ECAPA-TDNN WITH DUAL-BRANCH ATTENTION MECHANISM FOR SYNTHETIC SPEECH DETECTION

K. K. Kim[1], Sh. Sh. Abdullaev[2]

[1] Digital Technologies and Artificial Intelligence Development Research Institute, Tashkent, Uzbekistan

[2] Oriental University, Tashkent, Uzbekistan

https://doi.org/10.5281/zenodo.17767861

Abstract. *This study presents a modified ECAPA-TDNN model with a dual-branch attention mechanism for the task of synthetic speech detection aimed at countering audio spoofing. The architecture integrates two parallel spectral feature branches (LFCC and LogMel) with multi-level attention through Attentive Statistics Pooling followed by a fusion encoder. The proposed approach demonstrates a significant improvement in the accuracy of synthetic speech detection, achieving an EER (LA) of 0.08%, which confirms the effectiveness of the dual-branch design and the fine-tuning of hyperparameters to enhance detection performance.*

Keywords: *synthetic speech, ECAPA-TDNN, dual-branch attention, Attentive Statistics Pooling, audio spoofing.*

Аннотация. *В данном исследовании представлена модифицированная модель ECAPA-TDNN с dual-branch attention для задачи детектирования синтетической речи в целях защиты от аудиоспуфинга. Архитектура объединяет две параллельные спектральные ветки (LFCC и LogMel) с многоуровневым вниманием через Attentive Statistics Pooling и последующим fusion-энкодером. Предложенное решение демонстрирует значительное повышение точности выявления синтетической речи, достигая EER(LA) = 0.08 %, что подтверждает эффективность dual-branch подхода и целевую настройку гиперпараметров для повышения точности детектирования синтетической речи.*

Ключевые слова: *синтетическая речь, ECAPA-TDNN, dual-branch attention, Attentive Statistics Pooling, аудиоспуфинг.*

Introduction. Recent advances in speech synthesis and generative modeling have resulted in the creation of highly realistic audio simulations that can be exploited to bypass voice authentication systems or disseminate disinformation [1]. Consequently, detecting synthetic speech has become a critical research direction in safeguarding information security systems against audio spoofing.

This paper introduces a modified ECAPA-TDNN architecture incorporating a dual-branch attention mechanism that enhances spoofed speech detection accuracy by jointly processing two complementary spectral representations: LFCC and LogMel. The architecture employs Attentive Statistics Pooling to implement multi-branch attention and applies task-specific hyperparameter optimization tailored to anti-spoofing objectives.

The proposed model demonstrates strong performance on standard synthetic speech benchmarks, achieving an EER (LA) of 0.08%, which validates the effectiveness of the dual-branch attention design and highlights its potential for integration into intelligent information security systems.

Related work. Recent studies on synthetic speech detection highlight the growing

importance of attention mechanisms, which enable models to focus on the most informative acoustic representations. For instance, Kanwal et al. [2] enhanced the discriminative capability of the VGGish architecture by incorporating an embedded attention block. Wani et al. [3] proposed a cascaded capsule network, ABC-CapsNet, where attention regulates interlayer feature interactions. Mahum et al. [4] introduced the DeepDet model (YAMNet with a BAM module), demonstrating that attention effectively captures artifacts characteristic of TTS-generated speech.

Furthermore, Yaseen et al. [5] confirmed the versatility of attention mechanisms by showing their effectiveness in speaker identification tasks through multi-level bottleneck attention modules. Khan and Malik [6] developed SpoTNet, a transformer-based architecture in which attention facilitates contextual modeling for the detection of manipulated speech.

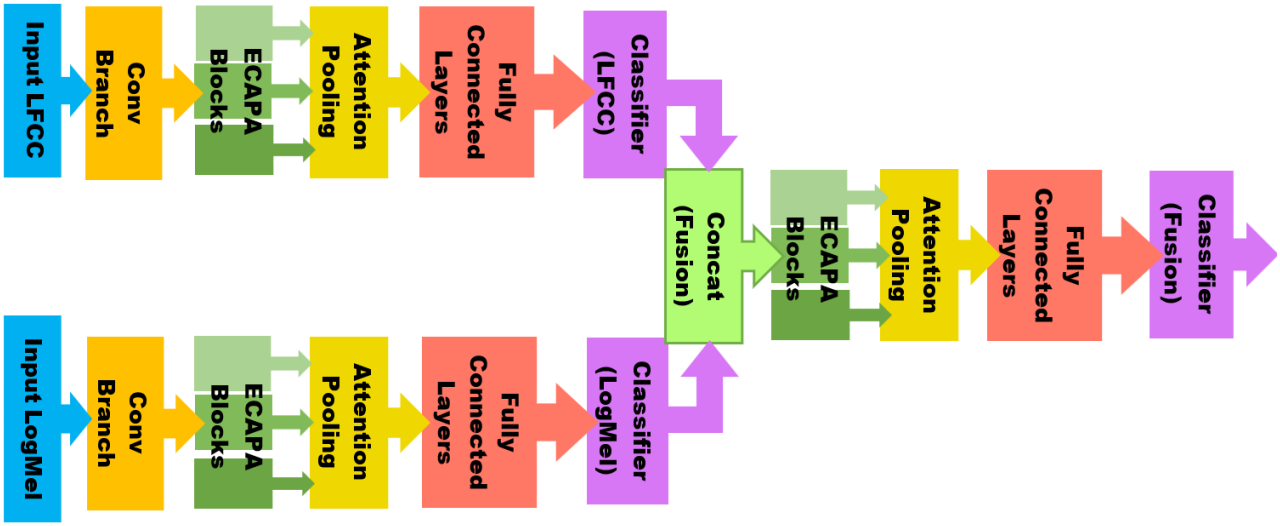
In contrast to these approaches, the ECAPA-TDNN model with a dual-branch attention mechanism proposed in this study integrates two complementary spectral representations: LFCC and LogMel and employs dual-branch attention to adaptively emphasize salient spectral regions, thereby improving the model's sensitivity to subtle distinctions between natural and synthetic speech.

Method. In this paper, we propose a modified Dual-Branch ECAPA-TDNN (Emphasized Channel Attention, Propagation, and Aggregation – Time Delay Neural Network) architecture that incorporates attention mechanisms to enhance the accuracy of synthetic speech detection. The model consists of two parallel processing branches: LFCC and LogMel, each extracting independent spectral representations and processing them through its own stack of convolutional and ECAPA blocks.

The proposed model was trained and validated on the ASVspoof 2019 Logical Access (LA) subset, which contains bonafide and spoofed utterances represented as LFCC and LogMel features.

As illustrated in Figure 1, the proposed architecture consists of two parallel branches that process LFCC and LogMel spectral representations independently. Each branch includes a convolutional block followed by ECAPA modules and an attention pooling layer that computes frame-wise attention weights, emphasizing the most informative time–frequency features. The resulting embeddings are passed through fully connected layers and local classifiers for each branch (LFCC and LogMel).

At the fusion stage, the embeddings from both branches are concatenated and further processed through an additional ECAPA block with attention pooling and fully connected layers. The final fused embedding is then passed to the classifier to distinguish between genuine and



synthetic speech.

Fig. 1. Architecture of the proposed ECAPA-TDNN model with a dual-branch attention mechanism.

At the statistical aggregation level, Attentive Statistics Pooling (ASP) is employed, performing frame-level weighting based on the informativeness of acoustic features. The attention weight coefficient for each time step *t* is calculated as:

$$\alpha_t = \frac{\exp\left(w^T \tanh(Wh_t+b)\right)}{\sum_\tau \exp\left(w^T \tanh\left(Wh_\tau+b\right)\right)}, \quad (1)$$

where α_t – the attention weight coefficient for the t-th time step, representing its relative importance in feature aggregation;

tanh() – the nonlinear activation function;

exp() – the exponential function used for attention normalization;

W, w, b – the trainable parameters of the attention layer;

T – the length of the input sequence (i.e., the total number of time frames);

h_t – the output of the ECAPA block;

τ – the summation index over all time steps.

The final representation of each branch is obtained through weighted statistical pooling, where the weighted mean (μ) and weighted standard deviation (σ) are computed as:

$$\mu = \sum_t \alpha_t \, h_t,$$
$$\sigma = \sqrt{\sum_t \alpha_t (h_t - \mu)^2} \quad (2)$$

These two vectors are concatenated into a single feature vector, forming the Attentive Statistics Pooling (ASP) layer, which passes the aggregated representation to the classifier.

Thus, the proposed architecture integrates two spectral streams: LFCC and LogMel through an attention block that adaptively emphasizes the most informative speech segments.

The attention mechanism enables the model to focus on artifacts characteristic of synthetic speech, while its weights are jointly optimized with the convolutional layer parameters to minimize a cross-entropy–based loss function.

Results. The performance of the proposed model was evaluated using the Equal Error Rate (EER) – a standard metric representing the point at which the probabilities of false acceptance and false rejection are equal. The developed model achieved an EER of 0.08%, demonstrating a strong capability to discriminate between synthetic and genuine speech.

Figure 2 presents the Receiver Operating Characteristic (ROC) curve, illustrating the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR). The curve indicates the classifier's stable behavior and consistently low error rate.

Additionally, the score distributions (CM scores) for bonafide and spoofed samples are shown. The clear separation between the distribution peaks confirms the model's ability to reliably distinguish synthetic signals from natural speech.

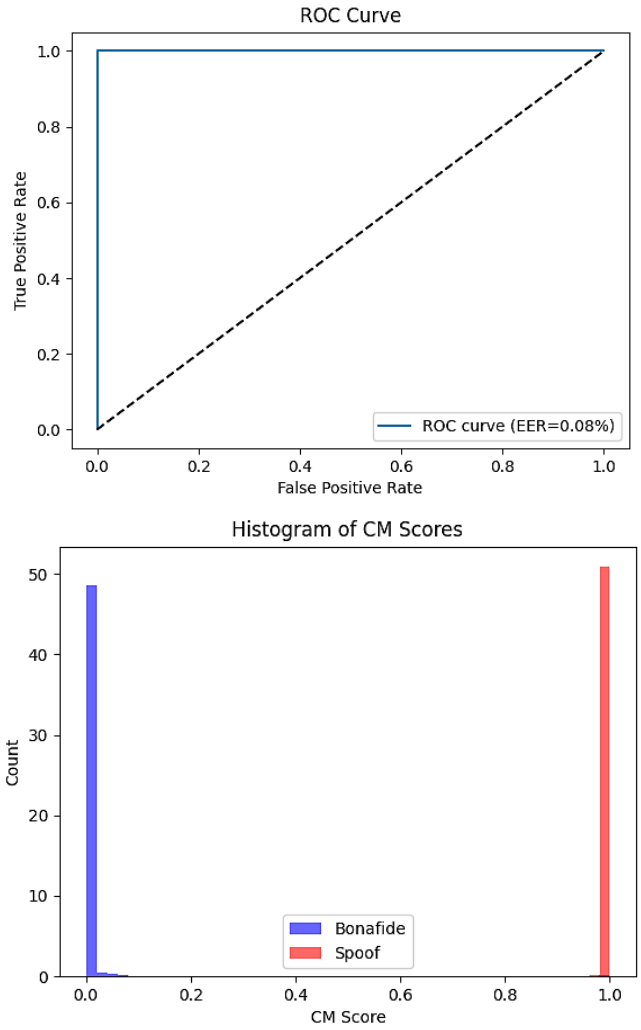


Fig. 2. ROC curve and CM score distributions for the proposed model.

Therefore, the proposed architecture with a multi-level attention mechanism demonstrates high accuracy and robustness in the task of synthetic speech anti-spoofing.

Conclusion. This paper presented a synthetic speech detection model based on a modified ECAPA-TDNN architecture incorporating dual-branch feature processing (LFCC and LogMel) and an attentive statistical pooling mechanism. The dual-branch structure allows the model to jointly exploit complementary spectral representations, enhancing its ability to capture subtle artifacts characteristic of synthetic speech. The attention-based pooling mechanism further refines temporal aggregation by adaptively weighting the most informative frames, contributing to the robustness of the learned embeddings.

Experimental evaluation on the ASVspoof 2019 Logical Access (LA) subset demonstrated high detection accuracy, achieving an EER of 0.08%, which confirms the model's strong discriminative capability and its potential for integration into practical anti-spoofing systems. The results indicate that combining multi-representation feature processing with attention-driven aggregation significantly improves system sensitivity to acoustic inconsistencies.

Future work will focus on further architectural optimization, such as refining the attention mechanism and exploring transformer-based extensions for long-term feature modeling. Additionally, research will be directed toward expanding the model's applicability to multimodal and cross-linguistic anti-spoofing scenarios, ensuring broader generalization and robustness across diverse real-world environments.

REFERENCES

1. *Malinka K., Firc A., Šalko M. et al.*Comprehensive multiparametric analysis of human deepfake speech recognition // J Image Video Proc. – 2024. – Vol. 24. – DOI: 10.1186/s13640-024-00641-4.
2. Kanwal T., Mahum R., AlSalman AM et al.Fake speech detection using VGGish with attention block // J. Audio Speech Music Proc. – 2024. – Vol. 35. – DOI: 10.1186/s13636-024-00348-4.
3. Wani TM, Gulzar R., Amerini I.ABC-CapsNet: Attention based Cascaded Capsule Network for Audio Deepfake Detection // IEEE/CVF Conf. Comput. Vis. Pattern Recogn. Workshops (CVPRW). – Seattle, WA, USA, 2024. – P. 2464–2472. – DOI: 10.1109/CVPRW63382.2024.00253.
4. Mahum R., Irtaza A., Javed A. et al.DeepDet: YAMNet with BottleNeck Attention Module (BAM) for TTS synthesis detection // J. Audio Speech Music Proc. – 2024. – Vol. 18. – DOI: 10.1186/s13636-024-00335-9.
5. Yaseen MU, Nasralla MM, Aslam F., Ali SS, Khattak SBAA Novel Approach Based on Multi-Level Bottleneck Attention Modules Using Self-Guided Dropblock for Person Re-Identification // IEEE Access. – 2022. – Vol. 10. – P. 123160–123176. – DOI: 10.1109/ACCESS.2022.3223426.
6. Khan W., Malik KMSpoTNet: A spoofing-aware Transformer Network for Effective Synthetic Speech Detection // Proc. 2nd ACM Int. Workshop Multimedia AI against Disinformation (MAD '23). – New York, NY, USA: ACM, 2023. – P. 10–18. – DOI: 10.1145/3592572.3592841.