5th Symposium on Security and Privacy in Speech Communication
16 August 2025, Delft, The Netherlands

Defence Against the Deepfake Arts : Improving Audio Deepfake Detection With Context Awareness

Ivanina Ivanova[1*], *Abhay Dayal Mathur*[1*], *Nicoline Nymand-Andersen*[1,2*], *Nils Holzenberger*[1]

[1]Telecom Paris, Institut Polytechnique de Paris, France
[2]Department of Computer Science, Technical University of Munich, Germany

ivanina.ivanova@ip-paris.fr, abhay.mathur@ip-paris.fr, nnymand-23@ip-paris.fr, nils.holzenberger@telecom-paris.fr

Abstract

The increasing use of generative AI models to create realistic deepfakes poses significant challenges to information security, particularly in the realm of audio manipulation. Current audio deepfake detection methods focus on the acoustic signal and learn to spot artifacts produced by specific deepfake generators. This makes them inherently brittle when deployed to real-world datasets. We start from the intuitive observation that, in audio deepfakes, the person heard on the recording did not author the content of the utterance. This leads us to leverage speaker and author representations for speech and text. We introduce a novel multimodal approach, Defence Against the Deepfake Arts (DADA), involving two independent models for speech and text trained using contrastive learning, which then feed into a classifier fine-tuned for deepfake detection. We show empirically how our method robustly transfers to multiple methods of deepfake generation, setting a new state-of-the-art on the EER metric on multiple benchmarks.

Index Terms: audio deepfake detection, countermeasures, robustness

1. Introduction

Highly realistic AI-generated content such as images, audio, and even videos can be created with open-source software [1]. Audio deepfakes use deep learning techniques to create voices that convincingly impersonate real people [2]. Distinguishing fake from real media becomes nearly impossible for the ordinary user. Deepfakes thus affect a range of domains, from political discourse to personal security. These deepfakes are not only used in creating political propaganda to manipulate public opinion [3] but have also become tools for conducting phone scams [4], posing significant threats to individual privacy and security. Yet robust detection mechanisms are still lacking [5]. This paper contributes to proactive measures in detecting audio deepfakes, focusing on public figures. Public figures are especially susceptible to the detrimental effects of audio deepfakes as their extensive online presence renders them prime targets for malicious actors seeking to disseminate misinformation.

Current state-of-the-art systems are typically benchmarked on the ASVspoof dataset [6], which comprises scripted recordings and spoofs generated using predefined synthesis techniques. A new dataset, In The Wild [5], was released recently to reflect the actual landscape of deepfake generation. This dataset contains real-world recordings and spoofs generated using a variety of techniques, including voice conversion and text-to-speech. Any deepfake countermeasure system therefore needs to be evaluated on both ASVspoof and In The Wild to assess its robustness in detecting real-world deepfakes. In particular, we are interested in assessing how models trained on ASVspoof transfer to In The Wild.

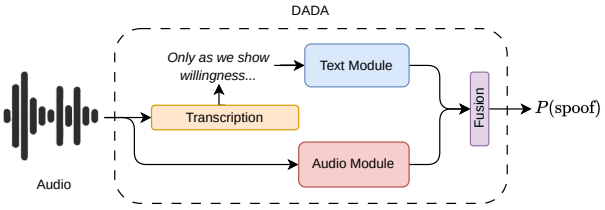

Figure 1: *Overview of the DADA pipeline. Audio files are first transcribed to obtain textual content. Features are then extracted from both the speech and text using separate models. Finally, these features are combined (using either mid fusion or late fusion strategies) to generate the final prediction.*

Aiming to increase robustness in deepfake detection, we propose a new pipeline that incorporates contextual information. Deepfakes can be leveraged to misattribute quotes to public figures. While a deepfake detector focused on audio might fail to find the spoof, a human would consider the mismatch between the content of the speech and the public figure's otherwise well-known opinion. Our approach, named Defence Against the Deepfake Arts (DADA) and summarized in Figure 1, takes this intuition into account to propose a pre-training objective for deepfake detection. We augment an audio-based deepfake classifier with an authorship attribution model. The latter learns to map the speech transcript to the most likely author. A final classification layer detects discrepancies between the audio-based embedding — trained to represent the speaker's features — and the text-based embedding — trained to represent the author of the speech's content.

Deepfake generation methods broadly fall into 6 categories [7], any of which may be used to create deepfakes of public figures. The first approach is imitation-based (voice conversion), wherein the original audio signal is modified to mimic another target voice. This technique is mainly used in show business and, therefore, is applied by a third person or deep learning software. The second approach is synthetic-based, producing artificially generated speech using text-to-speech techniques. This process involves three stages: text analysis, acoustic mapping, and vocoder. The third technique is centred on replaying existing recordings of the target speaker. Two techniques are notable here: cut-and-paste detection, where a victim's recorded segment is played through a phone handset for capture, and far-field detection, which involves crafting sentences for manipulation of a text-dependent system. Emotion fake alters the emotions of a speaker's voice while preserving the original words and speaker identity [8]. Scene fake involves

[*]Equal contribution

10.21437/SPSC.2025-11

modifying the background sounds of a recording, leaving the speaker and content untouched [9]. Finally, partial fakes involve replacing specific words within the original audio signal with either real or synthetic audio segments. The speaker remains the same throughout, but the content is altered [10].

Current automated deepfake detection methods typically use deep learning, and achieve high performance on benchmarks [6, 11]. Lightweight Convolutional Neural Networks (LCNN) [12] have been widely employed due to their efficiency in capturing spoofing artefacts in raw waveforms. RawNet2 [13], a variant of RawNet, leverages residual connections and attention mechanisms to improve feature extraction for deepfake detection. AASIST (Audio Anti-Spoofing using Integrated Spectro-Temporal features) [14] integrates spectro-temporal representations to enhance robustness against unseen attacks. RawGAT [15], a model combining raw waveform processing with Graph Attention Networks, has shown promising results in handling diverse spoofing techniques. More recently, and closer to our proposed approach, SLIM (Style-Linguistics Mismatch Model) [16] aims to leverage the mismatch between a speaker's style and linguistic content through self-supervised pretraining on real speech, to detect spoofing. We take an alternative approach, and focus solely on the identity of the speaker. We train author-attribution and speaker-identification models independently, and show how this information is sufficient to achieve competitive performance, with a new state-of-the-art on three evaluation benchmarks.

Recent advancements in authorship attribution using deep learning have showcased the effectiveness of various encoder-based architectures in capturing linguistic and stylistic patterns unique to individual authors. Multiple works explore BERT or T5-based models to identify authorship in literary texts, achieving notable accuracy [17, 18]. For instance, BertAA [19] is a fine-tuned pretrained BERT model for authorship attribution, demonstrating competitive performance on multiple datasets. Author identification framed as a classification problem for identifying authors of arXiv manuscripts has also been shown to achieve high attribution accuracy on a large-scale dataset with up to 2,000 authors [20]. Further, several methods use techniques from contrastive learning in creating discriminative embedding spaces [21] for short-text authorship attribution at scale, by leveraging triplet loss and Siamese networks [22], outperforming existing baselines on Twitter and Weibo datasets.

2. Method

In this section, we present our proposed architecture, DADA. It combines the traditional method of analyzing the audio of the speech sample, with contextual information extracted from the sample. Specifically, our approach seeks to determine not only whether a speech sample is a deepfake, but also the likelihood of the supposed speaker having made the given statement, thus embedding contextual awareness into the detection pipeline. An overview of the pipeline is shown in Figure 1.

Our method operates in three stages: transcription, feature extraction, and fusion. First, the audio files are transcribed using Whisper [23] to obtain the textual content. Manual inspection has shown that those transcriptions are highly accurate, whether speech is real or machine-generated. Following transcription, the architecture leverages two models trained independently: a text and a speech model, with the goal of learning speaker and author representations, respectively. The extracted speech and text features are then combined using one of two fusion strategies to obtain the final classification score.

2.1. Text Model

The main idea of the text model is to connect the textual surface form of the utterance with possible authorship. To do this, we fine-tune a pretrained encoder-based LLM for authorship attribution. To improve the model's ability to take into account stylistic and linguistic patterns unique to specific authors, we extracted the features from the last l hidden states of the encoder, instead of only the `CLS` token. Transformer models encode information hierarchically across layers. The lower layers capture phrase-level or segment-level interactions, where the intermediate layers represent syntactic information like grammar and vocabulary, and the higher layers encode semantic information. By using multiple layers we aim to capture a broader spectrum of features, combining syntactic and semantic information [24]. The extracted features are aggregated across subword units using a mean-pooling layer. This vector is passed to several fully-connected layers. To model the space of authors effectively and identify the characteristics of their speech, we employ contrastive learning using a triplet loss. This loss function helps create a well-defined embedding space, where texts attributed to the same author are closer together, while those from different authors are further apart [25]. The function is defined using triplets (A, P, N), where A is a randomly sampled anchor (reference point), P is a positive sample that has the same label as A, and N is the negative sample of a different random class than the anchor. Here, we use authors as labels, so that A and P are quotes sharing the same author, and N is a quote from a different author. The goal is to ensure that the distance between A and P is smaller than the distance between A and N by at least a certain margin λ. We define the loss function as follows:

$$\mathcal{L} = [d(f_\theta^A, f_\theta^P) - d(f_\theta^A, f_\theta^N) + \lambda]_+ \quad (1)$$

where $f_\theta(x)$ is the embedding of string x generated by the model, λ is the margin that enforces a minimum separation between similar and dissimilar pairs, $d(.,.)$ is the distance between the two embeddings, and $[\cdot]_+ = \max(\cdot, 0)$. We experiment with two different distance functions, squared Euclidean distance (L2-norm) and cosine (dis)similarity.

2.2. Speech Model

The speech model uses a Wav2Vec2 [26] backbone for feature extraction, which allows extracting representations directly from raw audio waveforms. To capture a comprehensive set of features, multiple layers from the model are pooled into a single feature vector. This pooling is achieved using a compression module that employs attentive mean pooling, permitting the model to focus on the most informative parts of the audio signal. Additionally, following [16], a non-linear bottleneck is applied to refine the feature vector, ensuring that it captures the essential characteristics needed for accurate deepfake detection.

Similar to the text module, speech features (denoted as f_ϕ) are trained using contrastive learning, with a triplet loss as in Equation 1, using cosine distance. The (A, P, N) triplets follow the same structure as the text module: A and P are speech samples from the same speaker, while N is a speech sample from a different speaker. To optimize the margin used in the triplet loss function, we employ an iterative process that involves three stages.

In the first stage, we train the model for a few epochs with the triplet loss function defined in Equation 1 using a margin λ_0, fixed *a priori*. This initial training helps achieve some separation between classes and provides a good starting point for

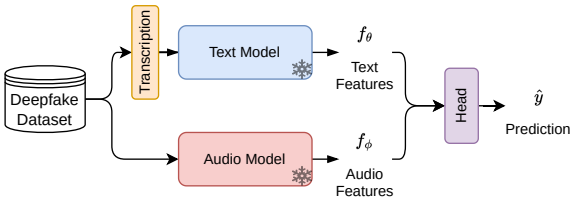

Figure 2: *Pipeline for Mid Fusion. Features from the text and speech modules are concatenated and passed through an MLP to produce the final classification.*

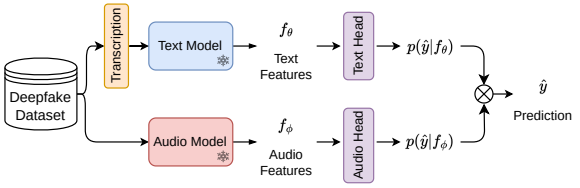

Figure 3: *Pipeline for Late Fusion. Predictions from the speech and text models are combined using a weighted mean to produce the final classification.*

further optimization. Then, in the second stage, we train the model for some additional epochs using the adaptive triplet loss introduced in [27]. This adds an additional soft constraint on the distance between the anchor and the negative sample using another margin λ' as follows:

$$\mathcal{L}_{ada} = [d(f_\phi^A, f_\phi^P) - d(f_\phi^A, f_\phi^N) + \lambda]_+ + [\lambda' - d(f_\phi^A, f_\phi^N)]_+ \quad (2)$$

In the third stage, we set the margin to the final margin obtained from the adaptive stage, λ_f, and train the model until convergence to achieve optimal performance. Empirically, we find that this method leads to greater separation among speech embeddings compared to training with the original triplet loss, and removes the need to tune the margin as a hyperparameter.

2.3. Fusion Strategies

After independently training the speech and text models, we freeze their parameters and combine their features using one of two fusion strategies: mid fusion or late fusion.

2.3.1. Mid Fusion

In mid fusion (Figure 2), the extracted features from the speech and text models are concatenated to create a joint representation. This combined representation is then passed through a Multi-Layer Perceptron (MLP). The MLP is trained on annotated examples of deepfake and bonafide speech and then learns to integrate the speech and text features effectively to make a final prediction. The training of the MLP is guided by the Binary Cross-Entropy loss.

2.3.2. Late Fusion

In this approach, the speech and text models are treated as independent branches. Each produces its own prediction score. We formulate this fusion strategy for the scenario where there exists a purported speaker, consistent with deepfakes for a specific individual or public figure. We add a binary classification head to both the speech and text models. The text model outputs the probability of the anticipated speaker having made the given statement, while the speech model predicts the likelihood of the speech being real or fake. The final prediction is obtained by taking the weighted mean of the scores from the two branches, as shown in Figure 3. This simple averaging method ensures that the final decision benefits from the independent contributions of both modalities while maintaining the interpretability of individual scores. Formally, we define the final prediction as:

$$\hat{y} = \alpha \cdot p(\hat{y}|f_\phi) + (1 - \alpha) \cdot p(\hat{y}|f_\theta) \quad (3)$$

Where $p(\hat{y}|f_\phi)$ and $p(\hat{y}|f_\theta)$ are the predictions from the speech and text models respectively, and α is the weight assigned to the speech model's prediction.

Table 1: *ABX accuracy of text models (in %) fine-tuned on Wikiquotes and VoxCeleb2 [28], on the validation sets. The parameter λ denotes the margin used in Eq. 1.*

Model	λ	Wikiquotes	VoxCeleb2
T5 Encoder [31]	0.44	77.6	76.8
Flan-T5 Encoder [32]	0.35	76.0	75.3
ModernBERT [33]	0.50	81.0	70.0
DeBERTa [34]	0.50	**84.1**	**86.5**

3. Experiments

3.1. Text Model

3.1.1. Datasets

For the authorship attribution task, we use two datasets, Wikiquotes[1] and VoxCeleb2 [28]. Both of them provide a rich corpus of quotes from various authors. We filter Wikiquotes, a repository of sourced quotations from notable individuals, to retain only the 58 authors present in the In The Wild dataset. The filtered dataset is then split into a development set (70%) and a validation set (30%), resulting in 5740 samples for development and 2461 for testing. VoxCeleb2, on the other hand, contains speeches from celebrities extracted from YouTube videos. We use a subset of the dataset, selecting 5,000 random samples for training and 3,000 for validation. This results in 256 unique authors in the training set and 343 in the validation set.

3.1.2. Metrics

ABX accuracy is a commonly used metric in speech and speaker representation learning, often used in tasks related to contrastive learning [29, 30]. Similar to triplet loss, we have three samples A, B and X, where A and B are from two different categories (in this case, two quotes from two different authors) and X is from one of the two categories (in this case, a third quote, from one of the two previous authors). The task is to determine whether X and A are from the same author, or whether X and B are. This is done by implementing a decision rule based on the learned embeddings: if $d(f_\theta(X), f_\theta(A)) < d(f_\theta(X), f_\theta(B))$ then A, else B. The accuracy measures the percentage of times the model correctly assigns X to its corresponding label (closer to A or B). A random decision rule would yield 50% accuracy.

3.1.3. Benchmarking

We benchmark the text model using state-of-the-art encoder-only architectures: DeBERTa [34], ModernBERT [33], and the encoder of T5 [31] and Flan-T5 [32]. Encoder-only models are

[1] https://en.wikiquote.org/wiki/Main_Page

well-suited for authorship attribution as they generate fixed-size representations, capturing stylistic patterns and semantic coherence. To identify the most effective architecture, we fine-tuned each model using Ray Tune [35], optimizing the triplet loss margin λ (second column in Table 1), the number of hidden layers for feature extraction and the number of fully connected layers after the aggregated features. These numbers differ from model to model, but for the best-performing model DeBERTa-v3-large, optimal results were achieved using only the last hidden layer, a two-layer MLP with dropout rate of 0.2, and a 512-dimensional feature output. All experiments were conducted on a single NVIDIA A100 GPU with 40GB vRAM. The results of the text model can be found in Table 1, and show that the learned representations distinguish authors effectively.

3.2. Speech Model

The speech model is pre-trained on the speaker representation learning task (Section 2.2) on an aggregation of the CommonVoice [36], RAVDESS [37] and VoxCeleb2 [28] datasets. Both CommonVoice and RAVDESS contain recordings of speakers reading predefined sentences. The aggregation contains 43 hours of real speech data across 1067 speakers.

We train the model using the adaptive-margin, contrastive learning strategy described in Section 2.2, and evaluate its performance with ABX accuracy. Our best model achieves an ABX accuracy of 92% on the aggregation.

3.3. Deepfake Detection

We further fine-tune both our fusion pipelines on the Logical Access (LA) subset of the ASVspoof 2019 Dataset [6]. This dataset contains a variety of real and spoofed speech samples designed to benchmark the performance of automatic speaker verification systems against spoofing attacks. We then evaluate the models' performance on the Deepfake (DF) subset of ASVspoof 2021, which contains spoofs generated from more sophisticated systems, and two out-of-domain datasets, In The Wild [5] and the English subset of MLAAD [38]. This helps us assess its generalization and robustness to real-world deepfakes.

We consider two metrics for our evaluations: F1-score and Equal Error Rate (EER). EER is a standard metric for audio deepfake detection systems, denoting the point at which the false acceptance rate and the false rejection rate are equal. A lower EER indicates better performance. As compared to F1, EER accounts for a possible tuning of the detection threshold, which would occur when deploying a deepfake detection system in practice.

4. Results

Our results can be found in Table 2. We consider four variants of our method for comparison and ablation: text-only, speech-only, late fusion, and mid fusion. The text-only model predicts the same label, `spoof`, regardless of the input, leading to infinitesmal F1 scores. This is expected, since the text modality entirely lacks the necessary information to make accurate predictions. The speech-only model outperforms baselines on ASVspoof and is competitive on the other two datasets. Interestingly, this is achieved with pre-training focused exclusively on speaker identification. Our speaker representations nonetheless contain relevant information for deepfake detection, as they achieve non-trivial performance. The late fusion strategy performs slightly better than the speech-only model, with the optimal text-module weight $\alpha = 0.61$, indicating that the text features provide relevant information but only marginal performance gains under this fusion strategy. The mid fusion model outperforms all DADA variants, achieving the highest F1 score on ASVspoof21 and the lowest EER across all datasets. Our best-performing mid fusion network is an MLP with a single hidden layer with 256 units, a Batch Normalisation operation after the first linear projection, and a dropout of 0.4. We observe a reduction of EER by 6.9%, 5.6% and 8.4% on ASVspoof 2019, In The Wild and MLAAD, respectively.

Table 2: *Performance (in %) of DADA compared to state-of-the-art baselines. Our mid fusion strategy outperforms all baselines on ASV Spoof 21 and achieves minimum EER for In The Wild and MLAAD[en].*

Method	ASVspoof21		In The Wild		MLAAD[en]	
	F1 ↑	EER ↓	F1 ↑	EER ↓	F1 ↑	EER ↓
LCNN [12]	19.7	25.5	37.3	65.6	65.4	37.2
Rawnet2 [13]	21.3	22.3	60.2	37.8	67.6	33.9
SLIM [16]	65.1	4.4	89.8	12.5	89.2	10.7
AASIST [14]	70.7	3.6	84.7	17.5	85.6	14.5
DADA Variants (Ours)						
Text-Only	0.8	46.1	0.6	47.7	0.0	50.1
Speech-Only	89.6	3.1	68.6	13.8	75.4	15.3
Late Fusion	89.2	2.9	70.1	13.3	84.0	14.8
Mid Fusion	**91.0**	**2.7**	72.1	**11.8**	79.7	**9.8**

Overall, these results support our motivating intuition: the speaker as identified by the audio, and the author features as identified by the content of the speech utterance, are valuable pieces of information in detecting audio deepfakes, across deepfake generation methods. Relying solely on acoustic artefacts for deepfake detection is in fact a source of overfitting, as those artefacts depend on the deepfake generation method. Incorporating context from transcripts is hence a method of reducing this over-reliance, by encouraging the deepfake detector to consider author information, as inferred from the content of the speech. SLIM makes a related but distinct assumption: that the audio signal carries information about the speaker and about the content of the speech, and that deepfake generation manipulates one type of information while leaving the other unchanged. Both SLIM and DADA attain competitive performance on multiple benchmarks. Both approaches leverage comparable amounts of bonafide data for pre-training, reducing the need for labeled deepfakes. However, both the linguistic and stylistic encoders in SLIM must be trained jointly, while the speech and text modules of DADA can be trained independently, on distinct data, as was done here. For deepfake detection, SLIM relies on learned style and linguistic embeddings, but also on the underlying speech encoder's features, whereas DADA only considers speaker and author embeddings.

5. Conclusions

In this paper, we introduced DADA, a novel approach for audio deepfake detection that leverages both speech and textual features to improve detection accuracy. Speech and text features are learned independently, on real data, reducing the reliance on annotated deepfake data. Our experiments demonstrate that DADA outperforms the state-of-the-art on multiple datasets, highlighting the potential of incorporating simple context awareness for robust deepfake detection. Future work will focus on further refining the fusion strategies and exploring additional datasets to validate the generalizability of our approach.

6. References

[1] K. Liu, I. Perov, D. Gao, N. Chervoniy, W. Zhou, and W. Zhang, "Deepfacelab: Integrated, flexible and extensible face-swapping framework," *Pattern Recognition*, 2023.

[2] P. Kawa, M. Plata, and P. Syga, "Defense against adversarial attacks on audio deepfake detection," in *Interspeech 2023*.

[3] G. Stănescu, "Informational war: Analyzing false news in the israel conflict," *Social Sciences and Education Research Review*, 2023.

[4] L. Yasur, G. Frankovits, F. M. Grabovski, and Y. Mirsky, "Deepfake CAPTCHA: A method for preventing fake calls," in *ASIA CCS 2023*.

[5] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar, and K. Böttinger, "Does audio deepfake detection generalize?" in *Interspeech 2022*.

[6] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. Kinnunen, M. Todisco, J. Yamagishi, N. Evans, A. Nautsch, and K. A. Lee, "ASVspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio Speech and Language Processing*, 2023.

[7] A. Dixit, N. Kaur, and S. Kingra, "Review of audio deepfake detection techniques: Issues and prospects," *Expert Systems*, 2023.

[8] Y. Zhao, J. Yi, J. Tao, C. Wang, and Y. Dong, "Emofake: An initial dataset for emotion fake audio detection," in *CCL 2024*.

[9] J. Yi, C. Wang, J. Tao, C. Zhang, C. Fan, Z. Tian, H. Ma, and R. Fu, "Scenefake: An initial dataset and benchmarks for scene fake audio detection," *Pattern Recognition*, 2024.

[10] J. Yi, Y. Bai, J. Tao, H. Ma, Z. Tian, C. Wang, T. Wang, and R. Fu, "Half-truth: A partially fake audio detection dataset," in *Interspeech 2021*.

[11] X.-M. Zeng, J.-T. Zhang, K. Li, Z.-L. Liu, W.-L. Xie, and Y. Song, "Deepfake algorithm recognition system with augmented data for ADD 2023 challenge," in *Deepfake Audio Detection and Analysis @ IJCAI 2023*.

[12] X. Wang and J. Yamagishi, "A comparative study on recent neural spoofing countermeasures for synthetic speech detection," in *Interspeech 2021*.

[13] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. W. D. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021*.

[14] J. Jung, H. Heo, H. Tak, H. Shim, J. S. Chung, B. Lee, H. Yu, and N. W. D. Evans, "AASIST: audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022*.

[15] H. Tak, J.-W. Jung, J. Patino, M. Kamble, M. Todisco, and N. Evans, "End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection."

[16] Y. Zhu, S. Koppisetti, T. Tran, and G. Bharaj, "SLIM: style-linguistics mismatch model for generalized audio deepfake detection," in *NeurIPS 2024*.

[17] R. M. M. Hicke and D. Mimno, "T5 meets tybalt: Author attribution in early modern english drama using large language models," in *Proceedings of the Computational Humanities Research Conference 2023, Paris, France, December 6-8, 2023*, ser. CEUR Workshop Proceedings, vol. 3558. CEUR-WS.org, 2023, pp. 274–302.

[18] K. Silva, I. Frommholz, B. Can, F. Blain, R. Sarwar, and L. Ugolini, "Forged-GAN-BERT: authorship attribution for llm-generated forged novels," in *EACL 2024: Student Research Workshop*, 2024.

[19] M. Fabien, E. Villatoro-Tello, P. Motlicek, and S. Parida, "BertAA: Bert fine-tuning for authorship attribution," in *ICON 2020*.

[20] L. Bauersfeld, A. Romero, M. Muglikar, and D. Scaramuzza, "Cracking double-blind review: Authorship attribution with deep learning," *Plos one*, vol. 18, no. 6, p. e0287611, 2023.

[21] Z. Hu, R. K.-W. Lee, L. Wang, E.-p. Lim, and B. Dai, "Deepstyle: User style embedding for authorship attribution of short texts," in *Web and Big Data: 4th International Joint Conference, APWeb-WAIM 2020, Tianjin, China, September 18-20, 2020, Proceedings, Part II 4*. Springer, 2020, pp. 221–229.

[22] C. Saedi and M. Dras, "Siamese networks for large-scale author identification," *Computer Speech & Language*, vol. 70, p. 101241, 2021.

[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[24] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?" in *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[25] C. Mao, Z. Zhong, J. Yang, C. Vondrick, and B. Ray, "Metric learning for adversarial robustness," *Advances in neural information processing systems*, vol. 32, 2019.

[26] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[27] K. Nguyen, H. H. Nguyen, and A. Tiulpin, "AdaTriplet: adaptive gradient triplet loss with automatic margin learning for forensic medical image matching," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part VIII*, ser. Lecture Notes in Computer Science, vol. 13438. Springer, 2022, pp. 725–735.

[28] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: deep speaker recognition," in *Interspeech 2018*.

[29] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline," in *Interspeech 2013*.

[30] M. Hallap, E. Dupoux, and E. Dunbar, "Evaluating context-invariance in unsupervised speech representations," in *Interspeech 2023*.

[31] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[32] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[33] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, and I. Poli, "Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference," *CoRR*, vol. abs/2412.13663, 2024.

[34] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," in *ICLR 2021*.

[35] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica, "Tune: A research platform for distributed model selection and training," *CoRR*, vol. abs/1807.05118, 2018.

[36] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *LREC 2020*.

[37] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLoS ONE*, vol. 13, 2018.

[38] N. M. Muller, P. Kawa, W. H. Choong, E. Casanova, E. Golge, T. Muller, P. Syga, P. Sperl, and K. Bottinger, "MLAAD: The multi-language audio anti-spoofing dataset," *IJCNN 2024*.