

SPEECH UNDERSTANDING

K.K.N Shyam Sathvik

Neermita Bhattacharya

B22EE036, B22CS092

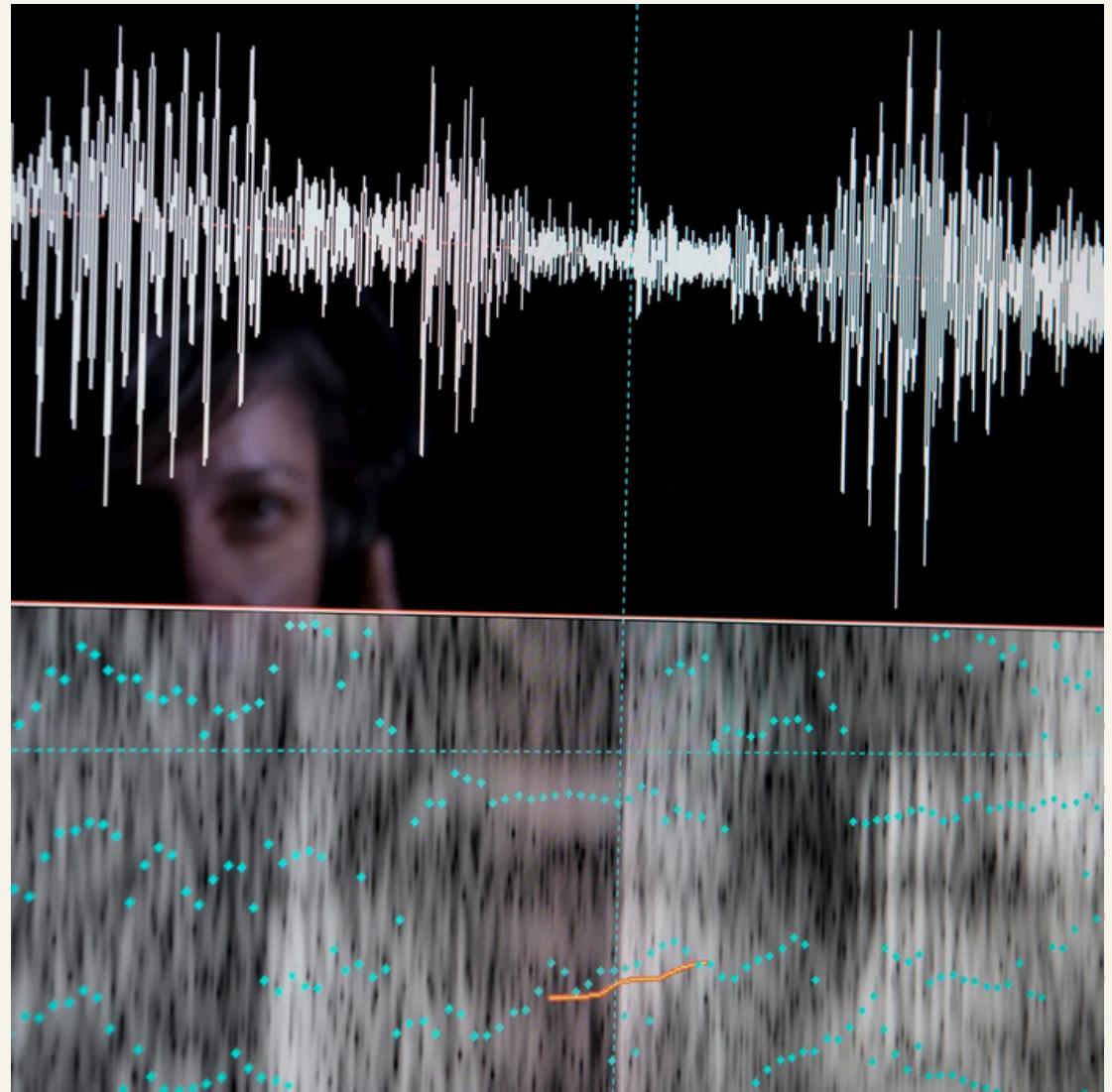
AUDIO FORENSICS

Audio forensics is the scientific analysis of audio recordings to determine their authenticity, improve intelligibility, and identify potential manipulations, often used in legal proceedings to present evidence as part of a criminal investigation or civil case. It involves using specialized techniques to verify if an audio recording is genuine and can be trusted as evidence.

INTRODUCTION

Audio forensics is a large field encompassing various tasks such as authentication, enhancement, speaker identification and recognition, lie detection, etc.

We discuss two aspects-
Lie Detection and Deepfake detection



IMPORTANCE

Lie Detection

Lie detection is crucial to solving crimes, ensuring public safety, and maintaining ethical standards.

Deep Fake Detection

DeepFake detection is crucial to prevent identity theft, fraud, misinformation, and security threats.

LIE DETECTION

What are the methods?

In general, we can detect deception from -

- Shortened length of speech
- Flushed face
- Changing frequencies
- Avoidant eye contact
- Change in pupil's diameter
- Increase in blood pressure
- Increased pause time

These are all a mix of physiological, behavioral and verbal analyses!

LIE DETECTION IN THE COURTROOM

A Presentation by Francis X. Shen, JD, PhD



Spotting Lies in Court



WHY IS VERBAL ANALYSIS SO POPULAR?

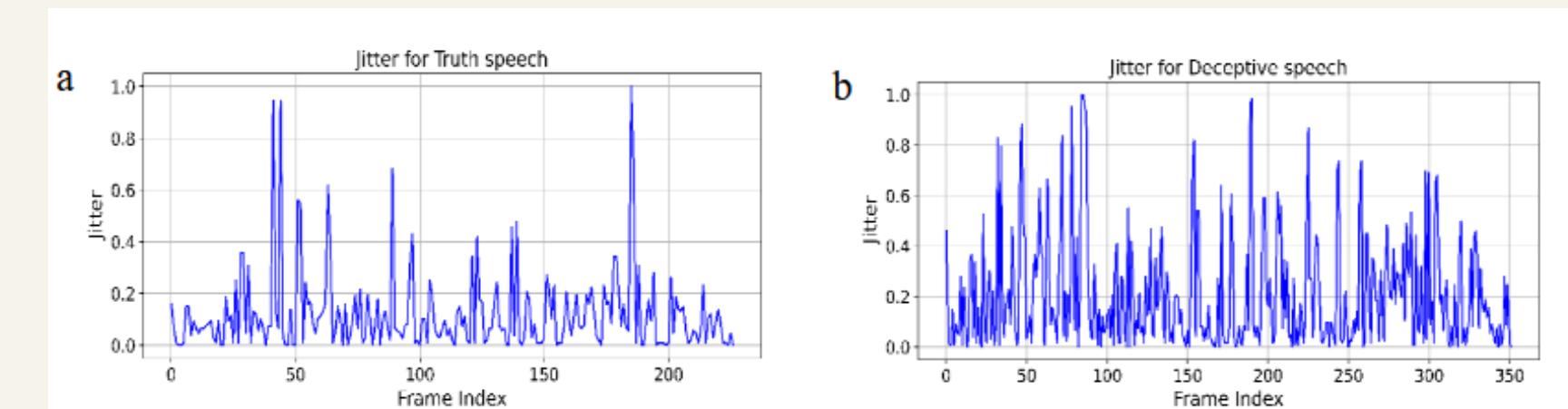
- Physiological analysis techniques are mainly invasive procedures. Polygraphs - utilize blood pressure cuffs, rubber tubes and metal plates on fingertips to measure the respiratory rate, blood pressure, electrodermal activity and heart rate.
- EEGs are also invasive, where electrodes are attached to the scalp to measure brain activities.
- People have found ways to cheat the system by using antiperspirants and sedatives.
- Tone of a person's voice expresses emotions very well. Changes in sound pressure, frequency, speed can be utilized non-invasively.

CURRENT STATE-OF-THE-ART METHODOLOGIES

Computer-based lie detection

- Support Vector Machines (SVM)
- Decision Trees
- Neural Networks
- Random Forests (RFs)
- Deep Neural Networks (DNN)
- Convolutional Neural Networks (CNN)
- LSTMs, RNNs

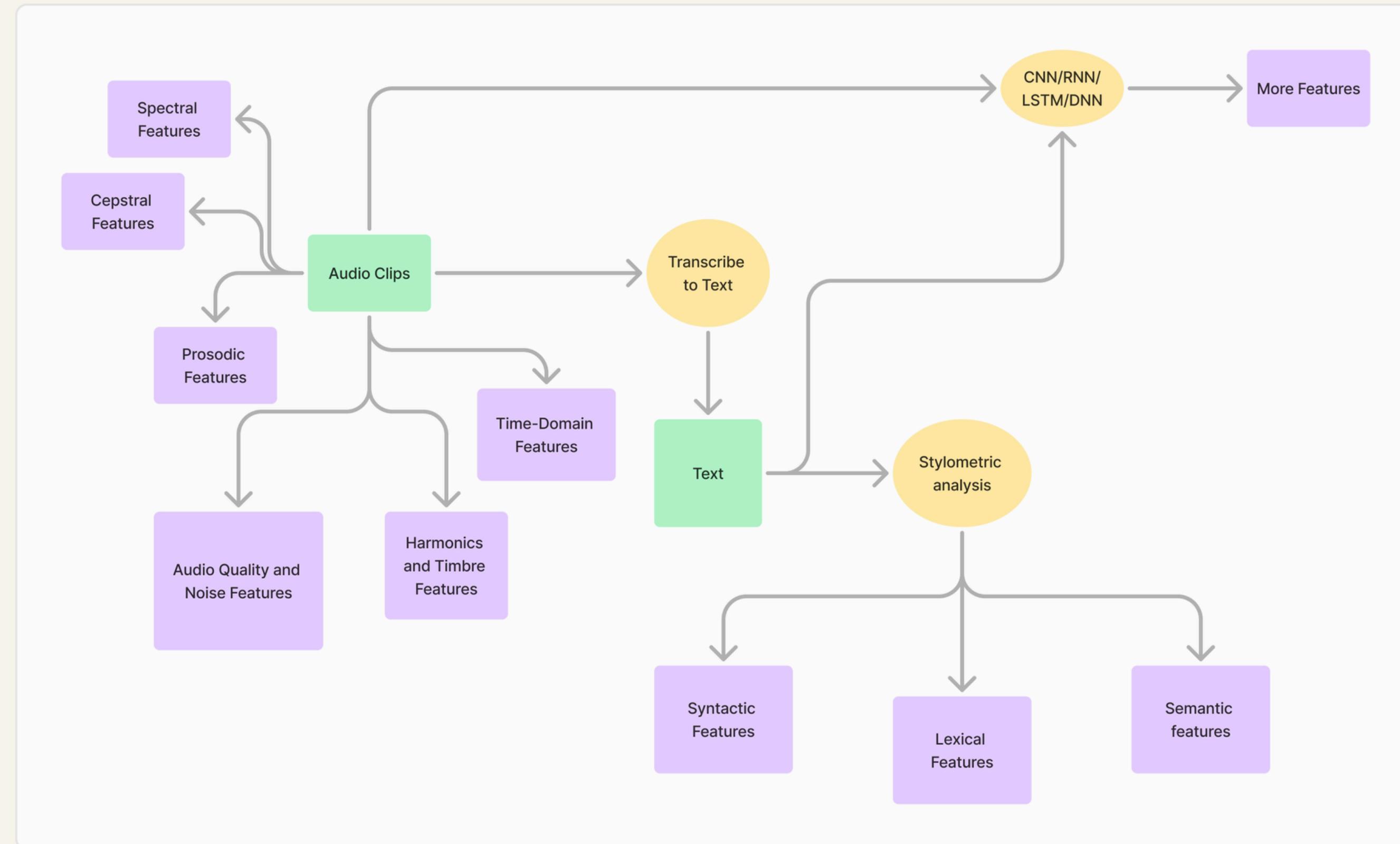
**They analyze features such as:
Speech patterns, Voice stress
Acoustic features in speech (mean
and std dev of pitch, MFCCs, jitter,
harmonic-to-noise ratio)**



METHODS

Study	Method	Dataset	Strengths	Limitation	Metric	Value
https://iopscience.iop.org/article/10.1088/1742-6596/1921/1/012028	MFCC and MFCC mean features, PCA + SVM	<u>RLDD</u>	Non-invasive, discriminative features	Only MFCC and MFCC mean as features	Accuracy	81% and 78% using polynomial and Gaussian kernels respectively.
https://www.scitepress.org/Papers/2020/90387/90387.pdf	MFCC sequence + padding, LSTM	Custom (interviews in Brazilian Portugese)	Used temporal information	Only focused on stress detection	Accuracy	72.5%
https://www.mdpi.com/1424-8220/22/3/1228	Spectrogram + handcrafted features, Hybrid CNN-MLP	<u>RLDD</u> , <u>RODeCAR</u>	Using spectrogram + handcrafted features + Kolmogorov-Smirnov test.	RODeCAR had poor audio quality, computational costs of CNN-MLP	Unweighted Accuracy	63.7% (RLDD), 62.4% (RODeCAR)
https://www.nowpublishers.com/article/Details/SIP-169	openSMILE + BERT + BLSTM + attention at word level and sentence level + late fusion	<u>Daily Deceptive Dialogues corpus of Mandarin</u>	Multimodal features + temporal information	Computational cost, Dataset imbalance in implicature labels	Unweighted Average Recall, Deception & Truth UAR, w-F1 score	80.61% UAR, DUAR 80.34, TUAR 80.87 79.95% w-F1 score

WHAT NEXT?



ON TO DEEPMODEL DETECTION!

DEEPCODEX DETECTION

What does it mean?

Why is it important in the real world?



- Fraud Prevention
- Misinformation Control
- Security
- Trust in Media
- Legal and Compliance
- Identity Theft prevention
- Intellectual Property Protection

Audio deepfakes use AI to synthesize human voices, mimicking speech patterns, tone, and emotions with high accuracy.

Why is deepfake detection important in the real world?

Background

The Washington Post

Tech Help Desk Artificial Intelligence Internet Culture Space Tech Poli

INNOVATIONS

They thought loved ones were calling for help. It was an AI scam.

Scammers are using artificial intelligence to sound more like family members in distress. People are falling for it and losing thousands of dollars.



By

March 5, 2023

CYBERSECURITY • EDITOR'S PICK
Fraudsters Cloned Company Director's Voice In \$35 Million Bank Heist, Police Find

How I Broke Into a Bank Account With an AI-Generated Voice

Banks in the U.S. and Europe tout voice ID as a secure way to log into your account. I proved it's possible to trick such systems with free or cheap AI-generated voices.

An 'easy win' for social media



The video of Putin has circulated for some weeks and was labelled as manipulated media by Twitter

Deepfake presidents used in Russia-Ukraine war



The深fake appeared on the hacked website of Ukrainian TV network Ukraine 24

A Voice Deepfake Was Used To Scam A CEO Out Of \$243,000

Jesse Damiani Contributor

I run the Reality Studies newsletter & Postreality Labs consultancy

Follow



How scammers likely used artificial intelligence to con Newfoundland seniors out of \$200K

Fake audio falsely claims to reveal private Biden comments

IIT Jodhpur | 2025

<https://www.washingtonpost.com/technology/2023/03/05/ai-voice-scam/>

<https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/?sh=1081d7802241>

<https://apnews.com/article/fact-check-biden-audio-banking-fake-746021122607>

DIFFERENT KINDS OF DEEPFAKES

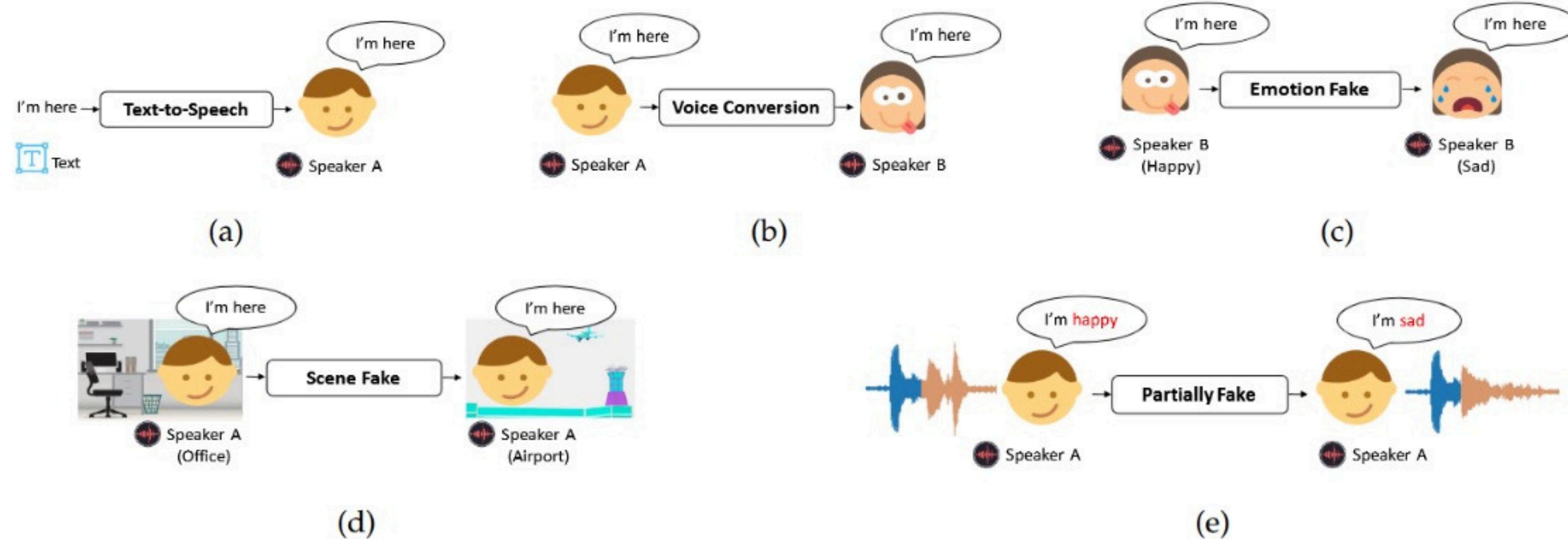


Fig. 2. Five kinds of deepfake audio: (a) text-to-speech, (b) voice conversion, (c) emotion fake, (d) scene fake, (e) partially fake.

SOURCES OF DEEPFAKES

- **Text-to-Speech (TTS):** Synthesizes natural-sounding speech from text using deep neural networks.
- **Voice Conversion (VC):** Clones a person's voice by altering the timbre and prosody while keeping the content intact.
- **Emotion Fake:** Alters the emotional tone of speech while keeping other attributes like speaker identity and content unchanged.
- **Scene Fake:** Manipulates the acoustic scene of the audio (e.g., changing a background from an office to an airport).
- **Partially Fake:** Modifies specific parts of an audio clip, such as changing a few words while keeping the rest of the audio intact.

DATASETS

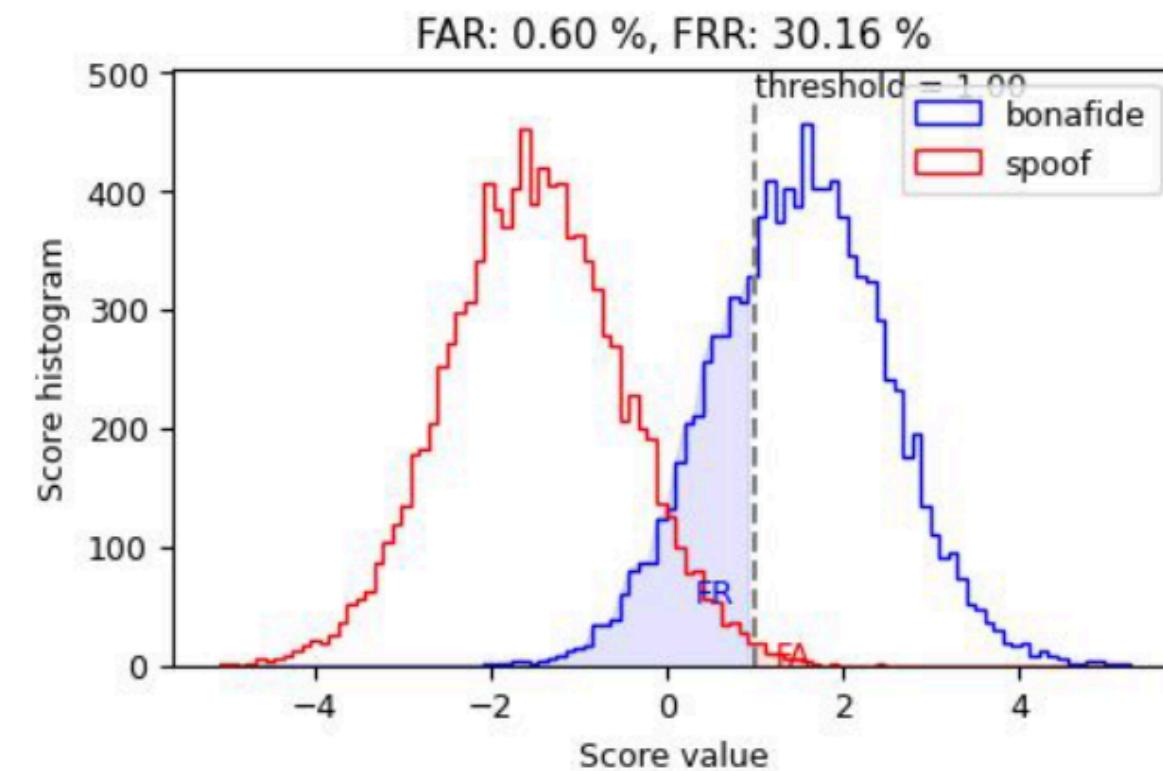
Generated using VC and TTS algorithms



- ASVspoof2019
- ASVspoof2021 : DF evaluation set is generated by more than hundreds of different TTS and VC spoofing attack algorithms
- WaveFake: 6 different GAN-based TTS algorithms across two languages, English and Japanese
- Chinese Fake Audio Detection (CFAD): Mandarin
- Multi-Language Audio Anti-spoofing (MLAAD): 54 models & 23 languages

PERFORMANCE METRICS

- False rejection rate (FRR)
- False acceptance rate (FAR)



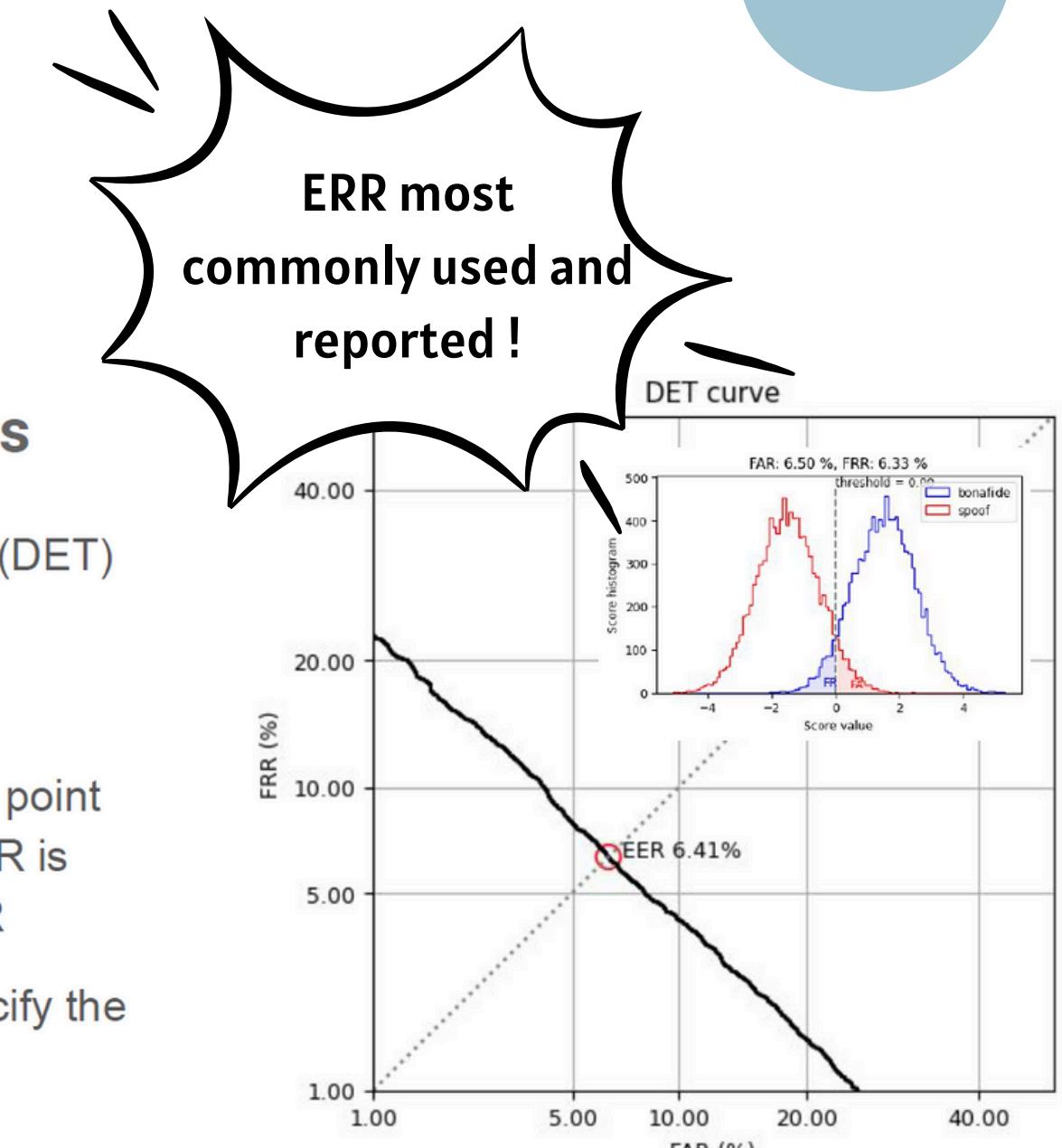
Evaluation metrics

Detection error trade-off (DET) curve (Martin 1997)

Equal error rate (EER)

- Choose the operating point (threshold) so that FAR is **roughly** equal to FRR
- We don't need to specify the threshold manually

EER measures discrimination, but not calibration. See more in [\(Van Leeuwen 2007\)](#)



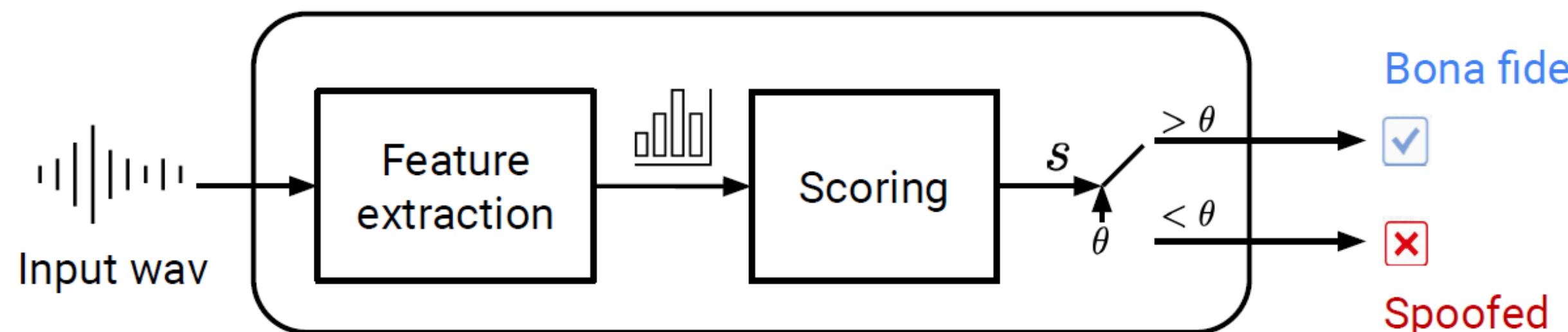
TASK DEFINITION

Feature based Classification

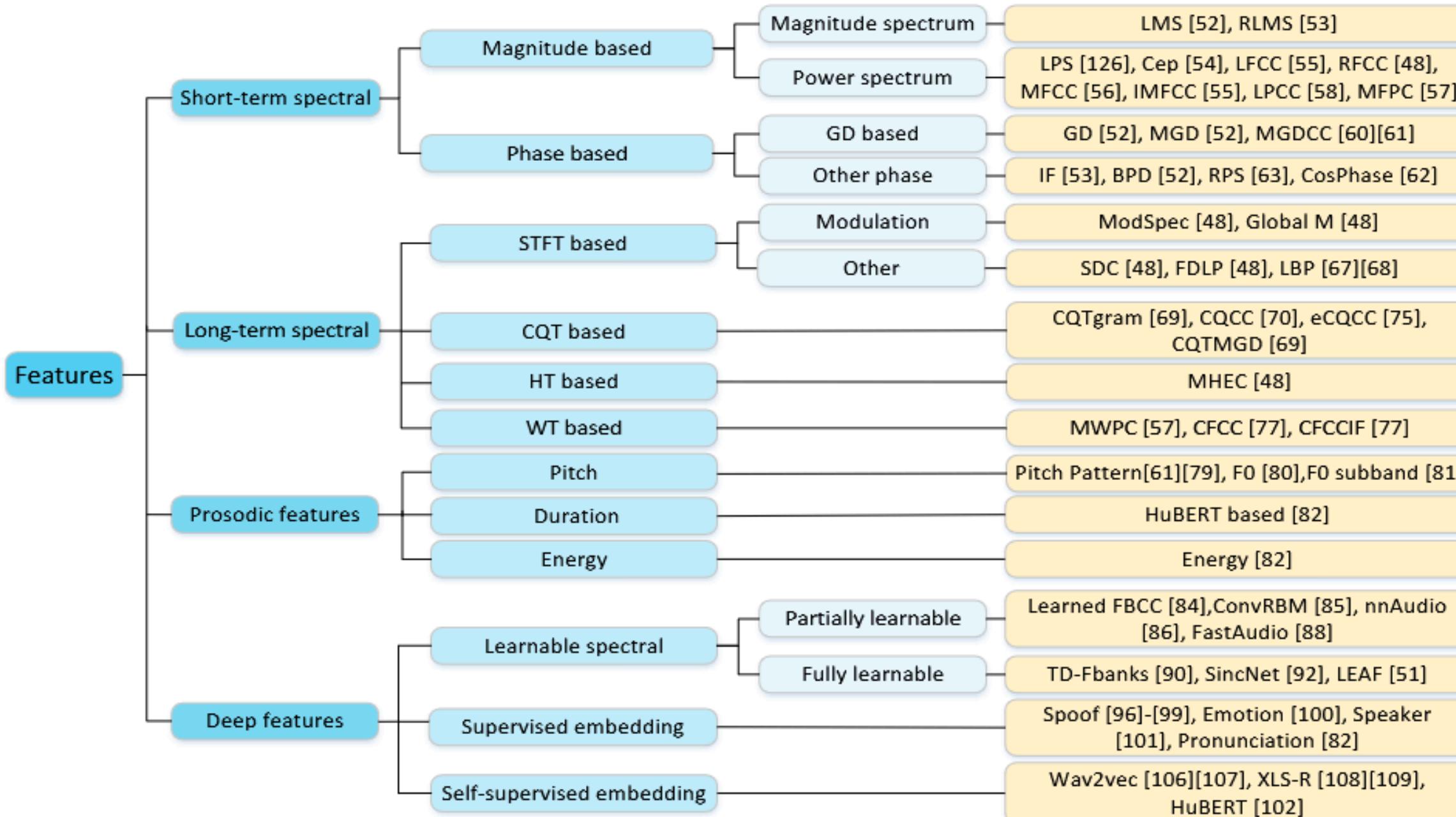
A binary classification task

- Feature extraction: front end
- Scoring: back end
- Decision

} $\text{Input wav} \rightarrow s \in \mathbb{R}$ How likely the input waveform is bona fide



DIFFERENT TYPES OF FEATURES



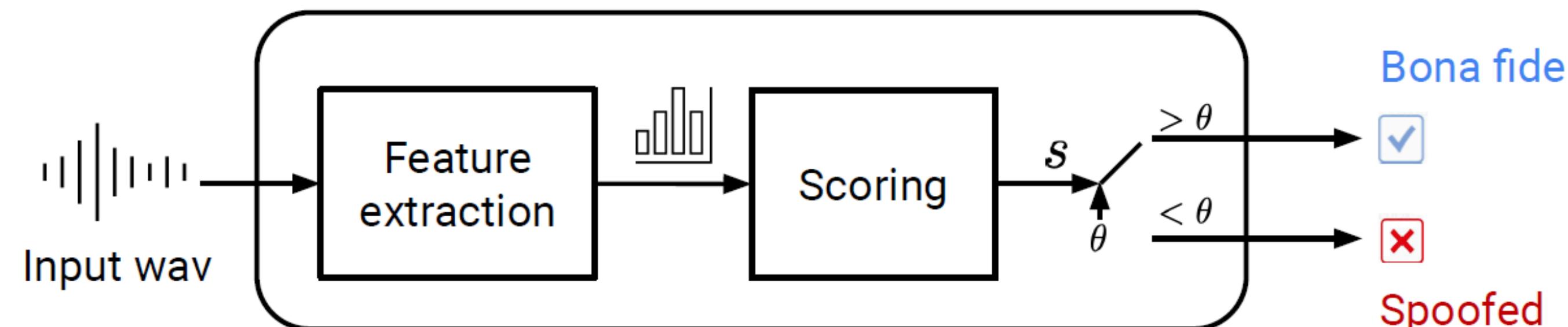
TASK DEFINITION

Feature based Binary Classification

A binary classification task

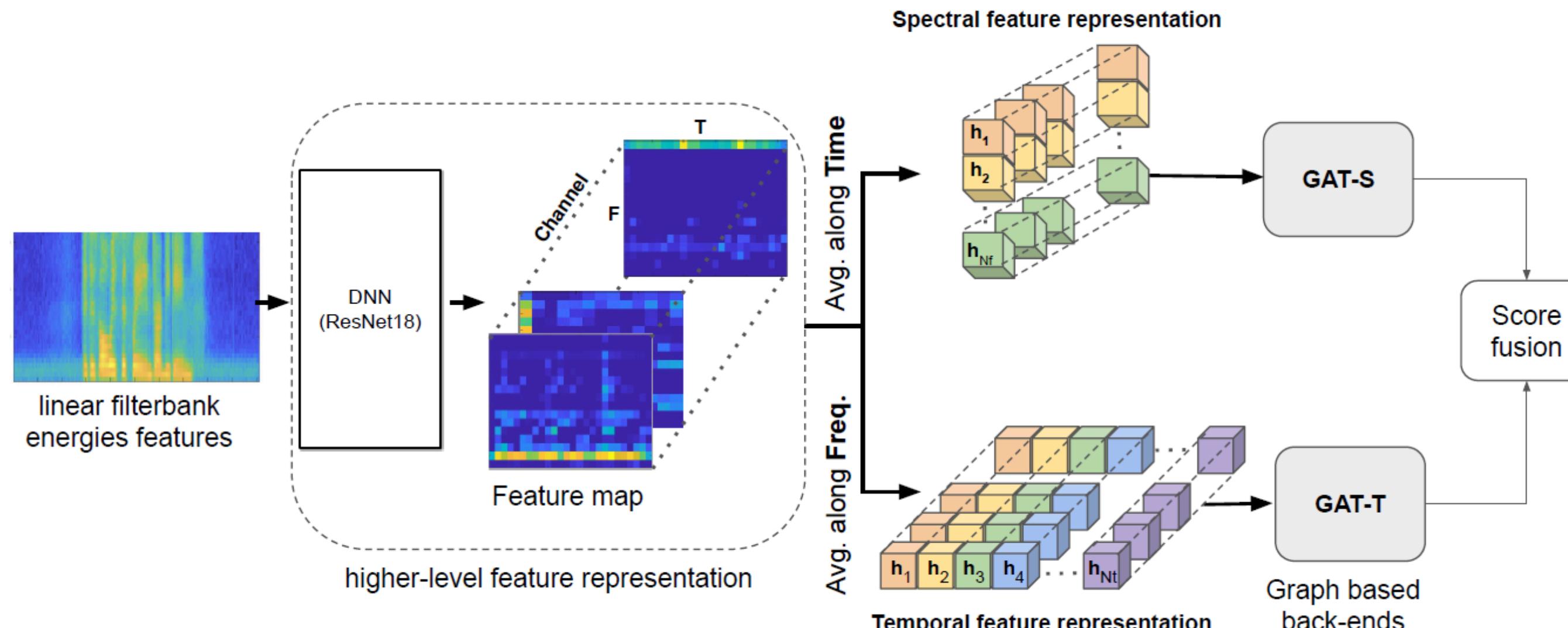
- Feature extraction: front end
- Scoring: back end
- Decision

} $\text{Input waveform} \longrightarrow s \in \mathbb{R}$ How likely the input waveform is bona fide



TASK DEFINITION 02

Graph attention network for anti-spoofing

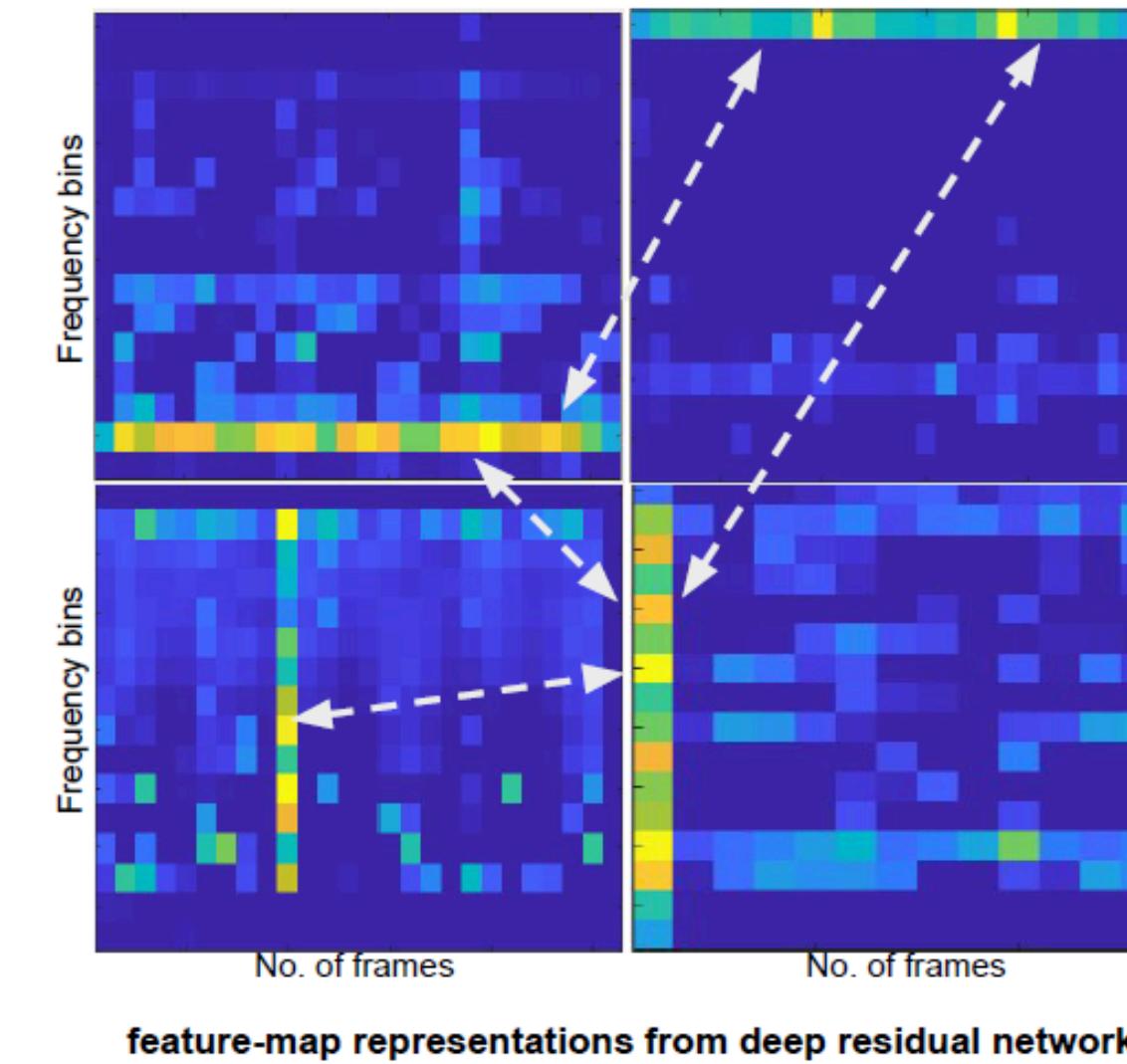


TASK DEFINITION 02

Graph attention network for anti-spoofing

GNN to model relationship between spectral and temporal representation

- Spoofing artefacts lie in **specific spectral subbands or temporal frames** [Yang 2019, Tak 2020, Chettri 2020, Tak 2021]
- modelling the relationship between the evidence spanning different **sub-bands and temporal intervals**
- to leverage the potential of **GNNs for modeling relationships** in spectral and temporal domain [Velickovic 2018]

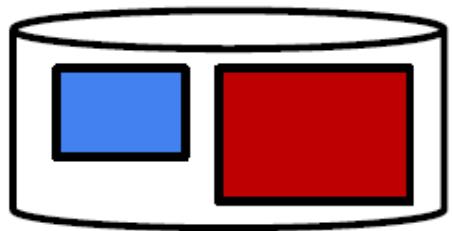


PROBLEMS

Why is anti-spoofing difficult

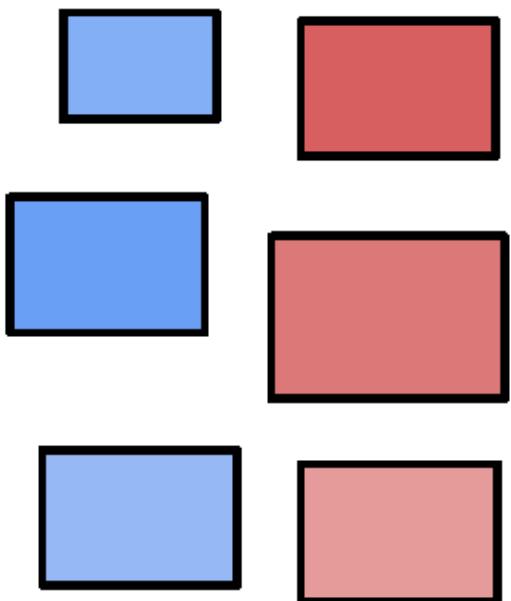
Unseen variabilities real world test data

Training data



We need more powerful tools to extract features and compute score

Real world



- Speakers
- Unseen attacks
- Channels
- Languages
- ...

Problem

To collect huge labelled data is time-consuming

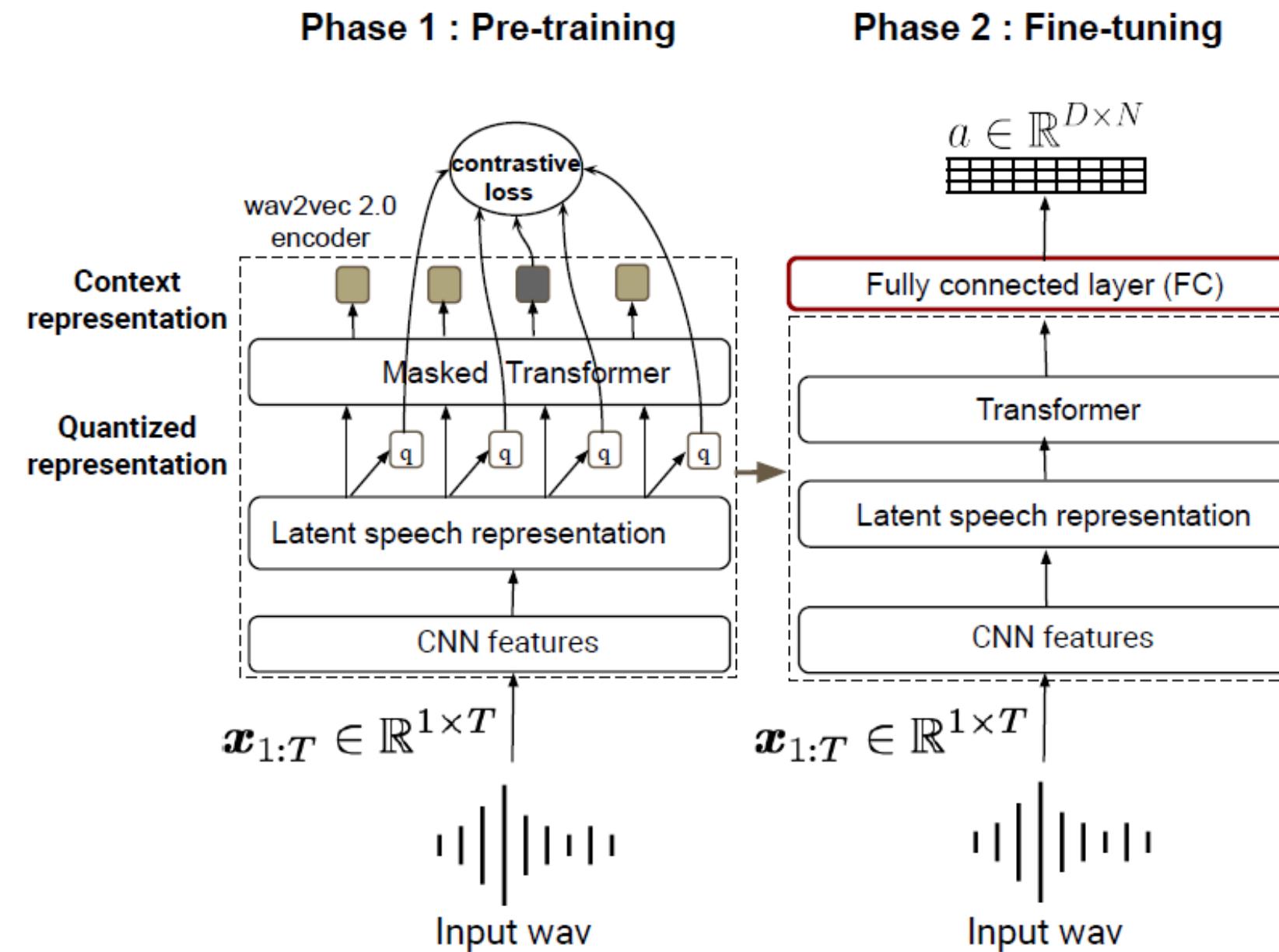
TASK DEFINITION 03

Self-supervised learning for speech anti-spoofing

SSL framework

Two stages in the framework:

1. Use SSL to pre-train an upstream model for general purpose task.
2. Downstream task uses the learned representation from a pre-trained model (fixed) or fine-tune the pre-trained model for specific downstream task.



CURRENT STATE-OF-THE-ART METHODOLOGIES

Publication	Data augmentation	Feature	Classifier	Loss funcion	# Params	ASVspoof			Access-ability
						19-LA	21-LA	21-DF	
[291]	INTERSPEECH'21	w/o	Mel-Spec on 0-4kHz	SE-ResNet-18	AM-Softmax	1.1M	1.14	-	No
[204]	INTERSPEECH'21	channel masking	RawNet2*	GAT	CE	440K	1.06	6.92	- Yes ¹
[71]	INTERSPEECH'21	channel masking	SincNet	Raw PC-DARTS	MSE	24.4M	1.77	6.43	- Yes ²
[87]	SPL'21	mix-up	E2E: CNN→ResNet→MLP		CE	350M	1.64	-	- Yes ³
[65]	ICASSP'22	w/o	FastAudio	ECAPA-TDNN	CE	Unknown	1.54	-	- Yes ⁴
[119]	DSP'22	w/o	L-VQT	DenseNet	CE	338K	2.19	-	No
[212]	ICPR'22	w/o	RawNet2+(CQT→ECAPA-TDNN)	CNN→MLP	CE	7.19M	1.11	-	- Yes ⁵
[114]	INTERSPEECH'22	w/o	wav2vec2.0-XLSR	MLP	CE	317M	0.31	-	No
[59]	INTERSPEECH'22	w/o	wav2vec2.0-960	MLP	CE	Unknown	0.40	-	No
[39]	INTERSPEECH'22	frequency masking	CQT-Spec	LCNN	CE	135K	1.35	-	No
[227]	ODYSSEY'22	w/o	wav2vec2.0-XLSR	Bi-LSTM →MLP	CE	317M	1.28	6.53	4.75 No
[209]	ODYSSEY'22	RawBoost	wav2vec2.0-XLSR	AASIST	CE	Unknown	-	0.82	2.85 Yes ⁶
[144]	DDAM'22	RawBoost	ImageNet + Jitter + Shimmer	MLP	AM-Softmax	Unknown	0.87	10.06	27.08 No
[217]	DDAM'22	w/o	wav2vec2.0-Large	DARTS	Unknown	Unknown	1.08	-	7.89 No
[92]	ICASSP'22	w/o	RawNet2	GAT	CE	297K	0.83	5.59	- Yes ⁷
[117]	SPL'22	w/o	LFCC	OCT	Focal loss	250K	1.06	-	No
[132]	APSIPA'22	adding noise, RIRs	wav2vec2.0	LCNN	CE	Unknown	0.24	-	No
[100]	MAD'23	w/o	Mel-spec + Spec-Env + Spec-Contrast	Transformer →CNN	CE	603K	0.95	-	- No
[218]	INTERSPEECH'23	w/o	Duration + pronunciation + wav2vec2.0-XLSR	LCNN →Bi-LSTM →MLP	CE	Unknown	1.58	-	- No
[151]	SPL'23	w/o	(LFCC →ResNet) + (CQT-Spec →ResNet)	GRL →MLP	CE	Unknown	0.80	-	- Yes ⁸
[139]	ICASSP'23	w/o	RawNet2	Rawformer	CE	370K	0.59	4.98	4.53 Yes ⁹
[158]	ICASSP'23	FIR filter	wav2vec2.0-XLSR	MLP	OC-Softmax	300M	-	3.54	6.18 No
[32]	ICASSP'23	time & frequency masking	LFB-Spec	GCN	CE	Unknown	0.58	-	- No
[276]	ALGORITHM'23	RawRoost	wav2vec 2.0	Transformer	CE	Unknown	-	1.18	4.72 No
[101]	ICASSP'24	FIR filter, codec, noises, shift	SDC + Bi-LSTM	Auto-encoder →SE-ResNeXT	CE	Unknown	0.22	3.50	3.41 No

ROOM FOR IMPROVEMENT

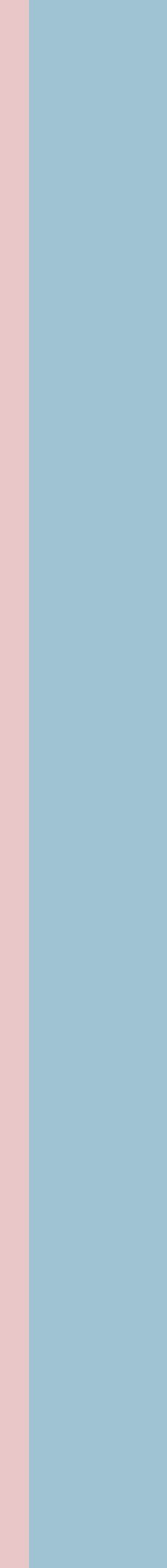
Best CM performance on unseen test set

Q : What is the Best CM performance on in the wild dataset^[Muller 2022] ?

In the wild dataset is more challenging

- more diverse acoustic characteristics
 - fake audios are generated using publicly available sources such as social networks and popular video sharing platforms
 - 50 English-speaking celebrities and politicians
- 10.46% EER on in the wild dataset

The best CM system is not well generalised to an unseen test set.



IIT Jodhpur | 2025

THANK YOU

