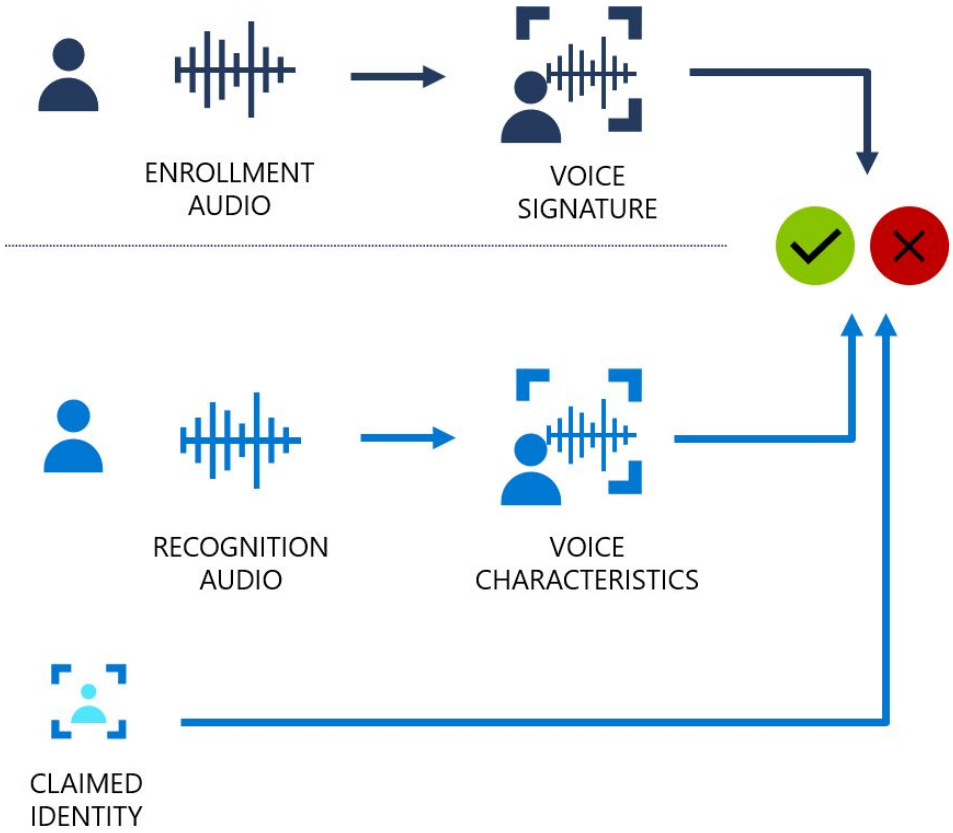
Enhancing Deepfake Detection: A Study Using WavLM and Advanced RawBoost Augmentation Techniques

Nhan Tri Do, Loi Nguyen Hoang, Phuong Ta Viet, Kien Phan Trung

VinBigData Joint Stock Company, Vietnam

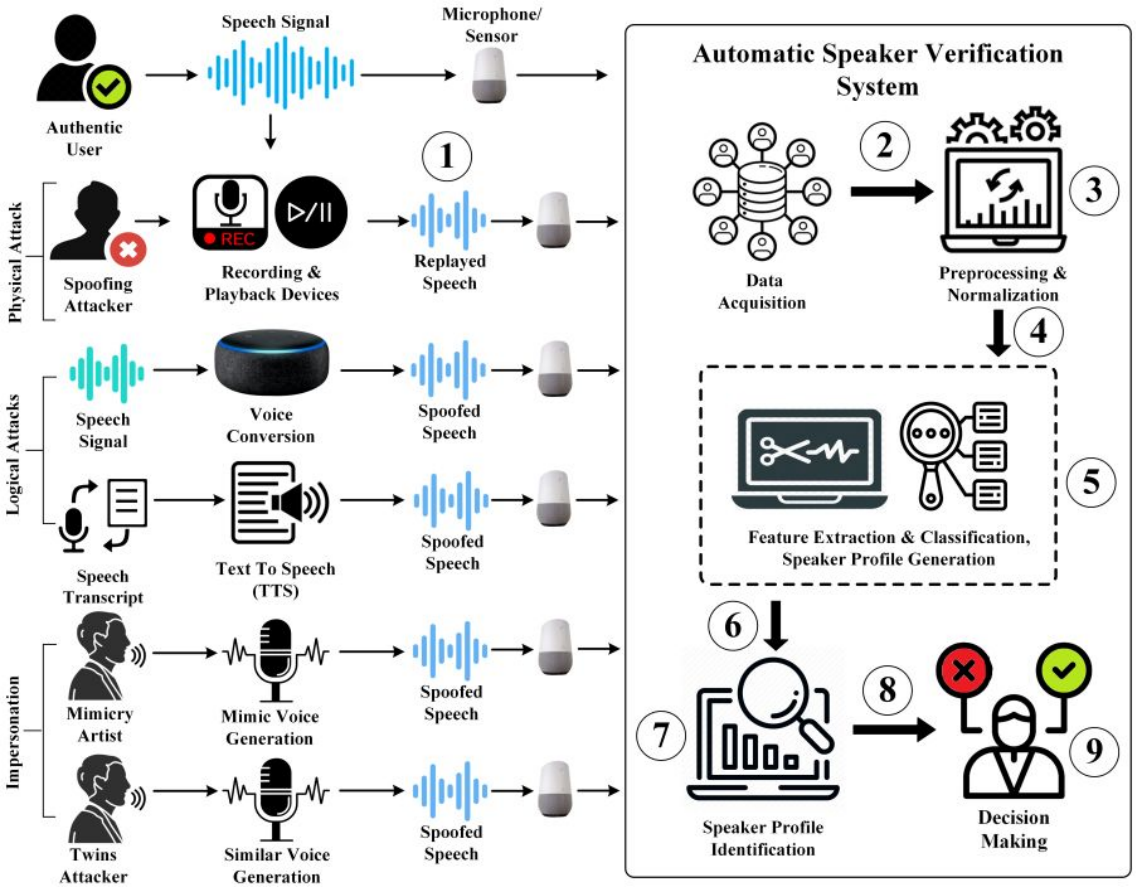
{dotrinhan99, hoangloi2001, tvphuong10, trungkien.it.98}@gmail.com

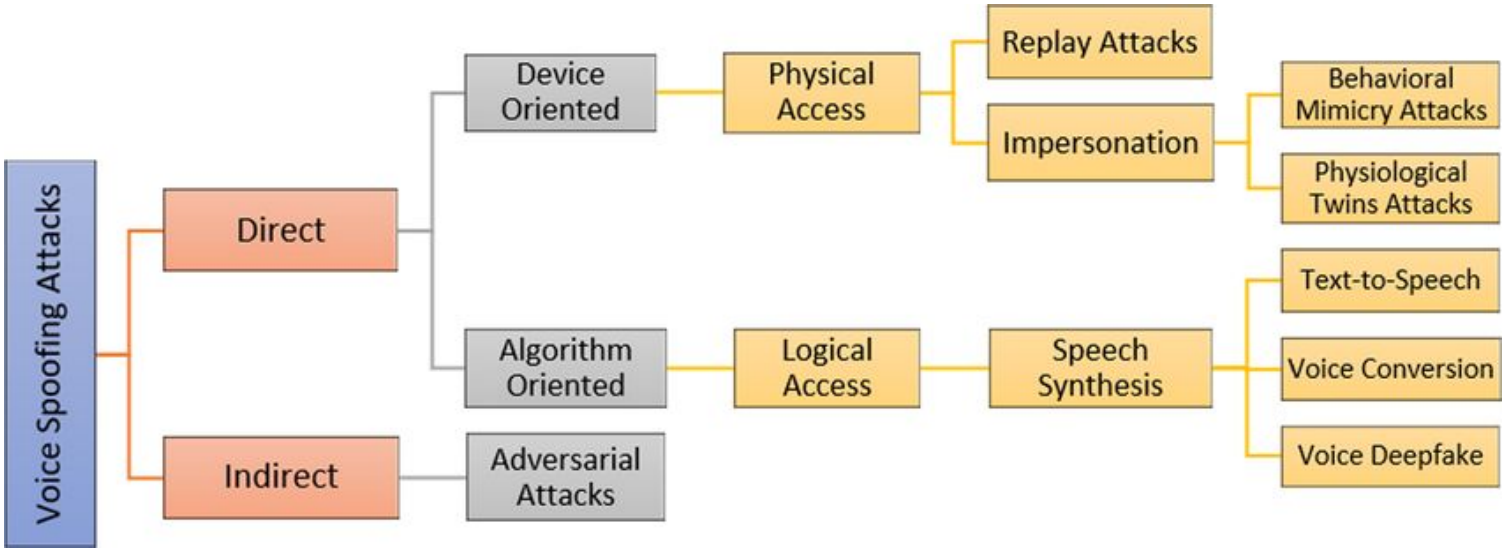
- **Speaker Verification**



Voice Spoofing Countermeasures

Threats to Automatic
Speaker Verification Systems



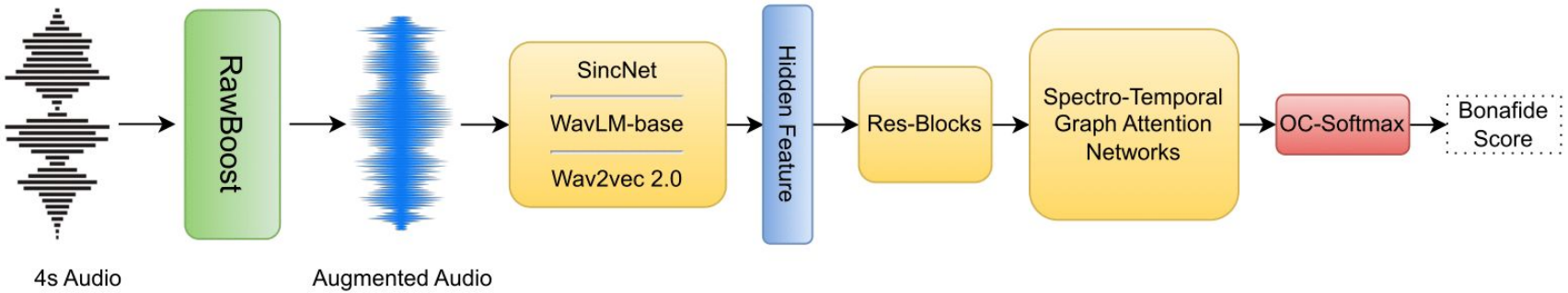

Voice Spoofing Attacks
Direct
Indirect
Device Oriented
Algorithm Oriented
Adversarial Attacks
Physical Access
Logical Access
Replay Attacks
Impersonation
Speech Synthesis
Behavioral Mimicry Attacks
Physiological Twins Attacks
Text-to-Speech
Voice Conversion
Voice Deepfake

Passive liveness detection

- Smartphone-based magnetometer → voice presentation attack → capture the use of loudspeaker by sensing the magnetic field which would be absent from human vocals→Magnetic field-based detection can be reliable for the detection of playback within 6-8 cm from the smartphone
- Distortion in human breath when it reaches a microphone
- Estimates dynamic sound source position (articulation position within the mouth) → capture the dynamics of time-difference-of-arrival (TDOA)
- Doppler effect to detect the replay attack → capture the articulatory gestures of the speakers when they speak a pass-phrase (Doppler radar and transmits a high-frequency tone at 20 kHz from the built-in speaker and senses the reflections using the microphone during the authentication process)
- POCO - Pop noise = sonic artifacts of plosive consonants [p][b][k][g], breath (the interactions between the airflow and the vocal cavities may result in a sort of plosive burst)→ real only→ Pop noise detector
- Void method → analyzes cumulative power patterns in acoustic spectrograms
- Deep Learning → Replay and Logical anti-spoofing model

Spectro-Temporal Graph Attention with WavLM Architecture Diagram

- Building on the AASIST framework
- Replaced the original Sinc-layer front-end with the WavLM model.
- Additionally, we also experimented with replacing Graph HS-GAL with a Conformer and integrated a Retention Network to enhance model complexity and improve inference speed.

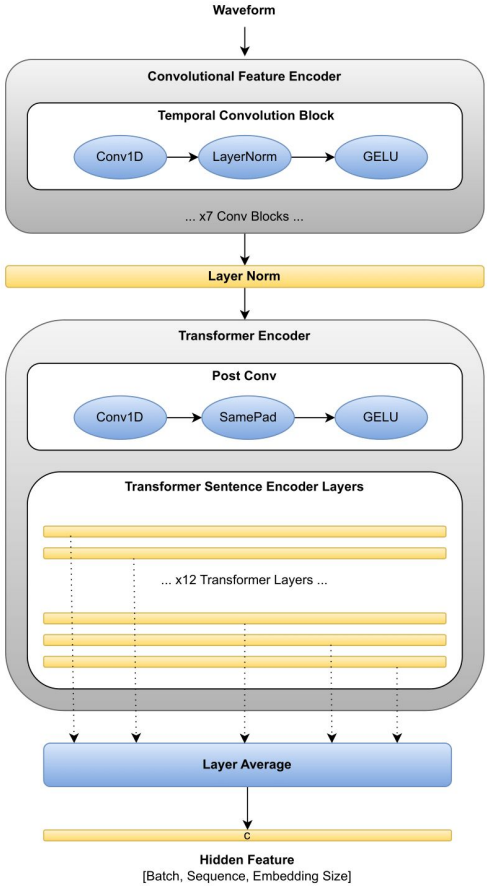


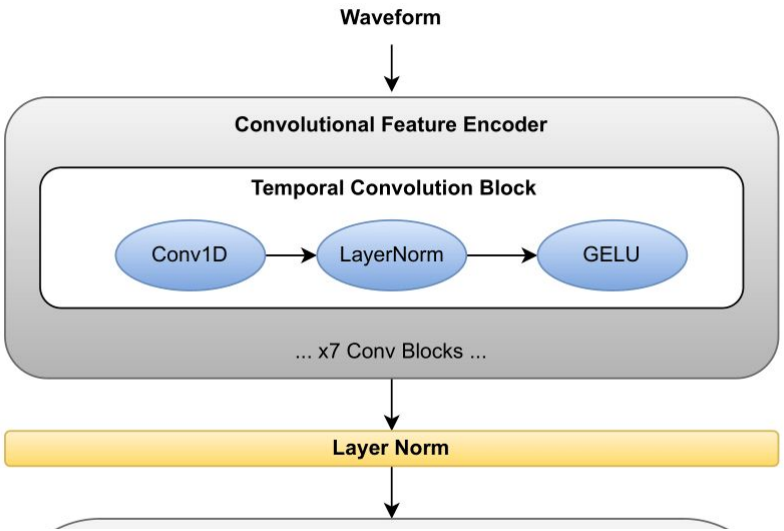
Self-supervised Feature Extractor

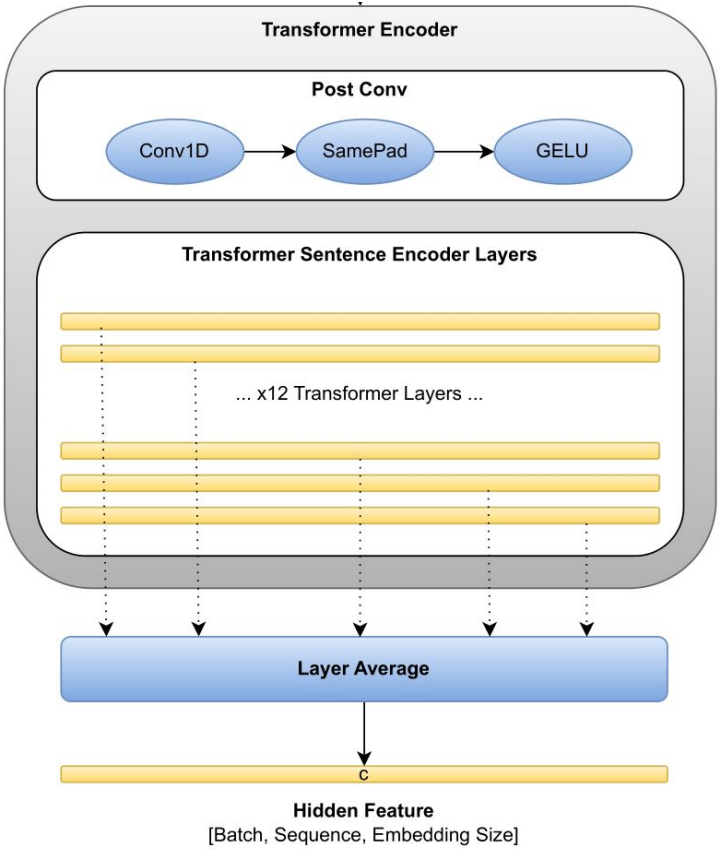
WavLM is versatile pre-trained model designed for robust speech processing

- WavLM is trained using a masked speech denoising and prediction task.
- Superior performance in noisy environments, compared to wav2vec 2.0.

-> Extract features from raw waveforms

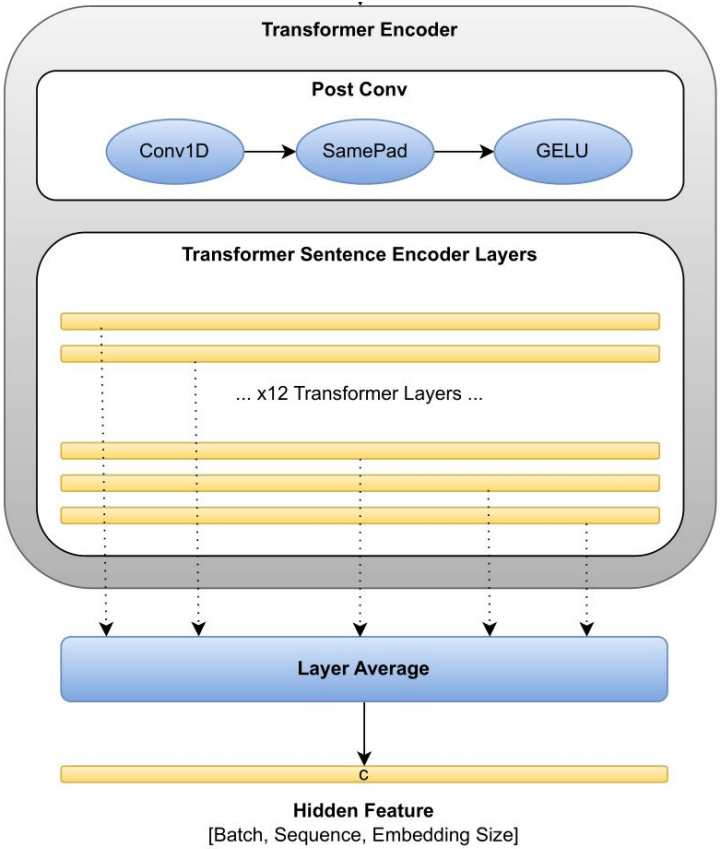



Waveform
Convolutional Feature Encoder
Temporal Convolution Block
Conv1D
LayerNorm
GELU
... x7 Conv Blocks ...
Layer Norm

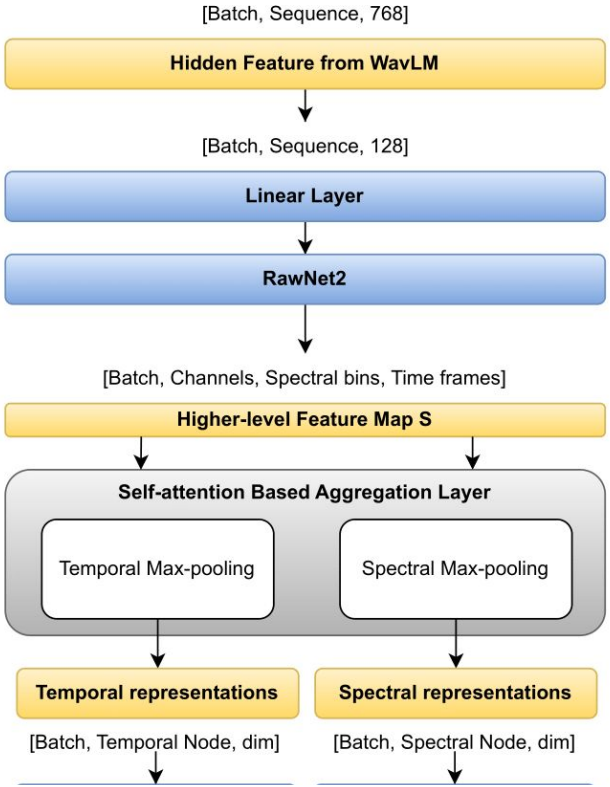

Transformer Encoder
Post Conv
Conv1D
SamePad
GELU
Transformer Sentence Encoder Layers
... x12 Transformer Layers ...
Layer Average
C
Hidden Feature
[Batch, Sequence, Embedding Size]

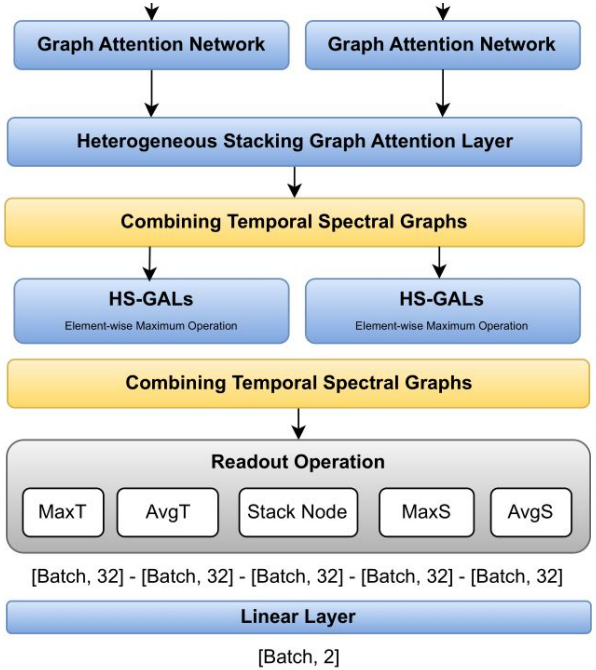
Self-supervised Feature Extractor

- Fine-tuned the WavLM base model using the LibriSpeech dataset
- Averaging the outputs of its layers to create feature vectors with a dimensionality of 768 as in Figure 2.

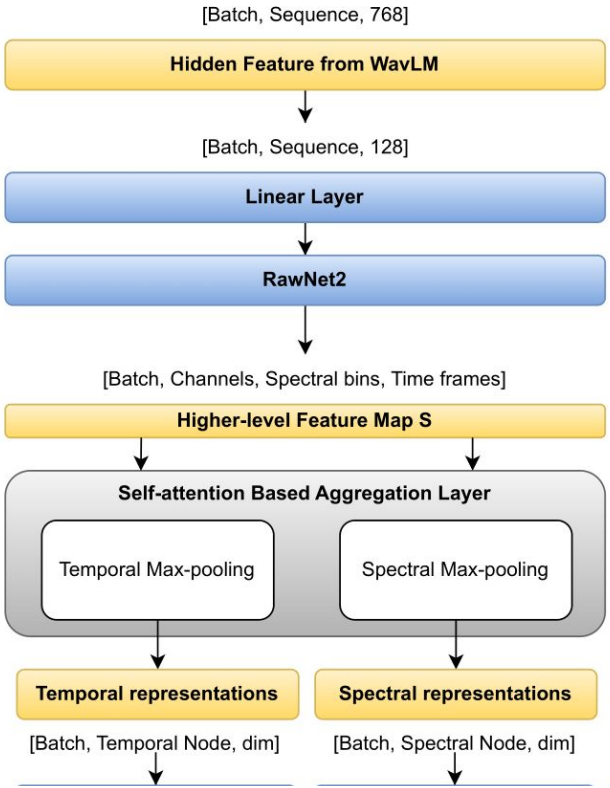


Spectrotemporal Graph Attention Network





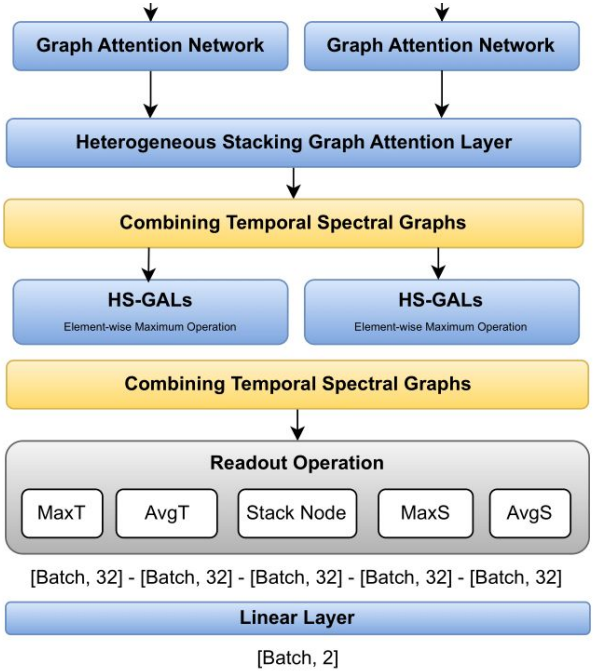
Spectrotemporal Graph Attention Network



- Takes the hidden features input extracted from the pretrained WavLM model.
- These features are passed through a linear post-processing layer to reduce the feature dimension before being fed into RawNet2
  - 6 residual blocks.
  - To learn high-level features that represent of channels, spectral, and time frames.
- Spectral and temporal representations are then created using max pooling functions and processed through a graph attention network.

Spectrotemporal Graph Attention Network

- Combined into a heterogeneous spectrotemporal graph using heterogeneous stacking graph attention layer - HSGAL.
- Fed into two parallel HSGAL modules to learn spoofing features before merging into a final graph.
- Readout is performed on nodes of this graph, including:
  - Node-wise maximum and Average for the spectral and temporal nodes, respectively.
  - Each operation produces a 32-dimensional feature, which then concatenated and results in a 160-dimensional vector.
- This vector is then passed through a fully connected layer to produce the two classes: bonafide and spoof.



Loss for AntiSpoofing

Our loss function strategy included:

- Weighted cross-entropy loss to address class imbalance between the Bonafide and Spoof classes
- The weights assigned to bonafide and spoof classes are 0.9 and 0.1, respectively

$$L_{\text{WCE}} = -\frac{1}{N}\sum_{n=1}^{N}\left[w_1 \cdot y_n \cdot \log(p_{n,1}) + w_0 \cdot (1 - y_n) \cdot \log(p_{n,0})\right]$$

$$p_{n,c} = \frac{\exp(x_{n,c})}{\exp(x_{n,0}) + \exp(x_{n,1})} \quad \text{for } c \in \{0, 1\}$$

Loss for AntiSpoofing

In practice, new voice attack methods are constantly being developed:
-> Leading to the emergence of unknown attacks.

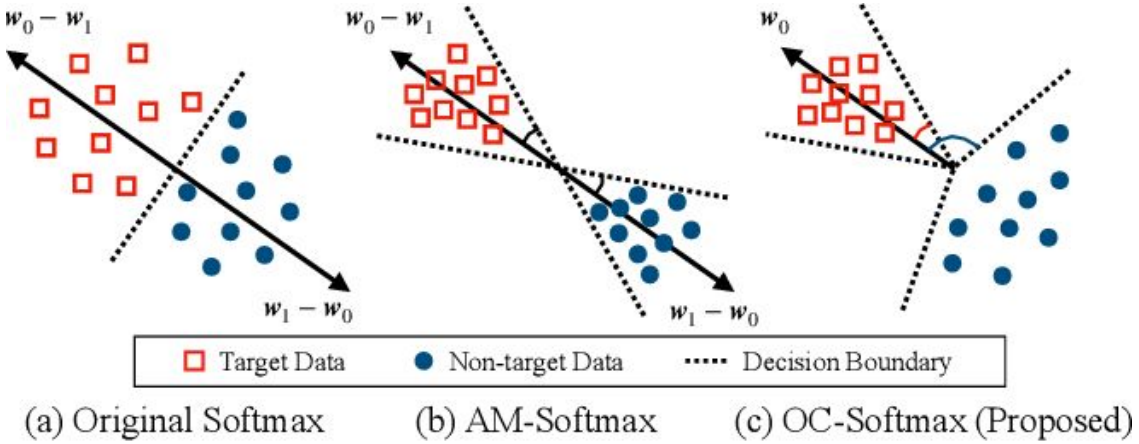
If the original Softmax loss for binary classification is used, the model may overfit to the attack methods present in the training set.

OCSoftmax (One-Class Softmax) is a specialized variant of the standard softmax function that focuses on detecting bonafide audio and isolating spoofing attacks.

$$L_{\text{OCS}} = \frac{1}{N} \sum_{i=1}^{N} \log \left(1 + e^{\alpha(m_{y_i} - \hat{w}_0 \cdot \hat{x}_i)(-1)^{y_i}} \right)$$

Loss for AntiSpoofing

OCSoftmax loss configured to focus on detecting bonafide cases more accurately



Rawboost Augmentation

- Using 4-second audio segments
- Preprocessed with random padding and silence trimming
- Augmented with RawBoost:
    - Linear and non-linear convolutive noise, impulsive signal-dependent additive noise
    - Stationary signal-independent additive noise.

Experiment

The experiment was conducted with:

- Batch size of 24
- Distributed training across two GeForce RTX 4090 GPUs and 10 CPU cores.

Feature Extractor	Backbone Model	EER	minDCF	actDCF	Cllr
Wav2Vec2	Graph Attention	4.26	0.1002	0.2255	0.3625
Wav2Vec2	Conformer	4.49	0.1255	0.1676	0.6403
WavLM	Graph Attention	2.85	0.0816	0.1227	0.5552
Fusion (Wav2Vec2 and WavLM)	Graph Attention	2.69	0.0764	0.1622	0.2440

Conclusion

- Leveraging advanced self-supervised learning models and robust augmentation techniques
- The fine-tuning of WavLM for feature extraction, combined with Spectro-Temporal Graph Attention Networks -> EER of 2.85% with WavLM and 2.69% with model fusion
- Future work will focus on optimizing model architecture with different hyper-parameters and exploring additional augmentation strategies

Market Products:

ID R&D ID R&D, a Mitek company Anti-Spoofing for Authentication Microsoft Azure Marketplace ID R&D, Inc.	• Hỗ trợ tích hợp iOS, Android, Docker Image, Azure App, ChatGPT, có sẵn demo: ID R&D, Inc. • App on Microsoft Azure → prevents speech conversion, replay attacks, and TTS attacks. • 1st at Logical Access Condition - ASVspoof Challenge 2019 + 1st at Replay Attack ASVspoof 2017 • Detect voice clones in the microphone channel (16 kHz), created by any of the leading clone tech providers • Detect replay, hardware- and software-based attacks • Requires just 3 seconds of audio Language independent • Telcos to bolster new subscriber fraud prevention efforts in the contact center
Borac Solutions	• Gaussian Mixture Models, Auto-encoders • Constant-Q-Cepstral Coefficients and linear frequency cepstral coefficients
Mobbeel	

Some problems in Real-life Application

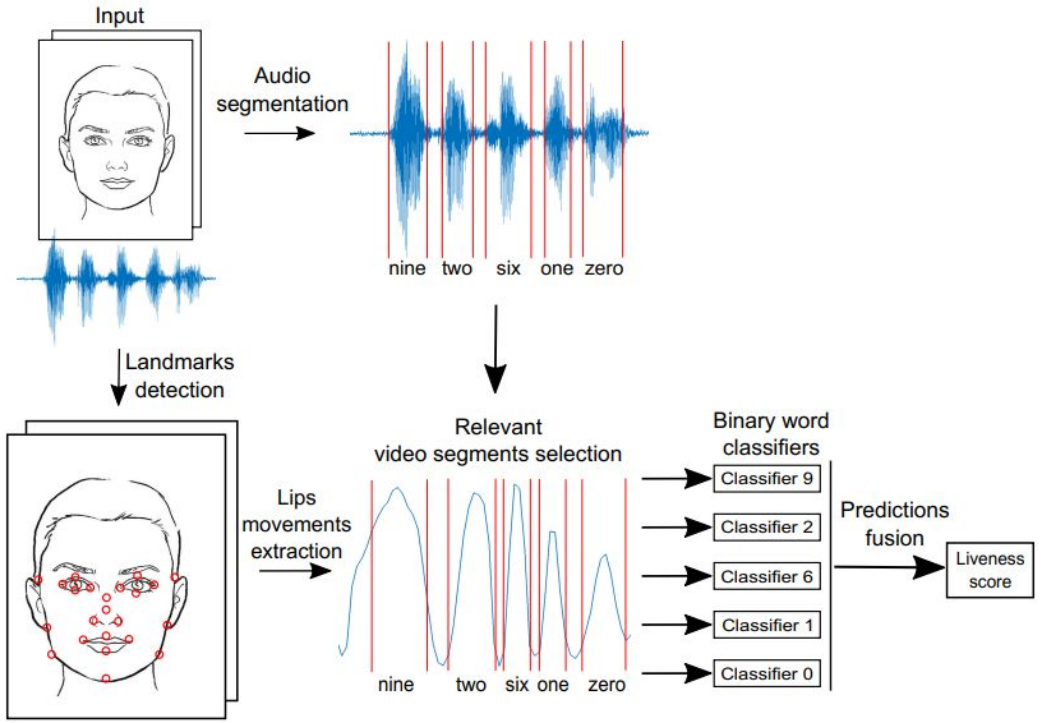
- The research was conducted through an experimental process to address the problem of detecting spoofing in voice-recorded contracts for insurance companies in Vietnam

The results dropped from 96% to 80% during real-world testing

- However, there are still many issues that require further analysis and adaptation as spoofing algorithms continue to improve, along with challenges related to robustness against diverse devices and real-world environments.

Active liveness detection

- Challenge-response: → Read text in real-time → ASR → Edit distance
- Multi-Factor Authentication (MFA) - combine with face



Open condition												
	#	ID	minDCF	actDCF	C_{llr}	EER	#	ID	minDCF	actDCF	C_{llr}	EER
•▲	1	T45	0.0750	1.0000	0.7923	2.59	18	-	0.1949	0.2438	0.7028	7.05
•▲	2	T36	0.0936	1.0000	0.8874	3.41	19	-	0.1966	1.0000	0.9327	6.80
•▲	3	T27	0.0937	0.1375	0.1927	3.42	•▲ 20	T33	0.2021	0.6028	0.5560	7.01
•▲	4	T23	0.1124	1.0000	0.9179	4.16	21	-	0.2148	1.0000	0.8124	7.43
•▲	5	T43	0.1149	0.5729	0.9562	4.04	•▲ 22	T51	0.2236	1.0000	0.8011	7.72
•▲	6	T13	0.1301	0.1415	0.3791	4.50	•▲ 23	T46	0.2245	1.0000	1.0308	9.36
•▲	7	T06	0.1348	0.2170	0.3096	5.02	24	-	0.2573	1.0000	0.9955	9.28
	8	-	0.1414	0.5288	0.6149	4.89	25	-	0.2642	0.7037	2.1892	10.32
○▲	9	T31	0.1499	0.2244	0.5559	5.56	•△ 26	T47	0.2660	0.3321	0.4932	9.18
•▲	10	T29	0.1549	0.2052	0.7288	5.37	27	-	0.2668	0.2923	0.6194	9.59
•▲	11	T35	0.1611	1.0000	1.0384	5.93	•▲ 28	T41	0.3010	0.3095	0.4773	10.45
	12	-	0.1665	0.1669	0.2351	5.77	29	-	0.4121	0.4266	0.7185	14.25
•▲	13	T21	0.1728	0.2392	0.9498	6.01	•▲ 30	T02	0.4845	1.0000	0.9332	17.08
○▲	14	T17	0.1729	1.0000	2.3217	5.99	○△ 31	T15	0.5112	0.6723	0.8858	22.24
○▲	15	T19	0.1743	0.3087	0.4757	6.06	32	-	0.6584	0.7451	1.1404	22.90
	16	-	0.1840	1.0000	0.8764	6.35	33	-	0.7969	1.0000	0.9920	35.72
	17	-	0.1933	1.0000	0.8342	6.67	○△ 34	T53	0.9744	1.0539	2.4977	44.94

Results							
#	User	Entries	Date of Last Entry	minDCF ▲	actDCF ▲	Cllr ▲	EER ▲
1	*Anonymous* <	1	07/22/24	0.0937 (3)	0.1375 (1)	0.1927 (1)	3.42 (3)
2	*Anonymous*	1	07/22/24	0.1301 (6)	0.1415 (2)	0.3791 (4)	4.50 (6)
3	*Anonymous*	1	07/22/24	0.1665 (12)	0.1669 (3)	0.2351 (2)	5.77 (11)
4	*Anonymous*	1	07/23/24	0.1549 (10)	0.2052 (4)	0.7288 (14)	5.37 (9)
5	*Anonymous*	1	07/22/24	0.1348 (7)	0.2170 (5)	0.3096 (3)	5.02 (8)
6	*Anonymous*	1	07/23/24	0.1499 (9)	0.2244 (6)	0.5559 (8)	5.56 (10)
7	*Anonymous*	1	07/22/24	0.1728 (13)	0.2392 (7)	0.9498 (25)	6.01 (14)
8	*Anonymous*	1	07/23/24	0.1949 (18)	0.2438 (8)	0.7028 (12)	7.05 (20)
9	*Anonymous*	1	07/23/24	0.2668 (27)	0.2923 (9)	0.6194 (11)	9.59 (26)
10	*Anonymous*	1	07/22/24	0.1743 (15)	0.3087 (10)	0.4757 (5)	6.06 (15)
11	*Anonymous*	1	07/23/24	0.3010 (28)	0.3095 (11)	0.4773 (6)	10.45 (28)
12	*Anonymous*	1	07/23/24	0.2660 (26)	0.3321 (12)	0.4932 (7)	9.18 (23)
13	*Anonymous*	1	07/22/24	0.4121 (29)	0.4266 (13)	0.7185 (13)	14.25 (29)
14	*Anonymous*	1	07/22/24	0.1414 (8)	0.5288 (14)	0.6149 (10)	4.89 (7)
15	*Anonymous*	1	07/22/24	0.1149 (5)	0.5729 (15)	0.9562 (26)	4.04 (4)
16	*Anonymous*	1	07/23/24	0.2021 (20)	0.6028 (16)	0.5560 (9)	7.01 (19)
17	*Anonymous*	1	07/23/24	0.5112 (31)	0.6723 (17)	0.8858 (20)	22.24 (31)
18	*Anonymous*	1	07/22/24	0.2642 (25)	0.7037 (18)	2.1892 (32)	10.32 (27)
19	*Anonymous*	1	07/23/24	0.6584 (32)	0.7451 (19)	1.1404 (31)	22.90 (32)
20	*Anonymous*	1	07/22/24	0.0750 (1)	1.0000 (20)	0.7923 (15)	2.59 (1)