on mel-spectrogram have shown promising results. Gong *et al.* and Koutini *et al.* have used mel-spectrograms and a transformer network for audio classification tasks (*e.g.*, environment classification [31–33]). Gong *et al.* created a self supervised framework for training the transformer [32]. Based on its high performance in audio classification tasks, we use this approach in our method for synthetic speech detection and attribution.

AlBadawy *et al.* perform bispectral analysis of speech signal using hand-crafted feature Bicohorence to attribute the speech signal [34]. Borrelli *et al.* estimate feature known as Short Term Long Term (STLT) for synthetic speech detection [20]. The method based on fusion of both of these features known as Bicoherence+STLT method outperform among all of them [20]. Hence, we used Bicoherence+STLT [20] as our baseline method and compared all our results with it.

Proposed Method

Our proposed method Synthetic Speech Attribution Transformer (SSAT) converts the time-domain speech signal to a mel-spectrogram [16] with 128 frequency bins as described in [31–33]. The mel-spectrogram is estimated using a 25 ms Hanning window with a shift of 10 ms. The height of the mel-spectrogram corresponds to the 128 mel frequency bins and the width of the mel-spectrogram corresponds to the duration of the speech signal. We fixed the width to 512 in all our experiments. Overall, we obtain a mel-spectrogram of dimension: 128×512. Figure 1 shows a mel-spectrogram of a speech signal and our proposed method SSAT.

As shown in Figure 1, the mel-spectrogram is divided into overlapping regions using a window of dimension 128×2 with a shift of 1. For each region, we find a corresponding vector representation $\boldsymbol{E}$ using linear projection. Let $\boldsymbol{E}_i$ be the vector representation corresponding to the i^{th} region. To $\boldsymbol{E}_i$, we add a positional encoding $\boldsymbol{P}_i$ as described in [32]. Using the transformer neural network [30] adapted from [32], we process the vector $\boldsymbol{E}_i + \boldsymbol{P}_i$ to obtain a 768-dimensional representation $\boldsymbol{O}_i$ corresponding to i^{th} region. We use this region-based approach because of its high performance in speech classification tasks [32]. The transformer is pretrained on the Audio Set dataset [35] and the Librispeech dataset [36] using a self-supervised learning framework [32]. Training a transformer typically requires availability of a large amount of data. Using a pretrained transformer facilitates transfer learning, and helps to counter our limited data availability. The vector representations for all the regions (*i.e.*,$\boldsymbol{O}_i$ for $i \in 1, 2, 3, ..., N$) are processed using a mean pooling operation to obtain a single 768-dimensional representation for a speech signal. Using a linear layer with SoftMax activation, we obtain a classification label and corresponding confidence score. For closed-set attribution, the classification label is either bona fide or one of the speech generation methods present in the training set. For open-set attribution, we threshold the confidence score. If the confidence score for the speech signal is lower than the threshold for bona fide class and all known generation methods, we classify the speech signal as generated from an unknown method. We investigate two more approaches for open-set attribution, as detailed in the Experiments and Results section. For all experiments, we use the Adam optimizer [37] for 50 epochs and batch size of 48. The initial learning rate is set to 2.5×10^{-4}. From the 6^{th} epoch, the learning rate decays by a ratio of 0.85 in every epoch. For evaluation, we select the model which achieves the highest accuracy on the validation set.

Experiments and Results

In this section, we describe the datasets used for the experiments and the results of our proposed Synthetic Speech Attribution Transformer (SSAT). As we mentioned earlier we used Bicoherence+STLT [20] as the baseline method and compared the performance of Synthetic Speech Attribution Transformer (SSAT) to it. We also qualitatively assess if SSAT can discriminate different speech generation methods by visualizing the latent representation in each of the experiment. For the visualization, we projected the 768-dimensional representation learned by SSAT to a 2-dimensional space using an unsupervised method known as t-distributed stochastic neighbor embedding (t-SNE) [38]. Finally, we also discuss robustness of SSAT against compressed speech

Table 1: Details of the ASVspoof2019 dataset.

ASVspoof2019 Dataset					
		D_{tr}	D_{dev}	D_{eval}	Category
Samples	**Bona fide**	2580	2548	7355	
	Synthetic	22800	22296	63882	
Speakers	**Bona fide**	20	10	48	
Synthetic	**A01**	✓	✓	×	NN
Methods	**A02**	✓	✓	×	VC
	A03	✓	✓	×	VC
	A04 = A16	✓	✓	✓	WC
	A05	✓	✓	×	VC
	A06 = A19	✓	✓	✓	VC
	A07	×	×	✓	NN
	A08	×	×	✓	NN
	A09	×	×	✓	VC
	A10	×	×	✓	NN
	A11	×	×	✓	NN
	A12	×	×	✓	NN
	A13	×	×	✓	NN
	A14	×	×	✓	VC
	A15	×	×	✓	VC
	A17	×	×	✓	VC
	A18	×	×	✓	VC

signal.

We use accuracy [39] and balanced accuracy [40] as our performance metrics for all experiments. Accuracy is defined as ratio of correct attribution to total number of attribution. For example, $Class\ A01\ Accuracy = {}^{TP_{A01}}/_{N_{A01}}$, where TP_{A01} is True Positives attribution of speech signal synthesized from synthesizer $A01$ and N_{A01} is the total number of speech signal synthesized from synthesizer $A01$. The balance attribution accuracy is average of accuracy for all classes.

ASVspoof2019 Dataset

In this section, we briefly describe the ASVspoof2019 dataset and detail our closed-set and open-set experiments.

Dataset

We use the ASVspoof2019 dataset which is described in [41, 42]. The dataset consists of speech signals for different tasks (*e.g.*, speech verification, spoofing detection and countermeasures to replay attacks). We consider only a part of the dataset which is relevant to synthetic speech detection and attribution. This subset is referred to as the Logical Access (LA) dataset in [41, 42]. It contains *bona fide* speech signals and *synthesized* speech signals.

Table 2: Confusion matrices showing closed-set results for baseline- Bicoherence+STLT and the proposed Synthetic Speech Attribution Transformer (SSAT) (in **bold**) for dataset $\boldsymbol{D_{dev}}$

True label \ Predicted label	BF	A01	A02	A03	A04	A05	A06
BF	0.85 **0.997**	0.01 **0**	0 **0**	0 **0**	0.10 **0**	0 **0.002**	0.04 **0**
A01	0 **0**	0.97 **0.995**	0 **0**	0 **0**	0.02 **0.005**	0 **0**	0 **0**
A02	0 **0**	0 **0**	0.99 **1**	0 **0**	0 **0**	0 **0**	0 **0**
A03	0 **0**	0 **0**	0.02 **0**	0.89 **1**	0 **0**	0.08 **0**	0 **0**
A04	0.09 **0**	0.01 **0**	0 **0**	0 **0.001**	0.85 **0.998**	0 **0**	0.05 **0.001**
A05	0 **0**	0 **0**	0 **0.002**	0.01 **0**	0 **0**	0.98 **0.998**	0 **0**
A06	0.02 **0**	0 **0**	0 **0**	0 **0**	0 **0.001**	0 **0**	0.98 **0.999**

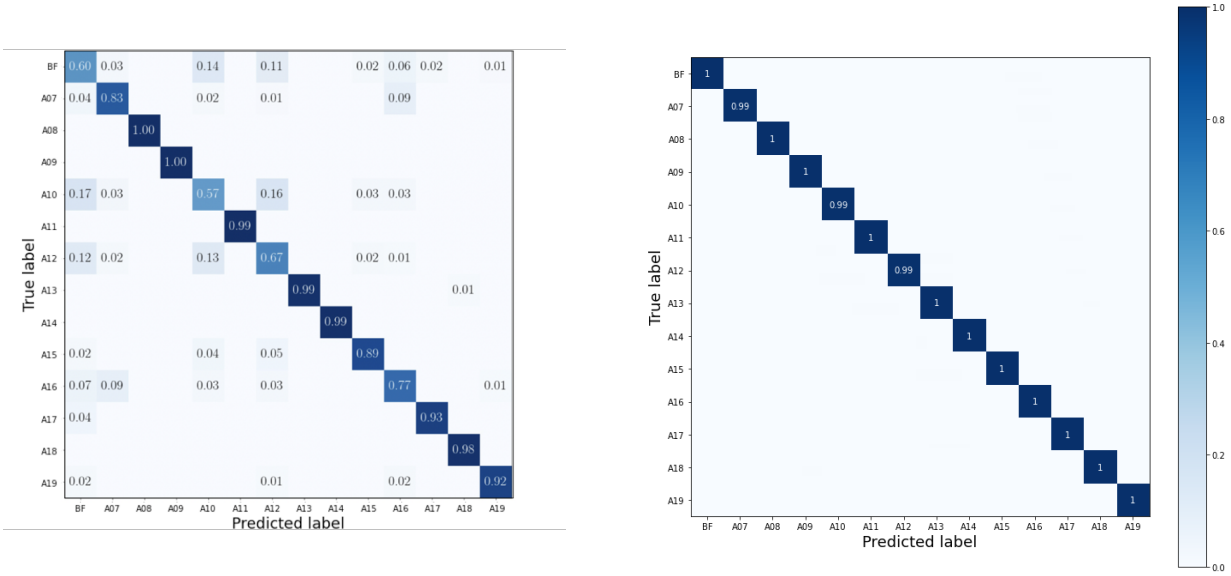
The bona fide speech signals are spoken by humans, and the synthesized speech signals are generated using neural networks (NN), vocoders (VC), and waveform concatenation (WC) [42]. The LA dataset is partitioned into a training set D_{tr}, a validation set D_{dev}, and a evaluation set D_{eval}. D_{tr} consists of bona fide speech from 20 speakers (8 male and 12 female) and synthetic speech generated from 6 methods (A01 to A06). D_{dev} consists of bona fide speech from 10 speakers (4 male and 6 female) and synthetic speech generated with A01 to A06 methods. D_{eval} consists of bona fide speech from 48 speakers (21 male and 27 female) and synthetic speech generated from 13 methods (A07 to A19). D_{eval} contains two synthetic speech generation methods, A16 and A19 which are same as A04 and A06 (in D_{dev} and D_{tr}), respectively. Each of the synthetic speech generation methods (A01 to A19) is described in [42]. D_{tr}, D_{dev}, and D_{tr} are disjoint in terms of speakers. All speech signals are sampled at 16KHz in identical recording conditions and encoded using Free Lossless Audio Codec (FLAC). Table 1 summarizes the details of the dataset.

Closed-Set Attribution

Closed-set attribution is a multi-class classification where all classes in the test set are the same as the classes in the training set. We consider the bona fide class to be a separate class and speech signal generated by different methods as different classes. Our goal is to classify a given speech signal either as bona fide or as synthetic speech generated using a known method from the training set.

Following [20], we did two experiments: Experiment 1 and Experiment 2. In Experiment 1, 80% of the speech signals in D_{tr} are used for training, and the remaining 20% are used for validation. D_{dev} is used for testing because D_{tr} and D_{dev} share the same synthetic speech generation (A01 to A06) methods. Table 2 shows the confusion matrix for Bicoherence+STLT [20] and our proposed method. Our method has balanced accuracy of 0.998 which is approximately 7% higher compared to the balanced accuracy of 0.930 obtained by Bicoherence+STLT [20].

For Experiment 2, we divide D_{eval} according to a 60:20:20 ratio for training, validation, and testing sets, respectively. Note



(a) Baseline Method: Bicoherence+STLT [20]

(b) Our Method: Synthetic Speech Attribution Transformer (SSAT)

Figure 2: Confusion matrices showing closed-set results for baseline- Bicoherence+STLT [20] and our method on dataset $\boldsymbol{D}_{eval}$.

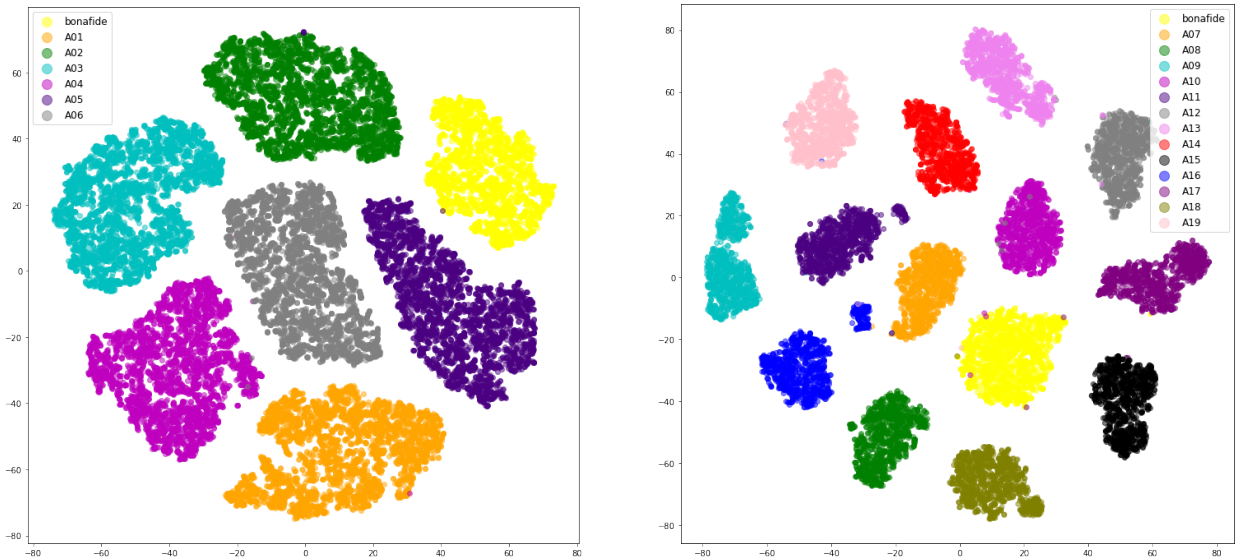


Figure 3: SSAT Latent space t-SNE visualization for ASVspoof2019 closed-set Experiment 1 (left) and closed-set Experiment 2.

that Bicoherence+STLT [20] is trained on 80% of D_{eval} and tested on 20% of D_{eval}. Hence, we are training on less data but testing on the same number of samples as Bicoherence+STLT. Figure 2 shows the confusion matrix for this experiment obtained using Bicoherence+STLT and our proposed SSAT. Our method shows significant improvement in the balanced accuracy. The balanced accuracy for our method is 0.998, which is approximately 14% higher than the balanced accuracy of Bicoherence+STLT at 0.866.

Overall, both Experiment 1 and Experiment 2 show that our method has very high balanced accuracy (*i.e.*, 99.8%) for closed-set attribution. Figure 3 also shows the t-SNE [38] visualization of the latent representation learned by our method for these experiments. We observe from the visualization plot that SSAT separates different generation methods in the latent space. Different generation methods form different clusters in the t-SNE visualization for both of the closed-set experiments.

Open-Set Attribution

Open-set attribution is a multi-class classification where all the classes in the test set do not overlap with the classes in the training set. All classes not present in the training set are combined into one class referred to as the unknown class. For open-set synthetic speech attribution, we train our method with samples from bona fide class and limited set of known method from all the available speech synthesizing methods. The goal is to evaluate if our method can classify the known classes as such and also detect synthetic speech generated from other unknown methods as unknown. The major challenge is to define the decision rule for unknown class. Note our main objective while defining the decision rule for unknown class should be that synthetic speech generated

from unknown methods should not be classified as bona fide.

We investigate open-set attribution with four different experiments: Experiment 3 to Experiment 6. In all experiments, we use D_{tr} split according to a 80:20 ratio for training and validation. For testing, we use the union of D_{dev} and D_{eval}. In Experiment 3 and Experiment 4, we use the same approach as in [20] and included an additional class label 'unknown' during training. We consider 4 out of the 6 synthetic speech classes and bona fide class present in D_{tr} as known classes. We consider two remaining synthetic speech methods in D_{tr} as known-unknown(KN-UNKN). The speech samples from known-unknown methods are used to train for unknown class. We assume that they are enough to model the unknown class. Similar to [20], we consider the pair (A02, A05) and pair (A04, A06) as KN-UNKN class for our first and second experiments, respectively. During evaluation, our method classifies speech sample into 6 classes: bona fide (BF), one of the 4 known synthetic methods, and unknown (UNKN). We separate methods A16 and A19 from the testing set as they should be recognised as A04 and A06, respectively. We also separate the known-unknown and known classes, as they should be recognized correctly. Table 3 and Table 4 show the confusion matrices for Experiment 3 and Experiment 4, respectively.

In both Experiment 3 and Experiment 4, our method outperforms the baseline in overall balanced accuracy. In Experiment 3, balanced accuracy of attribution is 0.853, and 39% of approximately 61.5K synthetic speech from unknown methods are detected as bona fide. This is an improvement of 10% over Bicoherence+STLT. In Experiment 4, balanced accuracy is higher than the baseline, our method has balanced accuracy of 0.788. However, in Experiment 4 our method classifies a higher percentage of samples from unknown synthetic speech method as bona fide.

Note that synthetic speech generation methods A16 and A19 are exactly same as methods A04 and A06, respectively (though they are trained using different data). For example, A04 is trained using CMUdict[1], while A16 is trained using VCTK data[2]. Both Experiment 3 and Experiment 4 are able to correctly classify A16 as A04 and A19 as A06. Hence, our method learns features attributing the underlying principle used in these methods for synthesizing speech and not the exact data used in these methods during their training. Additionally, it models the unknown class using synthetic speech samples from only two methods.

In Experiment 5, we threshold the confidence score. In Experiment 6, we use the representation from latent space.

In Experiment 5, we use 80% of D_{tr} to train our method on all the 7 classes in D_{tr}. Using 20% of D_{tr}, we found a threshold for the confidence score. The threshold is the mean of highest confidence score for each class such that 90% of samples in each class have confidence score higher than this. Any sample in our testing set(*i.e.*,union of D_{dev} and D_{eval}), classified with confidence score less than this threshold is assigned to unknown class.

In Experiment 6, we assume that SSAT cluster the 768-dimensional representations for speech signals from each known class inside a *'D'* radius hypersphere centered at *'C'*. Any representation falling outside this hypersphere is considered as sample from unknown class. For each class, we estimated *'C'* using the mean of representations for all speech signals inside that class. The *'D'* is the mean of radius for each class such that 95% of the samples in each class are within that radius. We use the Euclidean distance as our metric for the distance estimate.

Table 5 show the confusion matrix for Experiment 5 and Experiment 6 where the normalface is for Experiment 5 results and the boldface is for Experiment 6 results. In Experiment 5, balanced accuracy is 0.885 and 15% of approximately 54K synthetic speech from unknown methods are classified as bona fide. In Experiment 6, the balanced accuracy is 0.902 and only 11% of

Table 3: Confusion matrix showing open-set results for Experiment 3 with KN-UNKN = (A02, A05) for Bicoherence+STLT [20] and our method (in **bold**) on the union of $\boldsymbol{D}_{dev}$ and $\boldsymbol{D}_{eval}$.

True label \ Predicted label	BF	A01	A03	A04	A06	UNKN
BF	0.80 **1**	0.01 **0**	0 **0**	0.14 **0**	0.05 **0**	0 **0**
A01	0 **0**	0.97 **1**	0 **0**	0.03 **0**	0 **0**	0 **0**
A03	0 **0**	0 **0**	0.85 **1**	0 **0**	0 **0**	0.15 **0**
A04	0.11 **0**	0.01 **0**	0 **0**	0.82 **1**	0.06 **0**	0 **0**
A06	0.03 **0**	0 **0**	0 **0**	0.01 **0**	0.96 **1**	0 **0**
A16	0.12 **0**	0.04 **0**	0 **0**	0.77 **1**	0.06 **0**	0 **0**
A19	0.08 **0**	0 **0**	0 **0**	0.04 **0.02**	0.89 **0.97**	0 **1**
KN-UNKN	0 **0**	0 **0**	0.02 **0**	0 **0**	0 **0**	0.98 **1**
UNKN	0.49 **0.39**	0.09 **0.02**	0.02 **0.4**	0.11 **0.17**	0.05 **0**	0.24 **0.02**

Table 4: Confusion matrix showing open-set results for Experiment 4 with KN-UNKN = (A04, A06) for baseline- Bicoherence+STLT and our method (in **bold**) on the union of $\boldsymbol{D}_{dev}$ and $\boldsymbol{D}_{eval}$.

True label \ Predicted label	BF	A01	A02	A03	A05	UNKN
BF	0.92 **1**	0 **0**	0 **0**	0 **0**	0 **0**	0.08 **0**
A01	0.01 **0**	0.94 **1**	0 **0**	0 **0**	0 **0**	0.04 **0**
A02	0 **0**	0 **0**	0.98 **1**	0 **0**	0.01 **0**	0 **0**
A03	0 **0**	0 **0**	0.02 **0**	0.88 **1**	0.09 **0**	0 **0**
A05	0 **0**	0 **0**	0 **0**	0.02 **0**	0.97 **1**	0 **0**
A16	0.04 **0**	0.03 **0**	0 **0**	0 **0**	0 **0**	0.93 **1**
A19	0.04 **0**	0 **0**	0 **0**	0 **0**	0 **0**	0.96 **1**
KN-UNKN	0.03 **0**	0 **0**	0 **0**	0 **0**	0 **0**	0.97 **1**
UNKN	0.55 **0.69**	0.06 **0.01**	0.07 **0**	0.03 **0.14**	0.17 **0.14**	0.13 **0.02**

[1] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict/`
[2] `https://datashare.ed.ac.uk/handle/10283/2651/`

Table 5: Confusion matrix showing open-set results for Experiment 5(*i.e.*,thresholding confidence score) and Experiment 6 (*i.e.*,using 768-dimensional representation in the latent space). The Experiment 6 results are in **boldface**.

True label \ Predicted label	BF	A01	A02	A03	A04	A05	A06	UNKN
BF	0.91 **0.85**	0 **0**	0 **0**	0 **0**	0 **0**	0 **0**	0 **0**	0.09 **0.15**
A01	0 **0**	0.92 **0.99**	0 **0**	0 **0**	0 **0**	0 **0**	0 **0**	0.08 **0.01**
A02	0 **0**	0 **0**	0.98 **1**	0 **0**	0 **0**	0 **0**	0 **0**	0.02 **0**
A03	0 **0**	0 **0**	0 **0**	1 **0.99**	0 **0**	0 **0**	0 **0**	0 **0.01**
A04	0 **0**	0 **0**	0 **0**	0 **0**	0.94 **0.94**	0 **0**	0 **0**	0.06 **0.06**
A05	0 **0**	0 **0**	0 **0**	0 **0**	0 **0**	0.96 **0.89**	0 **0**	0.04 **0.11**
A06	0 **0**	0 **0**	0 **0**	0 **0**	0 **0**	0 **0**	0.77 **0.89**	0.23 **0.11**
UNKN	0.15 **0.11**	0 **0**	0 **0**	0.17 **0.11**	0.07 **0.05**	0 **0**	0 **0**	0.60 **0.72**

approximately 54K synthetic speech from unknown methods are classified as bona fide. These experiments show that naive thresholding of the confidence score and finding similarity in latent representation can significantly boost the detection rate for synthetic speech from unknown methods. As in Bicoherence+STLT, our Experiment 3, and Experiment 4 which model unknown class using samples from limited methods more than 50% of synthetic speech from unknown methods are classified as bona fide. While only 15% and 11% of such synthetic speech are classified as bona fide in Experiment 5 and Experiment 6, respectively.

We visualized the latent space of SSAT on speech signals from unknown speech speech generation methods A07-A19. The SSAT was trained on only bona fide speech signals and speech signals generated from A01 to A06 generation methods. Figure 4 shows the 2-dimensional t-SNE [38] visualization. We found that the SSAT cluster several unknown speech generation methods. This is very promising since SSAT was never trained on these unknown speech generation methods and is still able to differentiate speech signals from these unknown methods in different clusters. Overall, our experiments show that SSAT with an appropriate choice of decision rule such as using representations as in our Experiment 6 can attain balanced accuracy of more than 90% for open-set synthetic speech attribution. Our results also suggest that naive approaches such as thresholding the confidence score and estimating similarity in representation in the latent space are better approaches to model unknown class than using limited samples to train for unknown class. As the later inherently bias our unknown class decision rule.

DARPA SemaFor Audio Attribution Dataset

This section provide details about the DARPA SemaFor Audio Attribution dataset, closed-set and open-set experiments. We also visualize the latent space of Synthetic Speech Attribution Transformer (SSAT) trained on this dataset.

Dataset

The DARPA SemaFor Audio Attribution dataset has in total 17,000 synthetic speech signals from 11 different speech synthesizers as summarized in the Table 6.

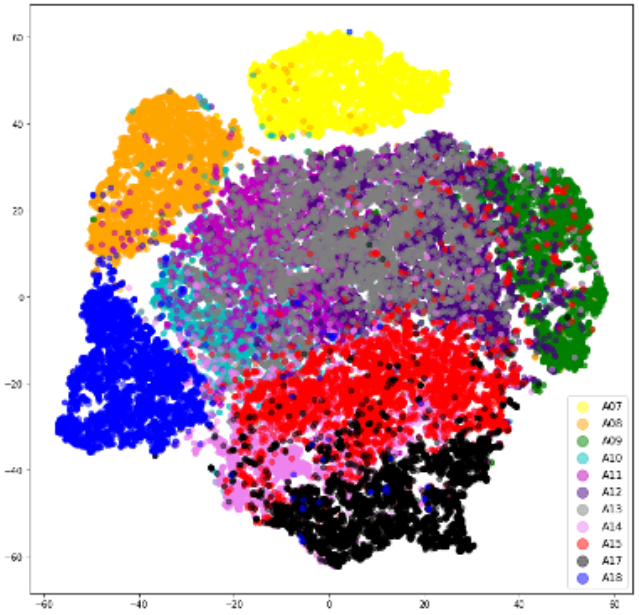


Figure 4: t-SNE visualization of latent representations for unknown speech generation methods A07-A19. SSAT is trained on bona fide and A01-A06 speech generation methods.

This dataset was generated for the Semantic Forensics (SemaFor) program organized by the Defense Advanced Research Projects Agency (DARPA) [43]. All the speech signals are Waveform Audio File Format (WAV) signals at a sampling rate of 16kHz. The dataset has training and testing splits. The training set consist of 8 different speech generation methods. The 8 speech generation methods are Fastpitch, Fastspeech2, Glowtts, Gtts, Riva, Tacotron, Tacotron 2 and Talknet. There are 1,000 speech signals for each speech generation method in the training set except for FastSpeech2 and Tacotron 2, each of them contain only 500 speech signals. The testing set contains speech signals from all the 8 speech generation methods in the training set and 3 additional methods, that are Mixertts, Speedyspeech and Vits.

Closed-Set Attribution

For closed-set attribution, we trained Synthetic Speech Attribution Transformer (SSAT) on synthetic speech samples from all the 8 different speech generation methods present in the training set. Since, the official split does not have validation split. We use K-fold training strategy (*i.e.*,we divide the training set into K parts randomly and then use one part for validation and remaining 4 parts for training). We used K=5. Hence, each of our models is trained on 5600 samples and validated on 1400 samples. From the 5 models from K-fold training, we select the model with highest accuracy on validation set for final evaluation. For closed-set evaluation we only test SSAT on speech samples generated from 8 different generation methods that are also present in the training set. Table 7 details our experimental results and compares them with other methods reported for the same dataset in [21]. We evaluate accuracy, precision, recall, and F-1 score [39]. Figure 5 shows a 2-dimensional t-SNE visualization of the latent space of SSAT trained for closed-set attribution on DARPA SemaFor Audio Attribution dataset.

Results show that our method outperforms all approaches re-

Table 6: Details of DARPA SemaFor Audio Attribution dataset.

Generation Method	Training Set	Testing Set	Total
Fastpitch	1,000	1,500	2,500
Fastspeech2	500	300	800
Glowtts	1,000	1,500	2,500
Gtts	1,000	1,500	2,500
Riva	1,000	1,000	2,000
Tacotron	1,000	1,500	2,500
Tacotron2	500	300	800
Talknet	1,000	1,500	2,500
Mixertts	-	300	300
Speedyspeech	-	300	300
Vits	-	300	300
Total	**7,000**	**10,000**	**17,000**

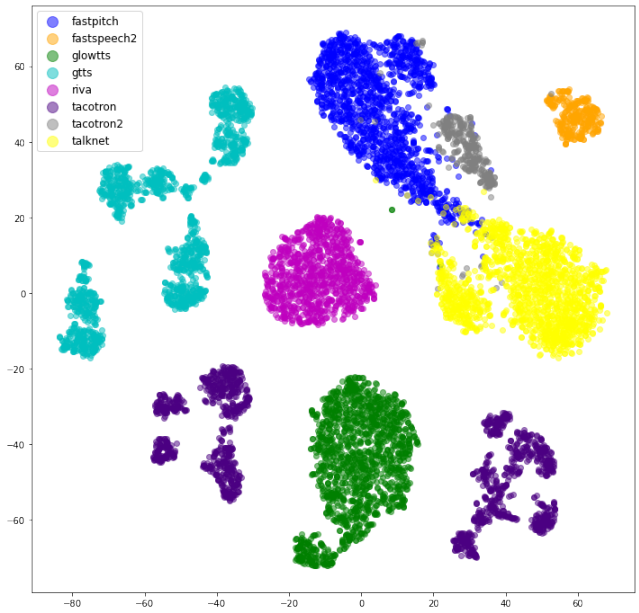


Figure 5: t-SNE visualization of latent representations for closed-set attribution on DARPA SemaFor Audio Attribution Dataset.

ported in [21]. Note that these are all machine learning-based approaches. Our visualization show that SSAT clusters synthetic speech signals generated from same generation method and forms different cluster for different speech generation methods.

Open-Set Attribution

For open-set attribution, we trained Synthetic Speech Attribution Transformer (SSAT) on synthetic speech samples from all the 8 different speech generation methods present in the training set. We use K-fold training strategy and divide the training set into 5 parts randomly and then use one part for validation and remaining 4 parts for training. From the 5 models from K-fold training, we select the model with highest accuracy on validation set for final evaluation. Speech signals from any of the 3 speech generation methods not present in the training set are classified in a single class referred as unknown. We used the latent representation for open-set attribution. If, for a given speech signal, its representation in the latent space is at a distance above a threshold from the mean representation for each class, we classify it as speech signal generated from unknown method. This approach for open-set attribution outperformed all approaches that we investigated on ASVspoof2019 dataset and is described in detail in open-set attribution Experiment 6 on ASVspoof2019 dataset. Table 8 details our experimental results and compares them with previous methods on the same dataset reported in [21]. When we visualize the latent representation (Figure 6) for unknown speech generation methods, we see that the three unknown speech generation method form three different clusters. Overall, our results and visualizations show that SSAT outperforms all previous work and is able to differentiate even unknown speech generation methods.

Table 7: Results of all methods on closed-set attribution on DARPA SemaFor Audio Attribution dataset. Performance of methods other than SSAT is taken from [21].

Method	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Baseline-Minority	3.30	0.11	3.30	0.21
Baseline-Majority	16.48	2.72	16.48	4.67
QDA	19.34	16.82	19.34	11.75
GP	21.62	68.80	21.62	12.82
AdaBoost	33.58	32.71	33.58	23.69
KNN	52.22	56.63	52.22	49.83
Naive Bayes	68.14	70.95	68.14	69.08
Decision Tree	69.02	71.08	69.02	69.93
MLP	78.91	77.06	78.91	77.68
Random Forest	81.04	79.90	81.04	79.10
Non-Linear SVM	81.13	81.35	81.13	81.05
Linear SVM	81.57	80.99	81.57	81.22
LogReg	90.68	88.29	90.68	89.43
CNN	91.99	90.21	91.99	90.88
CAT	92.53	90.37	92.53	91.27
SSAT	**93.38**	**91.37**	**93.38**	**91.62**

Table 8: Results of all methods on open-set attribution on DARPA SemaFor Audio Attribution dataset. Performance of methods other than SSAT is taken from [21].

Method	Accuracy(%)	Precision(%)	Recall(%)	F1-Score(%)
Baseline-Minority	3.00	0.09	3.00	0.17
Baseline-Majority	15.00	2.25	15.00	3.91
QDA	17.60	14.45	17.60	9.86
GP	9.00	0.81	9.00	1.49
AdaBoost	17.51	10.84	17.51	12.80
KNN	47.52	46.89	47.52	42.77
Naive Bayes	62.01	59.58	62.01	59.95
Decision Tree	62.66	60.01	62.66	60.77
MLP	55.97	57.03	55.97	53.36
Random Forest	47.74	80.92	47.74	46.29
Non-Linear SVM	65.41	73.24	65.41	66.41
Linear SVM	68.47	71.78	68.47	68.80
LogReg	83.85	81.44	83.85	81.62
CNN	83.56	77.26	83.56	79.67
CAT	84.10	82.37	84.10	83.00
SSAT	**88.45**	**89.01**	**88.45**	**87.59**

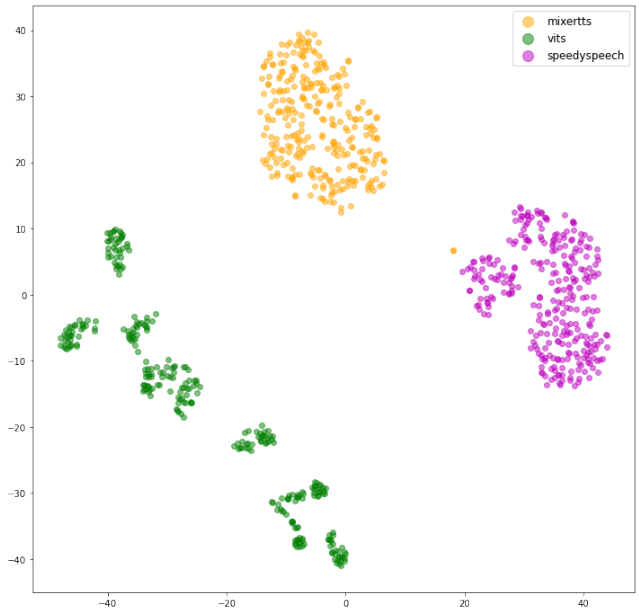


Figure 6: t-SNE visualization of latent representations for 3 unknown speech generation methods in DARPA SemaFor Audio Attribution Dataset.

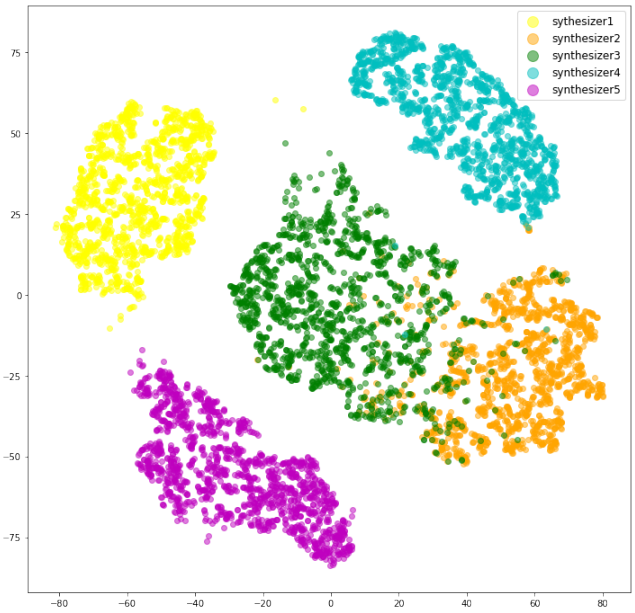


Figure 7: t-SNE visualization for closed-set attribution on IEEE SP Cup 2022 Synthetic Speech Attribution Open Competition 1 dataset.

IEEE SP Cup 2022 Synthetic Speech Attribution Dataset and Robustness to Compressed Speech Signal

In this section, we detail experiments using the IEEE SP Cup 2022 Synthetic Speech Attribution Open Competition 1 dataset[3] and robustness to compressed speech signal.

The IEEE SP Cup 2022 Synthetic Speech Attribution Open Competition 1 dataset consist of training and testing sets. We divided the total 5000 synthetic speech recordings generated from 5 different algorithms provided in the training set in 80:20 for our train and validation set, respectively. We evaluated on total 7500 synthetic speech signals present in the testing set generated from the 5 algorithms. We investigated the closed-set attribution performance of Synthetic Speech Attribution Transformer (SSAT) on this dataset as both the training and testing set have speech signals from same generation methods. Our method is able to accurately attribute 96.3% of these uncompressed synthetic speech signals. The Figure 7 shows the visualization of the latent space of SSAT trained on IEEE SP Cup 2022 Synthetic Speech Attribution Open Competition 1 dataset. We can see that speech signals from same generation method have similar representations and cluster in a region while speech signals from different generation methods have relatively different representation and part of different clusters.

Synthetic speech generated with malicious intent is shared on social platforms such as YouTube which result in lossy compression of these speech signals [44]. Hence, we also assess the robustness of our method on compressed speech signal. We encoded the speech signals in the testing set of IEEE SP Cup 2022 Synthetic Speech Attribution Open Competition 1 dataset to investigate the robustness of our method on compressed speech signal. We used Advanced Audio Coding (AAC) [45] for compression. AAC is successor to MP3 and one of the most popular audio codecs, used by Apple iTunes, and Youtube. We used bitrate of 126kbps, 64kbps, 32kbps and 16kbps. Figure 8 plot attribution accuracies of SSAT for different compression bitrates. We can observe that our method is robust and is able to accurately attribute speech signal AAC encoded at bitrate from 32kbps to 126 kbps. The performance of our method decrease drastically only for speech signal AAC encoded aggressively at bitrate of 16kbps.

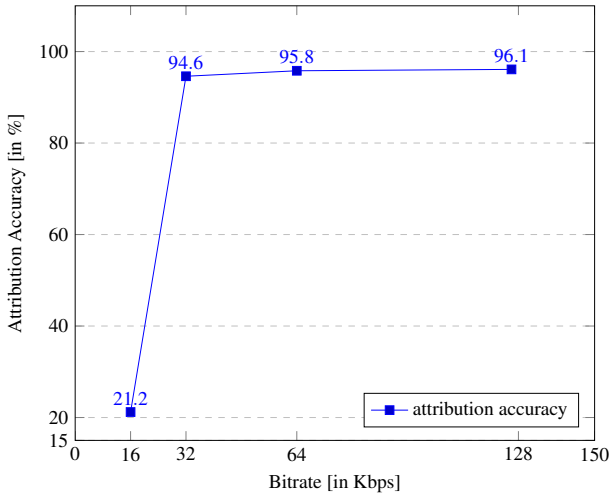


Figure 8: Attribution accuracy of our method on encoded speech signal

Overall, our results and visualization of the latent space show that our method is able to attribute synthetic speech signal in IEEE SP Cup 2022 Synthetic Speech Attribution Open Competition 1 dataset and our method is robust to speech signal AAC compressed at bitrate of 32kbps or higher.

[3] https://signalprocessingsociety.org/community-involvement/signal-processing-cup

Conclusion

In this paper, we proposed a method using transformer for synthetic speech attribution. We tested our method for closed-set attribution on different datasets. Our experimental results show that our method works for closed-set attribution and achieves high accuracy of 99.8% on ASVspoof2019 dataset, 96.3% on SP Cup dataset, and 93.4% on DARPA SemaFor Audio Attribution dataset. We also investigated open-set scenario to examine our method's ability to identify unknown speech generation methods. Our experiments highlight the challenges in open-set attribution. We propose different strategies for open-set attribution. We find that instead of modelling unknown class with samples from limited classes, thresholding confidence score, and using representation from latent space significantly improve attribution accuracy in open-set scenario. We obtained an open-set attribution accuracy of 90.2% on ASVspoof2019 dataset and 88.45% on DARPA SemaFor Audio Attribution dataset . Our method outperform both closed-set and open-set performance from existing methods on ASVspoof2019 dataset and DARPA SemaFor Audio Attribution dataset. We also find that our method is robust to AAC compression at data rates of 32kbps or larger. The transformer in our method has $\approx$ 87 M parameters. In future, we plan to reduce the computational complexity of our method and improve our accuracy for open-set attribution.

We also want to investigate robustness of our method against noise, mixup, and reverberation in speech signal. Also, although we can not know actual methods in unknown class, we are exploring ways to determine number of methods present in unknown class.

References

[1] J. Kim, J. Kong, and J. Son, "Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech," *Proceedings of the International Conference on Machine Learning*, vol. 139, pp. 5530–5540, July 2021, Virtual.

[2] T. Wang, R. Fu, J. Yi, J. Tao, Z. Wen, C. Qiang, and S. Wang, "Prosody and Voice Factorization for Few-Shot Speaker Adaptation in the Challenge M2voc 2021," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8603–8607, June 2021, Toronto, Canada.

[3] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech," *Proceedings of the International Conference on Machine Learning*, vol. 139, pp. 8599–8608, July 2021, Virtual.

[4] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and Unseen Emotional Style Transfer for Voice Conversion with A New Emotional Speech Dataset," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 920–924, June 2021, Toronto, Canada.

[5] X. Tian, S. W. Lee, Z. Wu, E. S. Chng, and H. Li, "An exemplar-based approach to frequency warping for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1863–1876, July 2017.

[6] Y. Gao, R. Singh, and B. Raj, "Voice impersonation using generative adversarial networks," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2506–2510, April 2018, Calgary, Canada.

[7] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," *Proceedings of the Advances in Neural Information Processing Systems*, vol. 31, p. 10019–10029, December 2018, Montreal, Canada.

[8] B. Smith, "Goldman Sachs, Ozy Media and a $40 Million Conference Call Gone Wrong," *The New York Times*, September 2021, https://www.nytimes.com/2021/09/26/business/media/ozy-media-goldman-sachs.html.

[9] B. Allyn, "Deepfake Video of Zelenskyy Could be 'Tip of the Iceberg' in Info War, Experts Warn," https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia, March 2022.

[10] F. Tesser, G. Paci, G. Sommavilla, and P. Cosi, "Open source voice creation toolkit for the mary tts platform," *Proceedings of the International Speech Communication Association (INTERSPEECH)*, pp. 3253–3256, August 2011, Florence, Italy.

[11] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[12] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1–15, May 2021, virtual.

[13] M. Sahidullah and G. Saha, "Design, Analysis, and Experimental Evaluation of Block Based Transformation in MFCC Computation for Speaker Recognition," *Speech Communication*, vol. 54, pp. 543–565, May 2012.

[14] B. Bogert, M. Healy, and J. Tukey, "The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Saphe Cracking," *Proceedings of the Symposium on Time Series Analysis*, vol. 15, pp. 209–243, June 1963, New York, NY.

[15] K. Bhagtani, A. K. S. Yadav, E. R. Bartusiak, Z. Xiang, R. Shao, S. Baireddy, and E. J. Delp, "An Overview of Recent Work in Multimedia Forensics," *Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval*, August 2022, Virtual.

[16] S. Stevens, J. Volkmann, and E. Newman, "A Scale for the Measurement of the Psychological Magnitude Pitch," *Journal of the Acoustical Society of America*, vol. 8, pp. 185–190, June 1937.

[17] E. R. Bartusiak and E. J. Delp, "Synthesized Speech Detection Using Convolutional Transformer-Based Spectrogram Analysis," *Proceedings of the IEEE Asilomar Conference on Signals, Systems, and Computers*, October 2021, Asilomar, CA.

[18] ——, "Frequency Domain-Based Detection of Generated Audio," *Proceedings of the IS&T Media Watermarking, Security, and Forensics Conference, Electronic Imaging Symposium*, pp. 273(1)–273(7), January 2021.

[19] G. Hua, A. B. J. Teoh, and H. Zhang, "Towards End-to-End Synthetic Speech Detection," *IEEE Signal Processing Let-*

ters, vol. 28, pp. 1265–1269, June 2021.

[20] C. Borrelli, P. Bestagini, F. Antonacci, A. Sarti, and S. Tubaro, "Synthetic speech detection through short-term and long-term prediction traces," *EURASIP Journal on Information Security*, vol. 2021, no. 1, p. 2, April 2021.

[21] E. R. Bartusiak and E. J. Delp, "Transformer-Based Speech Synthesizer Attribution in an Open Set Scenario," *Proceedings of the IEEE International Conference on Machine Learning and Applications*, December 2022, arxiv:2210.07546.

[22] X. Li, N. Li, C. Weng, X. Liu, D. Su, D. Yu, and H. Meng, "Replay and Synthetic Speech Detection with Res2Net Architecture," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6354–6358, June 2021, Toronto, Canada.

[23] T. B. Patel and H. A. Patil, "Cochlear Filter and Instantaneous Frequency Based Features for Spoofed Speech Detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 618–631, December 2017.

[24] M. Todisco, H. Delgado, and N. Evans, "Constant Q Cepstral Coefficients: A Spoofing Countermeasure for Automatic Speaker Verification," *Computer Speech & Language*, vol. 45, pp. 516–535, September 2017.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, June 2016, Las Vegas, NV.

[26] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*, 1st ed. USA: Prentice Hall Press, 2010.

[27] N. Subramani and D. Rao, "Learning Efficient Representations for Fake Speech Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 5859–5866, April 2020, New York, NY.

[28] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, "A Deep Learning Framework for Audio Deepfake Detection," *Arabian Journal for Science and Engineering*, vol. 47, p. 3447–3458, November 2021.

[29] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *Proceedings of the International Conference on Machine Learning*, vol. 97, pp. 6105–6114, June 2019, Long Beach, CA.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need," *Proceedings of the Neural Information Processing Systems*, December 2017, Long Beach, CA.

[31] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," *Proceedings of the ISCA Interspeech*, pp. 571–575, August 2021, Brno, Czech Republic.

[32] Y. Gong, C.-I. J. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-Supervised Audio Spectrogram Transformer," *arXiv*, October 2021.

[33] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient Training of Audio Transformers with Patchout," *arXiv*, October 2021.

[34] E. A. AlBadawy, S. Lyu, and H. Farid, "Detecting ai-synthesized speech using bispectral analysis," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*, pp. 104–109, June 2019, Long Beach, CA.

[35] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, New Orleans, LA.

[36] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, Queensland, Australia.

[37] D. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *Proceedings of the International Conference for Learning Representations*, May 2015, San Diego, CA.

[38] L. van der Maaten and G. Hinton, "Visualizing High-Dimensional Data Using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, November 2008.

[39] A. Tharwat, "Classification Assessment Methods," in *Applied Computing and Informatics*. Emerald Publishing Limited, December 2021, pp. 168–192.

[40] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The Balanced Accuracy and Its Posterior Distribution," *Proceedings of the 2010 20th International Conference on Pattern Recognition*, pp. 3121–3124, 2010, Istanbul, Turkey.

[41] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evans, T. Kinnunen, and K. A. Lee, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, September 2019, Graz, Austria.

[42] J. Yamagishi, M. Todisco, M. Sahidullah, H. Delgado, X. Wang, N. Evans, T. Kinnunen, K. Lee, V. Vestman, and A. Nautsch, "ASVspoof 2019: The 3rd Automatic Speaker Verification Spoofing and Countermeasures Challenge database," *University of Edinburgh. The Centre for Speech Technology Research*, March 2019.

[43] "Semantic Forensics (SemaFor)," https://www.darpa.mil/program/semantic-forensics.

[44] Google Inc., "YouTube Recommended Upload Encoding Settings," 2022. [Online]. Available: https://support.google.com/youtube/answer/1722171

[45] J. Herre and H. Purnhagen, "General Audio Coding," in *The MPEG-4 Book*, F. C. Pereira and T. Ebrahimi, Eds. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2002, pp. 487–544.

Author Biography

Amit Kumar Singh Yadav *received his B.Tech. major in Electrical Engineering and minor in Computer Science from Indian Institute of Technology - Gandhinagar. He worked in Computer Vision R&D team of Rakuten (RIT) and later joined Enphase Energy as an Embedded Software Engineer. He is doing Ph.D. in ECE from Purdue University. His research involves development of Speech Recognition, Image Processing, and Computer Vision methods for analyzing image, audio, and video signals for forensic applications.*

Emily R. Bartusiak *is a Ph.D. student in Electrical and Computer Engineering at Purdue University. She previously earned her B.S. and M.S. degrees in Electrical Engineering from Purdue with a minor in Management. She currently investigates machine learning and deep learning techniques for media forensics, aerospace, and biomedical research.*

Kratika Bhagtani *is a Ph.D. student in Electrical and Computer Engineering at Purdue University. She received her Bachelor of Technology degree with Honors in Electrical Engineering from Indian Institute of Technology - Gandhinagar. She worked at Enphase Energy as a Hardware Engineer. Her research interests include media forensics, image processing, video processing, speech forensics, and computer vision.*

Edward J. Delp *was born in Cincinnati, Ohio. He is Charles William Harrison Distinguished Professor of Electrical and Computer Engineering and Professor of Biomedical Engineering at Purdue University. His research interests include image and video processing and compression, computer vision, machine learning, multimedia security, medical imaging, and information theory. Dr. Delp is a Fellow of the IEEE, the SPIE, the Society for Imaging Science and Technology, and the American Institute of Medical and Biological Engineering.*