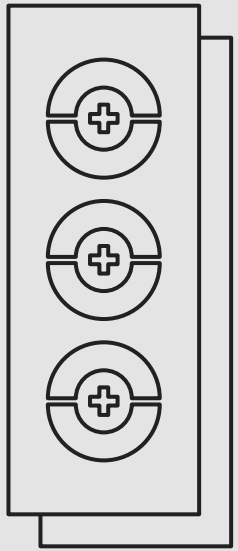
TTS VOICE CLASSIFICATION

Can you tell the difference?



By: Ryan Virgin

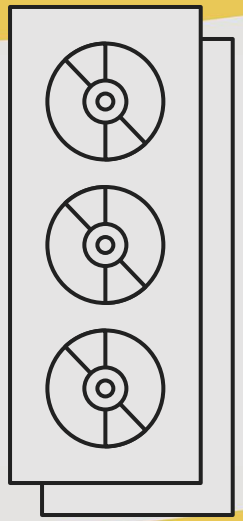
PROBLEM STATEMENT

Can a machine learning model tell the difference between human speech and computer generated speech?

WHY?

- Detecting fraud, preserving privacy.
- Enhancing service, improving synthesis.
- Applications in security, privacy.
- Addressing challenges, enabling possibilities.

TABLE OF CONTENTS



DATA

You are what you eat

(even in the metaverse)

DATA SUMMARY

01

HUMAN VOICE DATA

Recordings of humans reading a script

02

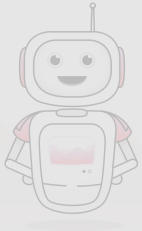
TTS SYNTHETIC VOICE DATA

Recordings of a trained TTS model's output given the same script

03

METADATA

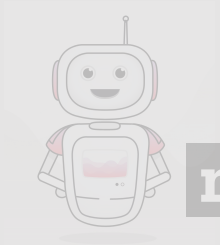
Information about the humans and TTS models behind the audio



moz://a

MOZILLA COMMON VOICE

- The Mozilla Common Voice (Moz) project is an open source initiative focused on creating a large, diverse voice dataset to improve speech recognition technology.
- The project relies on crowdsourced contributions from volunteers who record and validate audio clips in various languages and accents.

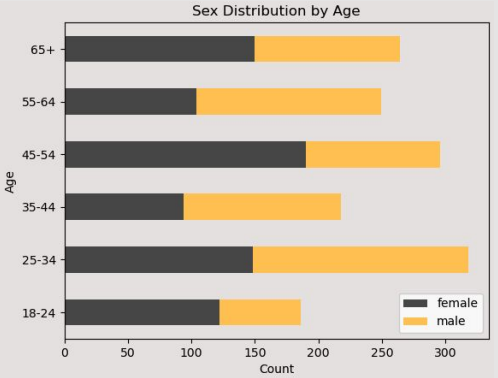


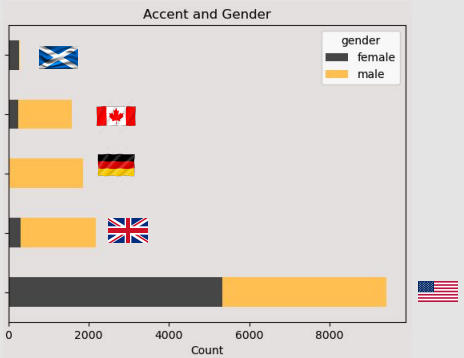
moz://a

OUR SUBSET OF MOZ

- GOAL: As much as possible, balance the demographic metadata for our subset of speakers
- Resultant demographic data represents the most balanced subset of individuals

This was accomplished through iteration. The resultant Mozilla voice dataset is balanced.

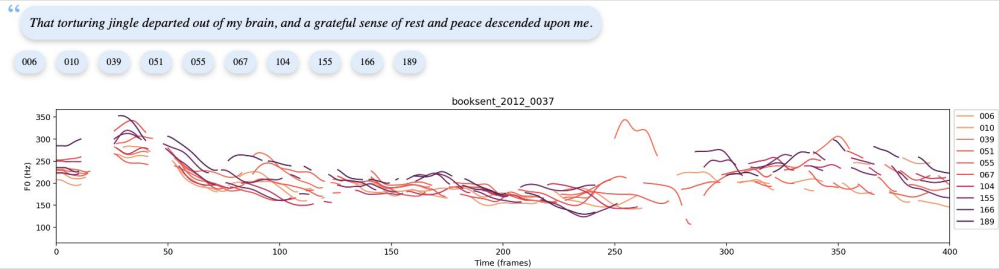




SAMSUNG

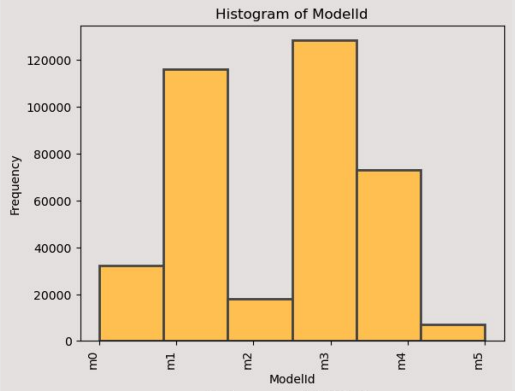
SOMOS

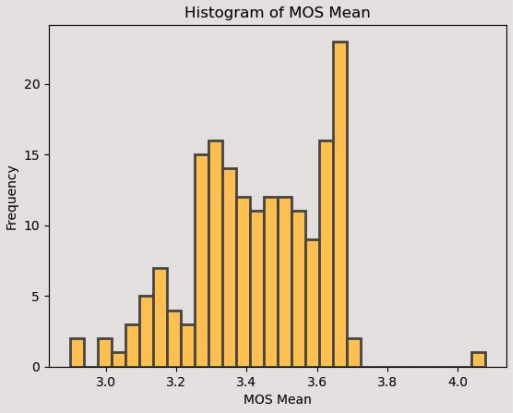
- Contains synthetic and natural utterances, and human-assigned scores to evaluate naturalness.
- The dataset includes over 250,000 utterances, each with a unique combination of system ID, utterance ID, listener's choice, and listener ID.
- The dataset was created by Samsung using the LJSpeech Dataset and is used to develop machine learning models for speech recognition



SOMOS BREAKDOWN

- This data is broken into two types: Human & Synthetic
- The source material for their TTS model is from LJSpeech and utilizes a single female speaker in her fifties
- The data has been validated with Amazon's Mechanical Turk
- This validation process provides us with a Mean Opinion Score (MOS) which measures the naturalness of the speech





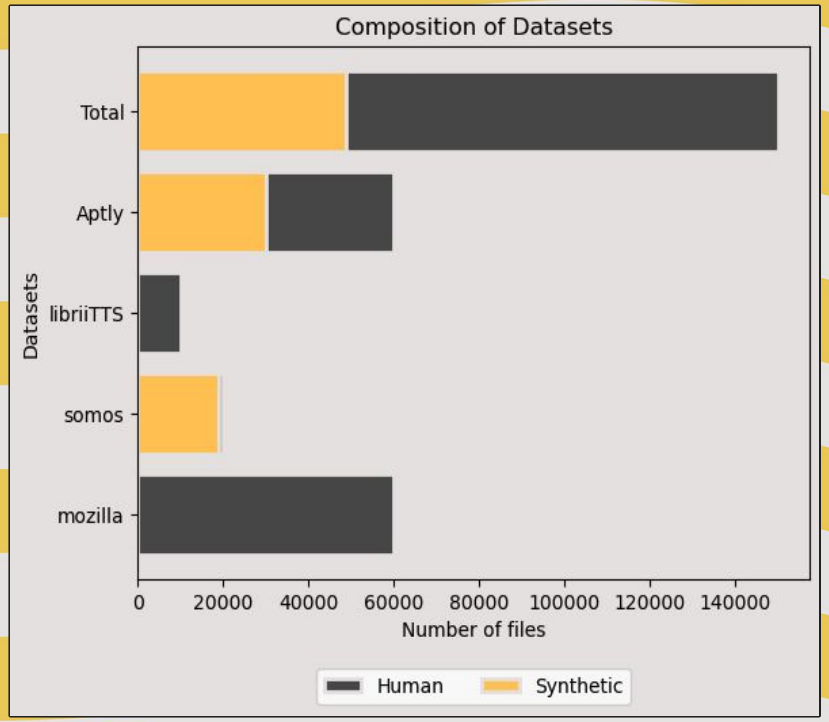
APTLY LABS (REAL OR FAKE)

- The Fake-or-Real (FoR) dataset contains approximately 60,000 utterances from real humans and computer generated speech. (equal split)
- Sourced text from the Arctic, LJSpeech, and VoxForge datasets
- Includes data from various TTS (Text-to-Speech) models such as Deep Voice 3 and Google Wavenet TTS
- There was no other metadata provided for this dataset

OpenSLR

LIBRI TTS DATASET

- The dataset is comprised of approximately 585 hours of English read speech from the LibriVox project, recorded in a professional studio environment.
- The dataset includes over 5,000 hours of speech from more than 2,500 human speakers and is annotated with speaker and book information.
- Each utterance is aligned with its corresponding text transcription, making it suitable for training and evaluating text-to-speech systems.
- The model will use about 1/3 of the total dataset.



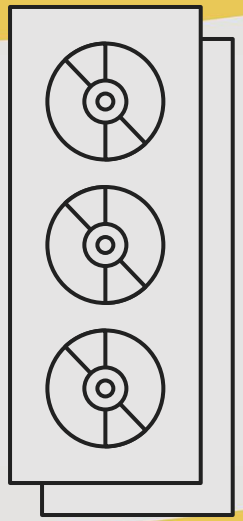
FINAL DATASET

This is what our final dataset looks like

CLEANING

What steps did we take to improve the quality of our data?

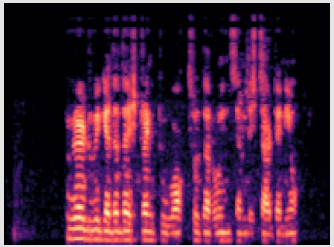
1. Removed low MOS value files
2. Normalized the audio using RMS norm
3. Removed files with no/low audio signal
4. Added background noise and variance
5. Applied several filters, such as:
    a. High Pass
    b. Pitch shifting
    c. Low level reverb



EXPLORING

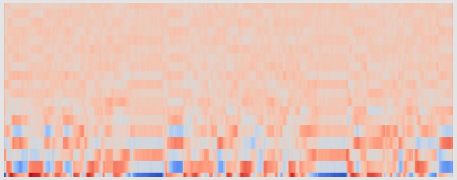
What can we find without computers?

WHAT ARE WE LOOKING FOR?



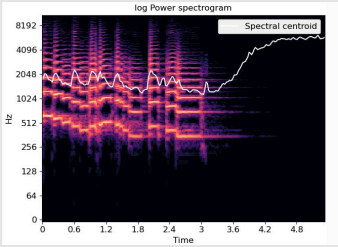
SPECTROGRAM

Representation of the frequency content of a sound signal over time



MFCCS

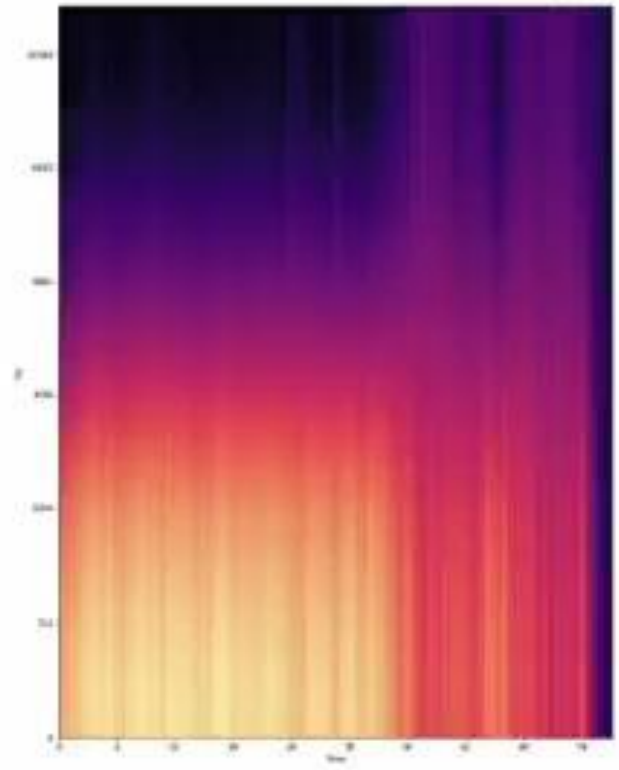
Mel-frequency cepstral coefficients represent the spectral envelope of a sound signal in a compressed form

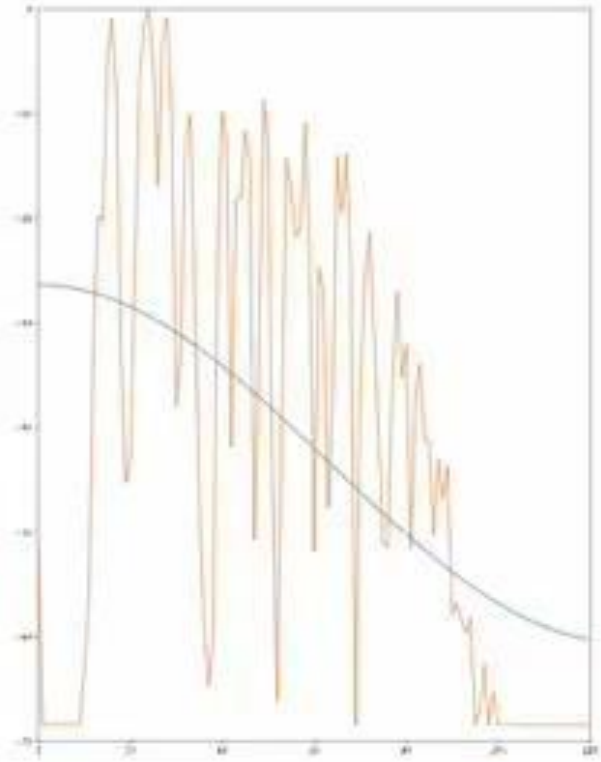


SPECTRAL CENTROIDS

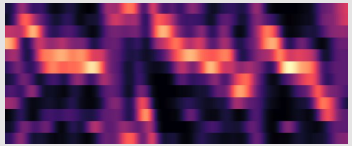
tell us about the "brightness" or "darkness" of a sound in a particular frequency range

2 MFCCs



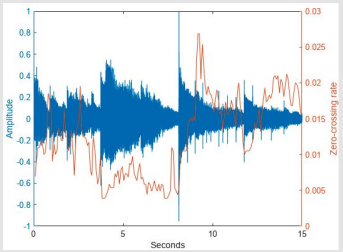


WHAT ARE WE LOOKING FOR?



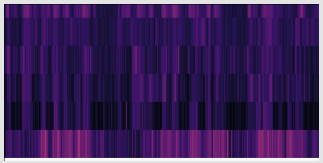
CHROMA

Helps to identify and distinguish between different notes and pitches in a person's voice as it changes over time



ZERO CROSSING RATE

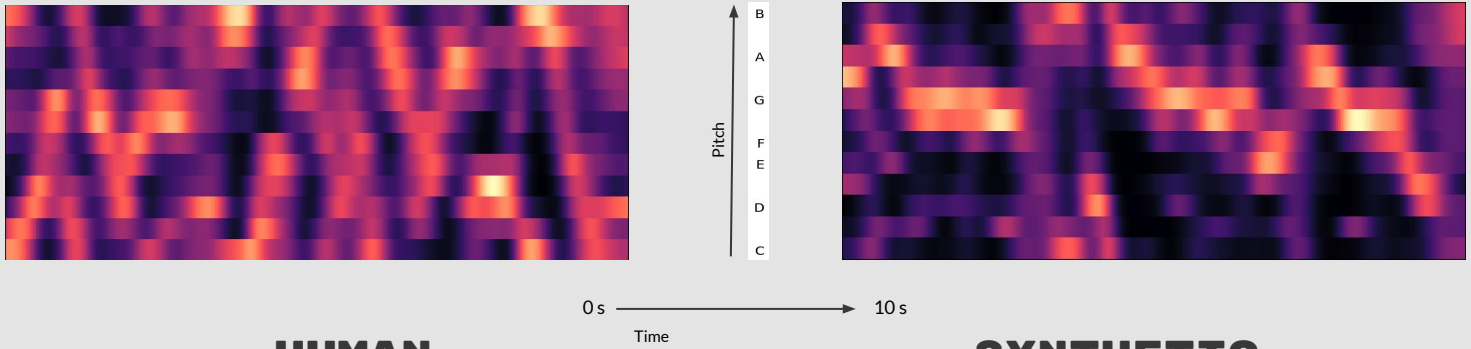
Refers to the rate at which the speech signal changes from positive to negative or vice versa



SPECTRAL CONTRAST

Measures the difference in magnitude between peaks and valleys in the frequency spectrum

WHAT WE CAN SEE

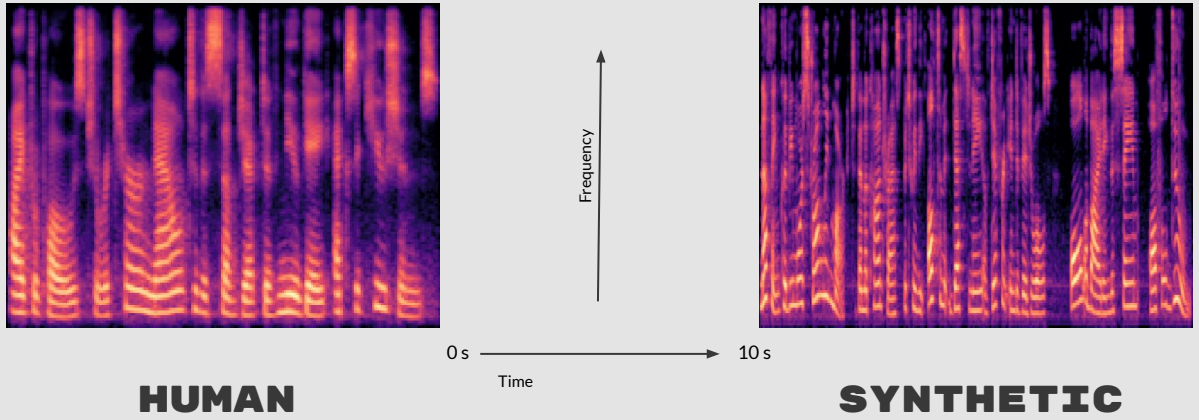


HUMAN **SYNTHETIC**

CHROMA

Provides a representation of the harmonic content of a person's voice

WHAT WE CAN SEE

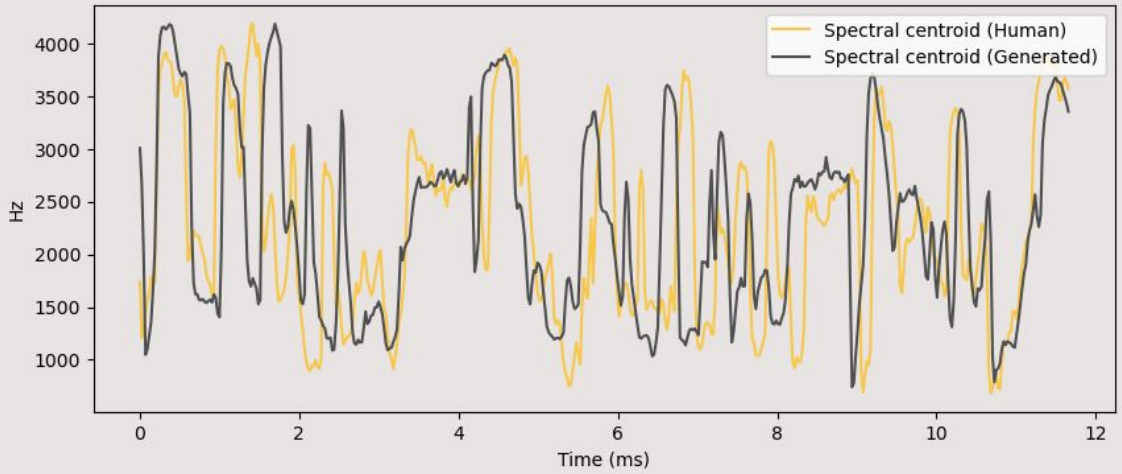


HUMAN **SYNTHETIC**

SPECTROGRAM

Representation of the frequency content of a sound signal over time

WHAT WE CAN SEE



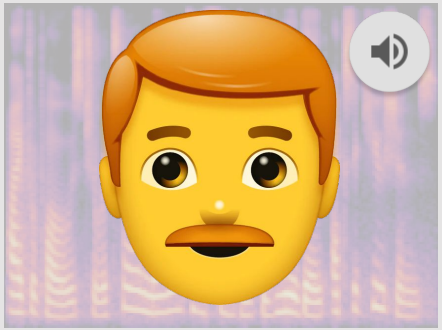
ZERO CROSSING

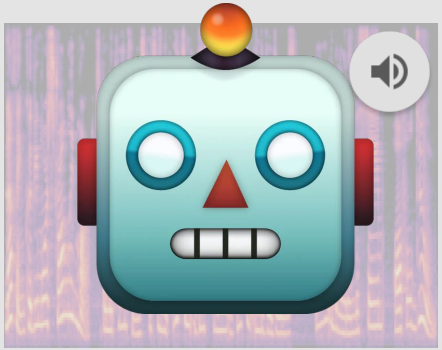
Refers to the rate at which the speech signal changes from positive to negative or vice versa

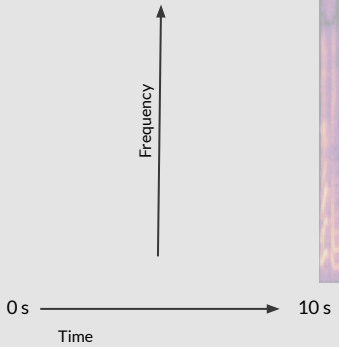
THE ANSWER

VERY LITTLE

WHAT WE HEAR?







HUMAN

SYNTHETIC

AUDIO

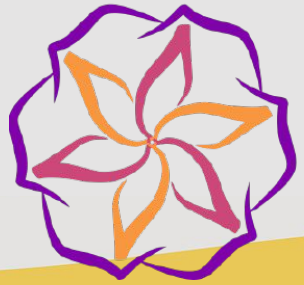
Both the TTS model and the human were given the same script

Source : Dall-E 2

ENTER: MACHINE LEARNING

What can our computers see that we can't?

HOW WE GOT OUR METRICS



Pydub

LIBROSA

Turns Audio into numbers

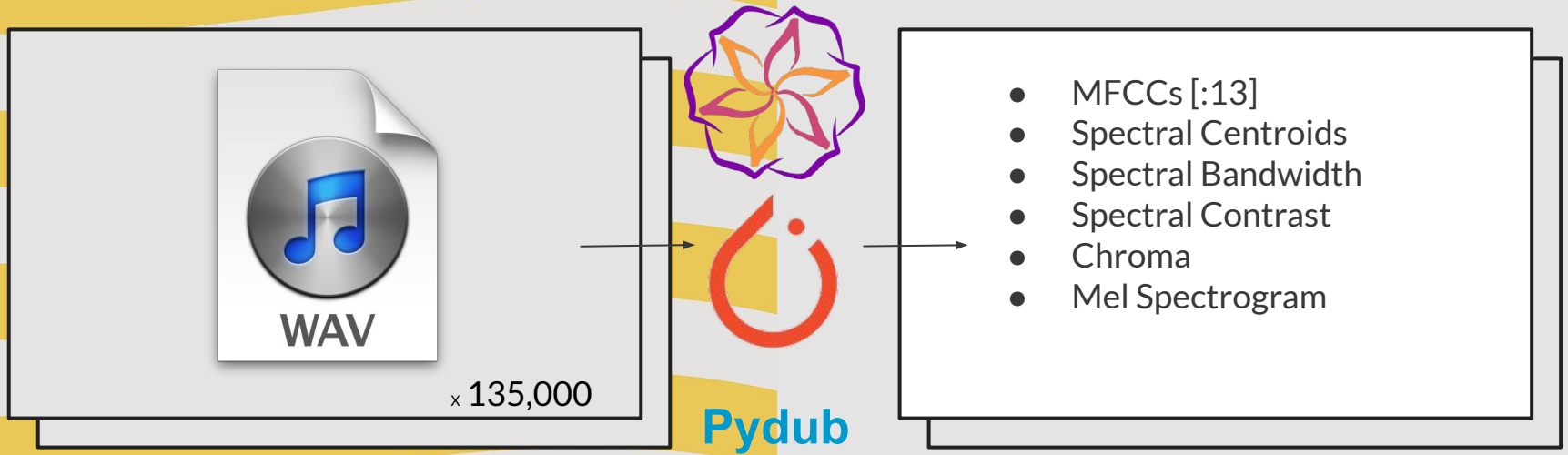
PYTORCH

Like Librosa, but lets you use your GPU

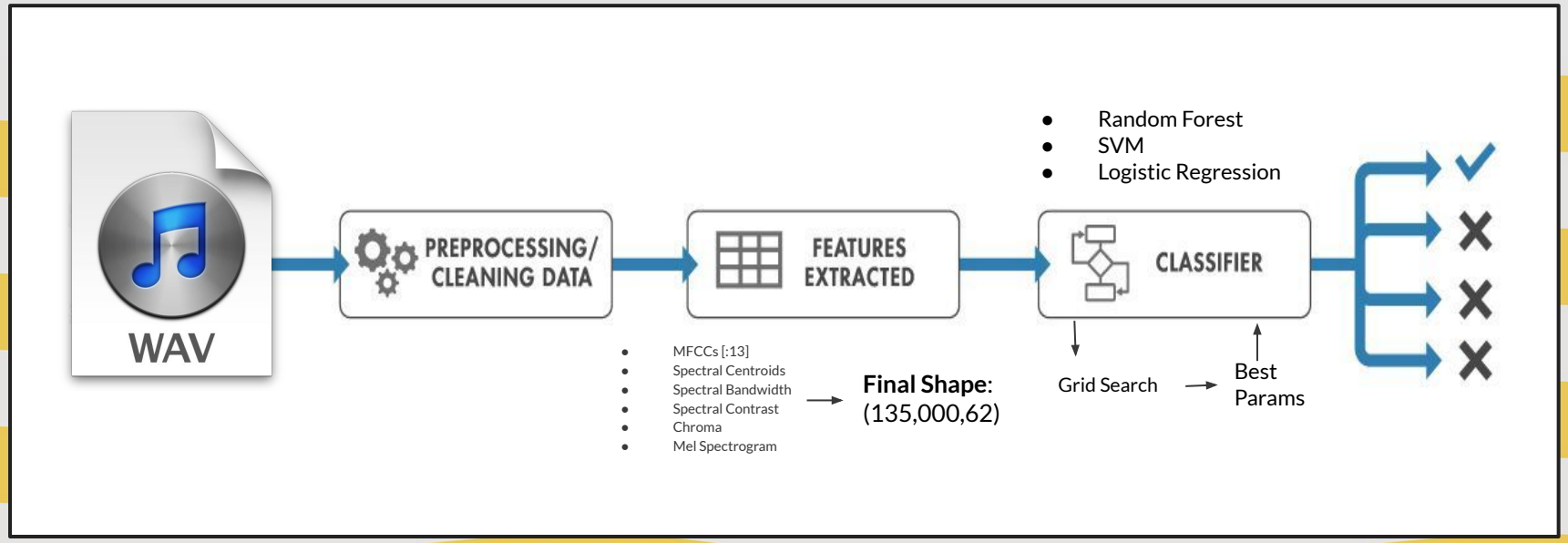
PYDUB

Pre-emphasis & manipulation

CONVERTING SOUND TO NUMBERS



CLASSICAL MM MODELS



SVM

Accuracy: 81%
Recall: 80%
F1: 78%

RANDOM FOREST

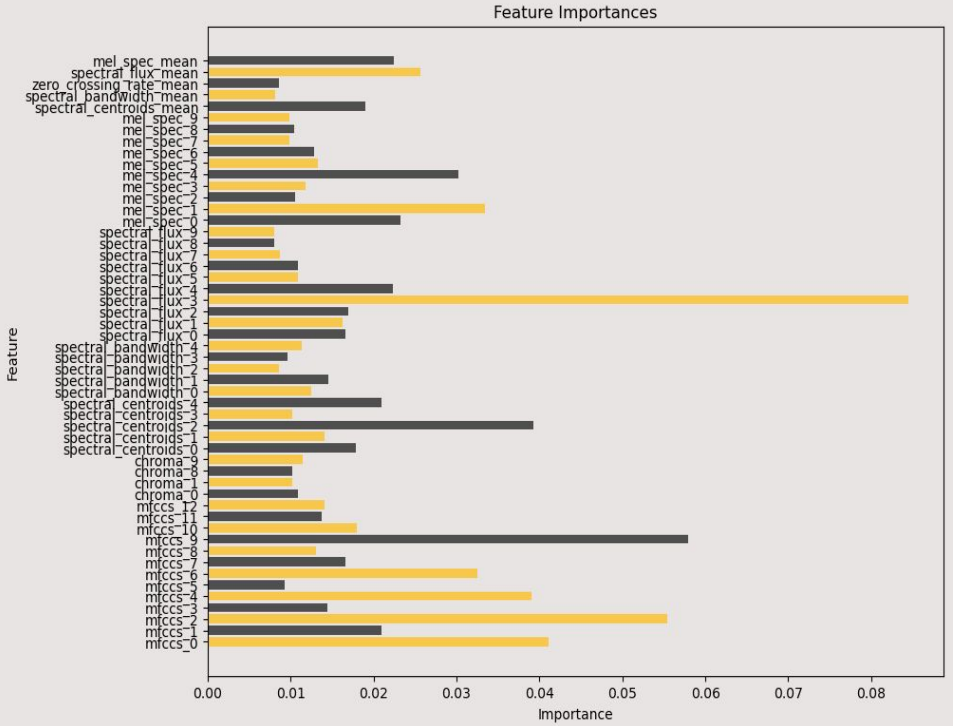
Accuracy: 90%
Recall: 90%
F1: 90%

LOGISTIC REGRESSION

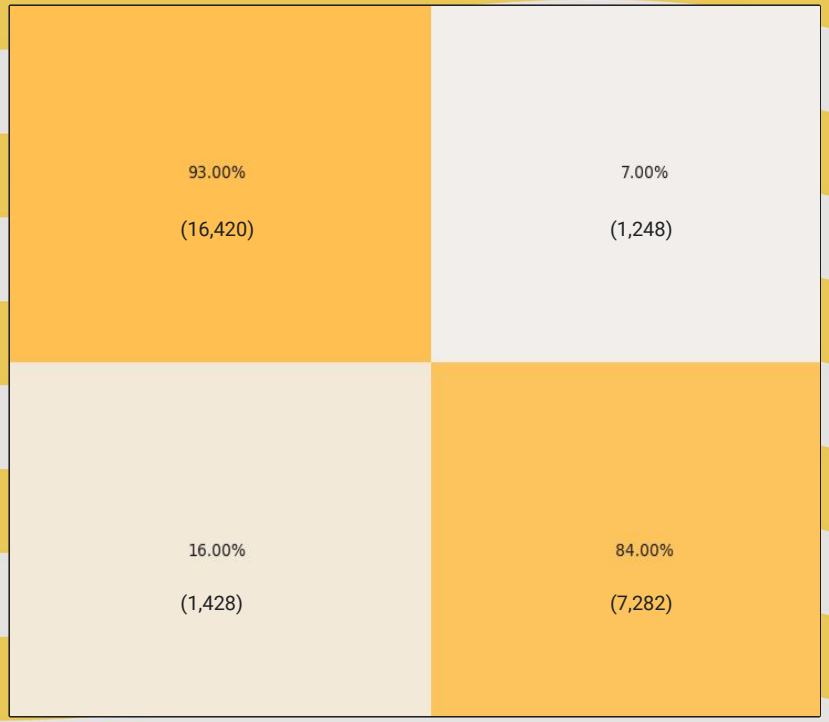
Accuracy: 79%
Recall: 78%
F1: 78%

Baseline: 67%

- MFCCs [:13]
- Spectral Centroids
- Spectral Bandwidth
- Spectral Contrast
- Chroma
- Mel Spectrogram



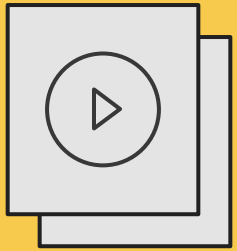
Random Forest –Confusion Matrix



RF- MODEL RESULTS

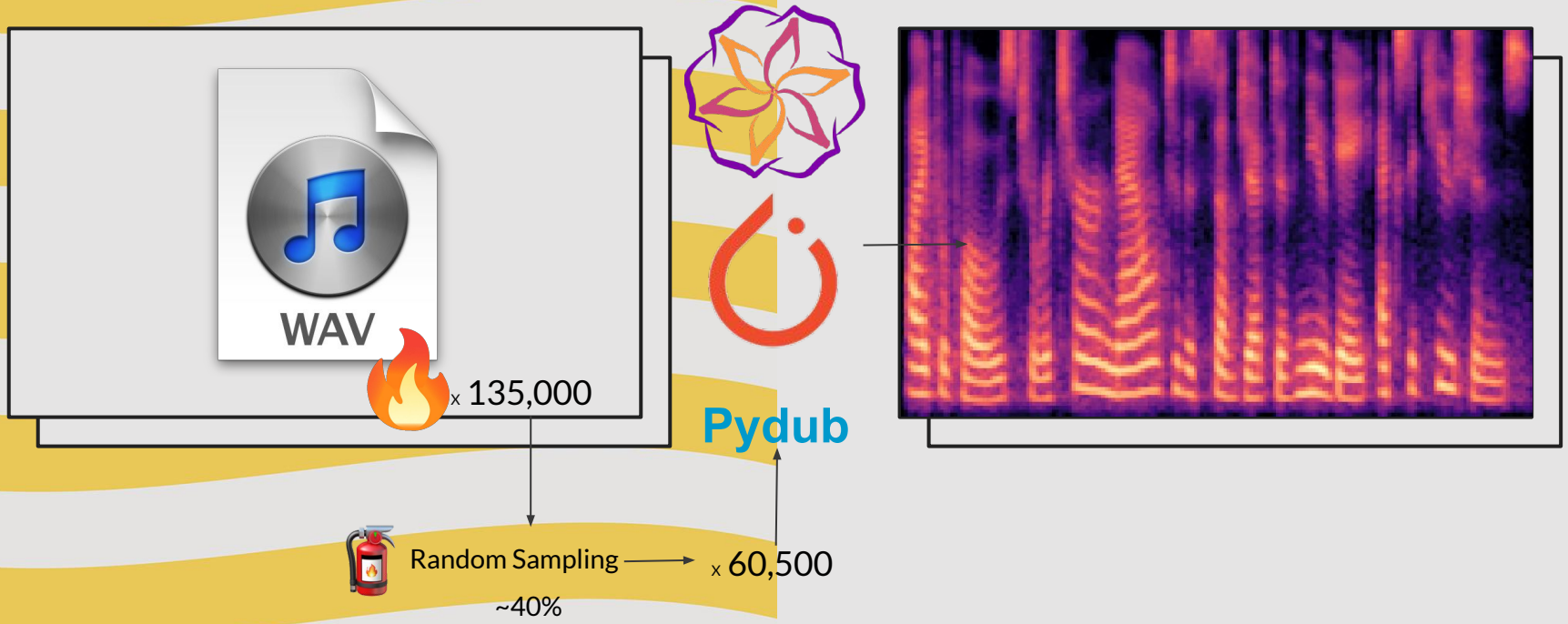
Grid Search Parameters:

```
param_grid = {
    'n_estimators': [100, 200, 300],
    'criterion': ['gini', 'entropy'],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'max_features': ['sqrt', 'log2']
}
```

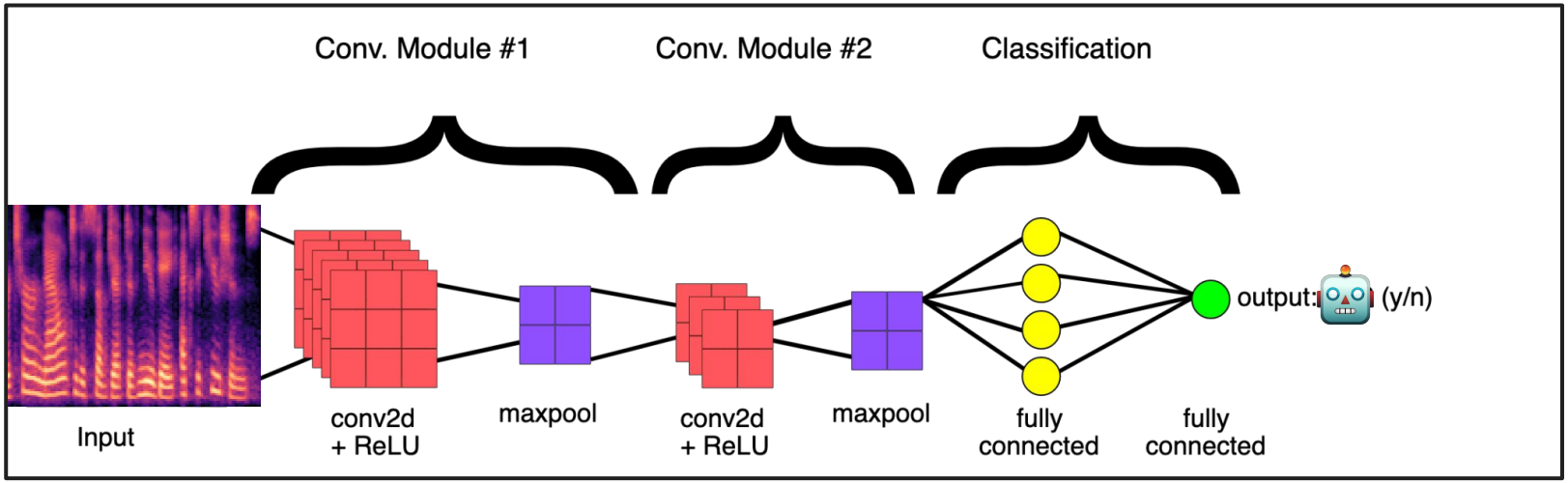


CONVOLUTIONAL NEURAL NET

CONVERTING SOUND TO IMAGES



CONVOLUTIONAL NEURAL NET



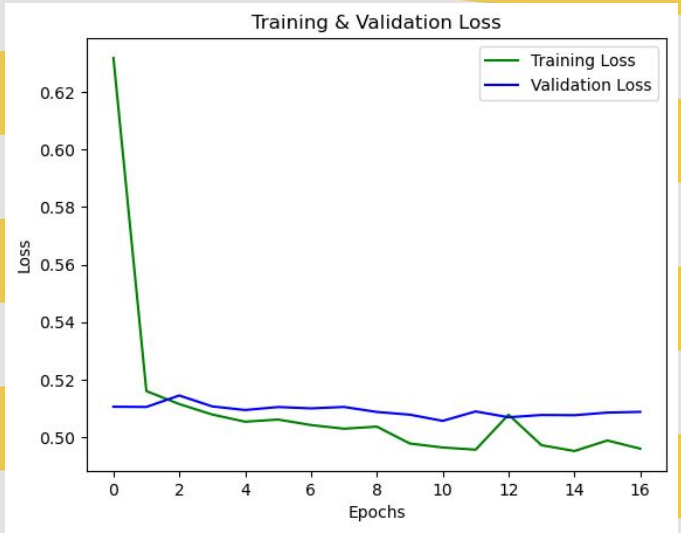
Source : Govinda Dumane

WHAT'S THE DIFFERENCE?

Can a CNN tell the difference?

YEAH, MOSTLY

Results:
Accuracy: 82%
Validation Loss: .52

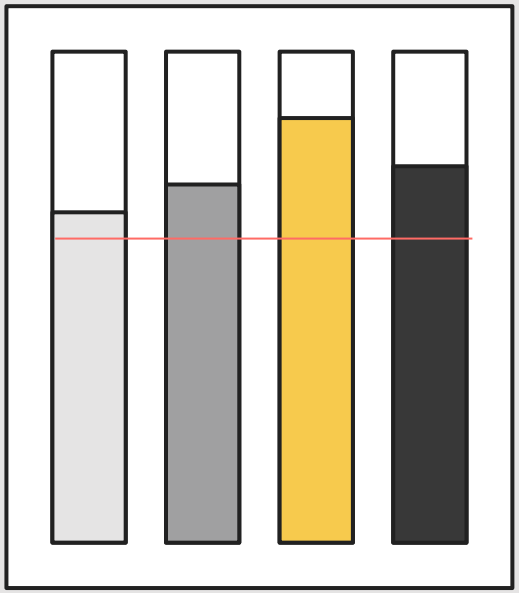



```
model = Sequential([
    Conv2D(32, (3, 3),
activation=LeakyReLU(alpha=0.05),
input_shape=(X_train.shape[1:])),
    MaxPooling2D(2),
    Conv2D(64, (3, 3),
activation=LeakyReLU(alpha=0.05)),
    MaxPooling2D(2),
    Flatten(),
    Dense(128, activation='relu'),
    Dropout(0.5),
    Dense(2, activation='sigmoid')
])
```

- Hardware limitations
  - Compressed PNGs
- Dataset quality
  - APTLY TTS not good
- Dataset diversity
  - Moz too dominant

WHAT MADE THE CNN WORSE?

SUMMARY OF MODELS



78%

LOGISTIC REGRESSION

Precision : 80%
F1: 78%

90%

RANDOM FOREST

Precision :90%
F1: 90%

81%

SVM

Precision: 80%
F1: 80%

82%

CNN

CONCLUSION

Is it possible?

Yes!

- Developing a machine learning model to discern between synthetic speech and human speech is entirely possible
- The models discussed previously all were better than the baseline
- Further tuning of the models could achieve better results
- Training a model on data from a specific situation would yield superior results

RECOMMENDATIONS

What could we do better?

1. Apply transfer learning from pre-trained models
2. Buy a better computer
3. Experiment with more complex neural network architectures.
4. Conduct more extensive hyperparameter tuning to optimize model performance.
5. Explore the impact of different feature extraction techniques, such as learning features directly from raw audio signals or using other types of acoustic features.

NEXT STEPS

What I plan to do next

1. Select only first 2 seconds of each clip
2. More aggressively trim the datasets
3. Train a TTS model and compare
4. Evaluate differences across demographics
5. How does text sentiment affect the result?

THANKS !

References:

1. Fayek, H. (2016). Speech Processing for Machine Learning: Filter banks, MFCCs, and What's In-Between.
2. McFee, B. et al. (2021). librosa: Audio and music signal analysis in Python.
3. Zero Crossing Rate. (n.d.). In ScienceDirect.
4. Fedden, L. (2019). Comparative audio analysis with WaveNet, MFCCs, UMAP, t-SNE, and PCA.
5. Mel Frequency Cepstral Coefficients. (n.d.). In Apple Machine Learning Research.
6. Burnwal, S. (2021). Speech Emotion Recognition.
7. Bhatia, K. (2019). Audio Signal Feature Extraction and Clustering.
8. Ambient Sounds. (n.d.). In Sound Jay.
9. English Speech Recognition and Text to Speech. (n.d.). In OpenSLR.
10. York Audio-Visual Speech Recognition Corpus. (n.d.). In York University.
11. Prosise, J. (2021). Audio Classification (CNN).
12. OpenAI. (2021). GPT-3.5.