UDC 004.4

A SURVEY OF ATTENTION-BASED ARCHITECTURES FOR SYNTHETIC SPEECH DETECTION

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

*kamilla.k.kim@gmail.com
[1]Oriental University, Tashkent, Uzbekistan.
[2]Digital Technologies and AI Development Research Institute, Tashkent, Uzbekistan.

Abstract. This paper presents an analytical review of contemporary architectures for synthetic speech detection, emphasizing the central role of attention mechanisms. Three representative models are examined: a modified VGGish enhanced with the CBAM, YAMNet augmented with BAM, and the cascaded capsule network, ABC-CapsNet. The analysis is based on mel spectrograms interpreted as two-dimensional images, where attention modules identify the most informative time–frequency regions. The discussion covers architectural design choices, feature extraction strategies, attention mechanisms, datasets, and performance metrics. Comparative evaluation reveals that YAMNet with BAM achieves the lowest EER with moderate computational complexity, while ABC-CapsNet exhibits strong robustness to data variability and generalizes well across heterogeneous datasets. In contrast, VGGish with CBAM provides a balanced trade-off between accuracy and implementation simplicity. The study underscores the pivotal role of attention mechanisms in enhancing the effectiveness of synthetic speech detection systems and outlines promising directions for further research and development.

Keywords: audio spoofing detection, synthetic speech, attention mechanisms, CBAM, cascaded capsule networks, BAM.

1 Introduction

The rapid advancement of speech synthesis technologies, including text-to-speech (TTS) and voice conversion (VC), has resulted in the generation of synthetic voices that are nearly indistinguishable from natural human speech. State-of-the-art algorithms can now produce artificial audio that even expert listeners and conventional verification systems struggle to differentiate from genuine recordings [1]. These technologies are widely applied across various domains – from voice interfaces and interactive assistants to media production, localization, and automated dubbing [2].

However, alongside their benefits, the misuse of speech synthesis has grown significantly, with fabricated audio being employed for deception, fraud, reputational harm, and the dissemination of misinformation [3].

In response to these challenges, automatic methods for synthetic speech detection have seen rapid development. Of particular interest are models incorporating attention mechanisms, which effectively emphasize salient spectrotemporal features and enhance robustness to signal variability and diverse attack types. Integrating attention modules with powerful feature extractors such as convolutional or capsule networks enables a deeper representation of time-frequency dependencies within the audio signal. Moreover, attention mechanisms have proven effective in hybrid architectures, including graph neural networks, as demonstrated in studies on dual-channel fake speech detection [4, 5].

This paper provides a comprehensive review of three representative attention-based architectures for synthetic speech detection: a modified VGGish enhanced with the Convolutional Block Attention Module (CBAM), YAMNet integrated with the Bottleneck Attention Module (BAM), and the Attention-Based Cascaded Capsule Network (ABC-CapsNet). The review examines architectural designs, input representations, attention mechanisms, datasets, and evaluation metrics, as well as the comparative advantages and limitations of each model. The findings highlight promising directions for advancing synthetic speech detection and inform potential applications in real-world verification systems.

2 Methods

This study conducts a comparative analysis of contemporary architectures developed for automatic synthetic speech detection, emphasizing the role of attention mechanisms. The main objective is to evaluate how various attention strategies influence model accuracy, robustness, and computational efficiency.

Three representative architectures were selected, each exemplifying a distinct approach to incorporating attention within audio processing frameworks:

- VGGish with CBAM – a modified convolutional neural network integrated with a Convolutional Block Attention Module;
- YAMNet with BAM – a lightweight model based on MobileNet, enhanced with a Bottleneck Attention Module;
- ABC-CapsNet – a cascaded capsule network augmented with an attention mechanism to reinforce inter-capsule relationships.

To assess model performance, both quantitative and qualitative criteria were applied. The comparison considered architectural design, feature extraction processes, attention configurations, and the datasets employed. Evaluation was carried out using standard metrics, including Equal Error Rate (EER), Minimum Tandem Detection Cost Function (min-tDCF), Precision, Recall and Accuracy, ensuring a comprehensive assessment of each model's performance.

This methodological framework enables an in-depth examination of the strengths and limitations of attention-based approaches and helps identify the most promising directions for further development and deployment of synthetic speech detection systems in practical scenarios.

3 Review of existing methods

This section presents three contemporary approaches to synthetic speech detection that employ attention mechanisms to extract the most informative features from mel spectrograms. The datasets, architectural designs, attention types, and key performance metrics used for each approach are described.

3.1 Datasets

For the comparative analysis of the examined models – VGGish with CBAM, YAMNet with BAM, and ABC-CapsNet – two of the most authoritative corpora in modern synthetic speech detection research were employed: ASVspoof 2019 and FoR (Fake or Real). These datasets provide extensive coverage of attack types and speech generation scenarios, enabling an objective evaluation of model performance under diverse conditions.

ASVspoof 2019

The ASVspoof 2019 corpus is the de facto standard in audio spoofing detection and is widely adopted for developing and benchmarking attention-based models. It was introduced as part of the international Automatic Speaker Verification Spoofing and Countermeasures Challenge 2019 [6].

The dataset comprises two independent scenarios: Logical Access (LA) and Physical Access (PA). The former represents attacks generated using TTS and VC systems, whereas the latter simulates replay attacks conducted through an acoustic environment.

Each scenario includes separate subsets for training, development, and evaluation, ensuring an unbiased assessment of model generalization capabilities.

All three architectures – VGGish with CBAM, YAMNet with BAM, and ABC-CapsNet – were evaluated on this dataset, as it offers a wide variety of attack types and a carefully balanced distribution of genuine and synthetic speech samples.

FoR (Fake or Real)

Additionally, the FoR (Fake or Real) dataset, introduced by Reimao and Tzerpos in "FoR: A Dataset for Synthetic Speech Detection" [7], was utilized. In contrast to ASVspoof 2019, FoR focuses on practical comparisons between synthetic and natural speech rather than purely simulated attacks.

It contains over 198,000 audio samples generated using both commercial and open-source TTS systems.

Speech samples are provided in several variants (for-original, for-norm, for-2seconds, and for-rerecorded), enabling the simulation of realistic recording and transmission conditions.

In this study, the FoR corpus was applied to assess the robustness of the examined architectures and to test the generalization capability of attention mechanisms across data differing in source, quality, and duration.

The combined use of these two corpora ensured the representativeness of the experiments and enabled an objective comparison of the models across tasks of varying complexity and types of synthetic speech.

3.2 VGGish Method with CBAM

The methodology proposed by Tahira Kanwal et al. [8] represents an enhanced version of the VGGish architecture, specifically adapted for the task of synthetic speech detection. The original VGGish model, derived from the VGG (Visual Geometry Group) visual framework, was modified by integrating a Convolutional Block Attention Module (CBAM) to improve the network's sensitivity to relevant time–frequency patterns in the audio signal.

The input data consist of mel spectrograms treated as two-dimensional images, providing compatibility with architectures originally developed for visual pattern recognition. The input spectrogram size used in the experiments was 96×64, representing a balance between computational efficiency and the preservation of key acoustic information.

The model retains the core structure of the VGG architecture, comprising sequential convolutional blocks with 3×3 kernels, followed by ReLU activation functions and max-pooling layers. The insertion of a CBAM module between convolutional blocks enables the network to dynamically emphasize the most informative regions of the spectrogram, thereby strengthening spatial and channel-wise feature dependencies (Fig. 1).

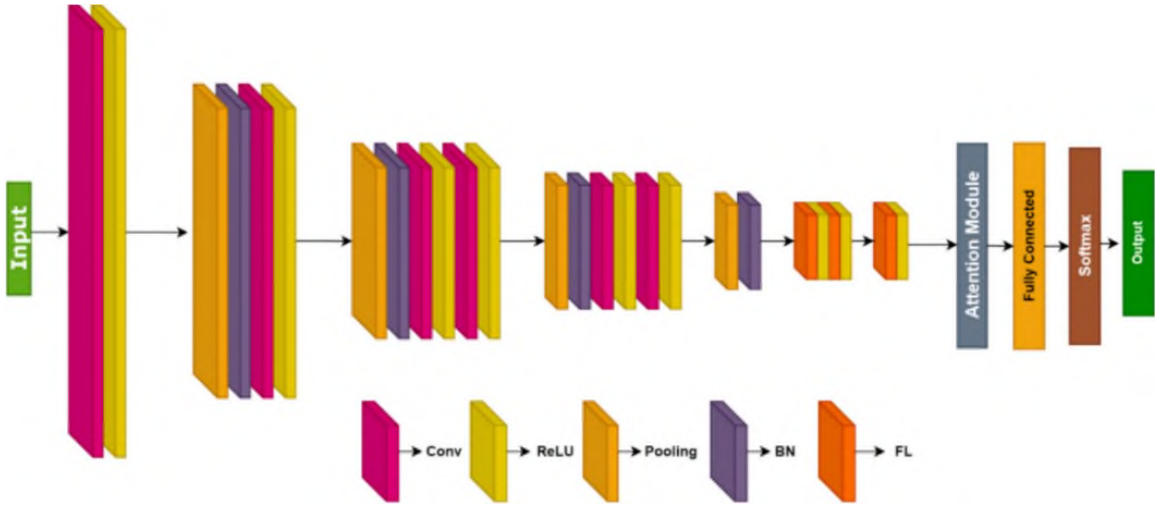


Fig. 1 Modified VGGish architecture (VGGish with CBAM)

The CBAM module is inserted before the fully connected layer and consists of two sequentially applied submodules: channel and spatial attention.

The channel attention submodule constructs a saliency map that identifies the most informative channels within the spectrogram. This is achieved by applying average and max pooling operations along the spatial dimensions, which extract complementary contextual information.

The resulting descriptors are passed through a shared two-layer MLP with a defined reduction ratio, after which the outputs are combined and normalized using a sigmoid activation function:

$$\text{or} \quad \begin{aligned} N_c(G) &= \sigma(MLP(Max-Pool(G)) + MLP(G))) \\ N_c(G) &= \sigma(W_1(W_0(G^c{}_{\max})) + W_1(W_0(G^c{}_{aug}))) \end{aligned}, \tag{1}$$

where G is the input feature tensor, MLP (Multi-Layer Perceptron) – multilayer perceptron, W_0, W_1 - training parameters, and σ - sigmoid function. This mechanism helps to identify the most significant feature channels.

The spatial attention module determines the regions of focus by analyzing the relationships among spatial locations in the feature map.

The average and max pooling projections across all channels are concatenated and processed through a convolutional layer with a 7×7 kernel, producing a two-dimensional attention map that is subsequently normalized using a sigmoid activation function:

$$\text{or} \quad \begin{aligned} N_s(G) &= \sigma(t^{7\times7}([Max-Pool(G); Avg-Pool(G)])) \\ N_s(G) &= \sigma(t^{7\times7}([G^s{}_{\max}; G^s{}_{avg}])) \end{aligned} \tag{2}$$

The schematic structure of both the channel and spatial attention modules is illustrated in the figures, showing the sequence of operations and the process of attention map formation (Fig. 2).

The model was trained and evaluated on the ASVspoof 2019 datasets (covering both logical and physical access attacks) as well as on ASVspoof 2021. Performance was assessed using standard metrics, including Accuracy, Precision, EER (LA) and min-tDCF.

Experimental results demonstrated EER = 0.07% for the Logical Access subset and EER = 0.52% for the Physical Access subset, confirming the high accuracy and robustness of the proposed synthetic speech detection approach.

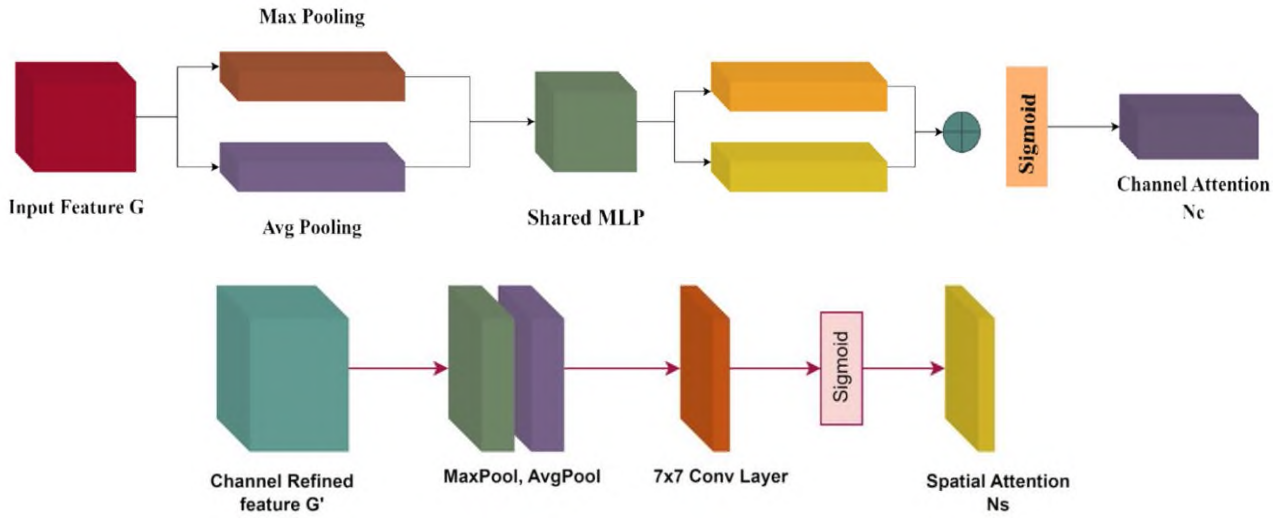


Fig. 2 CBAM modules of channel and spatial attention

The proposed approach is characterized by a streamlined architecture and an efficient integration of channel, spatial, and positional features through the CBAM attention mechanism. It demonstrates high recognition accuracy in the LA scenario. However, several limitations remain, including sensitivity to acoustic noise, increased computational demands when processing high-resolution spectrograms, and reduced performance against certain types of spoofing attacks. These challenges highlight promising directions for future research and model optimization.

3.3 ABC-CapsNet Method

n the study by Taiba Majid Wani et al. [9], the Attention-Based Cascaded Capsule Network (ABC-CapsNet) was proposed for synthetic speech detection. The model architecture consists of three primary components (Fig. 3):

- Feature extraction. At the first stage, a pre-trained VGG18 network is used to extract spectral-temporal features from 224×224×3 Mel spectrograms. Prior to feature extraction, audio signals are resampled to 16 kHz, denoised, amplitude-normalized, and stripped of silent segments to ensure consistency and quality of the input data.
- Attention mechanism. Attention is applied after the convolutional layers to emphasize the most relevant time-frequency regions of the spectrogram that may contain indicators of speech synthesis. This mechanism enhances the model's ability to capture subtle spectral artifacts typical of TTS and VC systems while suppressing the influence of background noise.
- Cascade of capsule networks. The classification stage employs capsule layers capable of retaining hierarchical and spatial relationships among features. The cascaded structure facilitates progressive information aggregation, improving the model's robustness and generalization across diverse audio sources.

The synergy between the attention mechanism and capsule-based architecture enables ABC-CapsNet to jointly exploit local and global feature representations, resulting in high accuracy and resilience to acoustic distortions in synthetic speech detection tasks.

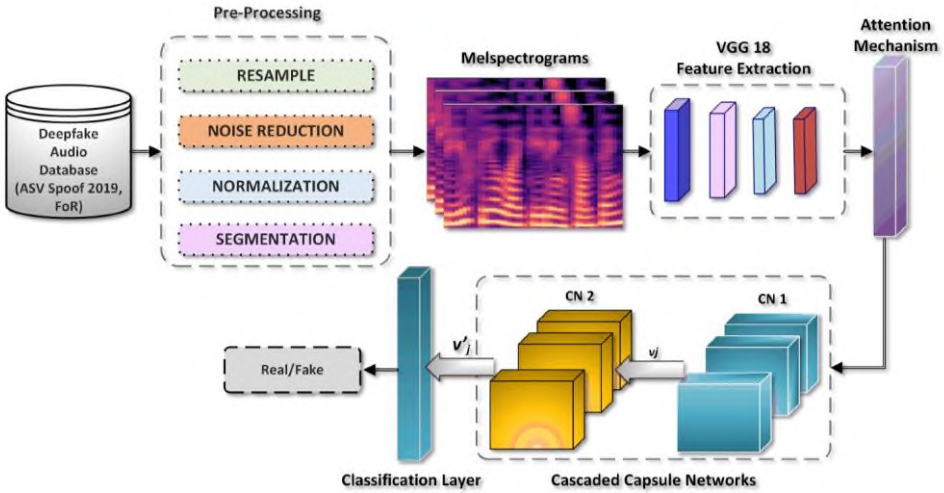


Fig. 3 ABC-CapsNet architecture for audio deepfake detection; CN1 is capsule network 1 and CN2 is capsule network 2.

In the ABC-CapsNet model, the attention mechanism operates on a set of features $F = \{f_1, f_2, \dots, f_n\}$, extracted by VGG18, where each f_i is a feature vector. For each vector, a weight w_i is calculated, reflecting its importance for the detection task. The attention output F' is calculated as a weighted sum:

$$F' = \sum_{i=1}^{n} w_i \cdot f_i \tag{3}$$

The weights w_i are defined via a softmax over the scalar scores $s(f_i)$ computed by the trainable attention layer (usually a fully connected layer with one output):

$$w_i = \frac{e^{s(f_i)}}{\sum_{j=1}^{n} e^{s(f_i)}} \tag{4}$$

Softmax guarantees that $\sum_{i=1}^{n} w_i = 1$, and the weights can be interpreted as feature significance probabilities. The proposed approach allows the model to adaptively focus on the most informative sections of the spectrogram that distinguish synthetic speech from real speech, thereby increasing sensitivity to various types of forgeries.

The cascaded capsule network enables deep modeling of feature interrelations, considering not only the presence of features but also their intrinsic parameters, which enhances the robustness of classification. The dynamic routing mechanism employed in capsule layers strengthens the

consistency between low-level and high-level feature representations, ensuring more reliable information flow through the hierarchy.

In experimental evaluations, the ABC-CapsNet model achieved an EER of 0.04% on the FoR dataset and 0.06% on ASVspoof 2019, outperforming earlier capsule-based approaches that relied on MFCC (Mel-Frequency Cepstral Coefficients) and CQCC (Constant-Q Cepstral Coefficients) features. Despite its high accuracy and resilience, the method demands considerable computational resources and extensive hyperparameter tuning, which constrains its deployment in real-time applications.

3.4 DeepDet Method (YAMNet with BAM)

The DeepDet (Deep Detection) method proposed by Rabbia Mahum et al. [10] integrates the YAMNet (Yet Another Mobile Network) architecture, built upon MobileNetV1, with the Bottleneck Attention Module (BAM). This approach combines the computational efficiency of depthwise separable convolutions with the capability of the attention mechanism to emphasize the most informative features.

A pre-trained YAMNet model, originally trained on the AudioSet dataset, is employed for feature extraction. The input data consist of 96×64 Mel spectrograms generated from pre-processed audio signals. The model produces spectral–temporal representations that are robust to noise and variations in voice characteristics.

In contrast to conventional convolutional neural network (CNN) approaches, DeepDet employs separable convolutions and a BAM attention module to highlight relevant regions within the spectrogram. The final classification is carried out by a binary classifier, which determines whether the analyzed speech is genuine or synthetic (Figure 4).

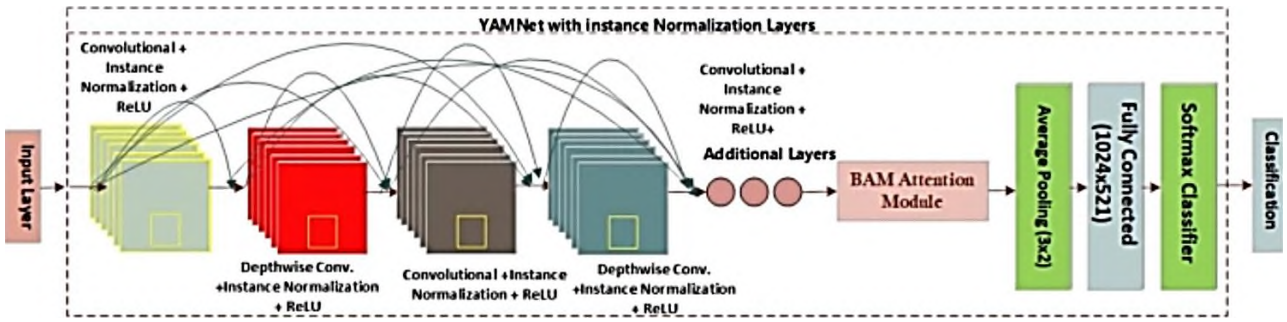


Fig. 4 Improved YAMNet architecture

BAM (Fig. 5) computes the final attention map $M(F)$ as the sum of the channel $M_c(F)$ and spatial attention $M_s(F)$:

$$M(F) = \sigma(resize\left(M_c(F)\right) + resize\left(M_s(F)\right)) \tag{5}$$

The final updated representation is calculated as:

$$F' = F \otimes M(F) + F, \tag{6}$$

where F'- updated feature tensor, $\otimes$- element-wise multiplication, and F input feature tensor (feature map) obtained after the previous network block (residual connection) [11].

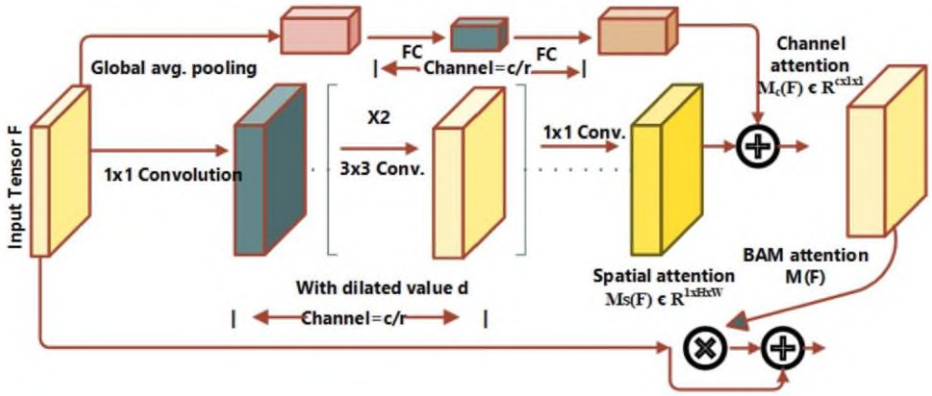


Fig. 5 Attention module BAM

In the DeepDet architecture, channel attention is generated through global pooling followed by a multi-layer perceptron (MLP) with a defined compression ratio, whereas spatial attention is computed via a separate branch that analyzes spatial feature projections. The BAM module is embedded within the network bottlenecks, enabling efficient feature refinement without a substantial increase in the number of parameters or computational cost.

Experimental evaluation on the ASVspoof 2019 dataset demonstrated that DeepDet achieves an EER of 0.042% for the Logical Access subset and 0.43% for the Physical Access (PA) subset, outperforming several state-of-the-art approaches while maintaining lower computational complexity.

The main advantages of the proposed approach include high detection accuracy, a compact and computationally efficient architecture, and suitability for real-time deployment. However, its limitations involve reduced performance for certain voice conversion attacks and sensitivity to normalization parameters, which can be mitigated by replacing Batch Normalization (BatchNorm) with Instance Normalization (InstanceNorm) to enhance robustness.

4 Analysis and development prospects

A comparative analysis (Table 1) of three contemporary attention-based methods for synthetic speech detection – VGGish with CBAM, ABC-CapsNet, and DeepDet (YAMNet with BAM) – demonstrates that all approaches achieve high accuracy on the ASVspoof 2019 and FoR datasets, while differing in their trade-offs between detection quality, computational efficiency, and robustness to various types of spoofing attacks.

A shared characteristic among these models is the use of Mel spectrograms as input features and the integration of attention mechanisms to emphasize informative representations: channel attention determines what is important, spatial attention determines where it is important, and trainable weight maps provide adaptability to new conditions and attack types [12].

Future research directions include the exploration of multimodal features (e.g., phase information, waveform, and vocal parameters), the development of lightweight and more robust attention modules incorporating the structural expressiveness of capsule networks, and the design of hybrid or transformer-based architectures capable of detecting subtle spectral and prosodic artifacts in synthetic speech [13].

Tab. 1. Comparative analysis of attention-based methods for synthetic speech detection

Comparison parameters	Research		
	VGGish with CBAM	**ABC-CapsNet**	**DeepDet (YAMNet with BAM)**
Architecture	Modified VGGish architecture: • Comprises approximately 29 layers, built on standard 3×3 convolutional kernels • Employs Batch Normalization (BatchNorm), ReLU activations, and max-pooling layers for feature extraction • Integrates a Convolutional Block Attention Module (CBAM) to emphasize salient channel and spatial features in Mel spectrograms	• Utilizes VGG18 for feature extraction via transfer learning. • Applies a soft-attention block to highlight the most relevant spectral-temporal regions. • Incorporates a cascaded capsule network, consisting of two or multiple stages of capsules, enabling hierarchical modeling of features for robust detection on Mel spectrograms.	• Based on the YAMNet model, which uses a MobileNetV1 backbone. • Incorporates a modified MobileNet as the feature extractor. • Employs the Bottleneck Attention Module (BAM) to enhance informative regions of the input. • Uses depthwise separable convolutions as the primary building block, providing a lightweight yet effective feature representation.
Features	• Mel-spectrograms (96×64); • CQCC; • logarithmic spectrum; • LC-GRNN (Light Convolutional Gated Recurrent Neural Network) is a neural network extractor.	• Mel-spectrograms (224×224×3); • VGG18 – feature extractor; • resampling of audio signal with a frequency of 16 kHz; • Noise reduction, normalization, removal of silence.	• Mel-spectrograms (96×64); • Features extracted using pre-trained YAMNet model
Attention mechanism	• Implemented consistently as channel attention followed by spatial attention. • Channel attention: computed using average and max pooling along spatial dimensions, followed by a shared MLP. • Spatial attention: computed using average and max pooling across channels, followed by a 7×7 convolution. • Enhances the most informative channels and spatial regions of the input features.	• A differentiable soft attention mechanism is applied, which adaptively weights the features extracted by VGG18. • Weighting is performed through a trainable fully connected layer followed by softmax normalization, allowing the model to focus on the most salient components of the audio data.	• The "bottleneck" attention is integrated into MobileNet blocks. • BAM functions as a guiding attention mechanism, enhancing the capture of relevant spectral-temporal patterns in the input features.
Dataset	**ASVspoof-2019 (LA and PA),** cross-validation at ASVspoof-2021	**ASVspoof-2019** and FoR	**ASVspoof-2019 (LA and PA)**
Metrics/ Results	Accuracy = 99.78% Precision = 98.86% Min-tDCF = 0.030 EER (PA) = 0.52% **Reported value:** EER (LA) = 0.07%	Accuracy = 98.1% FoR EER = 0.04% **Reported value:** EER (LA) = 0.06%	Accuracy = 99.92% Precision = 99.2% Recall = 99.76% EER (PA) = 0.43% **Reported value:** EER (LA) = 0.042%

Tab. 1. (continued) Comparative analysis of attention-based methods for synthetic speech detection

Comparison parameters	Research		
	VGGish with CBAM	**ABC-CapsNet**	**DeepDet (YAMNet with BAM)**
Advantages	• Simple and intuitive architecture. • Efficient fusion of spatial and channel attention via the CBAM module. • High accuracy on the ASVspoof 2019 dataset. • Reliable overall performance and stability, confirmed through cross-validation.	• Cascaded capsule networks enhance the depth and precision of feature analysis. • High detection accuracy with low EER values on multiple datasets. • Effective discrimination between real and synthetic audio regions. • Preserves hierarchical and spatial relationships among features.	• Very low EER on Logical Access attacks, demonstrating high detection accuracy. • Architecture optimized for computational efficiency using depthwise separable convolutions. • Robust feature extraction through a pre-trained YAMNet model, providing informative spectral-temporal representations.
Limitations	• Increased sensitivity to noisy audio recordings. • High computational requirements when processing large mel-spectrograms (authors reduced input resolution to 96×64 for optimization). • Reduced recognition performance on certain Voice Conversion (VC) attacks (A17–A19).	• Capsule networks have high computational complexity and are more difficult to train. • Scaling to very large datasets and real-time applications is challenging. • Model performance is sensitive to hyperparameter selection. • Input spectrogram size affects efficiency and computational cost.	• Sensitivity to attack type: high accuracy for TTS, lower for VC attacks. • Requires careful normalization adjustment (e.g., replacement of BatchNorm with InstanceNorm). • High hardware requirements for training on large datasets, which may limit scalability and real-time deployment.

Thus, the comparative analysis demonstrated that all three methods achieve high performance in synthetic speech detection, but differ in their priorities and limitations. VGGish with CBAM offers a balanced architecture and reliable performance, yet is sensitive to acoustic noise and certain types of attacks. ABC-CapsNet provides the highest robustness across heterogeneous datasets due to its capsule architecture and dynamic routing mechanism, but demands substantial computational resources. DeepDet (YAMNet with BAM) achieves minimal EER with low computational load, making it suitable for real-time applications, although its generalization across diverse attack types is somewhat limited.

The selection of a specific method depends on the operational context and the requirements for accuracy, speed, and robustness.

5 Conclusion

A comparative analysis of three modern synthetic speech detection methods – VGGish with CBAM, ABC-CapsNet, and DeepDet (YAMNet with BAM) – demonstrated that the integration of attention mechanisms substantially enhances the performance of fake voice detection systems. All models achieved high accuracy on standard datasets but differed in architectural complexity, robustness, and computational requirements. VGGish with CBAM provides a balanced trade-off between accuracy and computational efficiency, although it is sensitive to noise. ABC-CapsNet exhibits the highest robustness due to its capsule-based architecture, but demands significant computational resources. DeepDet achieves minimal error rates under low computational load, making it suitable for real-time applications.

Overall, these findings confirm that attention mechanisms play a crucial role in improving the accuracy and reliability of synthetic speech detection, and future research should focus on developing more adaptive, lightweight, and robust architectures.

References

[1] *Zahra K., Gabrielle W., Vandana P.* Audio deepfakes: A survey // Frontiers in Big Data. – 2023. – Vol. 5. – Article 1001063. – DOI: 10.3389/fdata.2022.1001063.

[2] *Groh M. et al.* Human detection of political speech deepfakes across transcripts, audio, and video // Nature Communications. – 2024. – Vol. 15. – Article 7629. – DOI: 10.1038/s41467-024-51998-z.

[3] *Malinka K. et al.* Comprehensive multiparametric analysis of human deepfake speech recognition // Journal of Image and Video Processing. – 2024. – Vol. 24. – DOI: 10.1186/s13640-024-00641-4.

[4] *Zhang B., Cui H., Nguyen V., Whitty M.* Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead // Sensors. – 2025. – Vol. 25. – Article 1989. – DOI: 10.3390/s25071989.

[5] *Tan Y., Weng X., Zhu J.* Dual-Channel Spoofed Speech Detection Based on Graph Attention Networks // Symmetry. – 2025. – Vol. 17. – Article 641. – DOI: 10.3390/sym17050641.

[6] *Todisco M. et al.* ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection // Proceedings of Interspeech 2019. – 2019. – P. 1008–1012. – DOI: 10.21437/Interspeech.2019-2249.

[7] *Reimao R., Tzerpos V.* FoR: A Dataset for Synthetic Speech Detection // Proceedings of 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD). – Timisoara, Romania, 2019. – P. 1–10. – DOI: 10.1109/SPED.2019.8906599.

[8] *Kanwal T. et al.* Fake speech detection using VGGish with attention block // Journal of Audio, Speech, and Music Processing. – 2024. – Vol. 35. – DOI: 10.1186/s13636-024-00348-4.

[9] *Wani T.M., Gulzar R., Amerini I.* ABC-CapsNet: Attention based Cascaded Capsule Network for Audio Deepfake Detection // IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). – Seattle, WA, USA, 2024. – P. 2464–2472. – DOI: 10.1109/CVPRW63382.2024.00253.

[10] *Mahum R. et al.* DeepDet: YAMNet with BottleNeck Attention Module (BAM) for TTS synthesis detection // Journal of Audio, Speech, and Music Processing. – 2024. – Vol. 18. – DOI: 10.1186/s13636-024-00335-9.

[11] *Yaseen M.U., Nasralla M.M., Aslam F., Ali S.S., Khattak S.B.A.* A 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.

[12] *Chaudhari S., Mithal V., Polatkan G., Ramanath R.* An attentive survey of attention models // ACM Transactions on Intelligent Systems and Technology. – 2021. – Vol. 12, No. 5. – Article 53. – DOI: 10.1145/3465055.

[13] *Khan W., Malik K.M.* SpoTNet: A spoofing-aware Transformer Network for Effective Synthetic Speech Detection // Proceedings of 2nd ACM International Workshop on Multimedia AI against Disinformation (MAD '23). – New York, NY, USA: ACM, 2023. – P. 10–18. – DOI: 10.1145/3592572.3592841.

SUN’IY NUTQNI ANIQLASH UCHUN DIQQAT MEXANIZMIGA ASOSLANGAN ARXITEKTURALAR SHARHI

[1]*Abdullayev Sh.Sh.*, [2*]*Kim K.K.*

*kamilla.k.kim@gmail.com
[1]Oriental Universiteti, Toshkent, Uzbekistan.
[2]Raqamli texnologiyalar va sun'iy intellektni rivojlantirish ilmiy-tadqiqot instituti, Toshkent, O‘zbekiston.

Annotatsiya. Ushbu maqola sintetik nutqni aniqlash uchun zamonaviy arxitekturaning analitik sharhini taqdim etadi va diqqat mexanizmlarining markaziy rolini ifodalaydi. Ishda uchta reprezentativ modellar ko‘rib chiqilgan: CBAM asosida yaxshilangan o‘zgartirilgan VGGish, BAM bilan to‘ldirilgan YAMNet va ABC-CapsNet kaskadli kapsula tarmog‘i. Tahlil ikki o‘lchovli tasvirlar sifatida talqin qilingan mel spektrogrammalariga asoslandi, bu yerda diqqat modullari eng informativ vaqt-chastota sohalarini aniqlaydi. Muhokamada arxitektura tanlovlarini, belgilarni ajratib olish strategiyalarini, diqqat mexanizmlarini, ma’lumotlar to‘plamini va samaradorlik ko‘rsatkichlarini qamrab oladi. Qiyosiy baholash shuni ko‘rsatdiki, BAM bilan YAMNet o‘rtacha hisoblash murakkabligi bilan eng past EERga erishdi, ABC-CapsNet esa ma’lumotlarning o‘zgaruvchanligiga kuchli mustahkamlikni namoyish etdi va turli xil ma’lumotlar to‘plamlarida yaxshi umumlashtirish qobiliyatini ko‘rsatdi. Bundan farqli o‘laroq, CBAM bilan VGGish aniqlik va amalga oshirishning soddaligi o‘rtasida optimal muvozanatni ta’minlaydi. Tadqiqot sintetik nutqni aniqlash tizimlarining samaradorligini oshirishda diqqat mexanizmlarining hal qiluvchi rolini ko‘rsatdi va keyingi tadqiqotlarni rivojlantirishning istiqbolli yo‘nalishlarini belgilaydi.

Kalit so'zlar: audio spufingni aniqlash, sintetik nutq, diqqat mexanizmlari, CBAM, kaskadli kapsula tarmoqlari, BAM.

ОБЗОР АРХИТЕКТУР, ОСНОВАННЫХ НА МЕХАНИЗМА ВНИМАНИИ ДЛЯ ОБНАРУЖЕНИЯ СИНТЕТИЧЕСКОЙ РЕЧИ

[1]*Абдуллаев Ш.Ш.*, [2*]*Ким К.К.*

*kamilla.k.kim@gmail.com
[1]Ориентал университет, Ташкент, Узбекистан.
[2]Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта, Ташкент, Узбекистан.

Аннотация. В данной статье представлен аналитический обзор современных архитектур для распознавания синтетической речи, подчёркивающий центральную роль механизмов внимания. Рассматриваются три репрезентативные модели: модифицированная VGGish, улучшенная с помощью CBAM, YAMNet, дополненная BAM, и каскадная капсульная сеть ABC-CapsNet. Анализ основан на mel-спектрограммах, интерпретируемых как двумерные изображения, где модули внимания выявляют наиболее информативные частотно-временные области. Обсуждение охватывает выбор архитектурного решения, стратегии извлечения признаков, механизмы внимания, наборы данных и показатели производительности. Сравнительная оценка показывает, что YAMNet с BAM достигает наименьшего EER при умеренной вычислительной сложности, в то время как ABC-CapsNet демонстрирует высокую устойчивость к изменчивости данных и хорошо обобщает результаты на гетерогенных наборах данных. В отличие от этого, VGGish с CBAM обеспечивает сбалансированный компромисс между точностью и простотой реализации. В исследовании подчёркивается ключевая роль механизмов внимания в повышении эффективности систем распознавания синтетической речи и намечаются перспективные направления для дальнейших исследований и разработок.

Ключевые слова: обнаружение аудиоспуфинга, синтетическая речь, механизмы внимания, CBAM, каскадные капсульные сети, BAM.