GAT BASED MODELS FOR SPEECH SPOOFING IN THE INDONESIAN LANGUAGE

Vincent Suhardi
Fakultas Ilmu Komputer, Universitas Indonesia
Depok, Jawa Barat
vincent.suhardi@ui.ac.id

Bryan Jeshua Mario Timung
Fakultas Ilmu Komputer, Universitas Indonesia
Depok, Jawa Barat
bryan.jeshua@ui.ac.id

Muh. Kemal Lathif Galih Putra
Fakultas Ilmu Komputer, Universitas Indonesia
Depok, Jawa Barat
muhkemal.lathif@ui.ac.id

ABSTRACT

Speech spoofing detection has become increasingly crucial with the advancement of speech synthesis technologies, particularly in languages with limited research coverage such as Indonesian. In this paper, we evaluate the performance of Graph Attention Network (GAT) based models—AASIST, AASIST-L, and RawGAT-ST—for detecting speech spoofing in the Indonesian language context. We develop a comprehensive dataset combining bonafide speech samples from the Indonesian News LVCSR dataset with synthetic speech generated using the VITS model through Coqui TTS. Our experimental results demonstrate both achievements and challenges in adapting state-of-the-art spoofing detection models to the Indonesian language. While all models achieved 0% Equal Error Rate (EER), further analysis through raw prediction scores revealed AASIST's superior discrimination capability, with the largest gap between bonafide (4.3964) and spoofed (-5.2115) classifications. AASIST also demonstrated the best loss convergence, achieving a final loss of 8.9×10^{-4}, followed by RawGAT-ST at 1.09×10^{-3} and AASIST-L at 1.64×10^{-3}. The study reveals important considerations for developing robust spoofing detection systems for Indonesian speech, particularly the need for more diverse synthetic speech data. While current results show promise, they also highlight the importance of expanding speaker diversity beyond binary gender representation in synthetic speech generation. This work provides a foundation for future research in Indonesian speech spoofing detection, contributing to the broader goal of securing voice-based applications in the Indonesian context.

[1] [2] [3]

***K*eywords** Speech spoofing, bonafide, AASIST, RawGAT-ST, RawNet2, Graph Attention Network (GAT), ASV, countermeasures

1 Introduction

The advancement of artificial intelligence (AI) technology has had significant impacts across various domains, particularly in speech recognition. The capability to generate synthetic voices indistinguishable from genuine human speech, especially through deep learning and Neural Network (NN) technologies continues to evolve rapidly. One notable negative consequence of this development is the emergence of spoofing attacks, which involve the manipulation or falsification of voice identities for fraudulent purposes. Indonesia, as the country with the fourth most social media users worldwide but the lowest literacy skills [1, 2], poses a huge risk in being fooled by AI-generated content or deepfakes and cause misinformation.

While various NN-based models have been employed for spoofing detection, their effectiveness often varies depending on the model architecture, data quality, and the target language [3, 4]. The primary challenges in spoofing de-

[1] Original AASIST Repository https://github.com/clovaai/aasist?tab=readme-ov-file

[2] Indonesian AASIST Fork https://github.com/InhumanlyInsane/aasist-id/tree/development

[3] Dataset & Model Checkpoints https://tinyurl.com/24l5da5z

tection encompass the diversity of techniques used for voice falsification, including replay, synthesis, and voice conversion [5, 6]. Although Automatic Speech Verification (ASV) technology has advanced significantly to address these issues, spoofing remains a substantial challenge. While extensive research has been conducted internationally, the majority of studies focus on globally dominant languages such as English and Mandarin [7, 8, 9, 10].

Research on spoofing detection within the Indonesian language context remains notably limited, potentially creating implementation challenges in Indonesia. This limitation is particularly problematic because the acoustic characteristics of the Indonesian language, which possess unique features, may affect the performance of spoofing detection models designed for other languages. Therefore, it is imperative to experiment on and adapt spoofing detection models to accommodate the distinctive characteristics of Indonesian speech to achieve optimal accuracy in detecting spoofing attempts in speech recognition systems.

This research proposes to address these challenges by developing a comprehensive and relevant spoofing dataset for the Indonesian language. The dataset will focus on synthesis spoofing, which involves the creation of synthetic voices using speech synthesis technology that converts text to speech through end-to-end NN models. Furthermore, this experiment will evaluate the performance of SOTA models that have achieved the best metrics on recent spoofing related studies. These models will be tested against the developed Indonesian language dataset, and the evaluation results will be used to provide an understanding towards the capabilities and effectiveness of each spoofing model in detecting spoofing attempts for the Indonesian language.

2 Related Studies

2.1 Graph Attention Networks (GAT)

Graph Attention Networks (GAT) represent a graph neural networks-based approach designed to model relationships between nodes in non-Euclidean data, such as audio signals that possess both spectral and temporal patterns. GAT employs an attention mechanism that enables the model to assign adaptive weights to inter-node relationships based on their relevance to specific tasks [11].

In the context of spoofing detection, GAT facilitates effective integration of spectral and temporal information. The AASIST model, as an evolution of RawGAT-ST, utilizes a heterogeneous stacking graph attention layer (HS-GAL) designed to project information from both spectral and temporal domains into the same latent space. This enables the model to capture spoofing artifacts present in both domains, thereby enhancing detection accuracy [12].

2.2 Automatic Speech Verification (ASV)

Speech spoofing tasks originate from Automatic Speech Verification (ASV) technology employed to verify whether a given voice originates from a legitimate speaker. However, the effectiveness of ASV systems can be compromised by spoofing attacks, including attack types like replay, synthesis, and conversion attacks. To address these threats, countermeasure mechanisms such as the AASIST model are designed to detect synthetic voices with high accuracy [13]. The performance evaluation of ASV typically employs two primary metrics:

- **Equal Error Rate (EER):** The operating point where the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR).

- **Tandem Detection Cost Function (t-DCF):** A more comprehensive metric for evaluating the impact of spoofing detection on ASV systems. This metric provides a holistic assessment of the system's ability to simultaneously handle both spoofing attacks and speaker verification tasks [14].

3 Methodology & Experiment

3.1 Experiment Flow

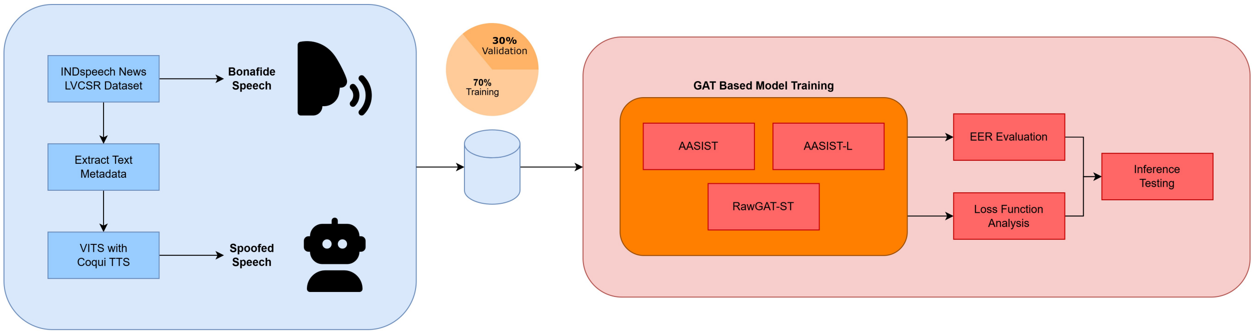

Figure 1. Experiment flow starting from dataset collection depicted by the left side and the color blue and the modeling phase depicted by the right side and the color red.

This research implements a comprehensive methodology to evaluate the effectiveness of speech spoofing detection models in the context of the Indonesian language. The research workflow is divided into two main phases: **data collection** and **data modeling**, as illustrated in the experimental flow diagram.

The data collection phase utilizes the Indonesian News LVCSR speech dataset (`INDspeech_NEWS_LVCSR`) from HuggingFace as the source of bonafide speech samples [15]. This dataset was selected for its representation of natural Indonesian speech variations from multiple speakers for both the male and female gender. Text metadata contained within the dataset is extracted and subsequently used as input for generating spoofed data through the Coqui TTS library, specifically using the VITS model. This process produces synthetic voice samples that maintain identical linguistic content to the bonafide data but are artificially generated. Both types of data are then combined to create a balanced and comprehensive dataset for model training and evaluation purposes. The combined dataset is strategically split into two portions: 70% for training and 30% for validation, ensuring robust model development and evaluation.

The modeling phase begins with a training process involving four GAT-based model architectures, as shown in the red section of the diagram. The evaluated models include AASIST and its lightweight version AASIST-L and its predecessor, RawGAT-ST. Each model is trained using the combined dataset with the objective of learning discriminative patterns that distinguish between bonafide and spoofed utterances. The models' performance is then evaluated using the Equal Error Rate (EER) metric, followed by comprehensive loss function analysis. The final stage then involves inference testing to analyze the models' ability to correctly classify audio samples as either bonafide or spoofed.

3.2 Dataset Collection

This research utilizes two distinct dataset sources to construct a comprehensive spoofing detection dataset with a 70% training and 30% validation split from the whole dataset. The primary dataset is the Indonesian News LVCSR dataset (INDspeech), available through the HuggingFace platform, which represents the first Indonesian speech corpus specifically developed for large vocabulary continuous speech recognition (LVCSR) [15]. This dataset encompasses over 40 hours of speech recordings collected from 400 speakers, making it a rich and diverse data source for Indonesian speech processing applications. In this research, the dataset serves as the source for bonafide speech samples and reference text for speech verification [13].

To generate synthetic speech samples (spoofed), this research employs the VITS model pre-trained on Indonesian speech data as a Text-to-Speech (TTS) model with an end-to-end architecture [16]. Inference of the VITS model is performed using Coqui TTS as a framework specifically designed to facilitate TTS tasks [17]. The text used for the VITS inference is sourced from the Indonesian News LVCSR dataset. To ensure dataset balance and consistency, voice selection in the VITS model is matched with the gender of speakers of the extracted bonafide audio files. This approach enables the generation of spoofed speech samples with identical duration and quantity to the bonafide samples while maintaining a balanced gender distribution within the dataset.

In the context of this research, the study focuses on the countermeasure task [13], where the system is focused on training to distinguish between genuine speech (bonafide) and artificially generated speech (spoofed). This approach aligns with the primary research objective of detecting spoofing attempts in the Indonesian language context, without performing speaker verification.

3.3 Spoofing Case

The ASVspoofing2019 dataset classifies various algorithms and synthetic speech generation systems into 19 cases, coded A01 to A19. This research specifically analyzes case A10, which represents NN-based end-to-end TTS systems with transfer learning implementation, particularly focusing on the use of pre-trained VITS models [12, 13].

The decision to focus on case A10 is based on the characteristics of the synthetic speech it produces, which demonstrates an exceptionally high degree of naturalness and similarity to genuine human speech. Given the sophistication of technology employed in case A10, this presents significant security risk potential, necessitating the development of effective anti-spoofing methods to identify and prevent the misuse of such technology. This forms the foundation for the importance of this research in investigating the characteristics and detection methods of spoofing in case A10.

3.4 Evaluation Metric & Loss Function

Drawing reference from the ASVspoofing2019 dataset, we conduct this research using only the Equal Error Rate (EER) metric to measure countermeasure detection performance. This decision was made because the t-DCF metric is more suitable for ASV-centric tasks involving speech matching rather than countermeasure tasks [14].

EER serves as the evaluation metric that measures the countermeasure system's performance in distinguishing between genuine and synthetic voice inputs. It can be formalized mathematically as follows:

$$\mathrm{EER} = \frac{\mathrm{FAR} + \mathrm{FRR}}{2} \tag{1}$$

From this equation, FAR represents the ratio of synthetic inputs incorrectly accepted as genuine, while FRR represents the ratio of genuine inputs incorrectly rejected as synthetic. A lower EER value indicates better system capability in detecting spoofing attempts as it indicates a lower error of the system.

Besides the EER metric, we apply the categorical cross-entropy (CCE) loss function for our spoofing task, as this is technically a classification task between bonafide and spoofed audio data. CCE quantifies the difference between two probability distributions: the true distribution (actual labels) and the predicted distribution (model outputs). The formula for CCE is given by:

$$L(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}) \qquad (2)$$

The value of cross-entropy loss ranges from 0 to infinity, with lower values indicating better model performance. A perfect model would have a cross-entropy loss of 0, meaning its predictions perfectly match the true labels. As predictions diverge from actual labels, the loss increases significantly, especially when predictions are confident but incorrect [18].

3.5 RawNet2 Encoder

RawNet2 was initially developed for speaker verification tasks and was later adapted for spoofing detection. Its architecture has become foundational in modern spoofing detection systems, including both RawGAT-ST and the state-of-the-art AASIST model, which incorporate its key components in their architectures [12, 19].

The RawNet2 encoder employs a sinc-convolution layer as its first layer, containing 70 filters. The representation from the sinc-convolution layer is interpreted as a 2-dimensional image with a single channel, similar to a spectrogram. Subsequently, six pre-activated residual blocks are used to extract high-level representations. Each residual block consists of batch normalization, 2-dimensional convolution, SeLU activation, and max pooling layers. The first two blocks utilize 32 filters, while the remaining four blocks employ 64 filters [20].

3.6 RawGAT-ST

RawGAT-ST represents a significant evolution in spoofing detection architecture, serving as the predecessor to the AASIST model and holding the state-of-the-art position before AASIST's introduction. The model builds upon the RawNet2 architecture by incorporating its encoder design, particularly the sinc-convolution filters, while introducing novel architectural elements for improved spoofing detection.

The key innovation of RawGAT-ST lies in its use of two parallel graphs to simultaneously model spectral and temporal information. These graphs are then combined using element-wise multiplication. Although this model achieved state-of-the-art performance at its time, there remained room for improvement due to the heterogeneous nature of the two graphs, which necessitated better integration techniques. This limitation was later addressed by AASIST through its introduction of more sophisticated graph integration mechanisms.

The RawGAT-ST architecture demonstrates the evolution of spoofing detection models, bridging the gap between the fundamental RawNet2 encoder and the more sophisticated AASIST model. Its success in combining spectral and temporal information through graph attention networks laid the groundwork for AASIST's further improvements in graph-based audio processing [19].

3.7 AASIST & AASIST-L

AASIST represents an evolution of RawGAT-ST, designed to detect various types of spoofing attacks without requiring score-level ensemble. The model introduces three key components to handle the heterogeneity of spectral and temporal graphs: heterogeneous stacking graph attention layer (HS-GAL), max graph operation (MGO), and a novel readout scheme.

HS-GAL consists of two components: heterogeneous attention and stack node. HS-GAL projects inputs into a latent space to provide uniform dimensions for both graphs using two fully-connected layers. The heterogeneous attention utilizes three different projection vectors to compute attention weights on edges connecting: (i) nodes within the spectral domain, (ii) nodes between spectral and temporal domains, and (iii) nodes within the temporal domain. Meanwhile, the stack node functions to accumulate heterogeneous information through unidirectional connections from all other nodes.

MGO implements a competitive mechanism using two parallel branches, where each branch consists of two HS-GALs and graph pooling layers. Element-wise maximum operations are applied to the outputs of both branches. This enables different branches to learn and prioritize different artifact groups, where elements containing artifacts survive after the maximum operation.

AASIST also employs a RawNet2-based encoder to extract high-level representations formulated as $F \in R^{C \times S \times T}$ directly from raw waveform inputs. Two graph modules then model the spectral and temporal domains in parallel. The results from both graph modules are subsequently combined using MGO, which includes four HS-GALs and four graph pooling layers.

AASIST also introduces AASIST-L, its lightweight variant optimized using population-based training algorithms to achieve a significantly reduced parameter count of only 85K parameters. Despite its substantially lower complexity, AASIST-L maintains competitive performance. The complete AASIST model achieves state-of-the-art performance with a minimum t-DCF of 0.0275 and EER of 0.83% on the ASVspoof2019 LA dataset, while AASIST-L achieves a minimum t-DCF of 0.0309% and EER of 0.99%. This architecture demonstrates significant improvements over its predecessors by effectively addressing the challenges of heterogeneous graph integration and providing efficient spoofing detection capabilities, even in its lightweight variant [12].

4 Results & Discussions

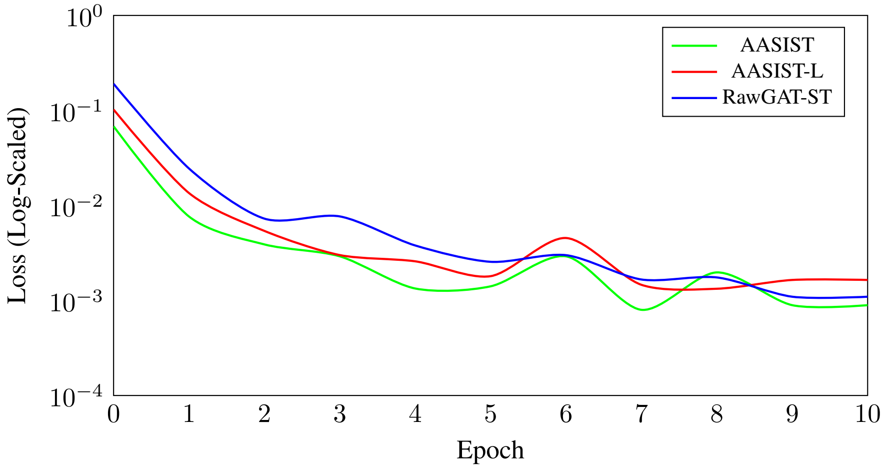

Figure 2. GAT-based model losses over 10 epochs

Model	EER (%)	Final Loss	Avg. Bonafide Score	Avg. Spoofed Score
AASIST, 297K	0.0	8.9×10^{-4}	4.3964	-5.2115
AASIST-L, 85K	0.0	1.64×10^{-3}	4.1610	-3.1450
RawGAT-ST, 437K	0.0	1.09×10^{-3}	2.1205	-2.3843

Table 1. Model Performance Evaluation

4.1 Results

The experimental results in both Figure 2 and Table 1. reveal several interesting findings regarding the performance of the GAT-based models. While the EER metric unexpectedly reached 0% across all models, this outcome warrants careful interpretation. Our investigation suggests this result stems from the limited speaker diversity in the synthetic speech generation process. Specifically, the spoofed audio samples were generated with only two speakers (one male, one female) from Coqui TTS, creating a distinct separation from the more diverse speaker characteristics present in the bonafide audio dataset.

Despite the limitations in EER assessment, the analysis of raw prediction scores provides more nuanced insights into model performance. Examining the average scores assigned to bonafide and spoofed audio samples (calculated across five samples for each category) reveals meaningful distinctions between the models. The AASIST model demonstrated the most pronounced discrimination capability, with the largest gap between bonafide (4.3964) and spoofed (-5.2115) scores. AASIST-L followed with a similarly strong differentiation pattern (4.1610 for bonafide, -3.1450 for spoofed), while RawGAT-ST showed the smallest separation (2.1205 for bonafide, -2.3843 for spoofed). This scoring pattern suggests that both AASIST variants developed more robust feature representations for distinguishing between authentic and synthetic speech.

The loss curves over the training epochs provide additional perspective on model behavior. RawGAT-ST exhibited the highest initial loss (approximately 10^{-1}) with a relatively slower convergence rate, though it ultimately achieved a final loss of 1.09×10^{-3}, outperforming AASIST-L's 1.64×10^{-3}. Notably, AASIST demonstrated superior performance throughout the training process, starting with a lower initial loss and converging to the lowest final loss of 8.9×10^{-4}. This consistent performance advantage, combined with its superior score differentiation, suggests that AASIST's architectural improvements effectively enhance its ability to detect spoofing attempts in Indonesian speech.

4.2 Interpretation & Discussion

The experimental results reveal an interesting paradox between architectural robustness and dataset limitations in applying GAT-based models to Indonesian speech spoofing detection. While the evaluated models, particularly AASIST, demonstrate sophisticated architectural design, their full potential appears constrained by the relatively shallow nature of our collected dataset. The uniform 0% EER across all models can be attributed to this limitation, where the spoofed audio samples were generated from only two speakers, creating an oversimplified distinction between authentic and synthetic speech patterns.

Despite these dataset constraints, alternative evaluation metrics provide valuable insights into the relative capabilities of each model. The raw prediction scores demonstrate AASIST's superior ability to differentiate between bonafide and spoofed audio, exhibiting the largest scoring gap. This pronounced differentiation suggests that AASIST's architectural innovations particularly its heterogeneous stacking graph attention layer and max graph operation, effectively capture distinctive features even within a limited dataset.

Although spoof related research has only been conducted on the using English language in the context of the latest AASIST model [12], the results provide valuable insights into the potential application of the evaluated models to the Indonesian language context. This adaptation of the AASIST model on a new language is crucial for developing effective speech spoofing detection systems within the local context, particularly considering the escalating risks of spoofing attacks across various voice-based applications not only in Indonesia, but the whole world. The model's architectural strength in processing both spectral and temporal features suggests its potential resilience in adapting to the phonological and prosodic characteristics specific to speech patterns.

4.3 Future Works

Several promising directions for future research emerge from our findings. First, addressing the dataset limitations observed in this study, future work should focus on developing more comprehensive and diverse spoofing datasets for the Indonesian language. This includes expanding the number of synthetic speakers beyond binary gender representation to better reflect the natural variation in human speech patterns. Such diversity would enable more robust evaluation of model performance and provide a more realistic assessment of spoofing detection capabilities.

Furthermore, while our study focused on the A10 case representing end-to-end neural network-based speech synthesis, future research should extend to other types of spoofing attacks. This includes investigating replay attacks, voice conversion techniques, and hybrid approaches that may pose different challenges to detection systems. Understanding model performance across various attack vectors is crucial for developing comprehensive anti-spoofing solutions for real-world applications.

Transfer learning presents another promising avenue for improvement. Given that the original AASIST model was trained on English language datasets, investigating the effectiveness of fine-tuning pre-trained models on Indonesian speech could potentially leverage the rich feature representations learned from larger English datasets while adapting to Indonesian-specific characteristics. This approach could be particularly beneficial given the current scarcity of large-scale Indonesian speech spoofing datasets.

Conclusion

This research presents initial efforts in developing and evaluating speech spoofing detection systems for the Indonesian language context. We have contributed to the field by creating a preliminary Indonesian speech dataset, which can be accessed at the footnote in the first page 3, comprising both bonafide and synthetic speech samples. While the dataset has its own limitations, particularly in speaker diversity for synthetic samples, we hope to give a foundation and understanding on how the dataset for the Indonesian language is built

The evaluation of GAT-based models demonstrated promising results, with all models achieving 0% EER on our dataset, mirroring their performance on the ASVspoof2019 dataset. However, a more nuanced examination through raw prediction scores revealed AASIST's superior capability in distinguishing between authentic and synthetic speech, followed closely by its lightweight variant, AASIST-L. The consistency in performance across different evaluation perspectives, including loss function trajectories and score differentiation, suggests the robustness of these architectures for speech spoofing detection tasks.

Our findings also highlight several important considerations for future development of speech spoofing detection systems in the Indonesian language context. The success of AASIST's architectural innovations, particularly its heterogeneous stacking graph attention layer and max graph operation, indicates the importance of effective spectral and temporal feature integration in spoofing detection. Additionally, the competitive performance of AASIST-L demonstrates the potential for deploying efficient anti-spoofing solutions in resource-constrained environments.

While this research represents a significant step toward developing robust speech spoofing detection for Indonesian language applications, it also underscores the need for more comprehensive datasets and evaluation frameworks. The models exhibit strong technical performance, but their real-world effectiveness will depend on continued development with more diverse and representative data. This work lays the groundwork for future research in Indonesian speech spoofing detection, contributing to the broader goal of securing voice-based applications in the Indonesian context.

References

[1] N. Anderson, "CEOWORLD magazine," *CEOWORLD magazine*, Feb 2024.

[2] E. L. NAPITUPULU, "Indonesian Students' Basic Literacy Skills Are Low," *kompas.id*, Oct 2023.

[3] K. Nugroho and E. Winarno, "Spoofing Detection of Fake Speech Using Deep Neural Network Algorithm," in *2022 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pp. 56–60, 2022.

[4] Q. Fu, Z. Teng, J. White, M. Powell, and D. C. Schmidt, “FastAudio: A Learnable Audio Front-End for Spoof Speech Detection,” 2021.

[5] X. Li, N. Li, and C. e. a. Weng, “Replay and Synthetic Speech Detection with Res2Net Architecture,” in *ICASSP*, 2021.

[6] G. Hua, A. B. J. Teoh, and H. Zhang, “Towards End-to-End Synthetic Speech Detection,” *IEEE Signal Processing Letters*, vol. 28, p. 1265–1269, 2021.

[7] J. Yamagishi, X. Wang, and M. e. a. Todisco, “ASVspoof 2021: Accelerating Progress in Spoofed and Deepfake Speech Detection,” *arXiv preprint arXiv:2109.00537*, 2021.

[8] N. Muller, P. Czempin, and F. e. a. Diekmann, “Does Audio Deepfake Detection Generalize?,” in *Interspeech*, pp. 2783–2787, 2022.

[9] J. Yi, R. Fu, and J. e. a. Tao, “ADD 2022: The First Audio Deep Synthesis Detection Challenge,” in *ICASSP*, pp. 9216–9220, 2022.

[10] J. Yi, J. Tao, and R. e. a. Fu, “ADD 2022: The Second Audio Deep Synthesis Detection Challenge,” *arXiv preprint arXiv:2305.13774*, 2023.

[11] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” in *Proc. ICLR*, 2018.

[12] J.-w. Jung, H.-S. Heo, H. Tak, H. j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, “AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks,” 2021.

[13] X. Wang, J. Yamagishi, M. Todisco, H. Delgado, A. Nautsch, N. Evans, M. Sahidullah, V. Vestman, T. Kinnunen, K. A. Lee, L. Juvela, P. Alku, Y.-H. Peng, H.-H. Hwang, Y. L. Oo, G.-H. Tan, L. M. Sebastian, E. Kucur, J. Yamagishi, M. Nagano, Y. Kamado, T. Ito, K. Nishizawa, H. Kameoka, J. Liu, C.-W. Wu, W.-C. Wu, T. Huang, K. Toda, H. Tanaka, H. Kameoka, S. Kaneko, J.-F. Zhao, Marouf, and Bonastre, “ASVspoof 2019: A Large-scale Public Database of Synthesized, Converted and Replayed Speech,” 2020.

[14] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, J. Yamagishi, and D. A. Reynolds, “t-DCF: a Detection Cost Function for the Tandem Assessment of Spoofing Countermeasures and Automatic Speaker Verification,” 2019.

[15] S. Sakti, E. Kelana, H. Riza, S. Sakai, K. Markov, and S. Nakamura, “Development of Indonesian Large Vocabulary Continuous Speech Recognition System within A-STAR Project,” in *Proceedings of the Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST)*, 2008.

[16] J. Kim, J. Kong, and J. Son, “Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech,” 2021.

[17] G. Eren and The Coqui TTS Team, “Coqui TTS,” Jan. 2021.

[18] S. Maheshkar, “What is Cross Entropy Loss - A Tutorial With Code,” *W&B*, 2024.

[19] H. Tak, J. weon 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,” 2021.

[20] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, “End-to-end anti-spoofing with rawnet2,” 2021.