# VoiceRadar: Voice Deepfake Detection using Micro-Frequency and Compositional Analysis

Kavita Kumari, Maryam Abbasihafshejani, Alessandro Pegoraro, Phillip Rieger, Kamyar Arshi, Murtuza Jadliwala, Ahmad-Reza Sadeghi

**NDSS 2025**



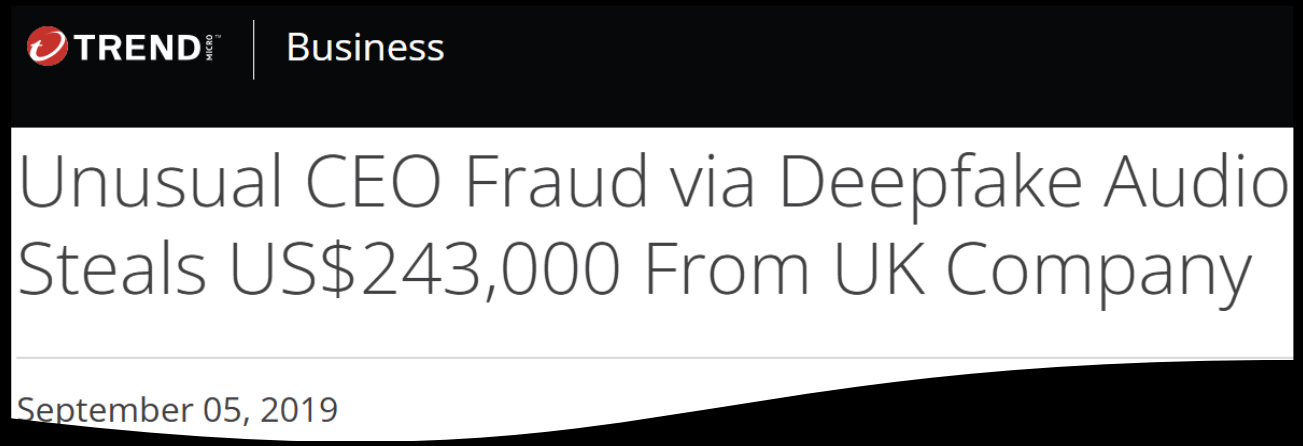
TECHNISCHE UNIVERSITÄT DARMSTADT

UTSA®

The University of Texas at San Antonio™

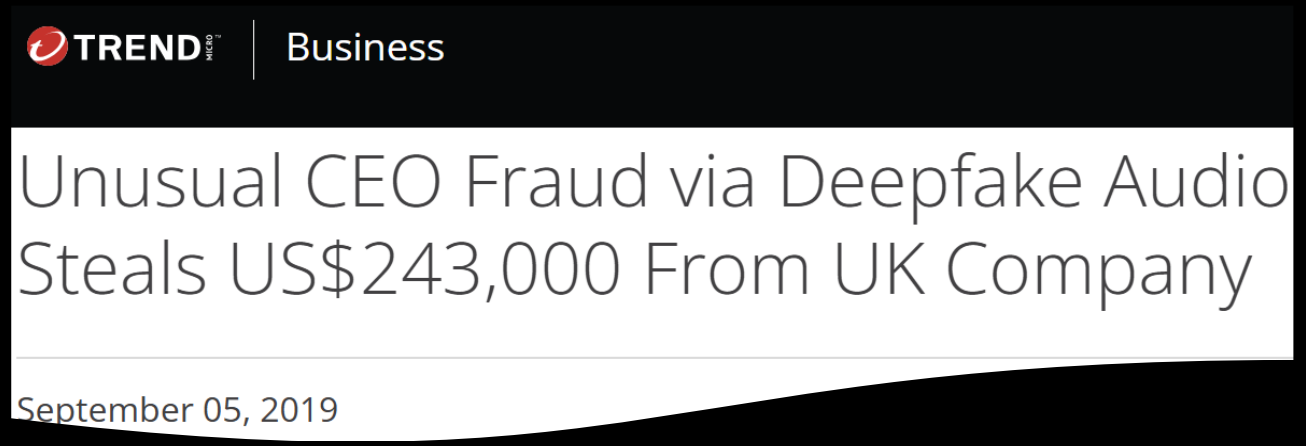
# Audio Deepfakes

## Fraud



# Audio Deepfakes

## Fraud




## Disinformation

# Audio Deepfakes

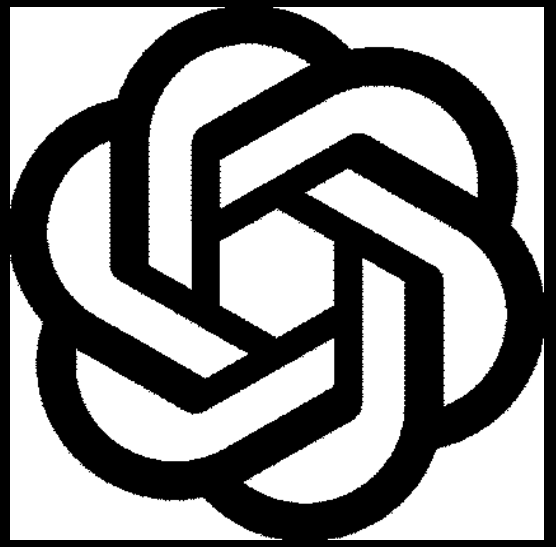
## Defamation

U.S. NEWS

**Deepfake of principal's voice is the latest case of AI being used for harm**

# The Audio Generation Arms Race

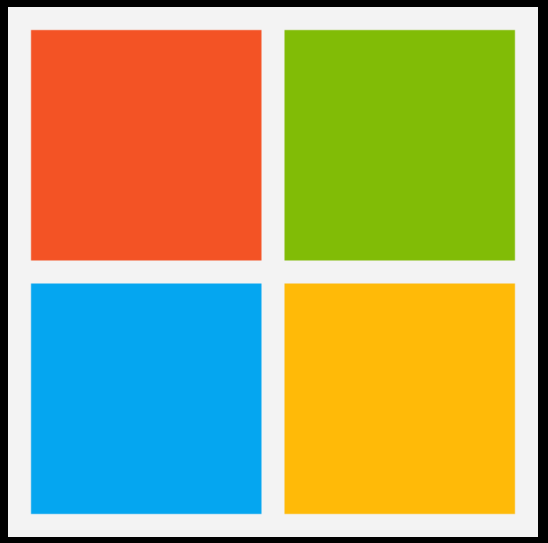
## OpenAI



### Voice Engine

(Voice Cloning with 15 sec of samples)
[Open AI Blog]

## Microsoft



### VALLE-X

(Text-to-speech voice generation using Language Modelling)
[Meng et al. arXiv 2024]

## Facebook/Meta

### Singing VC

(Singing Voice Conversion)
[Nachmani et al. Intespeech 2019]

## Google

### AudioLM

(Audio Generation using Language Modelling )
[Borsos et al. arXiv 2022]

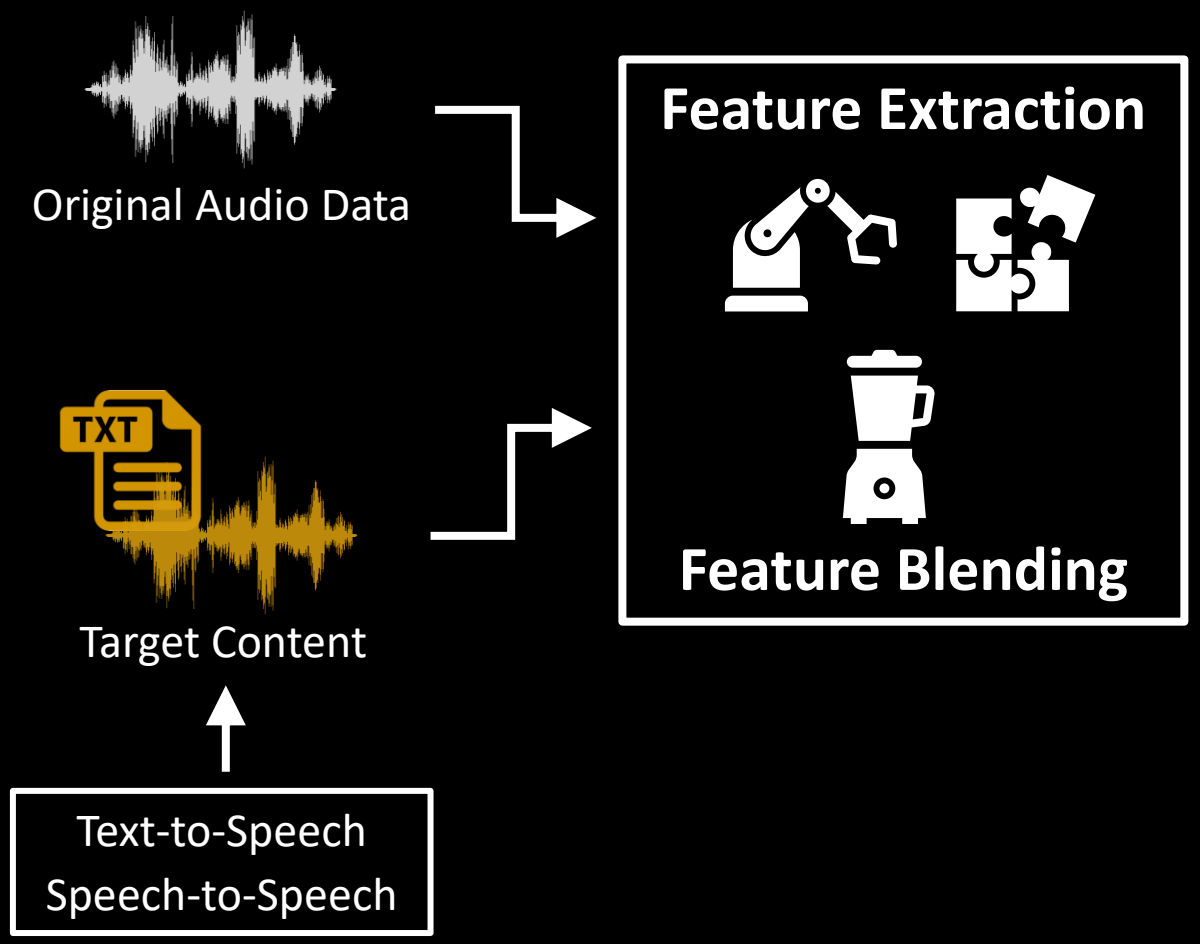
# General Principle of Deepfake

Original Audio Data

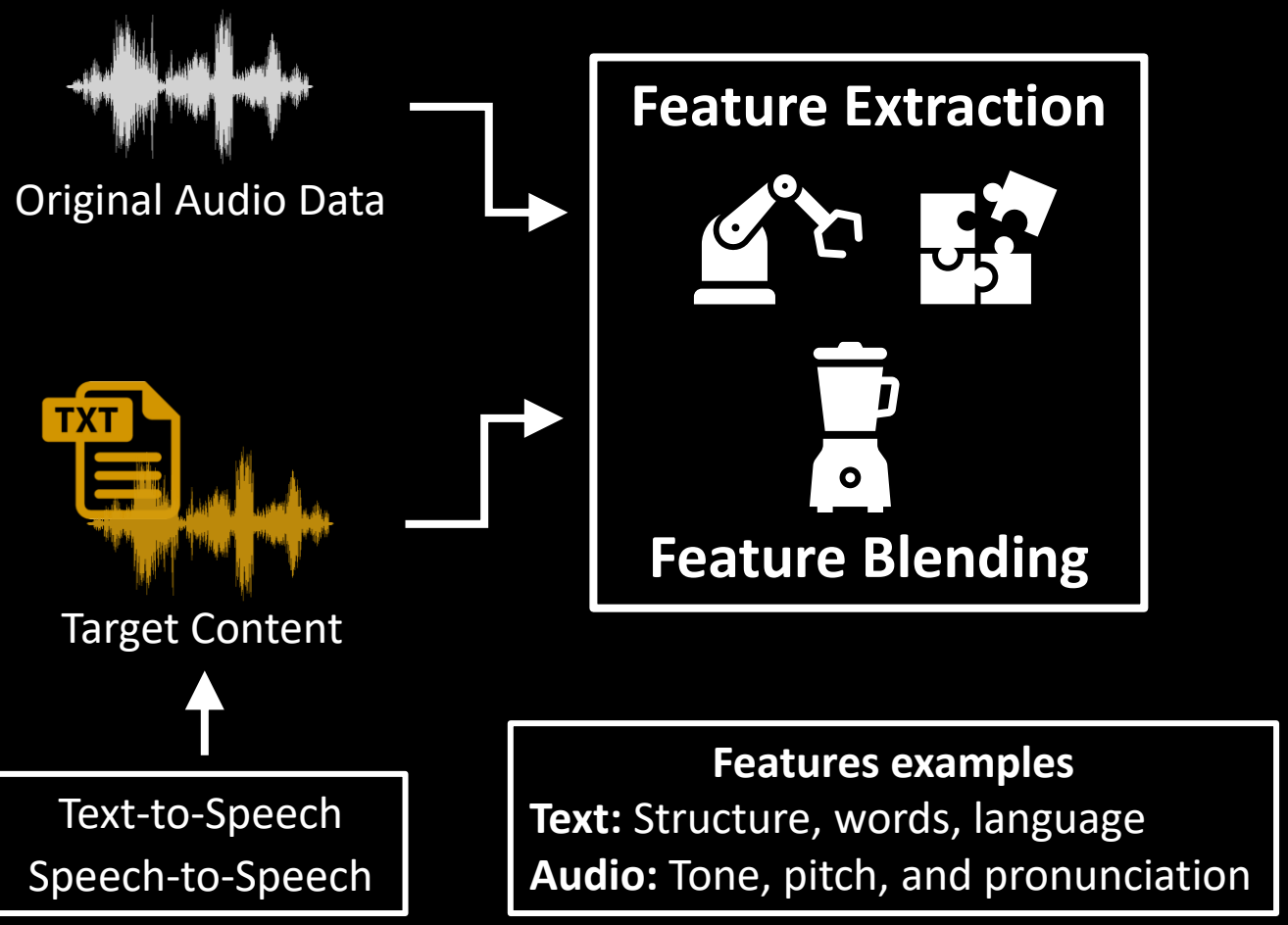

Target Content

Text-to-Speech
Speech-to-Speech

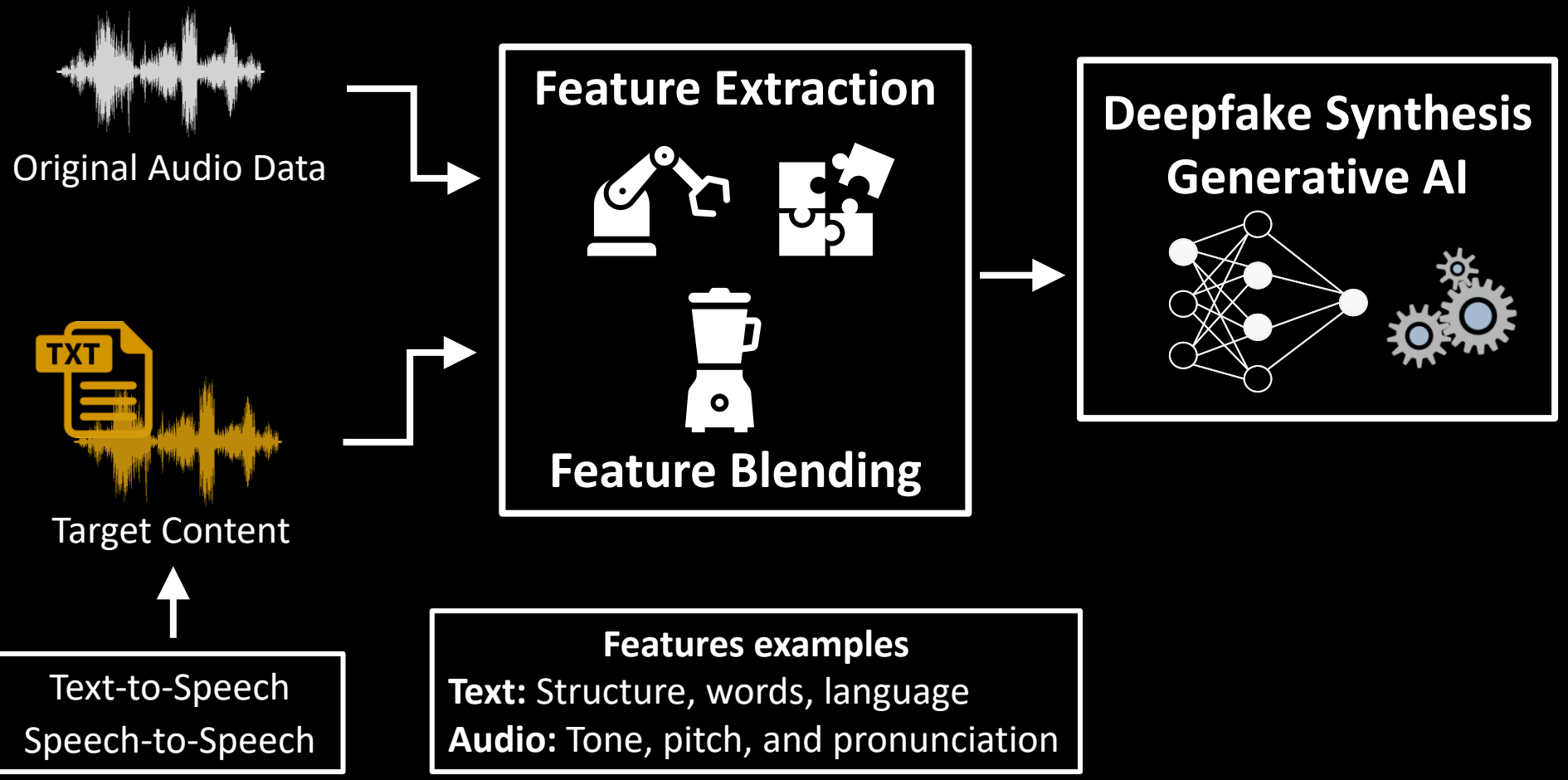
# General Principle of Deepfake



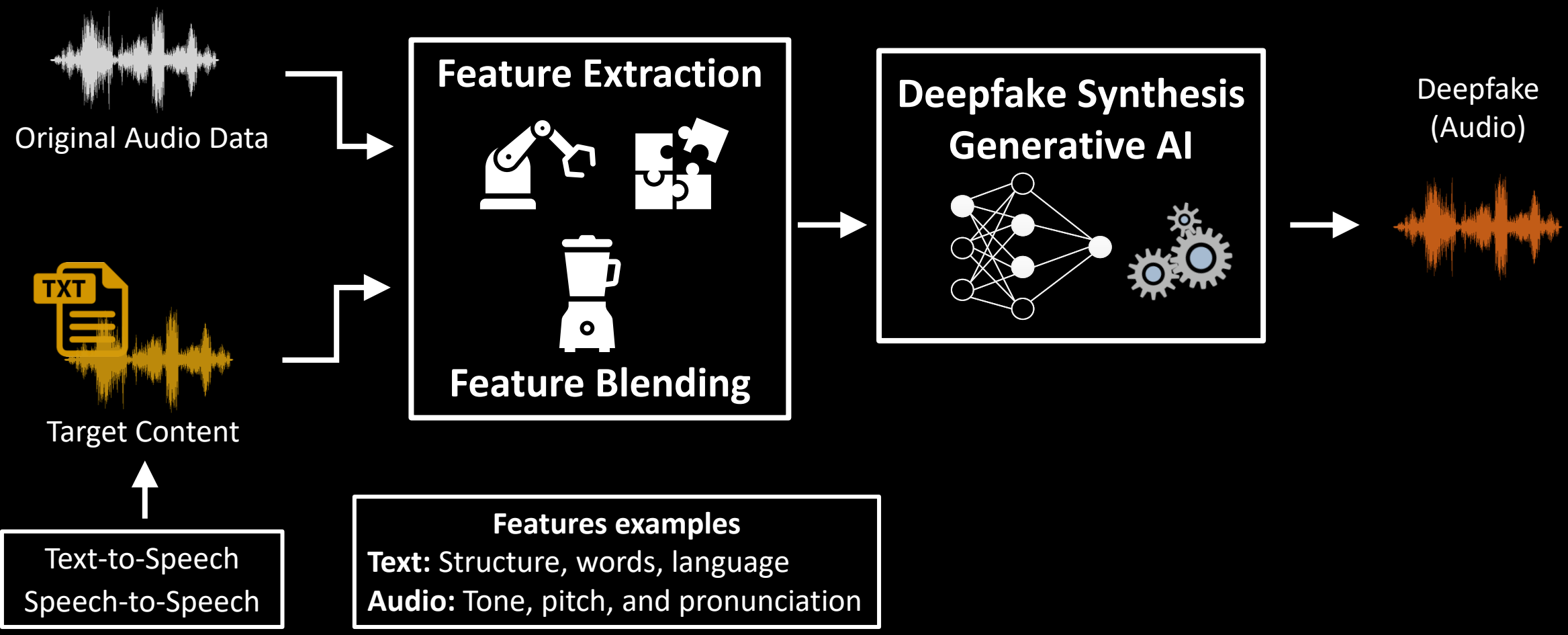
# General Principle of Deepfake



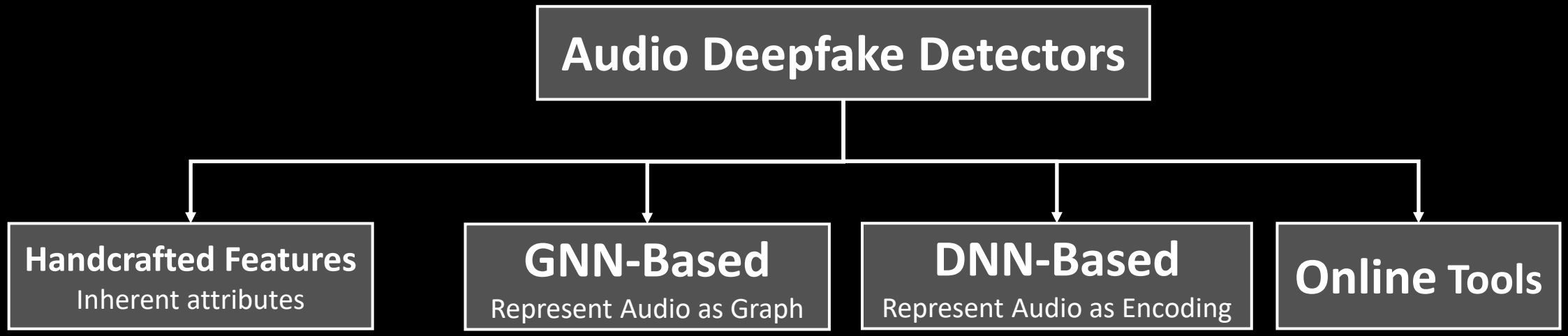
# General Principle of Deepfake



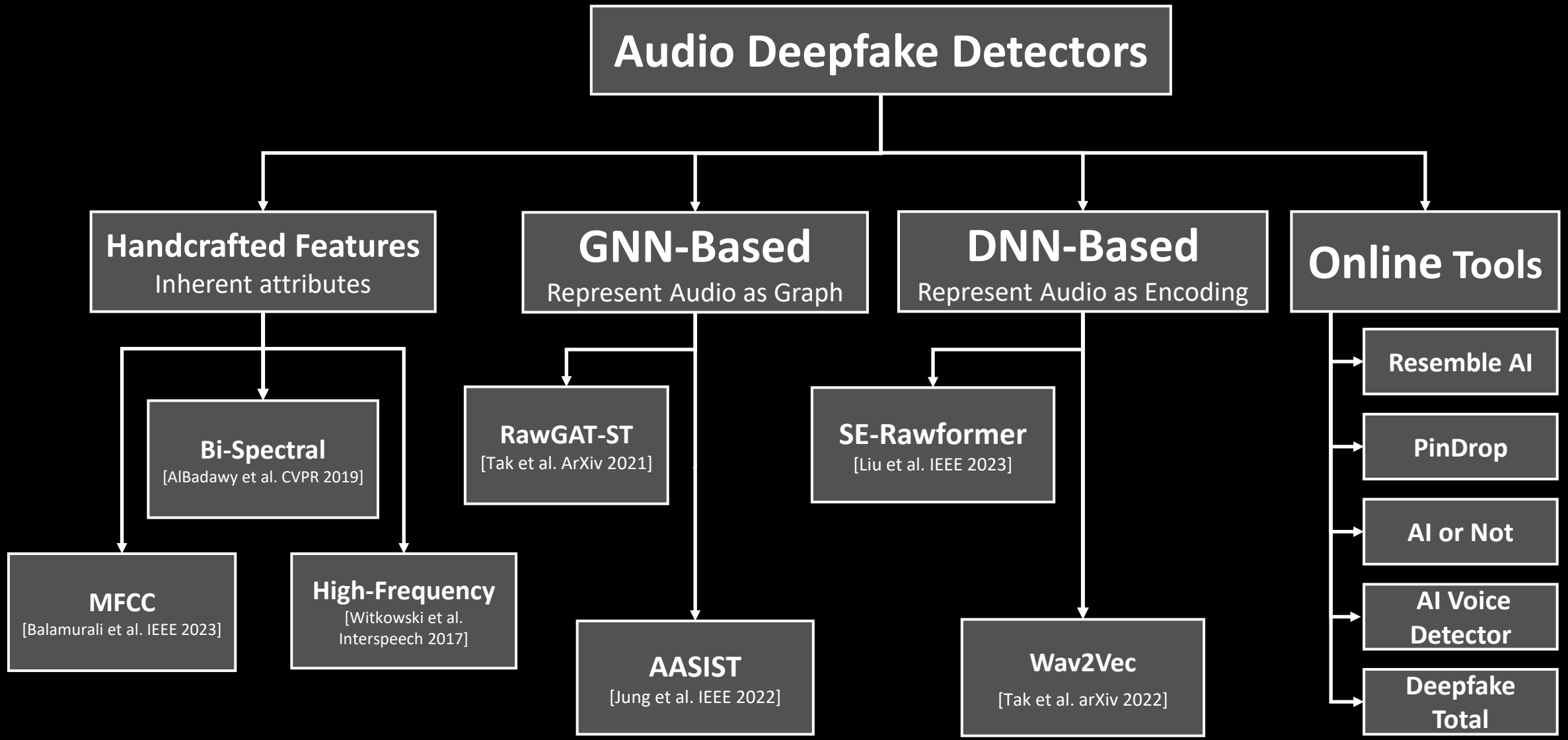
# General Principle of Deepfake



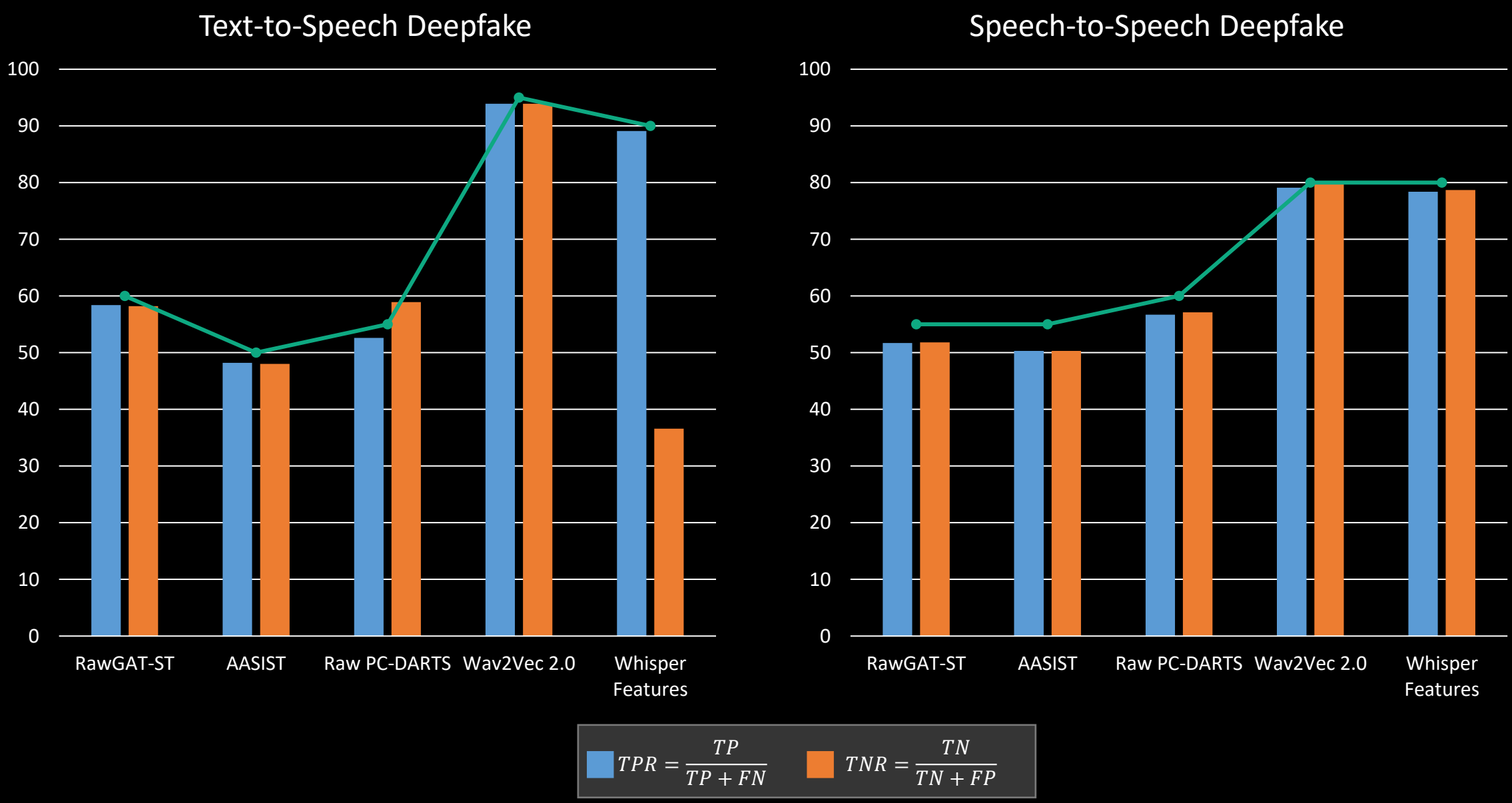
# Categorization of Audio Deepfake Detectors



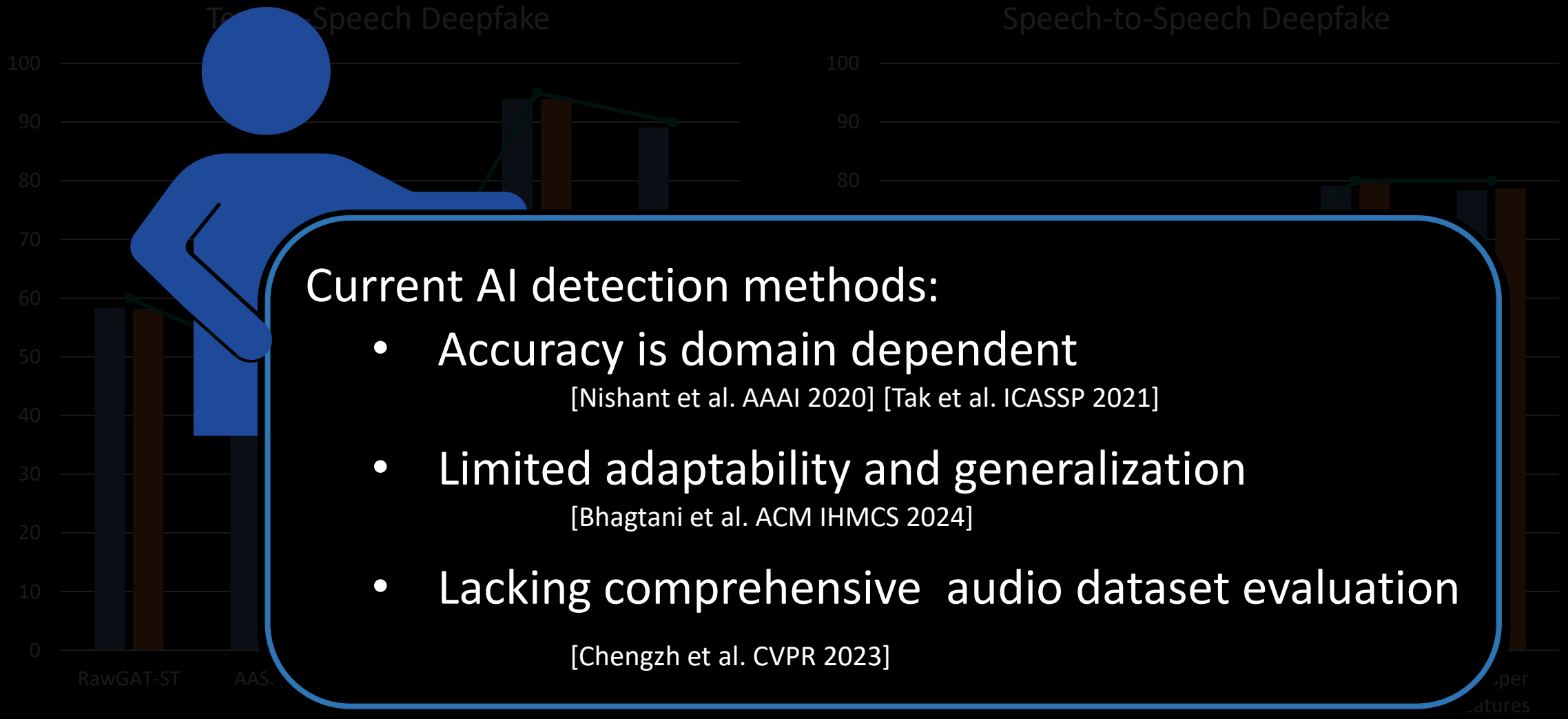
# Categorization of Audio Deepfake Detectors



# Evaluation of Deepfake Detectors



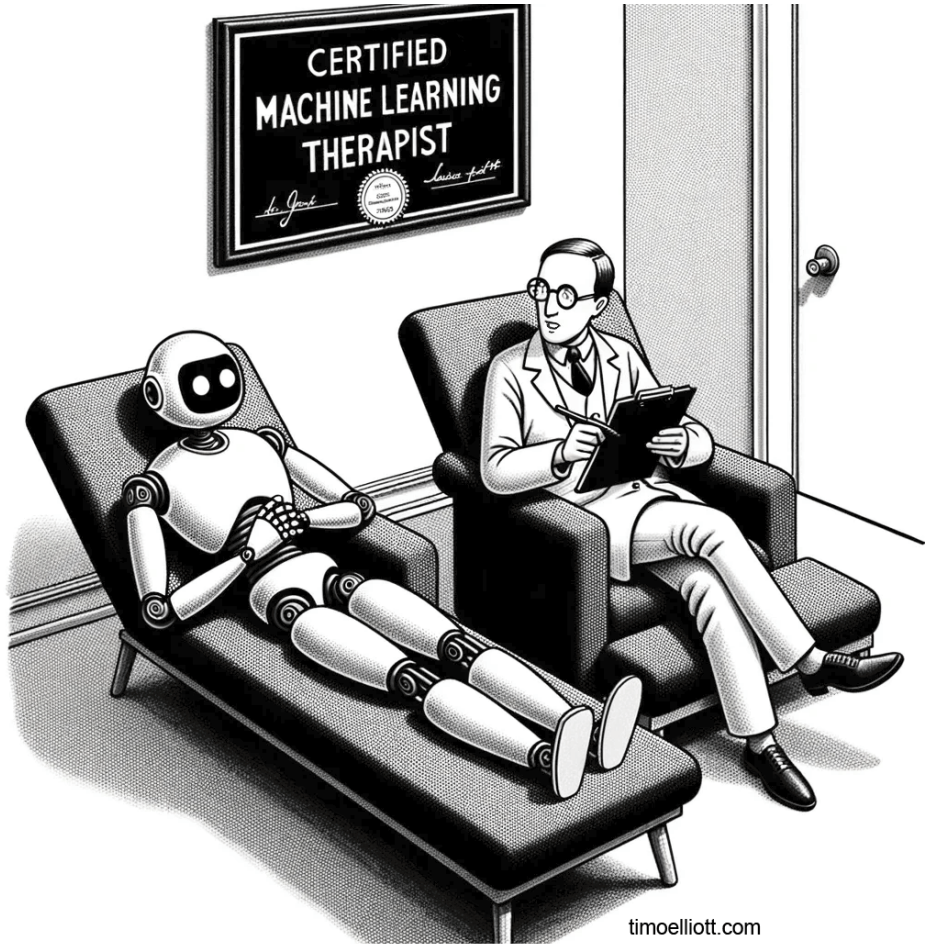
# Evaluation of Deepfake Detectors




Current AI detection methods:

- Accuracy is domain dependent
  [Nishant et al. AAAI 2020] [Tak et al. ICASSP 2021]
- Limited adaptability and generalization
  [Bhagtani et al. ACM IHMCS 2024]
- Lacking comprehensive audio dataset evaluation
  [Chengzh et al. CVPR 2023]

# Human vs. Machine




*"So, when did you first realize humans might have feelings?"*

# Intuition and Hypothesis

## Human interactions

- Affected by personal and environmental factors dependent on individuals involved

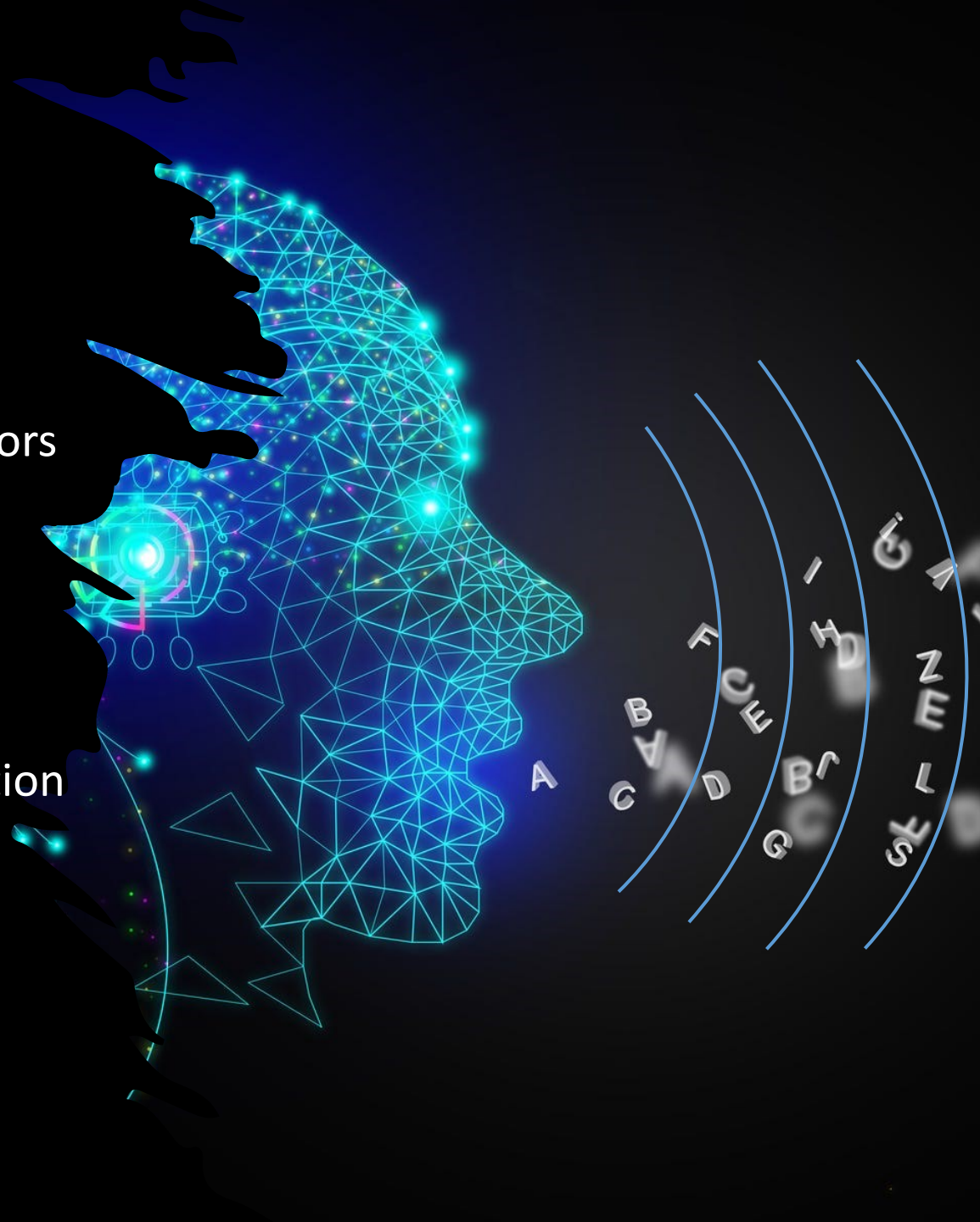
# Intuition and Hypothesis

## Human interactions

- Affected by personal and environmental factors dependent on individuals involved

## Machines interactions

- Follow predefined rules and pattern recognition
- Less dynamic than human interactions




VoiceRadar

# VoiceRadar: Motivation

Approximates the physical models

# VoiceRadar: Motivation

Approximates the physical models

To simulate the audio wave propagation

# VoiceRadar: Motivation

Approximates the physical models

To simulate the audio wave propagation

Extracts the subtle micro-frequencies

# VoiceRadar: Motivation

Approximates the physical models

To simulate the audio wave propagation

Extracts the subtle micro-frequencies

To differentiate synthetic audio

# VoiceRadar: Motivation

Approximates the physical models

To simulate the audio wave propagation

Extracts the subtle micro-frequencies

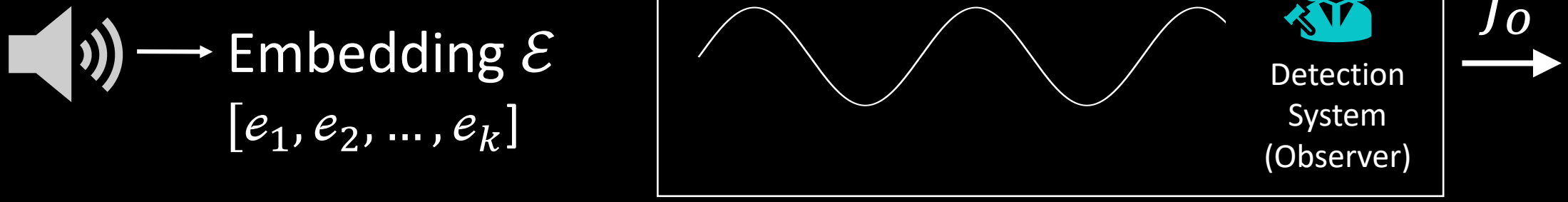
To differentiate synthetic audio

Incorporates compositional bias

# VoiceRadar: Motivation

| | | |
|---|---|---|
| Approximates the physical models | Extracts the subtle micro-frequencies | Incorporates compositional bias |
| ↓ | ↓ | ↓ |
| To simulate the audio wave propagation | To differentiate synthetic audio | To penalize mismatch of energies by the model |

# VoiceRadar: Micro-Frequency Analysis
## [Observer Frequency]



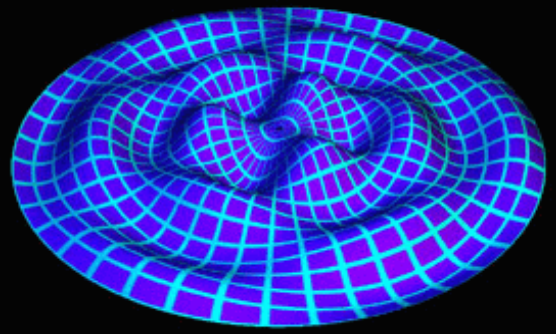

Physical models approximation

# VoiceRadar: Core Components

Audio Embedding

Vector $\mathcal{E}(\vec{x})$

**Drumhead vibrations**



$f_s$

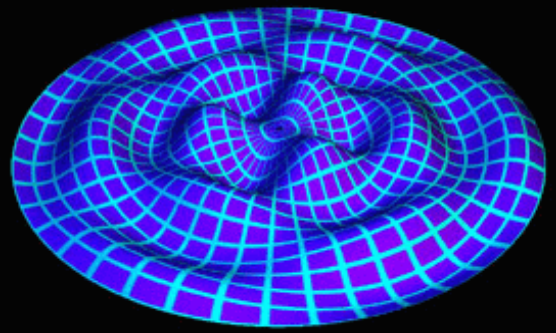
$$u(x,y,t) \approx \begin{cases} \Delta f_t & (y=0,x) \\ \Delta f_r & (y,x) \\ \Delta f_v & (y,x=c) \end{cases}$$

# VoiceRadar: Core Components

Audio Embedding

Vector $\mathcal{E}(\vec{x})$

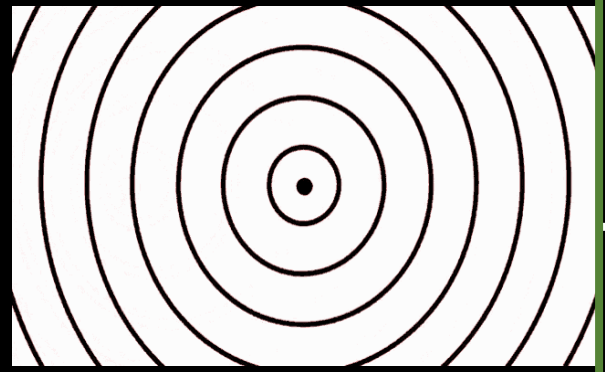
**Drumhead vibrations**



$$u(x,y,t) \approx \begin{cases} \Delta f_t & (y=0,x) \\ \Delta f_r & (y,x) \\ \Delta f_v & (y,x=c) \end{cases}$$

$f_s$

**Doppler Effect**



$$f_o = \frac{c_v + v_o}{c_v - v_s} * f_s$$

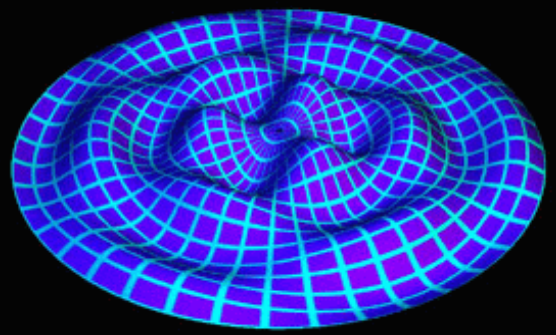
$f_o$

# VoiceRadar: Core Components

Audio Embedding

Vector $\mathcal{E}(\vec{x})$

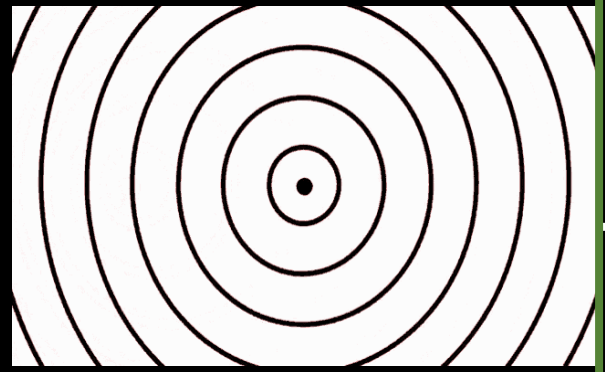
**Drumhead vibrations**



$$u(x,y,t) \approx \begin{cases} \Delta f_t & (y=0,x) \\ \Delta f_r & (y,x) \\ \Delta f_v & (y,x=c) \end{cases}$$

$f_s$

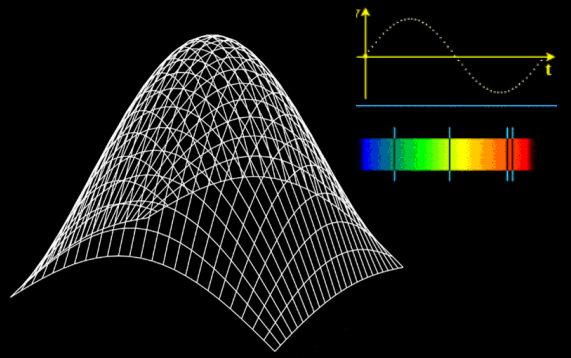
**Doppler Effect**



$$f_o = \frac{c_v + v_o}{c_v - v_s} * f_s$$

$f_o$

**Model Training with Physics Prior**




$$L = Binay\_cross\_entropy + f_o$$

$$W: argmin_W L(W, f_o)$$

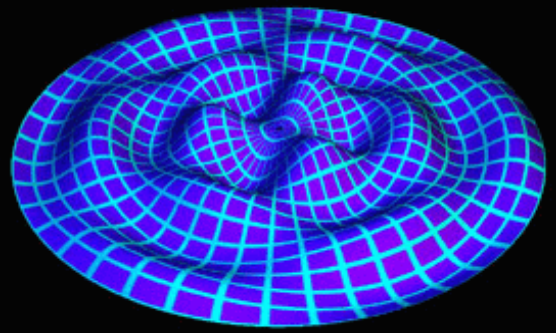
# VoiceRadar: Core Components

Embedding $\mathcal{E}(\vec{x}) \coloneqq [e_1, e_2, \dots, e_k]$

Audio Embedding

Vector $\mathcal{E}(\vec{x})$

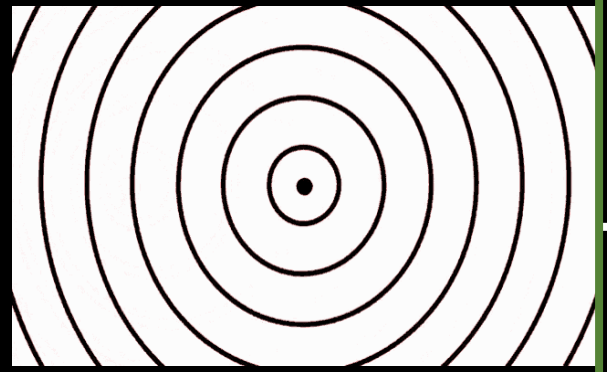
## Drumhead vibrations



$$u(x,y,t) \approx \begin{cases} \Delta f_t & (y=0,x) \\ \Delta f_r & (y,x) \\ \Delta f_v & (y,x=c) \end{cases}$$

$f_s$

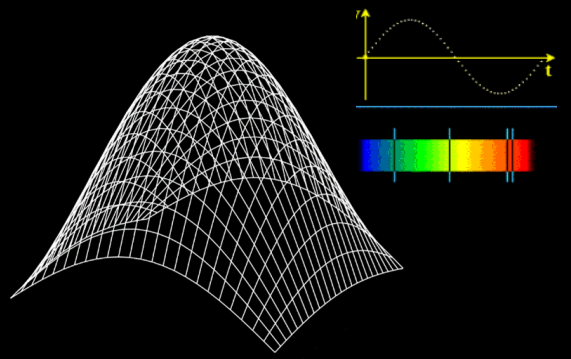
## Doppler Effect



$$f_o = \frac{c_v + v_o}{c_v - v_s} * f_s$$

$f_o$

## Model Training with Physics Prior




$$L = Binay\_cross\_entropy + f_o$$

$$W: argmin_W L(W, f_o)$$

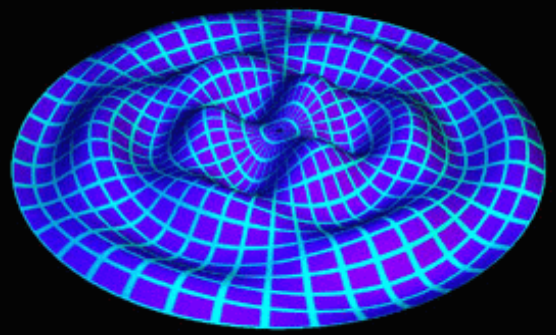
# VoiceRadar: Core Components

Embedding $\mathcal{E}(\vec{x}) \coloneqq [e_1, e_2, \dots, e_k]$

Audio Embedding

Vector $\mathcal{E}(\vec{x})$

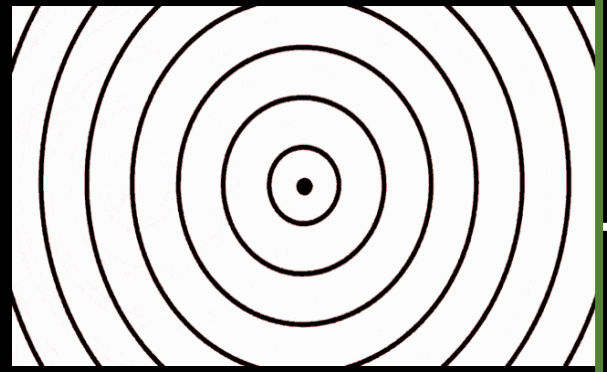
**Drumhead vibrations**



$$u(x,y,t) \approx \begin{cases} \Delta f_t & (y=0,x) \\ \Delta f_r & (y,x) \\ \Delta f_v & (y,x=c) \end{cases}$$

$f_s$

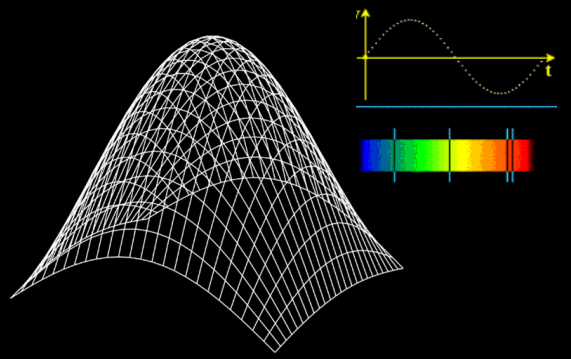
**Doppler Effect**



$$f_o = \frac{c_v + v_o}{c_v - v_s} * f_s$$

$f_o$

**Model Training with Physics Prior**




$$L = Binay\_cross\_entropy + f_o$$

$$W: argmin_W L(W, f_o)$$

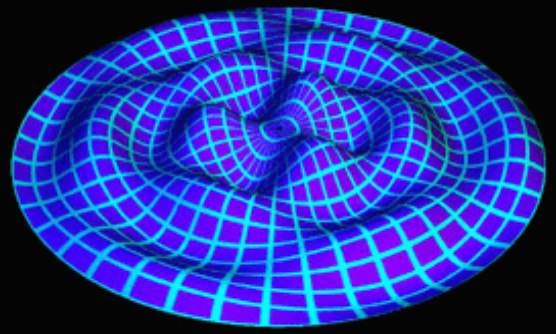
# VoiceRadar: Core Components

Embedding $\mathcal{E}(\vec{x}) \coloneqq [e_1, e_2, \dots, e_k]$

Audio Embedding

Vector $\mathcal{E}(\vec{x})$

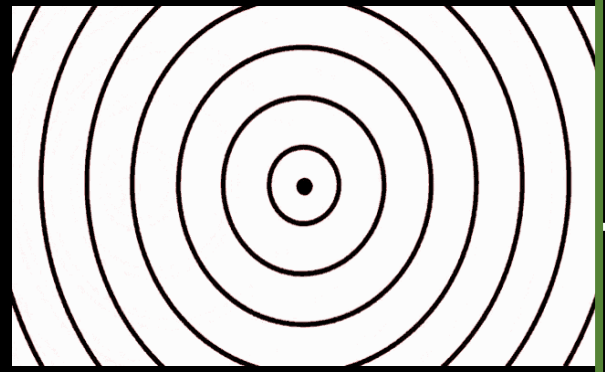
**Drumhead vibrations**



$$u(x,y,t) \approx \begin{cases} \Delta f_t & (y=0,x) \\ \Delta f_r & (y,x) \\ \Delta f_v & (y,x=c) \end{cases}$$

$f_s$

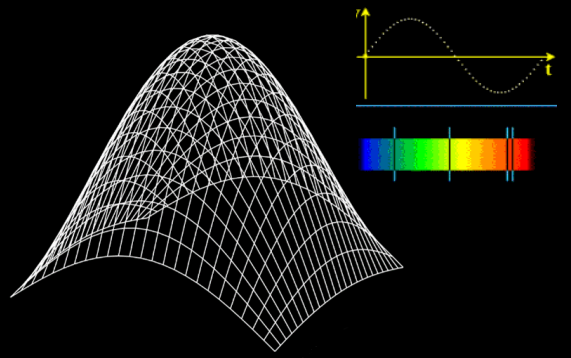
**Doppler Effect**



$$f_o = \frac{c_v + v_o}{c_v - v_s} * f_s$$

$f_o$

**Model Training with Physics Prior**

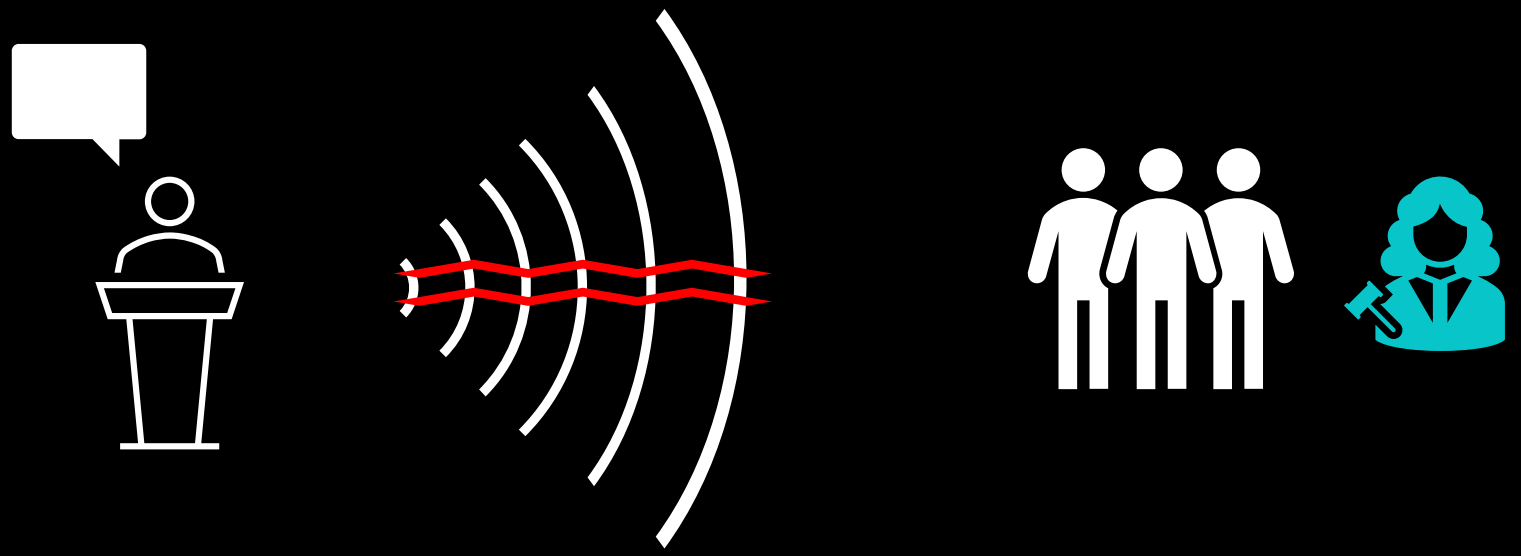



$$L = Binay\_cross\_entropy + f_o$$

$$W: argmin_W L(W, f_o)$$

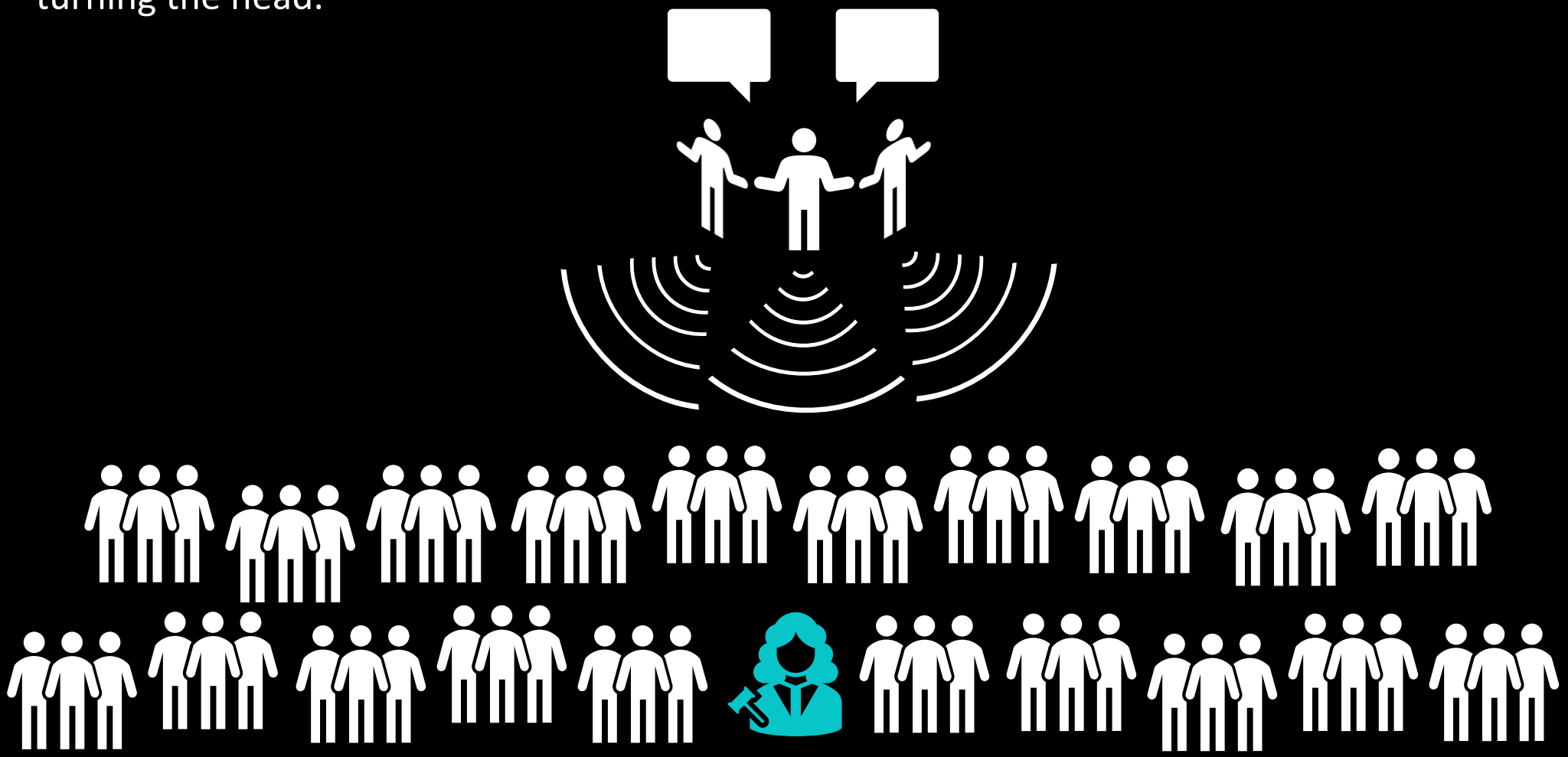
# VoiceRadar: Translational Frequency

- Wave propagation of speech in straight lines from speaker to observer.



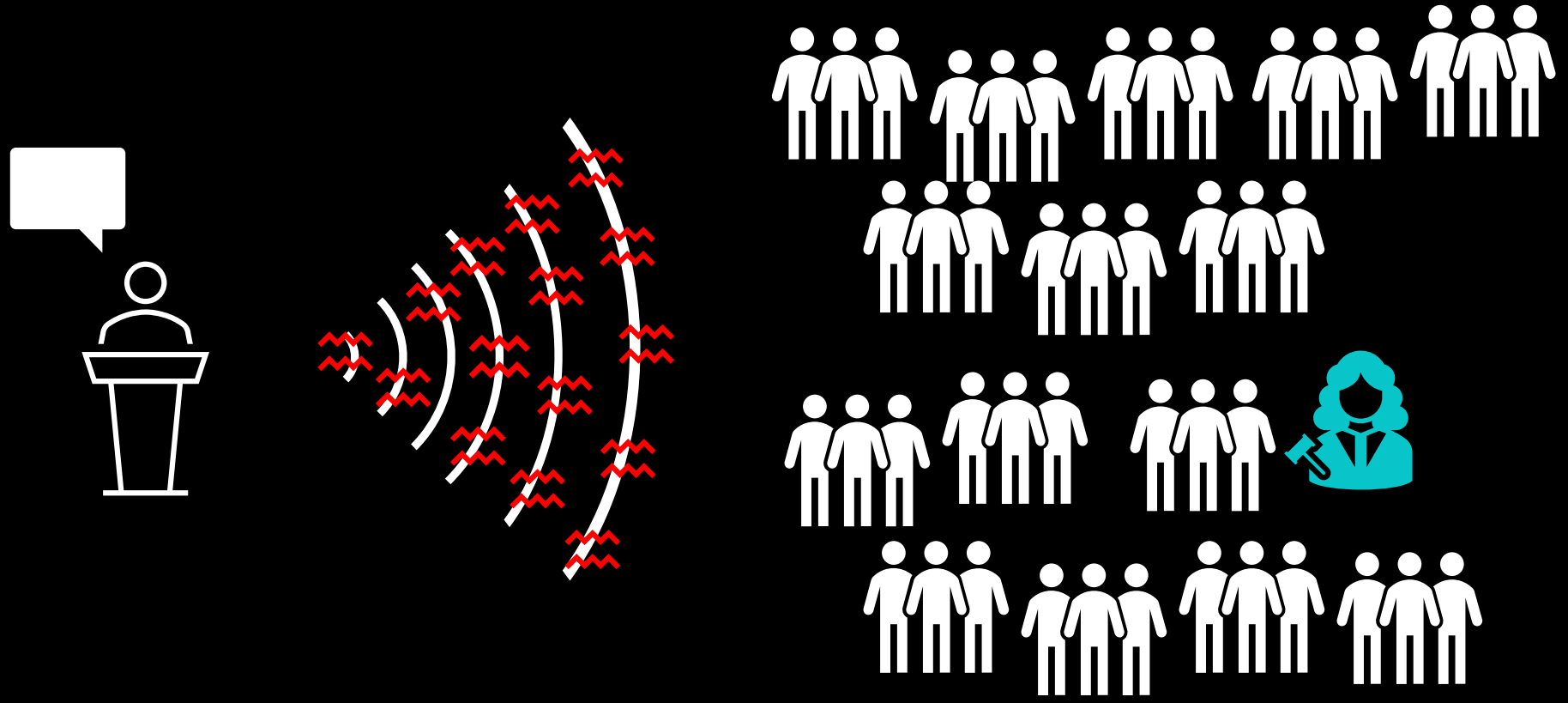
# VoiceRadar: Rotational Frequency

- Voice projection changes while speaking causing a rotational frequency shift, like when turning the head.



# VoiceRadar: Vibrational Frequency

- Speech has subtle vibrational variations
- Stemming from individual and environmental factors, causing tremors.



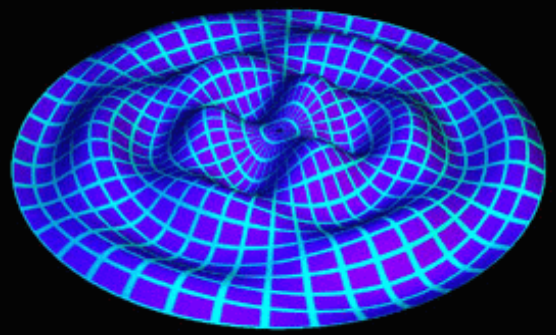
# VoiceRadar: Core Components

Embedding $\mathcal{E}(\vec{x}) \coloneqq [e_1, e_2, \dots, e_k]$

Audio Embedding

Vector $\mathcal{E}(\vec{x})$

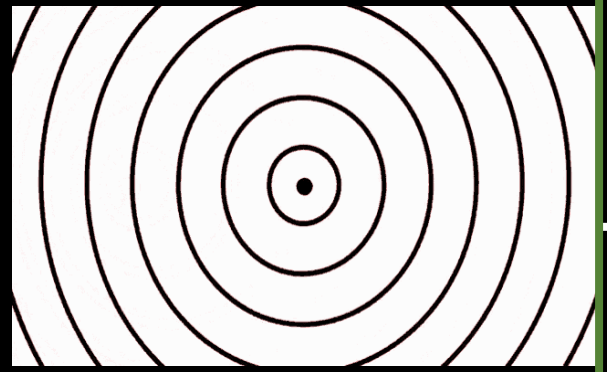
**Drumhead vibrations**



$$u(x,y,t) \approx \begin{cases} \Delta f_t & (y=0,x) \\ \Delta f_r & (y,x) \\ \Delta f_v & (y,x=c) \end{cases}$$

$f_s$

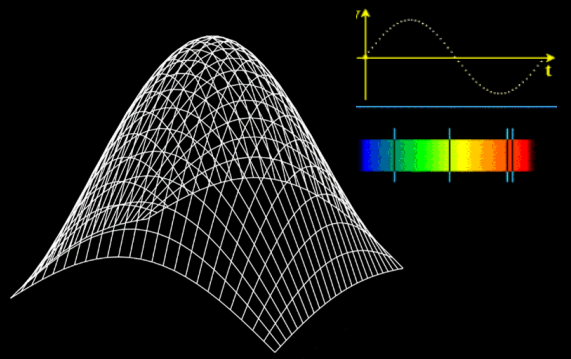
**Doppler Effect**



$$f_o = \frac{c_v + v_o}{c_v - v_s} * f_s$$

$f_o$

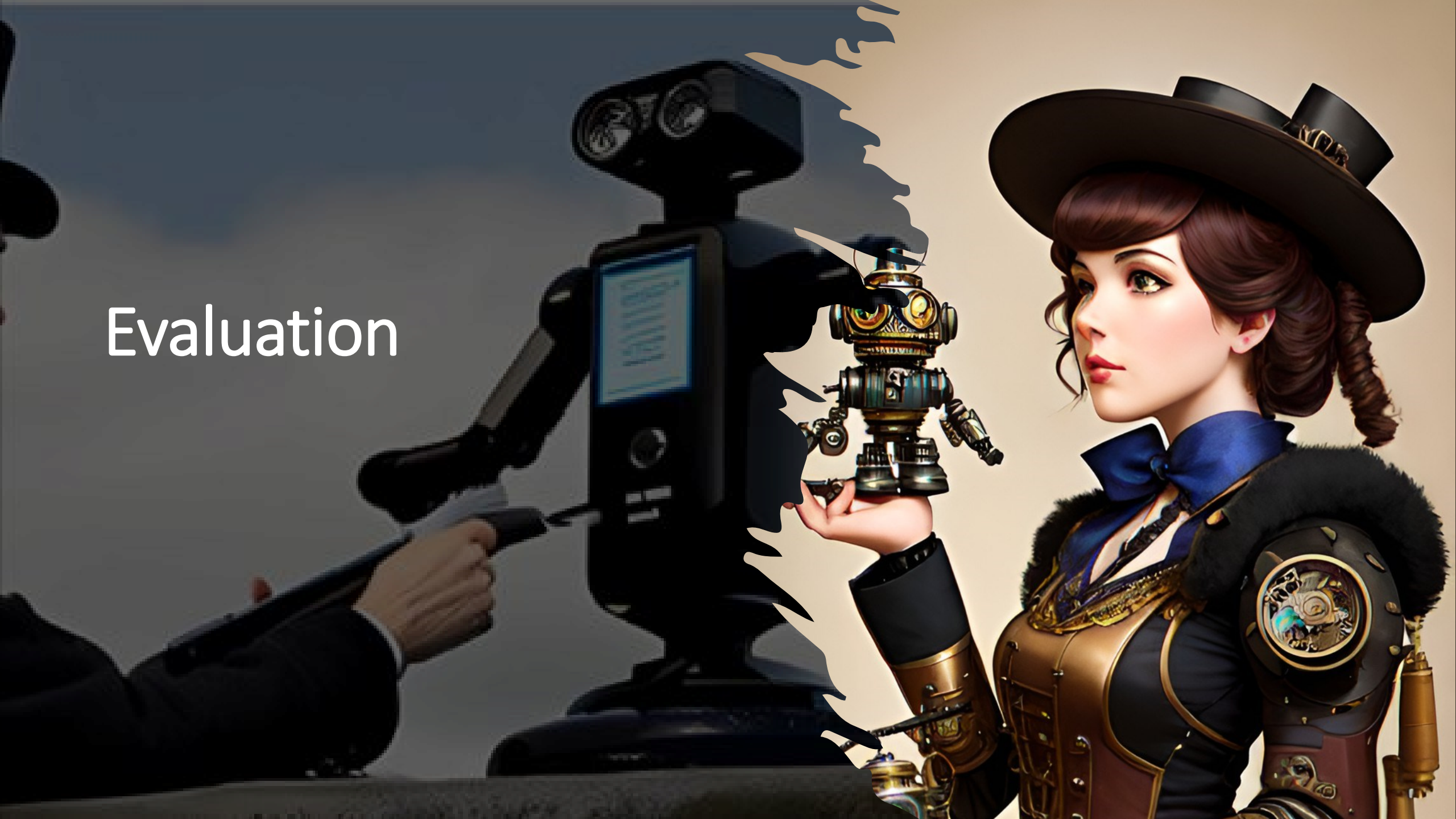
**Model Training with Physics Prior**




$$L = Binay\_cross\_entropy + f_o$$

$$W: argmin_W L(W, f_o)$$

# Evaluation



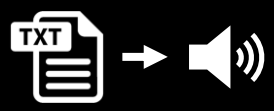
# Text-to-Speech (TTS) Dataset

Based on recordings from VCTK datasets

Used 8 most recent TTS approaches

40 Speakers

- Cover different distributions of: Gender, age, and regions

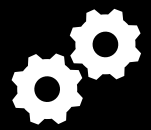



Generated 800 phrases for each speaker

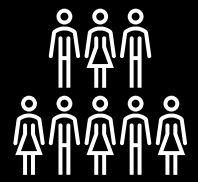
| Dataset | Records |
|---|---|
| VALL-E-X | 32 000 |
| Speech T5 | 32 000 |
| Bark | 32 000 |
| Style TTS2 | 32 000 |
| Jenny | 32 040 |
| Vits | 32 040 |
| XTTS | 32 040 |
| Tortoise | 32 000 |
| **Total** | 256 120 |

# Speech-to-Speech (STS) Dataset

Based on recordings from VCTK datasets

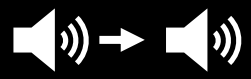


Used 4 most recent STS approaches



40 Speakers

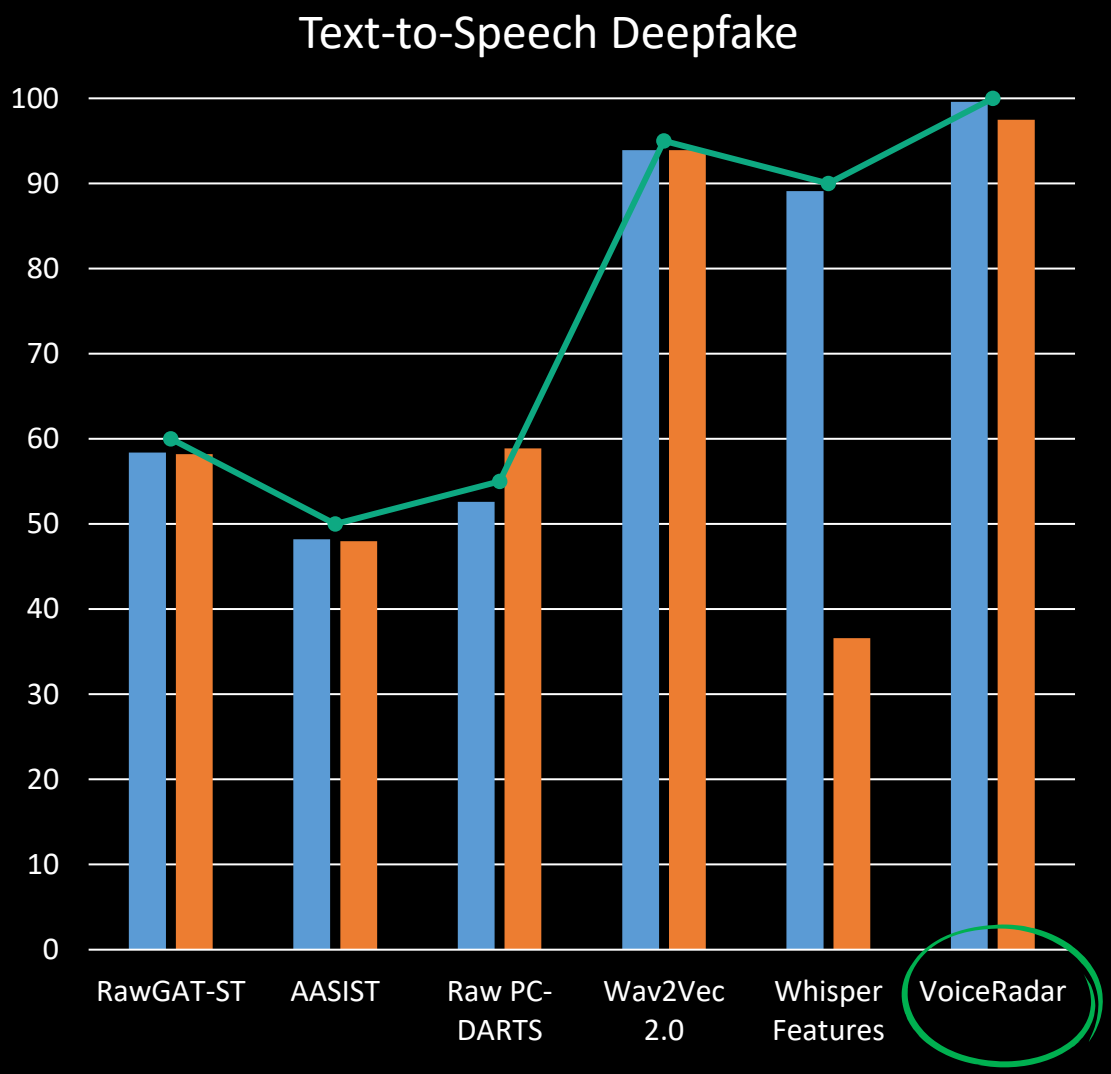
- Cover different distributions of: Gender, age, and regions

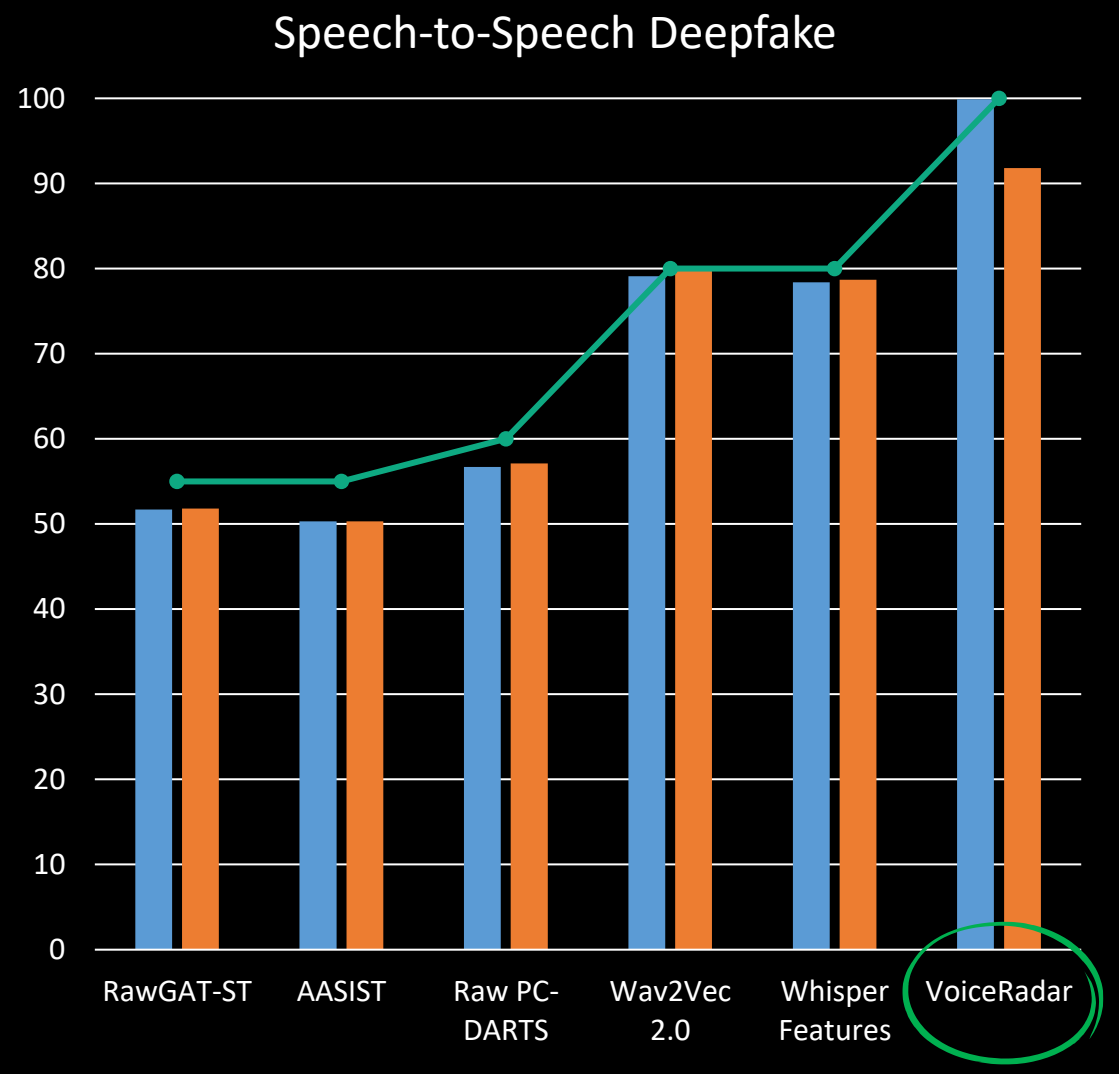


Generated for each speaker different combinations of STS

| Dataset | Records |
|---|---|
| DiffHierVC | 145 465 |
| DiffVC | 145 465 |
| HierSpeech++ | 145 483 |
| SpeechT5 | 145 483 |
| **Total** | 581 896 |

# Comparison of Audio Deepfake Detectors




$TPR = \frac{TP}{TP + FN}$ $TNR = \frac{TN}{TN + FP}$

# Conclusion

- Deepfakes are real threats to modern societies
- We addressed existing detectors' limitations
- We proposed VoiceRadar
  - Agnostic of Text-to-Speech or Voice-Conversion