Transformer Ensemble for Synthesized Speech Detection

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

***Abstract*—As voice synthesis systems and deep learning tools continue to improve, so does the possibility that synthesized speech can be used for nefarious purposes. Methods that determine if audio signals contain synthesized or authentic speech are needed. In this paper, we investigate three transformers to detect synthesized speech: Compact Convolutional Transformer (CCT), Patchout faSt Spectrogram Transformer (PaSST), and Self-Supervised Audio Spectrogram Transformer (SSAST). We show that each transformer independently detects synthesized speech well. Then, we propose an ensemble of transformers that can provide even better performance. Finally, we explore how much of an audio signal is needed for high synthesized speech detection. Evaluated on the ASVspoof2019 dataset, we demonstrate that our transformer ensemble detects synthesized speech from shorter segments of audio signals, even on a highly imbalanced dataset.**

***Index Terms*—deep learning, audio forensics, synthesized speech detection, transformers, mel spectrograms**

I. Introduction

Nowadays, it is common to interact with synthesized speech on a daily basis. We talk with Siri, Alexa, and Google Assistant in our homes, on our phones, and in our cars [1]–[4]. Virtual assistants answer customer service calls and help us with anything from scheduling appointments to paying bills to accessing bank accounts [5]–[7]. Often, we can easily tell that the voices of these virtual assistants are synthesized from their robotic tones. In some situations, though, it is more difficult to discern whether we are listening to synthesized or authentic human voices.

Social media platforms offer many tools that allow users to create new speech signals that sound realistic. Both Tiktok and Instagram provide text-to-speech (TTS) services, enabling users to generate new speech signals with custom messages [8], [9]. The platforms maintain libraries of voice styles, any of which can be used to deliver the messages. Tiktok also offers an option for users to upload an audio track of a specific voice style they would like to replicate. If users already have an audio track with a desired message, they can transform the message into a different voice style using Tiktok. Deep learning methods for speech synthesis and voice conversion systems can also generate realistic-sounding human speech [10]–[15]. Because many easy-to-use tools like these exist for modifying speech with high quality, the quantity of inauthentic speech is increasing rapidly [16]. Although all of these features can be used for comedic purposes, they can also easily be used with more nefarious consequences.

Attackers may generate new speech signals impersonating people with detrimental consequences. For example, they could create a new message for a public figure with the hope that the speech goes viral and disseminates misinformation. When inauthentic video accompanies inauthentic speech signals, such as in deepfakes, the potential to influence public opinion and current events is even higher [17], [18]. In 2022, a deepfake arose that showed Ukrainian President Volodymyr Zelensky surrendering to Russia [19], [20]. Although it was quickly debunked, this scenario shows the potential a realistic deepfake with inauthentic audio can have on war situations and other global matters. In another example, Goldman Sachs stopped a $40 million investment in 2021 after realizing they were conducting business with an impersonator using synthesized speech on a conference call [21]. As people conduct more and more business remotely, they must authenticate their virtual interactions more than ever [22]–[24].

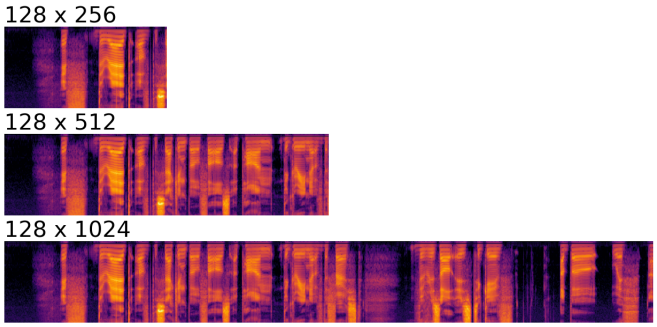

Fig. 1: Mel spectrograms showing the same speech signal cropped to different lengths of time. The vertical dimension is always 128, corresponding to 128 mel frequency bins. The horizontal dimension indicates the number of temporal windows of an audio signal used in the analysis.

To detect synthesized speech, we propose an ensemble of transformers. Transformers are neural networks that utilize an attention mechanism [25] and have achieved success on a variety of tasks in recent years. We investigate three transformers for synthesized speech detection: Compact Convolutional Transformer (CCT) [26]; Patchout faSt Spectrogram Transformer (PaSST) [27]; and Self-Supervised Audio Spectrogram Transformer (SSAST) [28]. Transformers typically require large datasets to train. In this paper, we explore their application on a smaller, imbalanced dataset for synthesized speech detection. We train each of the transformers on mel spectrograms [29] to authenticate speech signals. Examples of mel spectrograms are shown in Figure 1. Mel spectrograms are

2-D visual representations of audio signals showing frequency and intensity of audio signals over time [29]. The frequencies of the audio signals are in the mel scale, which is a perceptual scale based on the human auditory system [29]. We train all of the transformers from scratch on ASVspoof2019 [30]. The ASVspoof2019 dataset contains authentic and synthesized English speech signals from multiple speakers and synthesizers. We demonstrate that fusing the transformers in an ensemble achieves better and more robust detection than individual transformers. Finally, we explore how much of an audio signal is needed for synthesized speech detection. We show that individual transformers are more sensitive to different lengths of audio signals. When each transformer is trained on different mel spectrogram sizes (corresponding to different audio signal lengths), we show that detection results change. However, our transformer ensemble consistently and successfully detects synthesized speech from all mel spectrogram sizes considered.

II. Related Work

In audio analyses, it is important to determine a useful representation of an audio signal. There are many different ways to represent an audio signal: as a temporal waveform, as a spectrogram, or as a sequence of coefficients from a transform (*e.g.*, Constant Q Cepstral Coefficients (CQCCs) [31]). Chintha *et al*. use a Convolutional Neural Network Long Short-Term Memory network (CNN-LSTM) to detect synthesized speech [32] by analyzing audio waveforms. Hua *et al*. analyze audio waveforms with a convolutional network based on ResNet and Inception networks [33]. However, waveform-based methods struggle with longer audio signals because the inputs can be hundreds of thousands of samples long. Representing audio waveforms as sequences of transform coefficients, such as CQCCs [31], can create shorter-length inputs. Chen *et al*. use a Multilayer Perceptron Network (MLP), ResNet-based CNN, LSTM, Gated Recurrent Unit network (GRU), and Recurrent Neural Network (RNN) to detect synthesized speech represented as CQCCs [34], [35]. Other methods represent audio signals as 2-D arrays that contain information about an audio signal's frequencies and intensity over time. These 2-D arrays are known as spectrograms and can be treated as images. Bartusiak *et al*. analyze spectrograms of audio signals for synthesized speech detection using a CNN and a convolution transformer [36]–[40]. Spectrograms can represent frequencies according to the mel scale, which is a scale that represents pitches perceived to be equally distant according to the human auditory system [29]. These versions are known as mel spectrograms. Conti *et al*. analyze mel spectrograms to identify emotions and then use a Random Forest Classifier on the emotion features to detect synthesized speech [41]. Gong *et al*. and Koutini *et al*. use mel spectrograms to classify over 600 different types of audio signals, from bird noises to specific speech commands [27], [28], [42]. Based on the success of work with mel spectrograms, we use mel spectrograms for synthesized speech detection.

III. Method

A. Mel Spectrogram Creation

We convert speech waveforms into mel spectrograms by following a similar procedure as described in [27], [28], [42]. More specifically, we create mel spectrograms with 128 mel frequency bins. The mel spectrogram is computed using a 25 ms Hanning window with a shift of 10 ms. In other words, an audio waveform is divided into 25 ms "time frames" or "windows" with 10 ms overlap between consecutive windows. We orient the mel spectrograms so that the height of the mel spectrograms is 128 (corresponding to the 128 mel frequency bins) and the width of the spectrograms corresponds to the length of the audio signal (in terms of 25 ms time frames). We crop or zero-pad the mel spectrograms to each of the following dimensions: 128x256, 128x512, and 128x1024. Notice that only the second dimension (indicating the temporal length of the audio signal) is affected by the cropping and padding procedure. The first dimension (*i.e.*, 128) always remains the same. It corresponds to the number of mel frequency bins in the mel spectrograms. The second dimension corresponds to the number of 25 ms time frames of the spectrogram. Larger values indicate that more of an audio signal is included in the mel spectrogram, meaning that a longer-length audio signal will be analyzed. Figure 1 shows mel spectrograms of the same audio signal formatted as different sizes (*i.e.*, formatted to analyze different lengths of an audio signal). These mel spectrograms are used to train and evaluate the transformers and transformer ensemble.

B. Compact Convolutional Transformer

Compact Convolutional Transformer (CCT) [26] is a smaller transformer used for analyses on 2-D inputs, such as our mel spectrogram inputs. It uses a series of convolutional layers to extract features from the inputs. Then, the features are analyzed by the transformer block of the network. This approach ensures that the features provided to the attention mechanism capture information from all regions of an input. The convolutions also introduce inductive biases (*e.g.*, translational equivariance) into the network that normally require an extensive amount of data for a transformer to learn on its own. Convolution operations enable weight sharing, which decreases the size of the network and increases its computational efficiency. The CCT used in this work has approximately 405 thousand parameters, making it the smallest transformer we investigate by a significant margin. We train CCT from scratch with early stopping using a patience of 15 epochs. The AdamW optimizer [43] is used with an initial learning rate and weight decay of 10^{-4}. The batch size in our experiments depends on the size of the mel spectrograms analyzed because the size of the attention matrix scales quadratically with the increase in input size [26]. For mel spectrograms sized 128x256, 128x512, and 128x1024, we use a batch size of 16, 16, and 2, respectively.

C. Patchout faSt Spectrogram Transformer

Patchout faSt Spectrogram Transformer (PaSST) [27] is designed for analyzing mel spectrograms and considers frequency information explicitly. It has previously been used for audio classification [27] on Audio Set [44], which is a large-scale audio classification dataset with over 2 million sound clips. PaSST divides a 2-D mel spectrograms into smaller patches. Next, both frequency and time positional encodings are added to each patch so that the model knows the temporal and frequency ranges the patch represents. Finally, the patches are passed through the transformer. PaSST uses a technique called patchout to exclude certain patches during training. Patchout both shortens the input length (because fewer patches are analyzed) and regularizes the network (by forcing the network to succeed even when parts of the audio signals are missing). This reduces the network's reliance on certain frequency and temporal segments so that it can generalize to new audio signals better. The PaSST model used in our experiments has approximately 85.3 million parameters. We use an initial learning rate of 10^{-5} and weight decay of 10^{-4} with the AdamW optimizer [43] to train PaSST from scratch on the ASVspoof2019 dataset [30]. Training occurs for 51 epochs with a batch size of 12. We use a patch stride of 10 both in time and frequency.

D. Self-Supervised Audio Spectrogram Transformer

Self-Supervised Audio Spectrogram Transformer (SSAST) [28] is another transformer that has been used for audio classification, and it is modeled after the Audio Spectrogram Transformer [42], which is based on the first transformer used on images [45]. Similar to PaSST, SSAST divides a 2-D mel spectrogram into patches and uses patch-based dropout to regularize the network. Because large amounts of labeled data are not always readily available for a specific task, SSAST proposes a self-supervised learning stage to use unlabeled data to learn general characteristics of a certain data distribution. After SSAST completes the self-supervised learning stage, it is fine tuned on labeled data for a specific task. However, our experiments indicate that the self-supervision stage is not necessary for our task. We obtain better synthesized speech detection by training SSAST from scratch on our experimental dataset. We use an initial learning rate of 10^{-4} and weight decay $5 * 10^{-7}$ with the Adam optimizer [46] to train SSAST for 50 epochs with a batch size of 48, 48, and 12 for mel spectrograms sized 128x256, 128x512, and 128x1024, respectively. This network has approximately 87 million parameters.

E. Transformer Ensemble

After each of the transformers is trained for synthesized speech detection, we fuse the probabilities produced by each transformer to create the transformer ensemble. Since we train three transformers separately, we will fuse three output probabilities (one from each transformer) to create the transformer ensemble. In our experiments, the output probabilities indicate the liklihood that an audio signal is synthesized. We explore two fusion techniques: averaging and maximizing. In the averaging approach, the three output probabilities from each of the transformers are averaged to determine the final transformer ensemble probability. In the maximizing approach, the maximum probability of the three output probabilities is used as the final transformer ensemble probability. Results indicate that the averaging technique detects synthesized speech better than the maximizing technique, so we report the averaging results in this paper. We also explore an ensemble using only the two best-performing transformers. Results indicate that the two-transformer ensemble is better than the three-transformer ensemble, so we report those results in this paper. The two best-performing transformers used in the transformer ensemble are CCT and PaSST, so we refer to the transformer ensemble as "CCT-PaSST".

IV. Results

A. Dataset

We utilize the logical access (LA) portion of the ASVspoof2019 dataset [30] in our experiments. The dataset contains *authentic* speech signals spoken by humans as well as *synthesized* speech signals generated with neural acoustic models and deep learning methods. Some of the deep learning methods used to synthesize speech are LSTMs and Generative Adversarial Networks (GANs) [47] [48]. The ASVspoof2019 dataset contains significantly more synthesized speech signals than genuine speech signals (108,978 synthesized speech signals compared to 12,483 authentic ones). We utilize the official dataset split according to the challenge for training, validating, and testing our approach. Table I summarizes the details of the dataset.

TABLE I: Dataset used in our experiments, where each integer refers to number of audio tracks.

Subset	Synthesized	Authentic	Total
Training	22,800	2,580	25,380
Validation	22,296	2,548	24,844
Testing	63,882	7,355	71,237
Total	108,978	12,483	121,461

B. Evaluation Metrics

For all experiments, we report Receiver Operating Characteristic Area Under the Curve (ROC AUC) and Precision Recall Area Under the Curve (PR AUC) [49], [50]. For a binary classification problem (such as this synthesized speech detection task), a threshold must be selected to use as a cutoff to convert output probabilities to discrete categories (*i.e.*, *synthesized* or *authentic*). ROC AUC and PR AUC summarize detection performance using a full range of thresholds. Thus, they capture the trade-offs between the true positive rate and false positive rate (in the case of ROC AUC) and between true positive rate and the positive predictive value (in the case of PR AUC). Rather than reporting accuracy, precision, recall, and F1 metrics for a specific threshold, we wish to evaluate

all methods based on their "robustness", or independence to a specific threshold. If ROC AUC and PR AUC are high, the detection method is able to detect synthesized speech for a large range of thresholds.

ROC AUC indicates how skillful (*i.e.*, successful) a detection method is. In a way, it can be interpreted as an indication of accuracy for different thresholds. However, it does not account for class imbalances. If there are significantly more samples of one class in the evaluation set compared to the other class (as is the case with our dataset), it is possible for a detection method to have a high ROC AUC even if the detection method predicts the majority class all the time (but never predicts the minority class). For imbalanced datasets, it is important to consider other metrics, such as precision and recall, which measure how often the detection method correctly predicts the minority class. Thus, PR AUC is an important evaluation metric for our task because it will reflect the skill of all methods on our imbalanced dataset.

C. Experimental Results

Table II summarizes our experimental results. At first glance, results indicate that each transformer performs reasonably well on this task. Each individual transformer can achieve ROC AUC above 0.95 for at least one mel spectrogram size. CCT achieves the highest ROC AUC of the individual transformers for all mel spectrogram sizes, with all of its ROC AUC scores above 0.96. PaSST has the second-highest ROC AUC scores, which are all above 0.92. However, the ROC AUC scores fluctuate with different mel spectrogram sizes. For example, PaSST achieves 0.9258 ROC AUC for mel spectrograms sized 128x512, but it achieves a ROC AUC of 0.9590 for mel spectrograms sized 128x1,024. SSAST experiences the greatest fluctuation with respect to mel spectrogrm size. It has a ROC AUC of 0.7541 for mel spectrograms sized 128x512, but it can achieve a ROC AUC of 0.9536 for mel spectrograms sized 128x256. The CCT-PaSST ensemble achieves the highest ROC AUC scores, outperforming all individual transformer. It also achieves these results consistently for all mel spectrogram sizes. Notice that its ROC AUC scores are higher than 0.98 for all mel spectrogram sizes. Its ROC AUC scores vary less than 0.0015 for different mel spectrogram sizes. Thus, not only does the transformer ensemble (CCT-PaSST) achieve the best results in terms of ROC AUC, it also experiences the least fluctuation for different mel spectrogram sizes.

PR AUC scores have more variability compared to ROC AUC scores. Individual transformers achieve PR AUC scores ranging from 0.2213 to 0.7551. PaSST achieves the highest PR AUC of the individual transformers for all mel spectrogram sizes. Its PR AUC is greater than 0.70 for all mel spectrogram sizes. CCT has the second-highest PR AUC scores, which are all above 0.64 for all mel spectrogram sizes. The last transformer, SSAST, experiences the greatest fluctuation in terms of PR AUC. Its PR AUC scores range from 0.2213 to 0.6774 with different mel spectrogram sizes. The CCT-PaSST transformer ensemble achieves the best PR AUC scores overall and again outperforms all individual transformers. Its PR AUC is higher than 0.85 for all mel spectrogram sizes, which is higher than all PR AUC scores achieved with individual transformers by a significant margin.

TABLE II: Results for synthesized speech detection. Recall that the second dimension of values in the column titled "Mel Spectrogram Size" refers to the width of mel spectrograms used in the analysis, where the width corresponds to the length of the audio signals.

Method	Mel Spectrogram Size	ROC AUC	PR AUC
CCT	128 x 256	0.9668	0.6446
PaSST	128 x 256	0.9483	0.7332
SSAST	128 x 256	0.9536	0.6774
CCT-PaSST	128 x 256	**0.9810**	**0.8507**
CCT	128 x 512	0.9742	0.6876
PaSST	128 x 512	0.9258	0.7012
SSAST	128 x 512	0.7541	0.2213
CCT-PaSST	128 x 512	**0.9816**	**0.8508**
CCT	128 x 1,024	0.9718	0.7512
PaSST	128 x 1,024	0.9590	0.7551
SSAST	128 x 1,024	0.9064	0.4265
CCT-PaSST	128 x 1,024	**0.9801**	**0.8585**

Again, we observe that each method's results are impacted by mel spectrogram size. CCT PR AUC increases from 0.6446 to 0.7512 when the mel spectrogram size increases from 128x256 to 128x1,024. PaSST PR AUC ranges from 0.7012 to 0.7551 for different sizes. SSAST shows the most variability to mel spectrogram sizes with its PR AUC ranging from 0.2213 to 0.6774. However, the CCT-PaSST transformer ensemble achieves consistent results for all mel spectrogram sizes. Its PR AUC ranges from 0.8507 to 0.8585 for different mel spectrogram sizes, meaning that its results vary by the small value of 0.0078. From the results in Table II, we conclude that each individual transformer is sensitive to mel spectrogram size, but our transformer ensemble consistently and successfully detects synthesized speech from all mel spectrogram sizes.

Figure 2 shows the ROC and PR curves for all methods trained and evaluated on mel spectrograms sized 128x256. From this figure, we see that all detection methods achieve high ROC curves that are fairly similar. The PR curves are more distinct, though. Clearly from Figure 2, we see that the CCT-PaSST ensemble outperforms all individual transformers. Thus, from both Table II and Figure 2, we conclude that our proposed transformer ensemble achieves the best synthesized speech detection.

As mentioned earlier, PR AUC is an important metric for our task since our experimental dataset is highly imbalanced. The CCT-PaSST transformer ensemble can significantly increase synthesized speech detection in an imbalanced evaluation scenario, as evidenced by its high PR AUC scores. It is more robust to different thresholds used to differentiate between *synthesized* and *authentic* speech, which indicates that it is generally more confident of its detection capabilities. When it predicts that a speech signal is *synthesized*, it predicts a probability that is higher than the individual transformers'

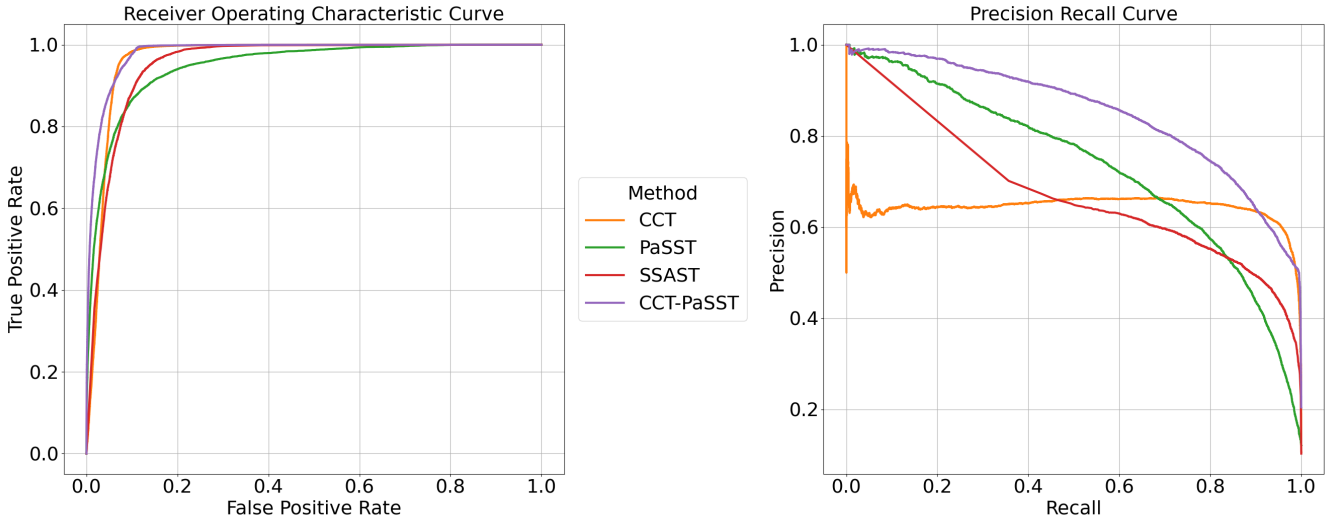

Fig. 2: ROC and PR curves of each evaluated method for mel spectrograms sized 128x256.

probabilities. This is a useful feature for forensics analysts because it requires less calibration for an operational scenario. Since the CCT-PaSST transformer ensemble is less sensitive to threshold selection, a forensics analyst can use a larger range of threshold values without concern of seeing drastically different detection results.

V. Conclusion

This paper investigates three transformers for synthesized speech detection and proposes a transformer ensemble to boost performance. We show that our proposed transformer ensemble achieves better synthesized speech detection than each of the individual transformers, especially considering the highly imbalanced nature of our experimental dataset. We also demonstrate that our transformer ensemble can achieve the same level of high success, even when analyzing less of a speech signal. Future work will focus on other tasks in addition to detection, such as speech synthesizer attribution and localization of synthesized speech within full audio signals.

Acknowledgment

This paper is based on research sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under agreement number FA8750-20-2-1004. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, AFRL, or the U.S. Government.

Address all correspondence to Edward J. Delp at ace@ecn.purdue.edu.

References

[1] S. Malodia, N. Islam, P. Kaur, and A. Dhir, "Why Do People Use Artificial Intelligence (AI)-Enabled Voice Assistants?" *IEEE Transactions on Engineering Management*, pp. 1–15, December 2021.

[2] Apple, "Siri Does More than Ever. Even Before You Ask." 2021. [Online]. Available: https://www.apple.com/siri/

[3] Google, "Google Assistant in Your Car," 2021. [Online]. Available: https://assistant.google.com/platforms/cars/

[4] Amazon, "What Is Alexa?" 2021. [Online]. Available: https://developer.amazon.com/en-US/alexa

[5] H. Mandlekar, S. Purohit, P. Kadam, and H. Tigaiya, "AVA - A Cloud-based Banking Virtual Assistant," *Proceedings of the IEEE International Conference on Intelligent Technologies*, pp. 1–6, June 2021, Hubli, India.

[6] Y. Zhu, Z. C. Gan, and Y. Huang, "The Virtual Sales Assistant in SaaS," *Proceedings of the IEEE International Conference on Advanced Computer Theory and Engineering*, vol. 5, pp. V5–238–V5–241, August 2010, Chengdu, China.

[7] A. Kongthon, C. Sangkeettrakarn, S. Kongyoung, and C. Haruechaiyasak, "Implementing an Online Help Desk System Based on Conversational Agent," *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, pp. 450–451, October 2009, Lyon, France.

[8] A. Hutchinson, "Instagram Adds New Text-to-Speech and Voice Effect Options in Reels," November 2021. [Online]. Available: https://www.socialmediatoday.com/news/instagram-adds-new-text-to-speech-and-voice-effect-options-in-reels/609925/

[9] K. Hatton, "How to Use TikTok's Text-to-Speech feature," July 2021. [Online]. Available: https://www.theverge.com/22594929/tiktok-text-to-speech-how-to

[10] 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.

[11] 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.

[12] 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.

[13] 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, pp. 324–329, August 2022, Virtual.

[14] 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.

[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 Media Forensics: Methods and Threats," *arXiv:2204.12067*, pp. 1–17, April 2022.

[16] H. Ajder, G. Patrini, F. Cavalli, and L. Cullen, "The State of Deepfakes: Landscape, Threats, and Impact," *Deeptrace Lab*, September 2019. [Online]. Available: https://regmedia.co.uk/2019/10/08/deepfake_report.pdf

[17] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1–11, August 2019, Seoul, Korea.

[18] R. Toews, "Deepfakes are Going to Wreck Havoc on Society. We are Not Prepared." *Forbes*, May 2020. [Online]. Available: https://www.forbes.com/sites/robtoews/2020/05/25/deepfakes-are-going-to-wreak-havoc-on-society-we-are-not-prepared/#67f7615d7494

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

[20] T. Telegraph, "Deepfake Video of Volodymyr Zelensky Surrendering Surfaces on Social Media," March 2022. [Online]. Available: https://www.youtube.com/watch?v=X17yrEV5sl4

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

[22] L. Yang, D. Holtz, S. Jaffe, S. Suri, S. Sinha, J. Weston, C. Joyce, N. Shah, K. Sherman, B. Hecht, and J. Teevan, "The Effects of Remote Work on Collaboration Among Information Workers," *Nature Human Behavior*, vol. 6, no. 1, pp. 43–54, September 2021.

[23] B. Wang, Y. Liu, J. Qian, and S. K. Parker, "Achieving Effective Remote Working During the COVID-19 Pandemic: A Work Design Perspective," *Applied Psychology*, vol. 70, no. 1, pp. 16–59, October 2020.

[24] S. Lund, A. Madgavkar, J. Manyika, and S. Smit, "What's Next for Remote Work: An Analysis of 2,000 Tasks, 800 Jobs, and Nine Countries," *McKinsey Global Institute*, November 2020. [Online]. Available: https://www.mckinsey.com/featured-insights/future-of-work/whats-next-for-remote-work-an-analysis-of-2000-tasks-800-jobs-and-nine-countries

[25] 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 International Conference on Neural Information Processing Systems*, pp. 1–11, December 2017, Long Beach, CA, USA.

[26] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the Big Data Paradigm with Compact Transformers," *arXiv:2104.05704*, pp. 1–18, April 2021.

[27] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient Training of Audio Transformers with Patchout," *Proceedings of the ISCA Interspeech Conference*, pp. 1–5, September 2022, Incheon, Korea.

[28] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-Supervised Audio Spectrogram Transformer," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 10 699–10 709, February 2022, Virtual.

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

[30] J. Yamagishi, M. Todisco, M. Sahidullah, H. Delgado, X. Wang, N. Evans, T. Kinnunen, K. A. 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*, 2019.

[31] B. P. Bogert, M. R. Healy, and J. W. 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, USA.

[32] A. Chintha, B. Thai, S. J. Sohrawardi, K. M. Bhatt, A. Hickerson, M. Wright, and R. Ptucha, "Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection," *The IEEE Journal of Selected Topics in Signal Processing*, pp. 1024–1037, June 2020.

[33] G. Hua, A. B. J. Teoh, and H. Zhang, "Towards End-to-End Synthetic Speech Detection," *IEEE Signal Processing Letters*, vol. 28, pp. 1265–1269, June 2021.

[34] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and Model Fusion for Automatic Spoofing Detection," *Proceedings of the Conference of the International Speech Communication Association*, pp. 102–106, August 2017, Stockholm, Sweden.

[35] Z. Chen, W. Zhang, Z. Xie, X. Xu, and D. Chen, "Recurrent Neural Networks for Automatic Replay Spoofing Attack Detection," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2052–2056, April 2018, Calgary, Canada.

[36] E. R. Bartusiak and E. J. Delp, "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, Virtual.

[37] H. Hao, E. R. Bartusiak, D. Güera, D. Mas, S. Baireddy, Z. Xiang, S. K. Yarlagadda, R. Shao, J. Horváth, J. Yang, F. Zhu, and E. J. Delp, "Deepfake Detection Using Multiple Data Modalities," in *Handbook of Digital Face Manipulation and Detection - From DeepFakes to Morphing Attacks, Series on Advances in Computer Vision and Pattern Recognition*. Springer, January 2022, pp. 235–254.

[38] 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*, pp. 1426–1430, October 2021, Asilomar, CA, USA.

[39] E. R. Bartusiak, "Machine Learning for Speech Forensics and Hypersonic Vehicle Applications," Ph.D. dissertation, Purdue University, West Lafayette, IN, USA, December 2022.

[40] 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*, pp. 329–336, December 2022, Nassau, Bahamas.

[41] E. Conti, D. Salvi, C. Borrelli, B. Hosler, P. Bestagini, F. Antonacci, o. Sarti, M. C. Stamm, and S. Tubaro, "Deepfake Speech Detection Through Emotion Recognition: A Semantic Approach," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8962–8966, May 2022, Singapore.

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

[43] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *Proceedings of the International Conference on Learning Representations*, pp. 1–19, May 2019, New Orleans, LA, USA.

[44] J. F. Gemmeke, D. P. W. 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*, pp. 776–780, March 2017, New Orleans, LA, USA.

[45] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *Proceedings of the International Conference on Learning Representations*, pp. 1–22, May 2021, Virtual.

[46] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *Proceedings of the International Conference for Learning Representations*, pp. 1–15, May 2015, San Diego, CA, USA.

[47] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computing*, vol. 9, no. 8, p. 1735–1780, November 1997.

[48] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," *Proceedings of the International Conference on Neural Information Processing Systems*, vol. 27, pp. 1–9, December 2014, Montréal, Canada.

[49] A. Tharwat, "Classification Assessment Methods," in *Applied Computing and Informatics*. Emerald Publishing Limited, Brighton, United Kingdom, July 2020, pp. 168–192.

[50] J. Davis and M. Goadrich, "The Relationship between Precision-Recall and ROC Curves," *Proceedings of the International Conference on Machine Learning*, pp. 233–240, June 2006, Pittsburgh, Pennsylvania, USA.