Trustworthy AI Systems

-- Audio Recognition

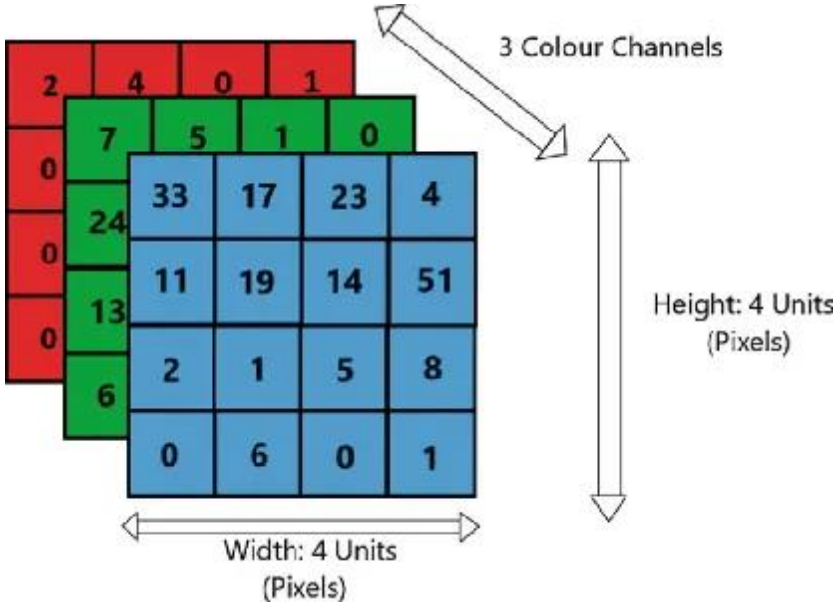
Instructor: Guangjing Wang

guangjingwang@usf.edu

Last Lectures

Using the **Image** modality as a target to introduce deep learning-based

- Classification
- Object Detection
- Generative Models
  - Generative Adversarial Networks
  - Variational Autoencoders
  - Diffusion Models



Computer Vision Topics

- 3D from multi-view and sensors
- 3D from single images
- Adversarial attack and defense
- Autonomous driving
- Biometrics
- Computational imaging
- Computer vision for social good
- Computer vision theory
- Datasets and evaluation
- Deep learning architectures and techniques
- Document analysis and understanding
- Efficient and scalable vision
- Embodied vision: Active agents, simulation
- Event-based cameras
- Explainable computer vision
- Humans: Face, body, pose, gesture, movement
- Image and video synthesis and generation
- Low-level vision
- Machine learning (other than deep learning)
- Medical and biological vision, cell microscopy
- Multimodal learning
- Optimization methods (other than deep learning)
- Photogrammetry and remote sensing
- Physics-based vision and shape-from-X
- Recognition: Categorization, detection, retrieval
- Representation learning
- Computer Vision for Robotics
- Scene analysis and understanding
- Segmentation, grouping and shape analysis
- Self-, semi-, meta- and unsupervised learning
- Transfer/ low-shot/ continual/ long-tail learning
- Transparency, fairness, accountability, privacy and ethics in vision
- Video: Action and event understanding
- Video: Low-level analysis, motion, and tracking
- Vision + graphics
- Vision, language, and reasoning
- Vision applications and systems

Source: https://cvpr.thecvf.com/Conferences/2025/CallForPapers

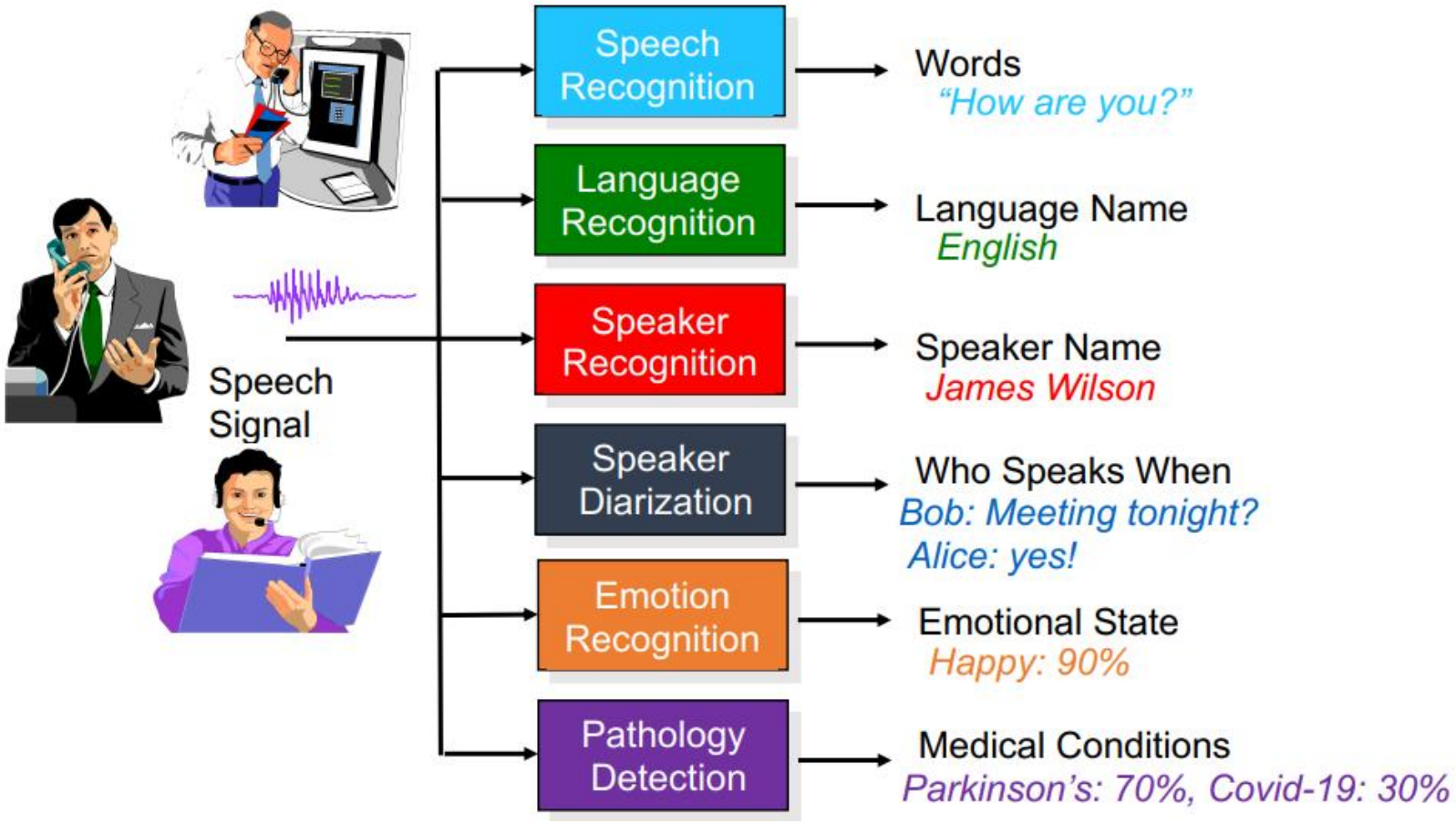
This Lecture

- Speech Recognition

- Speaker Recognition
  - Speaker Identification
  - Speaker Verification

- Humans as Deepfake Audio Detectors

Extracting Information from Audio

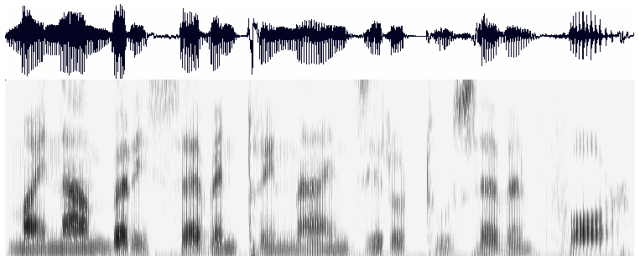


Speech Recognition

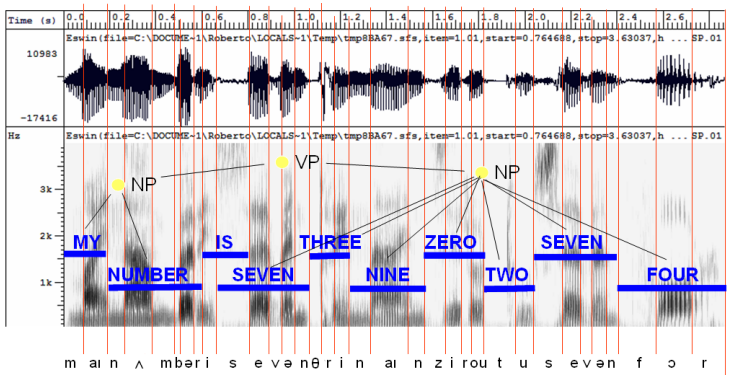
Air pressure over the time

Acoustic waveform

Acoustic signal



Pattern recognition



Speech recognition

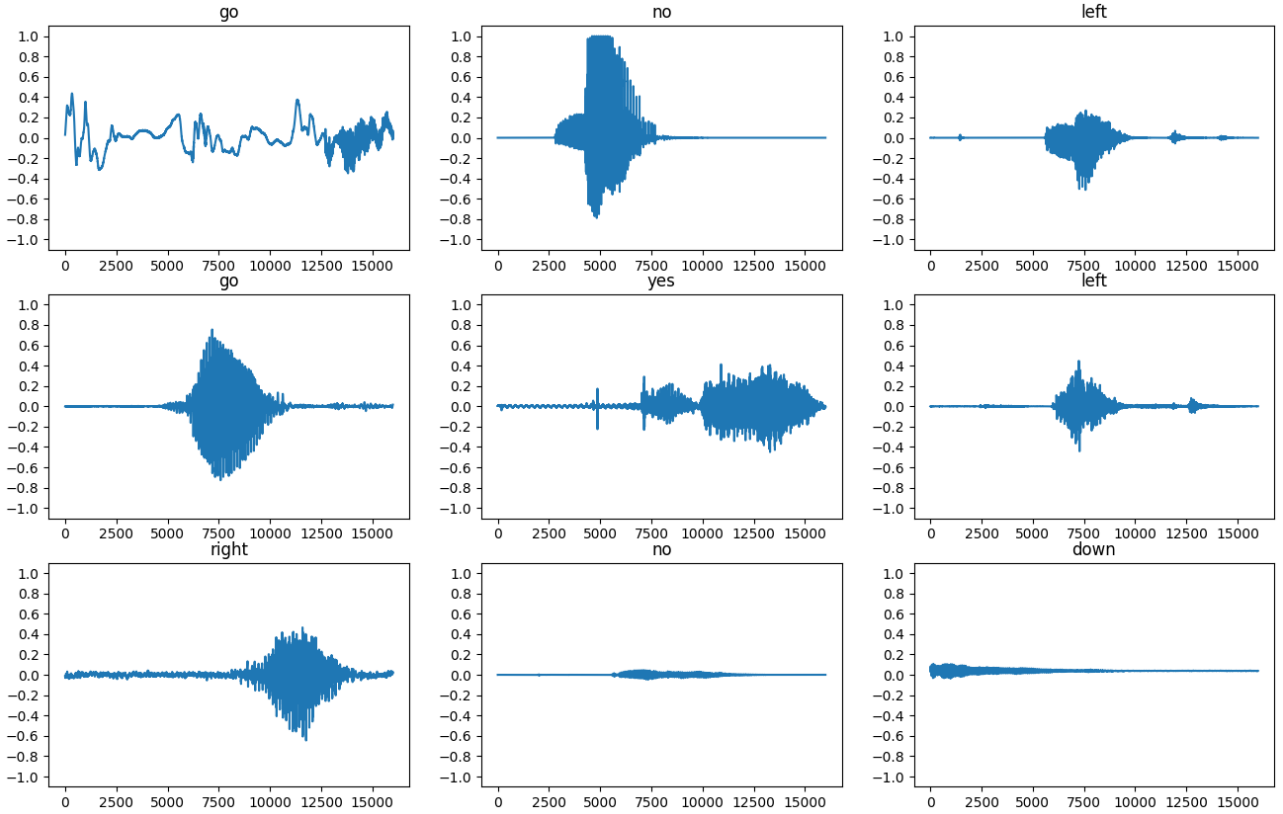
- Digitization
- Acoustic analysis of the speech signal
- Linguistic interpretation

https://personalpages.manchester.ac.uk/staff/harold.somers/LELA30431/Automatic%20Speech%20Recognition.ppt

Audio Samples: Waveform

The frequency is 16kHz, what does this mean?

If the audio clip is 1 second or less, then we use padding; if longer than 1 second, then we trim longer ones.



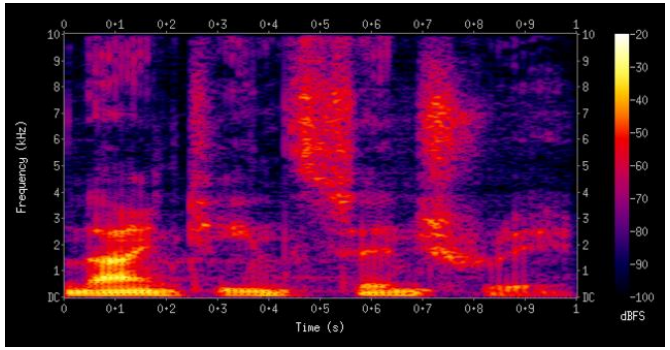
https://www.tensorflow.org/tutorials/audio/simple_audio

Fourier Transform and Short-Time Fourier Transform

- A Fourier transform converts a signal to its component frequencies but loses all time information.

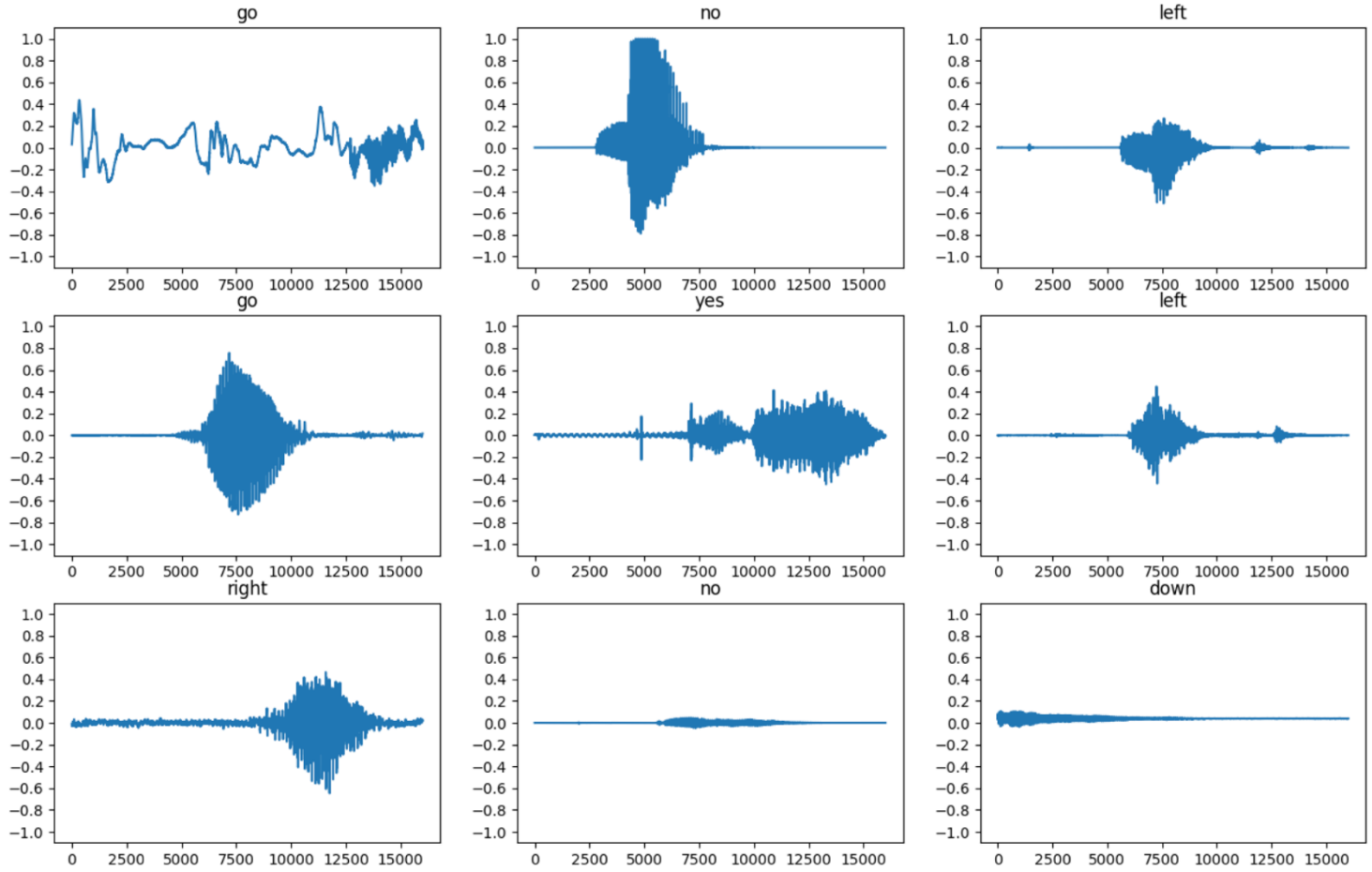
- Short-Time Fourier Transform (STFT) splits the signal into windows of time and runs a Fourier transform on each window, preserving some time information, and returning a 2D tensor.

Audio Samples: Spectrogram

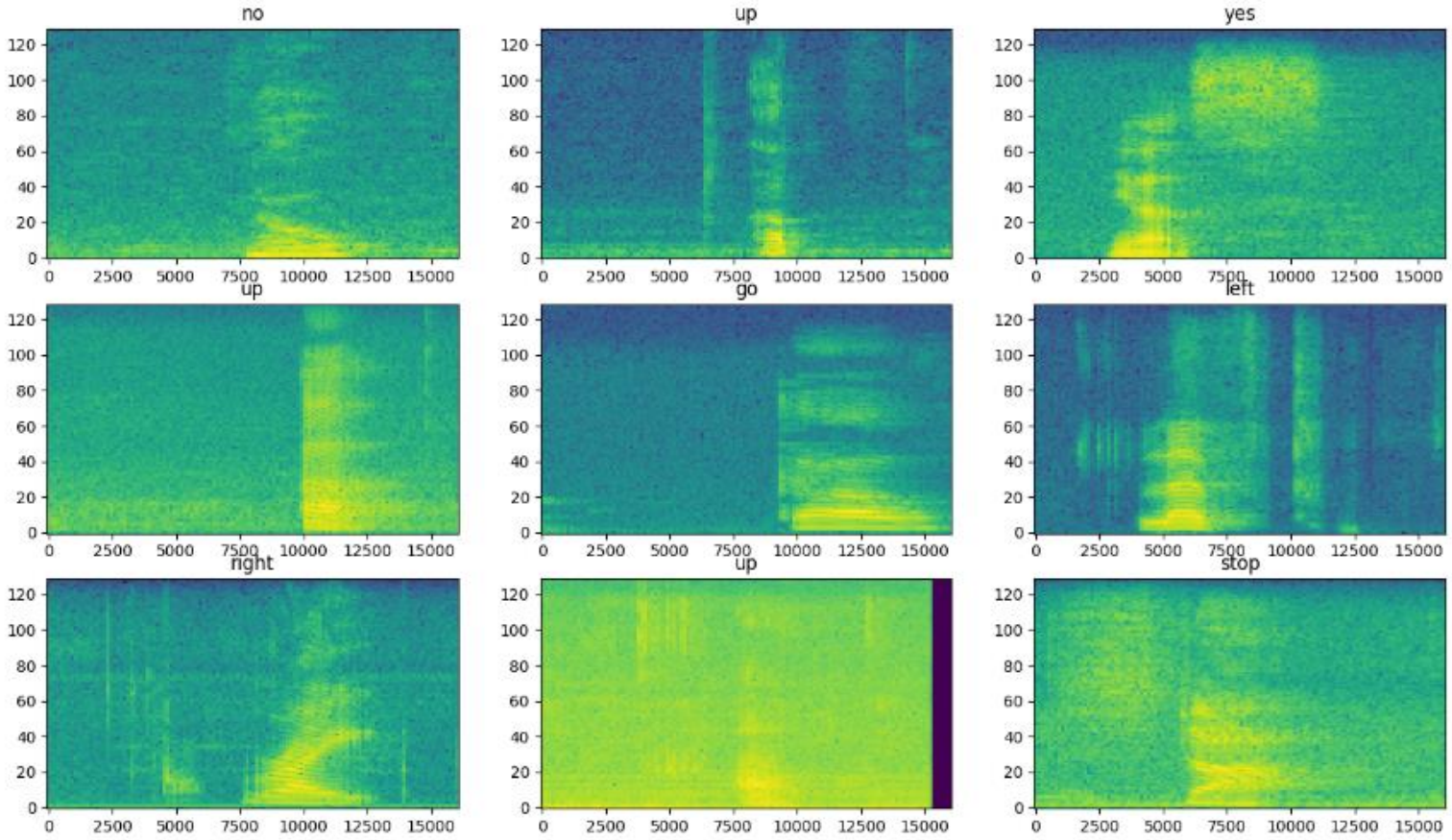


- Time-Frequency domain signals:
  - STFT to convert waveform to spectrograms
- Spectrograms show frequency changes over time and can be represented as 2D images
- Feed the spectrogram images into your neural network to train the model

Signal in Time Domain

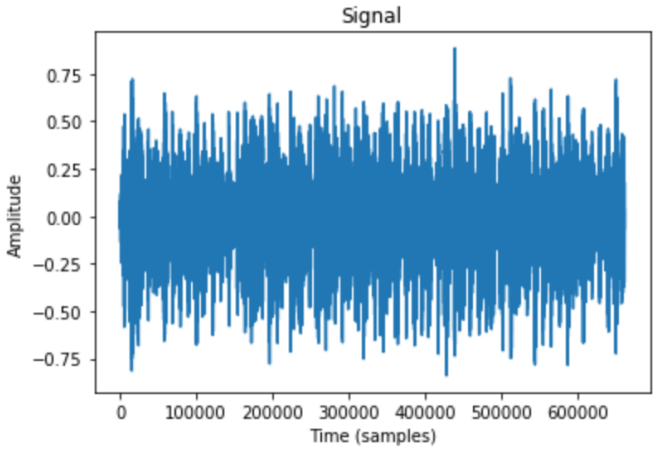


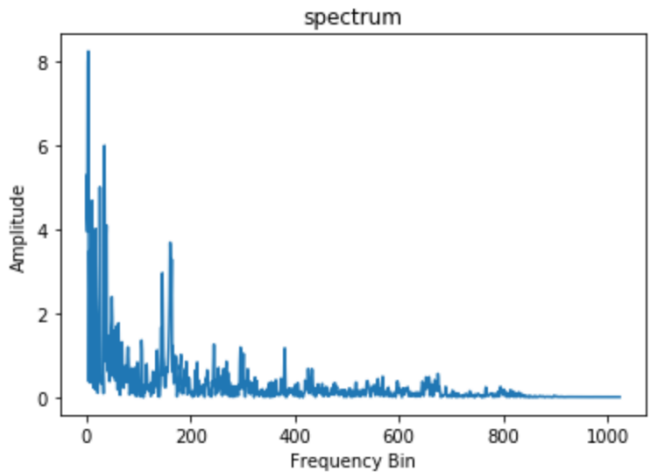
STFT Spectrogram

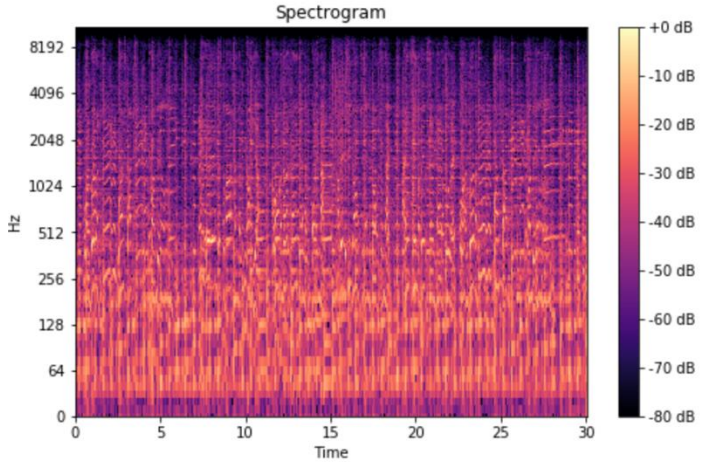


https://www.tensorflow.org/tutorials/audio/simple_audio

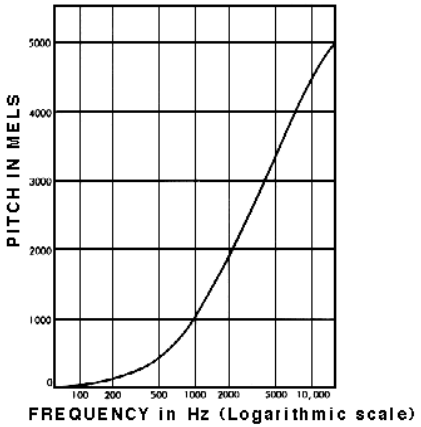
Some Key Concepts

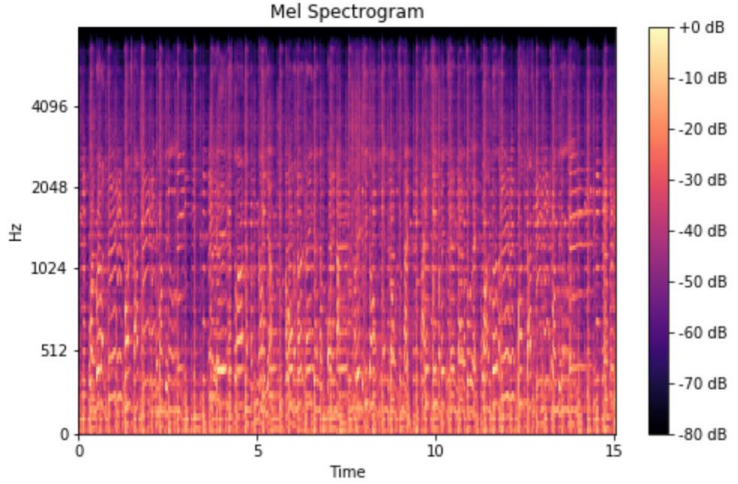






Mel Scale: humans do not perceive frequencies on a linear scale. A unit of pitch such that equal distances in pitch sounded equally distant to the listener.





Time domain to Time-frequency domain

- The waveforms need to be of the same length, so that when you convert them to spectrograms, the results have similar dimensions.

- The STFT produces an array of complex numbers representing magnitude and phase. We can get the absolute value of the complex numbers for the magnitude information

  e.g., 3.90625, -0.134048+0.027221j, -0.005211+0.008368j, -0.000981-0.000033j

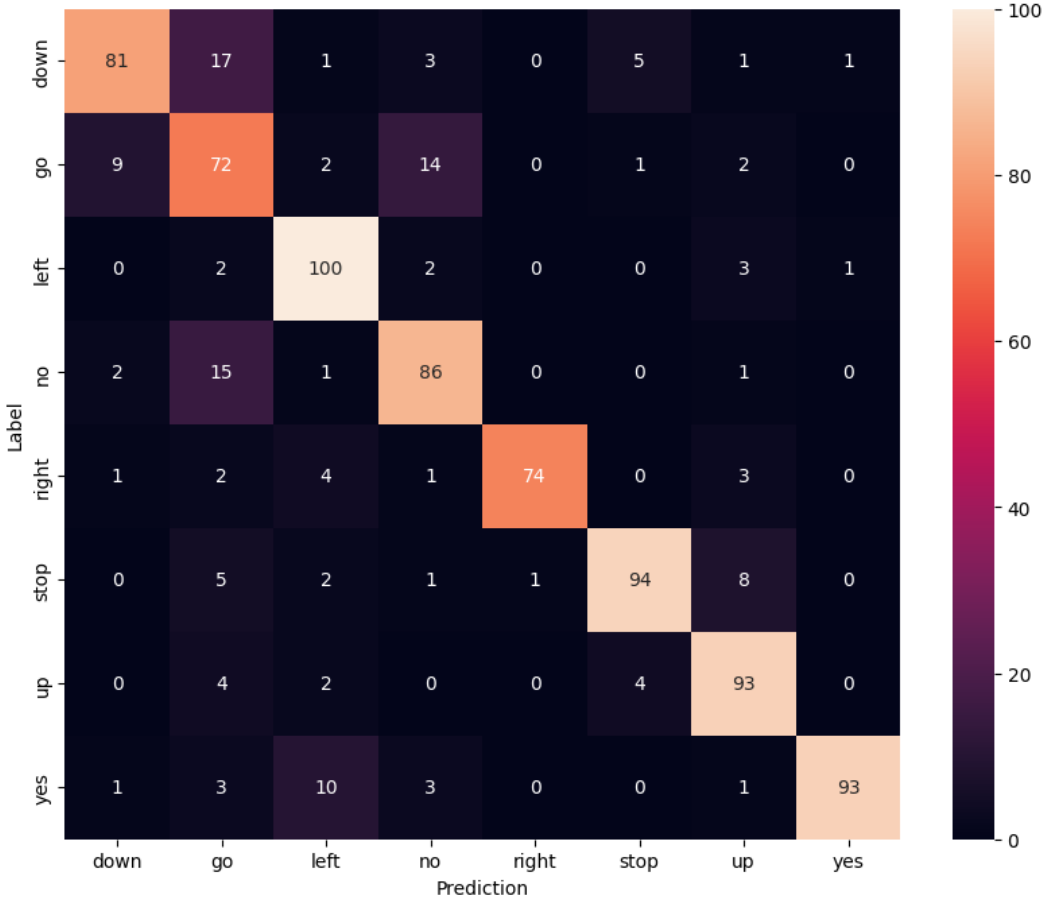
  Real part + Imaginary part

A Simple Speech Recognition Example

```
input_shape = example_spectrograms.shape[1:]
print('Input shape:', input_shape)
num_labels = len(label_names)

# Instantiate the `tf.keras.layers.Normalization` layer.
norm_layer = layers.Normalization()
# Fit the state of the layer to the spectrograms
# with `Normalization.adapt`.
norm_layer.adapt(data=train_spectrogram_ds.map(map_func=lambda spec, label: spec))

model = models.Sequential([
    layers.Input(shape=input_shape),
    # Downsample the input.
    layers.Resizing(32, 32),
    # Normalize.
    norm_layer,
    layers.Conv2D(32, 3, activation='relu'),
    layers.Conv2D(64, 3, activation='relu'),
    layers.MaxPooling2D(),
    layers.Dropout(0.25),
    layers.Flatten(),
    layers.Dense(128, activation='relu'),
    layers.Dropout(0.5),
    layers.Dense(num_labels),
])
```



https://www.tensorflow.org/tutorials/audio/simple_audio

This Lecture

- Speech Recognition

- Speaker Recognition
  - Speaker Identification
  - Speaker Verification

- Humans as Deepfake Audio Detectors

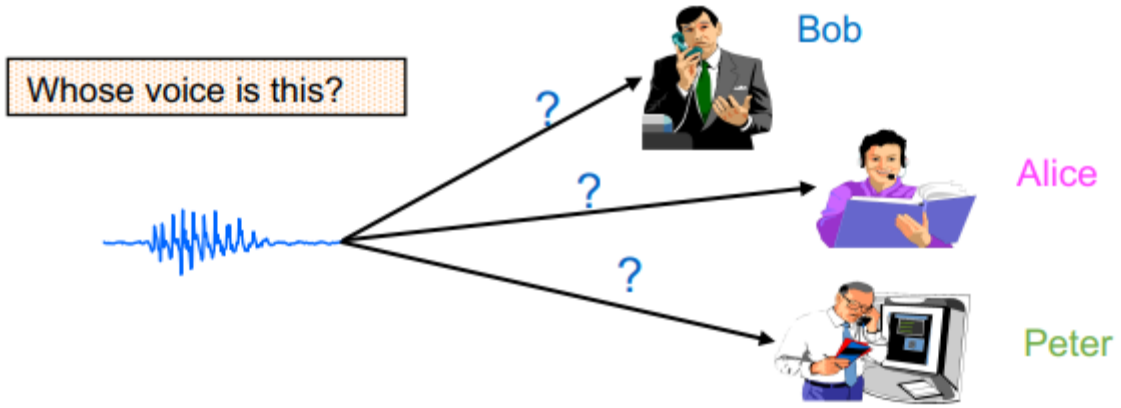
Speaker Recognition

Biometric modality consisting in recognizing people from the characteristics of their voices

- Properties of speech influenced by:
  - Anatomy:
    - Shape and size of voice production organs (vocal track, larynx, nasal cavity)
  - Behavioral patterns (Manner of Speaking):
    - Accent, rhythm, intonation style, pronunciation pattern, vocabulary
- Advantages:
  - Easy to use, speech is a natural way of communication
  - Non-intrusive, well accepted by users

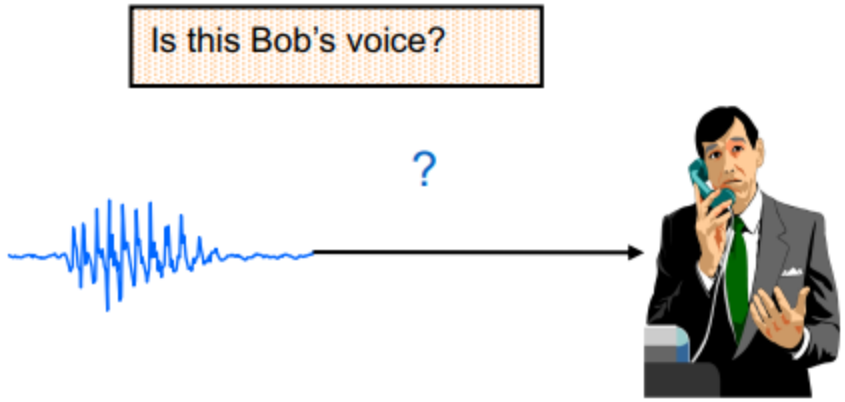
Speaker Identification

- Determine whether a test speaker matches one of a set of known speakers
  - Referred as closed-set identification.



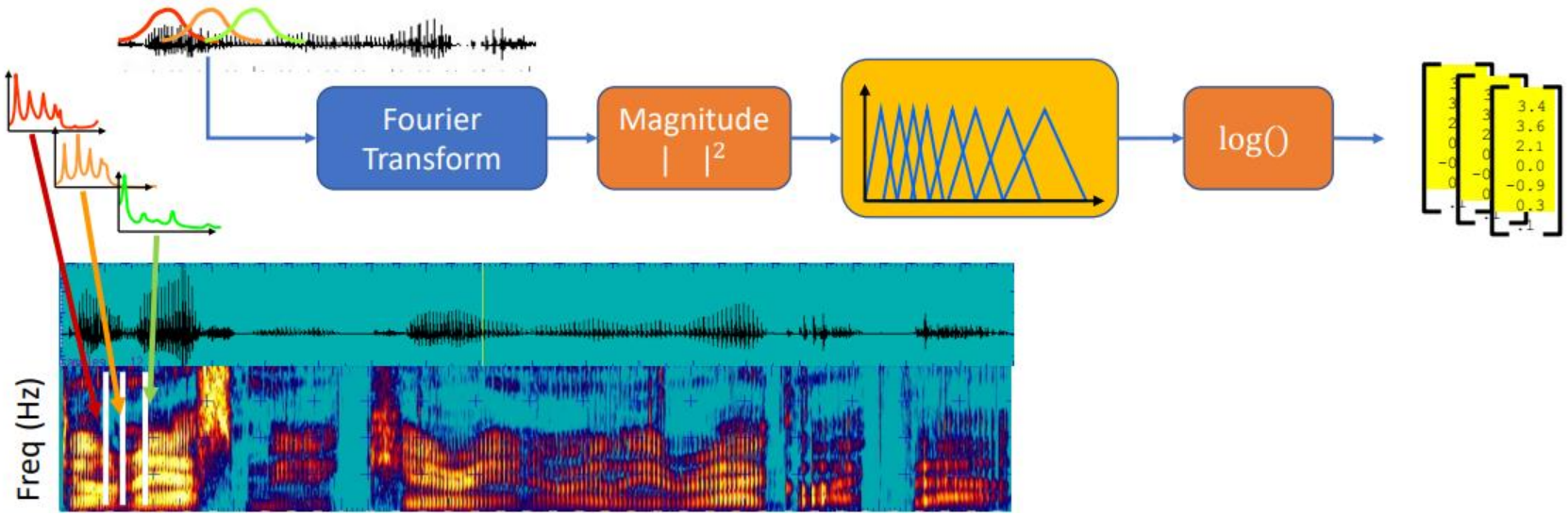
Speaker Verification

- Determine whether a test speaker matches a specific target speaker
- Unknown speech may come from a large set of unknown speakers – referred as open-set verification
- This is most common task in speaker recognition, close to real application.



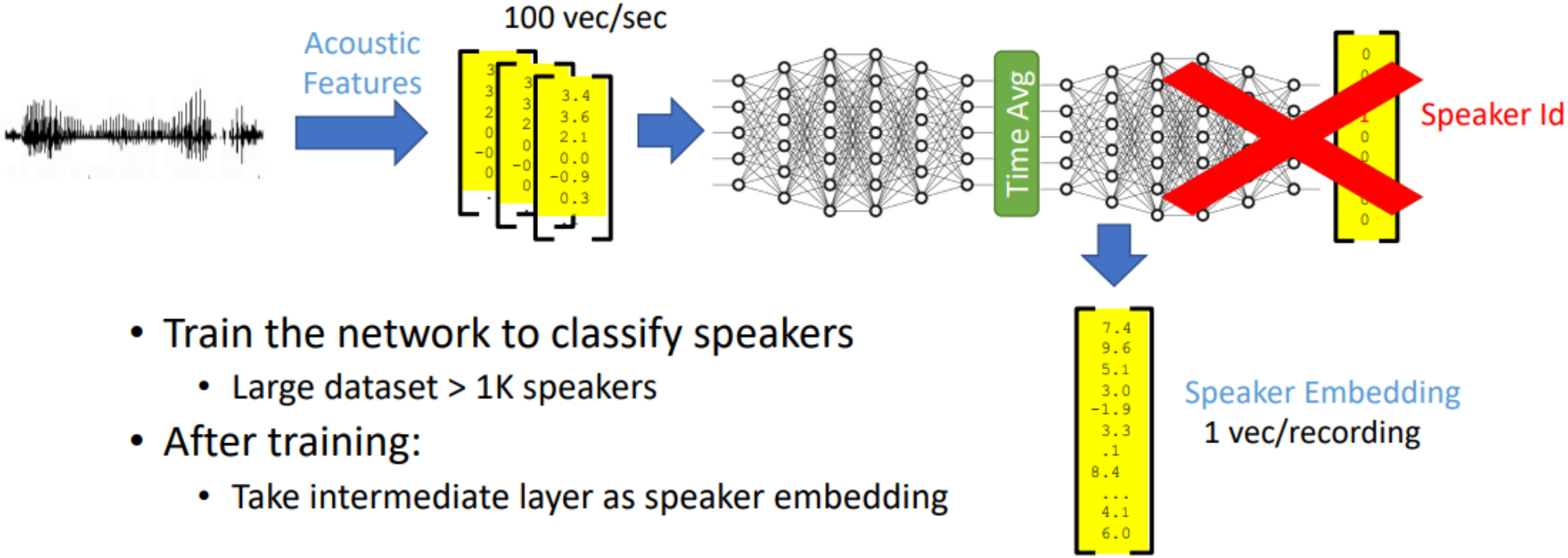
Acoustic Features

- Time sequence of acoustic features is needed to extract the speech information
  - Time-frequency representation of the signal
  - Filter bank in log Mel scale (Mel filtered spectrogram)



Speaker Embedding

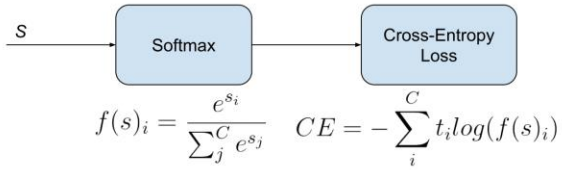
- Modern solution: **Speaker Embeddings**
  - Transform variable length recording into a single vector - Embedding
  - Embedding retains the speaker identity information



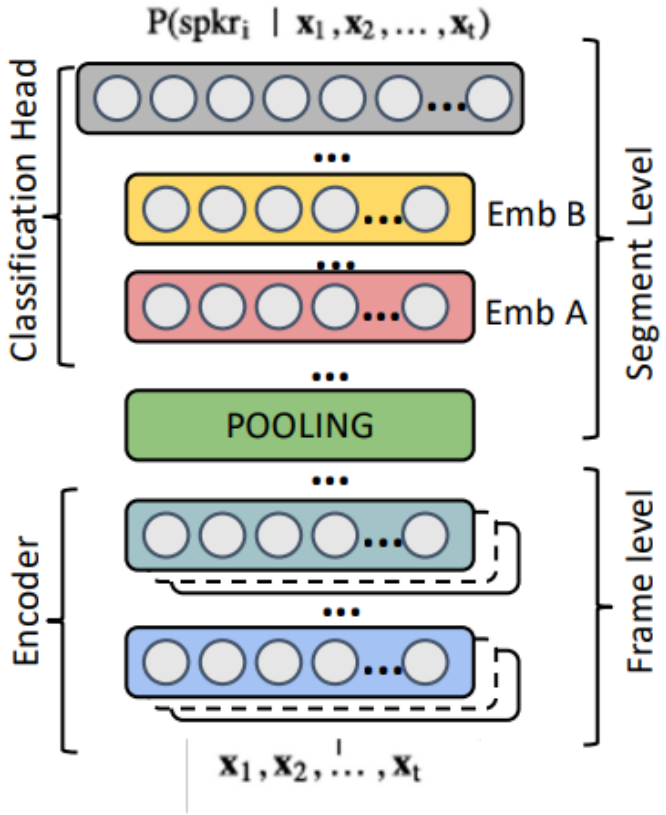
- Train the network to classify speakers
  - Large dataset > 1K speakers
- After training:
  - Take intermediate layer as speaker embedding

Speaker Embedding: X-vectors

- X-Vector network has three parts:
  - Encoder:
    - Input: Acoustic features log-Mel spectrogram.
    - Output: frame level hidden representations.
  - Pooling:
    - Summarizes representations into a single vector / utt.
    - Mean, Mean+Stddev, …
  - Classification Head:
    - Predicts posterior probabilities for the training speakers.
    - Embedding extracted from middle layer.
- Categorical cross-entropy loss:



  - Compares each utterance against all the speakers in the training data.
  - Does not need hard negative sampling.



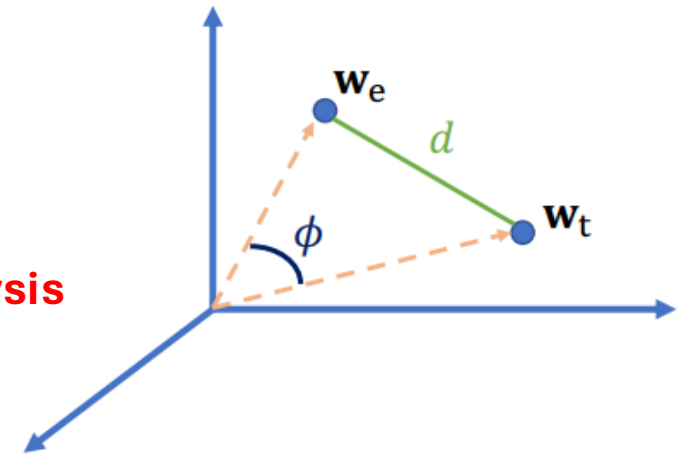
X-vectors: Robust dnn embeddings for speaker recognition. ICASSP 2018

Metric

- Assume that:
  - $\mathbf{w}_e$ spk. embedding from enrollment utterance of speaker **X**
  - $\mathbf{w}_t$ spk. embedding from test utterance of person that claims to be speaker **X**

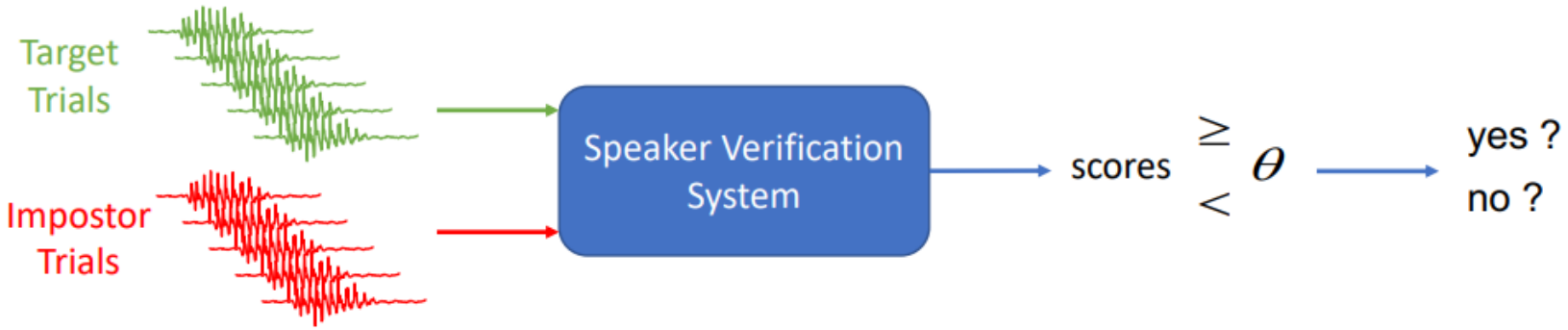
- The Metric compares enrollment and test embeddings $\mathbf{w}_e$, $\mathbf{w}_t$
  - Cosine scoring:

$$s = \cos(\phi) = \frac{\mathbf{w}_e^T \mathbf{w}_t}{\|\mathbf{w}_e\|_2 \|\mathbf{w}_t\|_2}$$

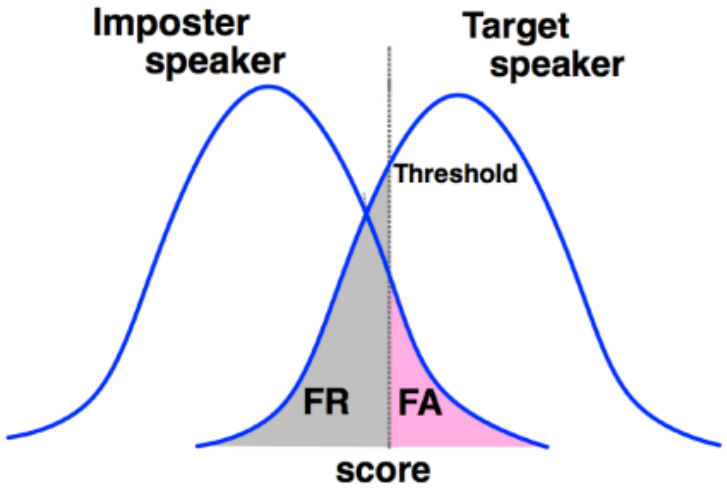


  - PLDA **Probabilistic Linear discriminant Analysis**
    - Map to low dimensional subspace
    - Inter-class covariance larger
    - Intra-class covariance smaller
    - PLDA allows to make inference about the classes not present during training

Choosing Threshold

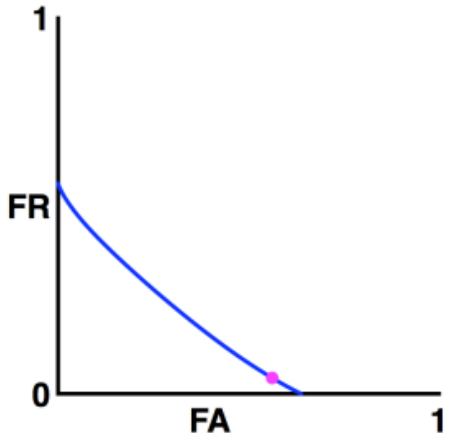


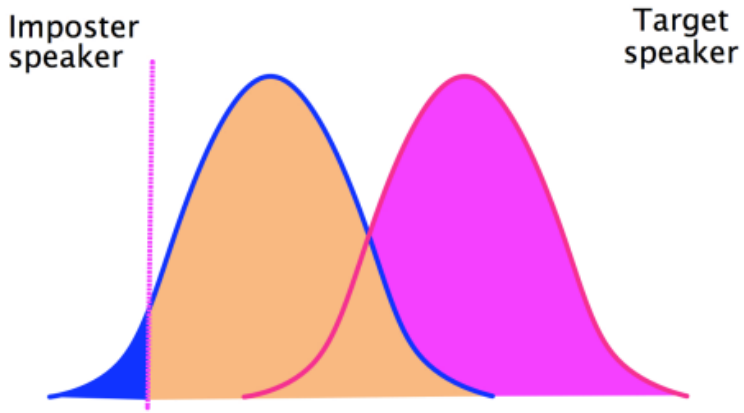
- Put >1k of target and impostor trials into the systems and count the errors
- Types of Errors:
  - Miss/False rejection:
    - True speaker is classified as impostor
    - Metric: Miss rate P_{Miss}
  - False alarm:
    - Impostor is classified as the true speaker
    - Metric: False alarm rate P_{FA}



Performance Metric

- Detection Error Trade-off (DET)





- True positive
- True Negative
- Miss, false rejected
- False Accepted

- Equal Error Rate (EER)

$$P_{\text{Miss}}(\theta_{\text{EER}}) = P_{\text{FA}}(\theta_{\text{EER}})$$

θ = decision threshold

- Detection Cost Function (DCF)

$$C_{\text{Det}}(\theta) = P_{\text{Miss}}(\theta) + \beta P_{\text{FA}}(\theta) \quad \text{with } \beta = \frac{1-P_{\text{target}}}{P_{\text{target}}}$$

$$\text{Minimum } C_{\text{Det}} = \min_{\theta} C_{\text{Det}}(\theta)$$

Voice Cloning (Take a break)

https://www.youtube.com/watch?v=vhArHsfsLAQ

Audio Deepfake Detection

Factors in real audio or fake audio

- Airflow pressure
- Time-difference-of-arrival of phoneme sequences
- The pop sound made by a breath
- The attributes of the airwaves
- The movement or structure of the human vocal anatomy
- Subtle spectral differences

A Large-Scale Evaluation of Humans as Audio Deepfake Detectors

Contribution

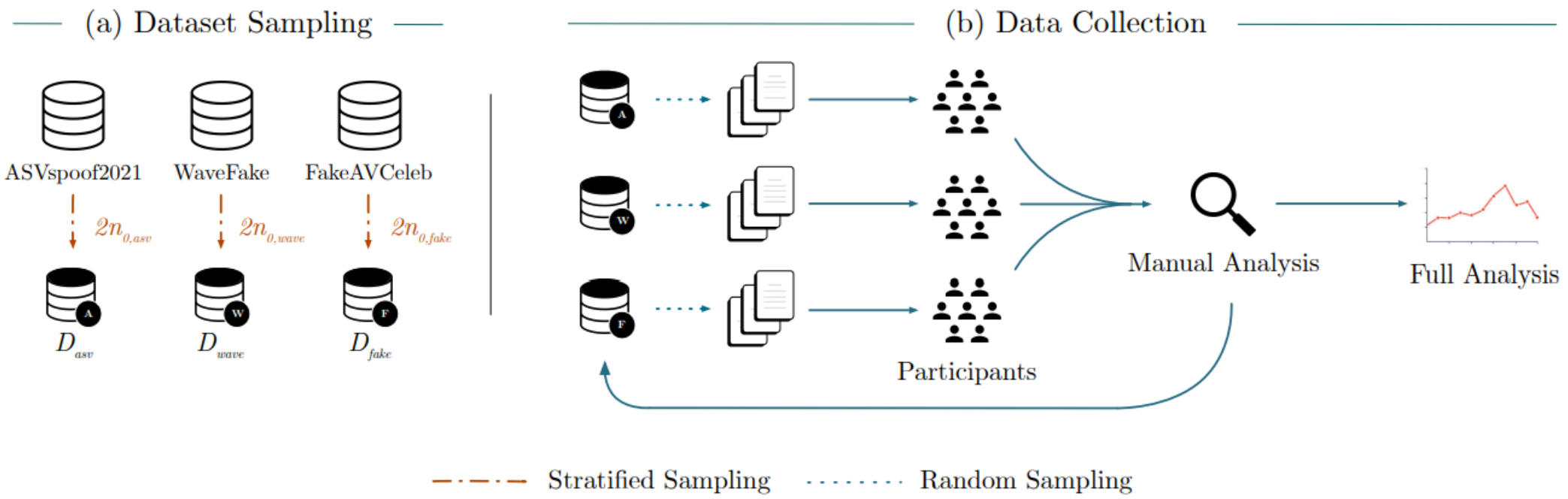
- Largest user study on audio deepfake detection
  - 1200 participants
  - Three datasets: Wavefake, ASVspoof2021, and FakeAVCeleb
- Qualitative study identifying decision factors
- Comparative analysis on human and ML performance

"Better Be Computer or I'm Dumb": A Large-Scale Evaluation of Humans as Audio Deepfake Detectors CCS 2024

Research questions

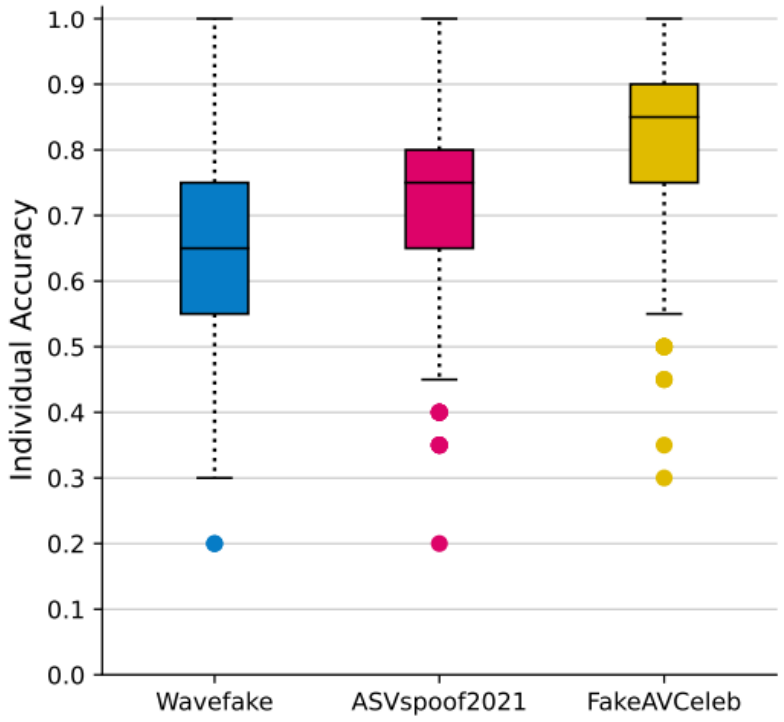
- What are the performance metrics for humans used in audio deepfake detection?
- What are the common themes affecting how humans classify audio deepfake samples as real or fake?
- Is there a demonstrable difference in audio deepfake detection capability between humans and ML models?

Data Preparation for Survey



For each participant, we randomly select 20 samples from one of these populations. Each sample is listened to by at least three unique participants

Individual User Performance

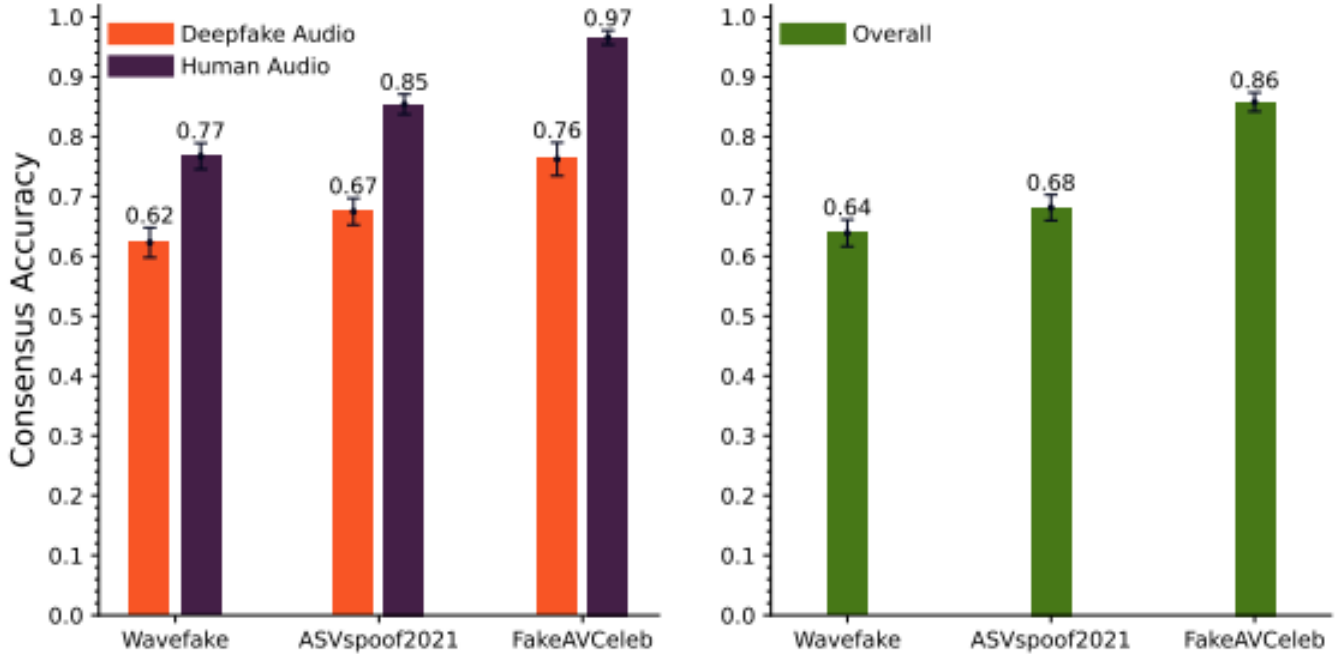


Individual user accuracy on the 20 samples given to each participant

Human recognize audio deepfakes:

- Each dataset had at least one person score a perfect accuracy, however, the average performance varied from dataset to dataset.
- On average, participants performed better on FakeAVCeleb and worse on the Wavefake dataset.

Application of Voting Scheme



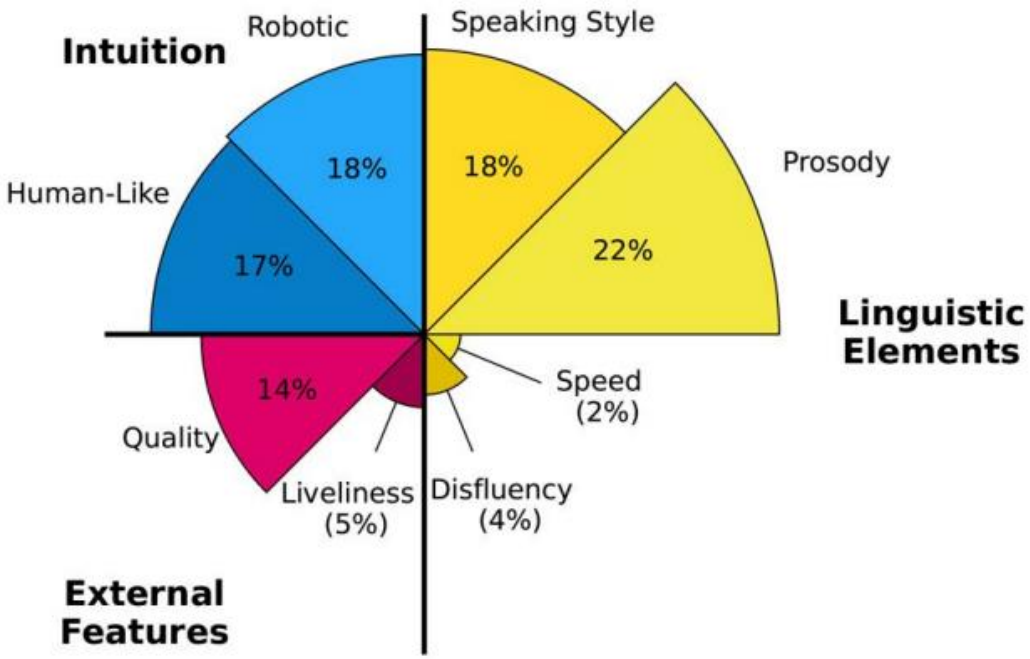
User accuracy based on consensus voting on human audio versus deepfake audio, and human accuracy on the audio overall.

Thematic Analysis

Theme	Code	Keywords
Linguistic Elements	Speaking Style	Accent, List, Articulation, Specific Word Choice
	Prosody	Tone, Inflections, Cadence, Pitch, Monotone, Raspy, Emotion
	Disfluency	Pauses, Filler Words
	Speed	Fast, Slow, Rushed
External	Quality	Background Noise, Microphone, Recording, Clipping
	Liveliness	Breathing, Mouth Noises, Nasal
Intuition	Human-Like	Natural, Human
	Robotic	Robotic, Glitchy, Mechanical

The codebook for categorizing responses from participants in the user study. The authors analyze each response using eight unique codes with corresponding keywords, then group those codes into three major themes.

Reasoning Themes



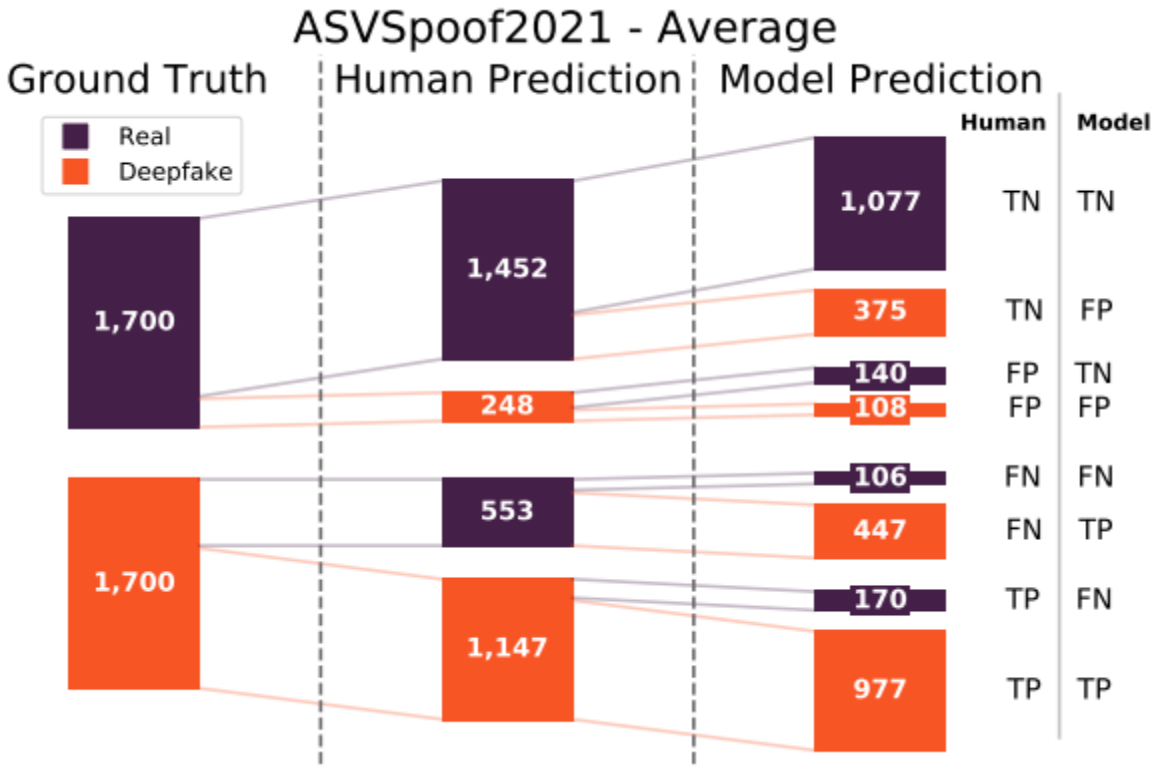
Appearance rate of the eight codes used in the thematic analysis.

It demonstrates that Prosody is the most common factor that contributes to classification decisions by people, while Speed is the least common.

Key Findings

- Participants have **pre-conceived ideas** of what computer voice generation is capable of, which impacts how they reason about detecting deepfake audio.
- **Audio artifacts** play a key role in how participants discriminate on deepfake audio, which could easily be manipulated by deepfake generators.
- While not as prevalent as linguistic features, participants still heavily rely on intuition when discriminating on deepfake audio.
- Humans misclassify fake samples which exhibit organic features and real samples that sound robotic at high rates.
- Humans perform well on real and fake samples that primarily feature sentence mistakes, odd speed, and quality issues.

Comparison with ML Detectors

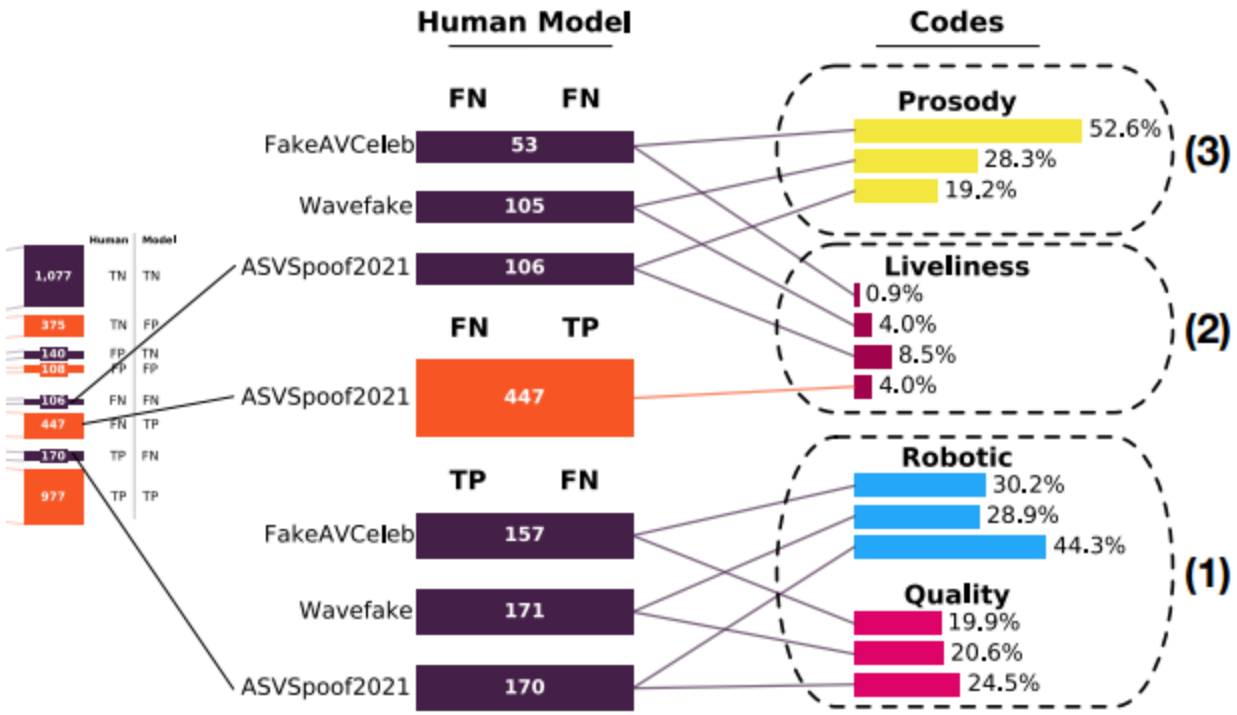


The classification breakdown for the average human and average model performance on the ASVspoof2021 samples D_{asv} .

The average human performance is more prone to false negatives while the average model is more prone to false positives.

Humans attain 76% accuracy compared to the average models' accuracy of 78%

Thematic Analysis in Models



Three cases (up2down):

- when humans and models both misclassify deepfakes.
- when models correctly predict deepfakes that humans miss
- when humans correctly predict deepfakes that models miss

Key Findings

- Many additional factors impact the way humans classify including a distrusting environment, recently heard audio for comparison, audio content, alternative reasoning for faults and audio sample construction.

- Models do not strictly perform better than humans, but rather there is a significant difference in the way that humans and models classify audio samples.

- Humans are prone to false negatives while models are prone to false positives.

References

- https://www.tensorflow.org/tutorials/audio/simple_audio

- https://jhu-intro-hlt.github.io/slides/speaker-id-2022.pdf