1. Research

1. Voice Radar: Voice Deepfake Detection using Micro-Frequency and Compositional Analysis

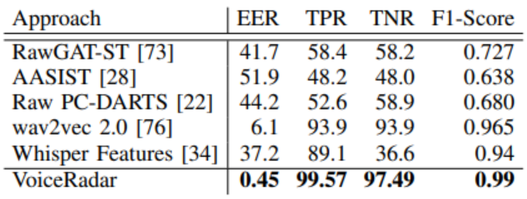
➢ Key Technical Innovation:

Contributions of this paper can be summarized as:

- This paper presents the design and implementation of a novel detection technique, VoiceRadar to accurately distinguish between audio deepfakes and human-generated audio.
- VoiceRadar's approach approximates the physical models of the Doppler effect and drumhead vibrations to capture inherent audio signal variations. By integrating the frequency signatures of audio waves' translation, vibration, and rotation frequencies, it models the observed frequency within the Doppler effect framework. This captures translational changes, rhythmic patterns, and subtle nuances within the audio signal. The observed frequency is then used in the loss function of a supervised learning algorithm to classify audio signals accurately.
- This technique creates a comprehensive benchmark dataset comprising over 500,000 audio samples, generated using state-of-the-art TTS generators and VC modules.
- VoiceRadar is extensively evaluated using this new benchmark dataset, demonstrating the robustness and effectiveness of the approach in detecting audio deepfakes. It is also compared with existing state-of-the-art deepfake detection frameworks and shows superior detection performance.

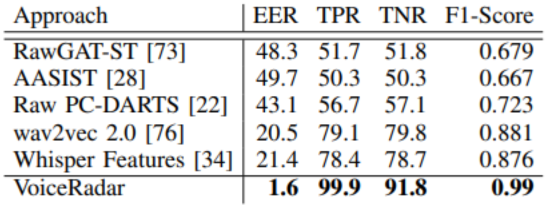
Reported Performance Metrics:

True Positive Rate (TPR) indicates the tool's sensitivity in detecting fake audio samples.
True Negative Rate (TNR), also called specificity, indicates the tool's ability to correctly recognize human-voice samples.
Equal Error Rate (EER) is a combination of the False Positive Rate (FPR = 1 − TNR) as well as the False Negative Rate (FNR = 1 − TPR).
• F1-Score: The F1-score considers both precision (the ratio of TP to the total number of positive predictions) and recall (the ratio of TP to the total number of actual positive instances).
Comparison of VoiceRadar with existing deepfake detection approaches for text-to-speech (TTS) generated fake samples.



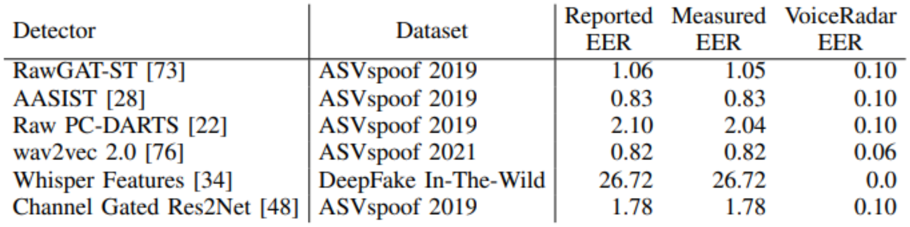
Approach	EER	TPR	TNR	F1-Score
RawGAT-ST [73]	41.7	58.4	58.2	0.727
AASIST [28]	51.9	48.2	48.0	0.638
Raw PC-DARTS [22]	44.2	52.6	58.9	0.680
wav2vec 2.0 [76]	6.1	93.9	93.9	0.965
Whisper Features [34]	37.2	89.1	36.6	0.94
VoiceRadar	**0.45**	**99.57**	**97.49**	**0.99**

Comparison of VoiceRadar with existing deepfake detection approaches for Voice Conversion (VC) generated fake samples.



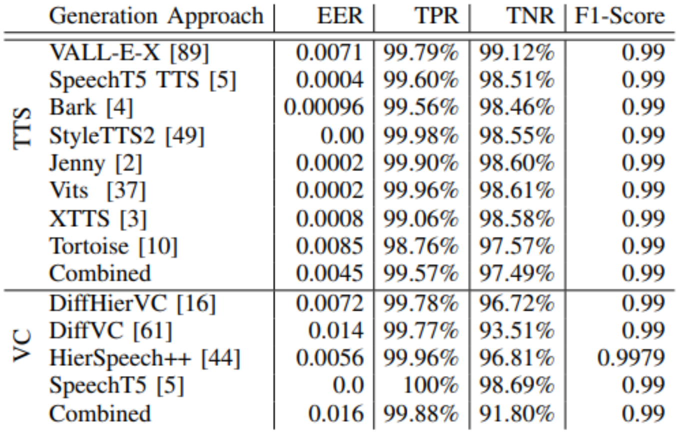
Approach	EER	TPR	TNR	F1-Score
RawGAT-ST [73]	48.3	51.7	51.8	0.679
AASIST [28]	49.7	50.3	50.3	0.667
Raw PC-DARTS [22]	43.1	56.7	57.1	0.723
wav2vec 2.0 [76]	20.5	79.1	79.8	0.881
Whisper Features [34]	21.4	78.4	78.7	0.876
VoiceRadar	**1.6**	**99.9**	**91.8**	**0.99**

Comparison for state-of-the-art detection tools of performance reported in the paper, the performance we reproduced for respective dataset and VoiceRadar on the dataset.



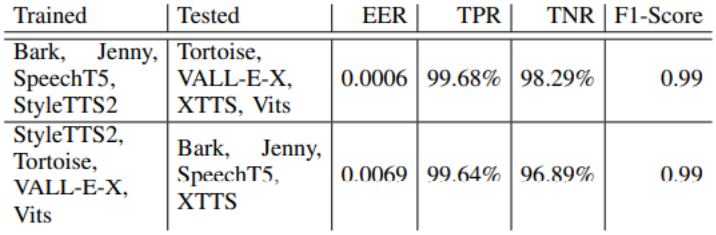
Detector	Dataset	Reported EER	Measured EER	VoiceRadar EER
RawGAT-ST [73]	ASVspoof 2019	1.06	1.05	0.10
AASIST [28]	ASVspoof 2019	0.83	0.83	0.10
Raw PC-DARTS [22]	ASVspoof 2019	2.10	2.04	0.10
wav2vec 2.0 [76]	ASVspoof 2021	0.82	0.82	0.06
Whisper Features [34]	DeepFake In-The-Wild	26.72	26.72	0.0
Channel Gated Res2Net [48]	ASVspoof 2019	1.78	1.78	0.10

Effectiveness of VoiceRadar for the individual textto-speach (TTS) and voice conversion (VC) approaches.



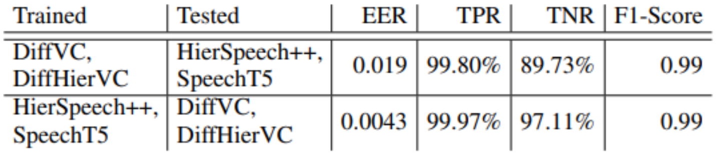
	Generation Approach	EER	TPR	TNR	F1-Score
TTS	VALL-E-X [89]	0.0071	99.79%	99.12%	0.99
	SpeechT5 TTS [5]	0.0004	99.60%	98.51%	0.99
	Bark [4]	0.00096	99.56%	98.46%	0.99
	StyleTTS2 [49]	0.00	99.98%	98.55%	0.99
	Jenny [2]	0.0002	99.90%	98.60%	0.99
	Vits [37]	0.0002	99.96%	98.61%	0.99
	XTTS [3]	0.0008	99.06%	98.58%	0.99
	Tortoise [10]	0.0085	98.76%	97.57%	0.99
	Combined	0.0045	99.57%	97.49%	0.99
VC	DiffHierVC [16]	0.0072	99.78%	96.72%	0.99
	DiffVC [61]	0.014	99.77%	93.51%	0.99
	HierSpeech++ [44]	0.0056	99.96%	96.81%	0.9979
	SpeechT5 [5]	0.0	100%	98.69%	0.99
	Combined	0.016	99.88%	91.80%	0.99

Cross-evaluation of VoiceRadar for TTS datasets.



Trained	Tested	EER	TPR	TNR	F1-Score
Bark, Jenny, SpeechT5, StyleTTS2	Tortoise, VALL-E-X, XTTS, Vits	0.0006	99.68%	98.29%	0.99
StyleTTS2, Tortoise, VALL-E-X, Vits	Bark, Jenny, SpeechT5, XTTS	0.0069	99.64%	96.89%	0.99

Cross-evaluation of VoiceRadar for VC datasets.



Trained	Tested	EER	TPR	TNR	F1-Score
DiffVC, DiffHierVC	HierSpeech++, SpeechT5	0.019	99.80%	89.73%	0.99
HierSpeech++, SpeechT5	DiffVC, DiffHierVC	0.0043	99.97%	97.11%	0.99

➢ Why you find this approach promising for our specific needs.

- Detecting AI-generated human speech- Performance metrics shows this approach is promising for this particular need.
- Potential for real-time or near real-time detection-

For feature extraction lightweight or distilled speech embedding models like Wav2Vec2-light, Whisper-tiny, or MobileNet-based speech models for real time use cases.

Once the model is trained, inference using the observed frequency (f_0) and audio embeddings can be performed efficiently, making it suitable for real-time applications. However, for deployment on mobile or edge devices, an active internet connection is typically required, as inference is performed on the cloud. This ensures that computational resources are optimized and model updates can be seamlessly integrated.

- Analysis of real conversations-Real conversations vary in accent, tone, and background. A diverse training set (500,000+ samples) ensures the model can still catch deepfakes without being over-sensitive. The model captures subtle nuances like emotion in a person's voice .A tiny tremor or breath, a small change in pitch or tone which can enhance analysis of real conversations.

  Use cases- Call Centers: Monitor customer calls in real time to flag potential deepfake impersonations of VIP clients.

  Video Conferencing: Integrate into Zoom/MS Teams to alert when a participant's voice is synthetic.

➢ Potential limitations or challenges:

The full **HuBERT base model** is large (~95M+ parameters) can be slow on edge devices or in constrained environments.

Real-time implementation of VoiceRadar on edge devices may be limited by the need for cloud-based inference (adding operational costs) or the computational requirements of estimating physical motion parameters such as rotation angles and velocities. Efficient approximations or heuristics are necessary for deployment on low-power devices.

2. SafeEar: Content Privacy-Preserving Audio Deepfake Detection

➢ Key Technical Innovation:

- SafeEar is the first framework to investigate and validate the feasibility of achieving audio deepfake detection while preserving speech content privacy.
- This model proposes a novel privacy-preserving deepfake detection framework that devises a neural audio codec into a semantic-acoustic information decoupling model, ensuring content privacy. It further develops an advanced detector that achieves effective deepfake detection with only acoustic information.
- SafeEar constructs CVoiceFake and establishes a comprehensive benchmark focusing on the deepfake detection and content privacy preservation tasks. Experiments demonstrate the effectiveness of SafeEar in detecting deepfake audio under various impact factors and in thwarting multiple content recovery attacks.

➢ Reported Performance Metrics: Tandem Detection Cost Function (t-DCF)

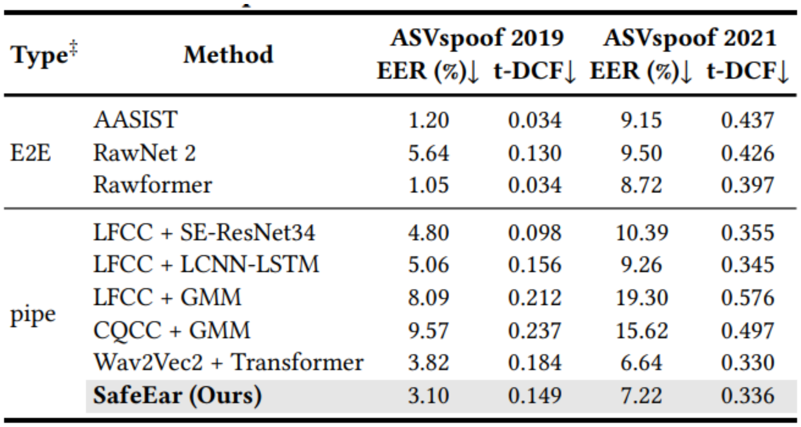
It quantifies the trade-off between two types of errors:

1. False Acceptance – when a deepfake (spoof) is wrongly accepted as real.
2. False Rejection – when a real (bonafide) sample is wrongly rejected as fake.

A lower t-DCF means the system makes fewer costly mistakes.

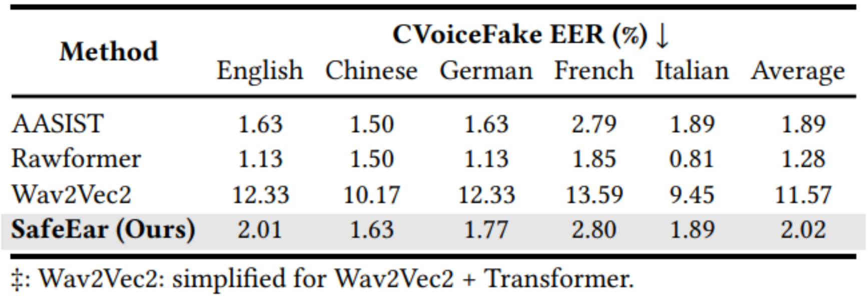
Word/Character Error Rate (WER/CER): they measure the accuracy of content recovery from processed audio by indicating the proportion of words or characters incorrectly transcribed by an ASR system. A higher WER/CER denotes a better privacypreserving ability against content recovery attacks.

Overall Performance of SafeEar compared with baselines on ASVspoof 2019 & 2021 datasets.



Type‡	**Method**	**ASVspoof 2019**		**ASVspoof 2021**	
		EER (%)↓	**t-DCF↓**	**EER (%)↓**	**t-DCF↓**
E2E	AASIST	1.20	0.034	9.15	0.437
	RawNet 2	5.64	0.130	9.50	0.426
	Rawformer	1.05	0.034	8.72	0.397
pipe	LFCC + SE-ResNet34	4.80	0.098	10.39	0.355
	LFCC + LCNN-LSTM	5.06	0.156	9.26	0.345
	LFCC + GMM	8.09	0.212	19.30	0.576
	CQCC + GMM	9.57	0.237	15.62	0.497
	Wav2Vec2 + Transformer	3.82	0.184	6.64	0.330
	SafeEar (Ours)	3.10	0.149	7.22	0.336

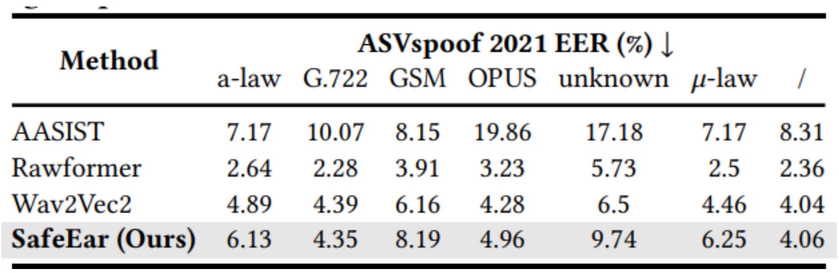
Overall Performance of SafeEar compared with baselines on the CVoiceFake dataset.



Method	CVoiceFake EER (%) ↓					
	English	Chinese	German	French	Italian	Average
AASIST	1.63	1.50	1.63	2.79	1.89	1.89
Rawformer	1.13	1.50	1.13	1.85	0.81	1.28
Wav2Vec2	12.33	10.17	12.33	13.59	9.45	11.57
SafeEar (Ours)	2.01	1.63	1.77	2.80	1.89	2.02

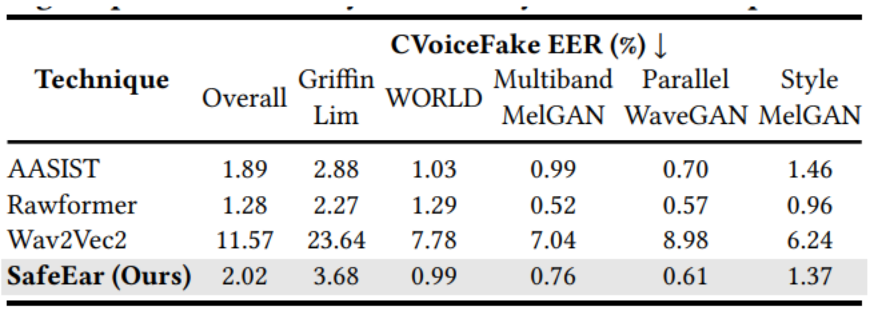
‡: Wav2Vec2: simplified for Wav2Vec2 + Transformer.

Comparison of SafeEar and baselines in detecting deepfakes transmitted via different channels.



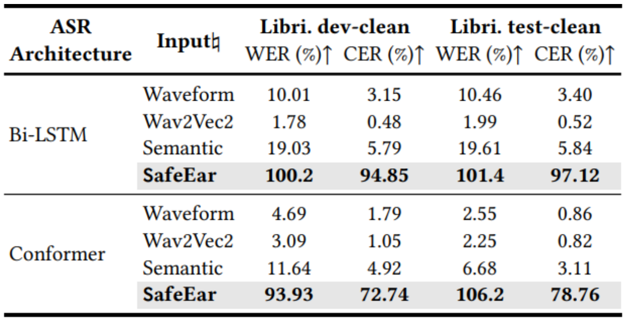
Method	ASVspoof 2021 EER (%) ↓						
	a-law	G.722	GSM	OPUS	unknown	μ-law	/
AASIST	7.17	10.07	8.15	19.86	17.18	7.17	8.31
Rawformer	2.64	2.28	3.91	3.23	5.73	2.5	2.36
Wav2Vec2	4.89	4.39	6.16	4.28	6.5	4.46	4.04
SafeEar (Ours)	6.13	4.35	8.19	4.96	9.74	6.25	4.06

Comparison of SafeEar and baselines in detecting deepfakes created by different synthetic techniques.



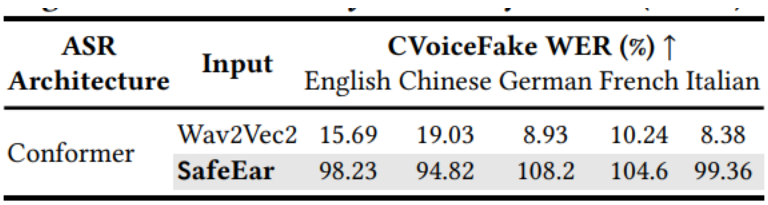
Technique	CVoiceFake EER (%) ↓					
	Overall	Griffin Lim	WORLD	Multiband MelGAN	Parallel WaveGAN	Style MelGAN
AASIST	1.89	2.88	1.03	0.99	0.70	1.46
Rawformer	1.28	2.27	1.29	0.52	0.57	0.96
Wav2Vec2	11.57	23.64	7.78	7.04	8.98	6.24
SafeEar (Ours)	2.02	3.68	0.99	0.76	0.61	1.37

English (Seen language) content protection against naive adversary’s recovery attacks (CRA1).



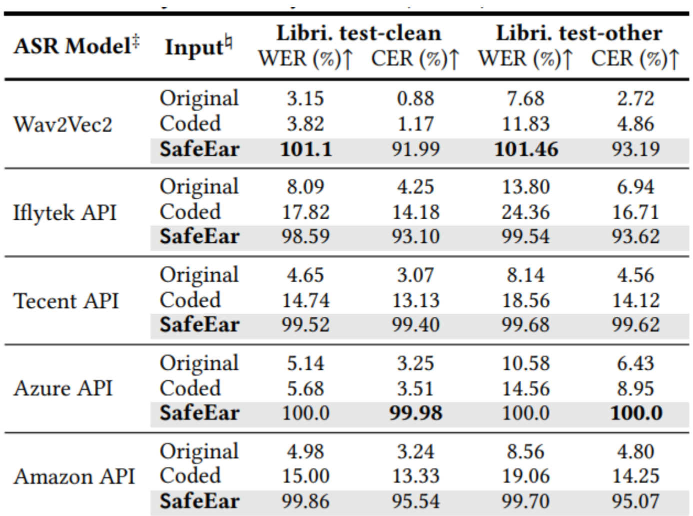
ASR Architecture	Input♮	Libri. dev-clean WER (%)↑	Libri. dev-clean CER (%)↑	Libri. test-clean WER (%)↑	Libri. test-clean CER (%)↑
Bi-LSTM	Waveform	10.01	3.15	10.46	3.40
	Wav2Vec2	1.78	0.48	1.99	0.52
	Semantic	19.03	5.79	19.61	5.84
	SafeEar	**100.2**	**94.85**	**101.4**	**97.12**
Conformer	Waveform	4.69	1.79	2.55	0.86
	Wav2Vec2	3.09	1.05	2.25	0.82
	Semantic	11.64	4.92	6.68	3.11
	SafeEar	**93.93**	**72.74**	**106.2**	**78.76**

Multilingual (Unseen language) content protection against naive adversary’s recovery attacks (CRA1).



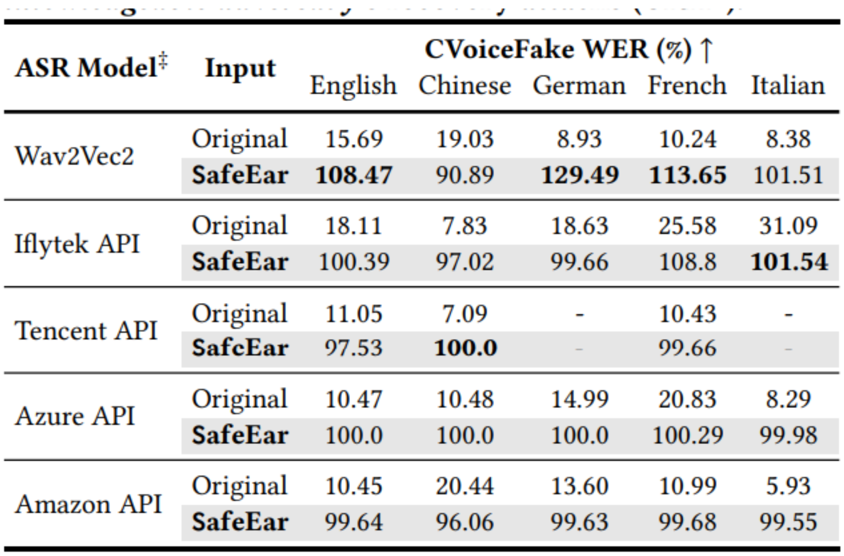
ASR Architecture	Input	CVoiceFake WER (%) ↑ English	Chinese	German	French	Italian
Conformer	Wav2Vec2	15.69	19.03	8.93	10.24	8.38
	SafeEar	98.23	94.82	108.2	104.6	99.36

English content protection against knowledgeable adversary’s recovery attacks (CRA2).



ASR Model‡	Input♮	Libri. test-clean WER (%)↑	Libri. test-clean CER (%)↑	Libri. test-other WER (%)↑	Libri. test-other CER (%)↑
Wav2Vec2	Original	3.15	0.88	7.68	2.72
	Coded	3.82	1.17	11.83	4.86
	SafeEar	**101.1**	91.99	**101.46**	93.19
Iflytek API	Original	8.09	4.25	13.80	6.94
	Coded	17.82	14.18	24.36	16.71
	SafeEar	98.59	93.10	99.54	93.62
Tecent API	Original	4.65	3.07	8.14	4.56
	Coded	14.74	13.13	18.56	14.12
	SafeEar	99.52	99.40	99.68	99.62
Azure API	Original	5.14	3.25	10.58	6.43
	Coded	5.68	3.51	14.56	8.95
	SafeEar	100.0	**99.98**	100.0	**100.0**
Amazon API	Original	4.98	3.24	8.56	4.80
	Coded	15.00	13.33	19.06	14.25
	SafeEar	99.86	95.54	99.70	95.07

Unseen-language content protection against knowledgeable adversary's recovery attacks (CRA2).



ASR Model[‡]	Input	CVoiceFake WER (%) ↑				
		English	Chinese	German	French	Italian
Wav2Vec2	Original	15.69	19.03	8.93	10.24	8.38
	SafeEar	**108.47**	90.89	**129.49**	**113.65**	101.51
Iflytek API	Original	18.11	7.83	18.63	25.58	31.09
	SafeEar	100.39	97.02	99.66	108.8	**101.54**
Tencent API	Original	11.05	7.09	-	10.43	-
	SafcEar	97.53	**100.0**	-	99.66	-
Azure API	Original	10.47	10.48	14.99	20.83	8.29
	SafeEar	100.0	100.0	100.0	100.29	99.98
Amazon API	Original	10.45	20.44	13.60	10.99	5.93
	SafeEar	99.64	96.06	99.63	99.68	99.55

➢ Why you find this approach promising for our specific needs.

- Detecting AI-generated human speech- The performance metrics EER shows that this model is effective in detecting AI-generated human speech effectively.
- Potential for real-time or near real-time detection-

  If a deepfake audio clip is being used in a live conversation, SafeEar can continuously monitor and alert about any deepfake activity in real-time. Voice authentication systems could use SafeEar for verifying users' identities while ensuring that deepfake audio cannot bypass the system.
- Analysis of real conversations-

  SafeEar is designed to be robust in analyzing real-world conversations, even when the audio is degraded or altered due to transmission through real communication systems like phones or messaging apps.

  Emotion and Sentiment Detection- Detecting customer frustration in call centers or monitoring mental health through vocal stress pattern.

➢ Potential limitations or challenges:

Latency: SafeEar needs to be optimized for low-latency processing, ensuring that the feature extraction, tokenization, and detection process happens in real-time.

High-performance hardware like GPUs might be required to speed up both the frontend feature extraction and backend detection models.

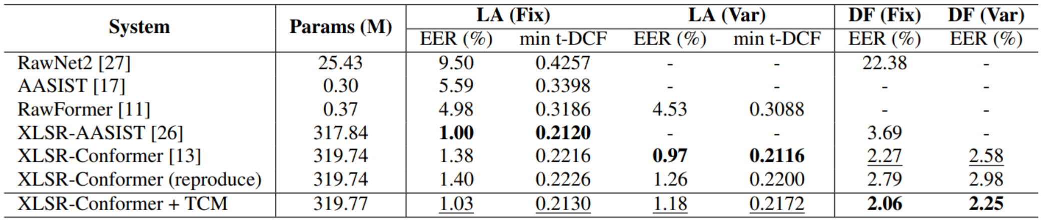
3. Temporal-Channel Modeling in Multi-head Self-Attention for Synthetic Speech Detection

➢ Key Technical Innovation:

- Identified limitations of CNNs and standard MHSA in modeling channel-wise dependencies in SSD tasks.
- Proposed the TCM module to better model temporal-channel interactions.
- Introduced a modified head token mechanism to facilitate richer information flow in attention.
- Achieved improved performance over the previous state-of-the-art on ASVspoof 2021 with minimal parameter overhead.
- Provided empirical analysis to validate the impact of temporal and channel components.

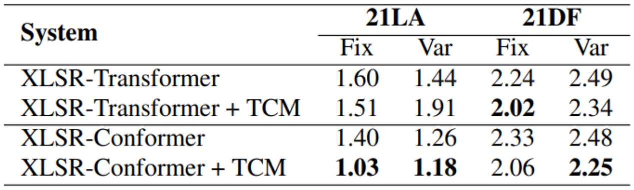
➢ Reported Performance Metrics:

Performance comparison with the state-of-the-art systems on the ASVspoof 2021 eval set with fixed-length (Fix) and variablelength (Var) utterance evaluation



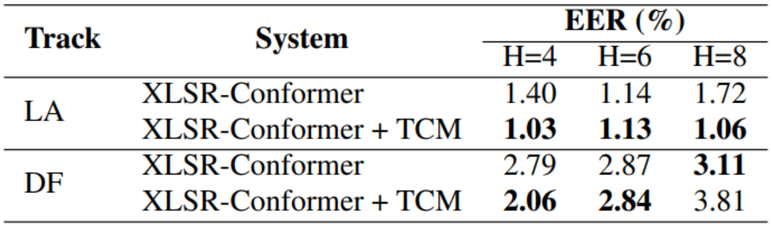
System	Params (M)	LA (Fix)		LA (Var)		DF (Fix)	DF (Var)
		EER (%)	min t-DCF	EER (%)	min t-DCF	EER (%)	EER (%)
RawNet2 [27]	25.43	9.50	0.4257	-	-	22.38	-
AASIST [17]	0.30	5.59	0.3398	-	-	-	-
RawFormer [11]	0.37	4.98	0.3186	4.53	0.3088	-	-
XLSR-AASIST [26]	317.84	**1.00**	**0.2120**	-	-	3.69	-
XLSR-Conformer [13]	319.74	1.38	0.2216	**0.97**	**0.2116**	2.27	2.58
XLSR-Conformer (reproduce)	319.74	1.40	0.2226	1.26	0.2200	2.79	2.98
XLSR-Conformer + TCM	319.77	1.03	0.2130	1.18	0.2172	**2.06**	**2.25**

EER (%) results to evaluate the robustness of TCM for the Transformer and Conformer Block.



System	21LA		21DF	
	Fix	Var	Fix	Var
XLSR-Transformer	1.60	1.44	2.24	2.49
XLSR-Transformer + TCM	1.51	1.91	**2.02**	2.34
XLSR-Conformer	1.40	1.26	2.33	2.48
XLSR-Conformer + TCM	**1.03**	**1.18**	2.06	**2.25**

EER(%) with different numbers of heads on ASV2021 LA & DF eval set



Track	System	EER (%)		
		H=4	H=6	H=8
LA	XLSR-Conformer	1.40	1.14	1.72
	XLSR-Conformer + TCM	**1.03**	**1.13**	**1.06**
DF	XLSR-Conformer	2.79	2.87	**3.11**
	XLSR-Conformer + TCM	**2.06**	**2.84**	3.81

➢ Why you find this approach promising for our specific needs.

- Detecting AI-generated human speech-

Transformer models have demonstrated strong effectiveness in deepfake audio detection due to their ability to capture long-range dependencies and complex contextual relationships in speech. The Temporal-Channel Modeling (TCM) module enhances the multi-head self-attention (MHSA) capability for capturing temporal-channel dependencies. Experimental results confirm that with minimal additional parameters, the TCM module significantly improves the performance of the XLSR-Conformer model on deepfake detection tasks.

- Potential for real-time or near real-time detection-

  The current model is accurate but relatively heavy due to:

  - XLSR (wav2vec 2.0 variant): large, transformer-based SSL model.
  - Conformer architecture with TCM: adds complexity and computation.

  Pytorch's quantization can be used to reduce size and speed up inference with little accuracy loss.

- Analysis of real conversations-

  Emotion and Sentiment Analysis:

  - Temporal: Tracks shifts in emotional tone (e.g., anger to calm) across dialogue turns.
  - Channel: Identifies acoustic correlates of emotions (e.g., high pitch for excitement, spectral flux for stress).

  Deception Detection:

  - Temporal: Flags anomalies in speech rate or hesitation patterns.
  - Channel: Correlates micro-expressions (e.g., vocal tremors) with stress.

  Intent and Dialogue Act Recognition

  - Temporal: Detects conversational structure (e.g., questions vs. statements).
  - Channel: Links lexical and prosodic cues to intent (e.g., rising pitch for questions).

➢ Potential challenges

- Increased computational head
- Hyperparameter Sensitivity-Too fews heads may under-represent channel dependencies and too many heads may affect performance.
- Dependence on Pretrained SSL Models