**SOLUTION DESIGN DOCUMENT**

**AI Voice Impersonation Detection Real/Fake Design**

A diagram of a dropout layer

Description automatically generated

We use the Fake-or-Real Dataset - <https://bil.eecs.yorku.ca/datasets> to train our Fake/Real Model. The Fake-or-Real (FoR) dataset is a collection of more than 195,000 utterances from real humans and computer-generated speech. The dataset can be used to train classifiers to detect synthetic speech.

For the model, we use deep learning AI ML to classify. We do Feature Extraction of important voice features from the huge York University audio dataset and feed to a neural/deep network to create a model to identify real/deep faked audios.

The Features are saved in TRAINING and TEST CSV’s used as Input to create the MODEL.

1. INPUT: We have wave files from our dataset and we do preprocessing.

2. Feature Extraction: We extract the important speech features using PYTHON LIBROSA package.

The main features we relied on were **RMSE (E for energy), Zero crossing rate, Spectral centroid, Spectral roll off, Chroma, MFCCs (we took**

**20).**

* + Root Mean Square Energy (RMSE): Represents the energy of the signal, and shows how loud the signal is.
  + Zero crossing rate: It indicates the number of times the value of the signal changes between positive and

negative and vise versa. It is also used to measure the noise in a signal, and it usually gives high value in case

of a noisy signal.

* + Spectral centroid: It is a feature based on frequency which indicates the location of the center of mass of the

spectrum. In audios it is known as a good predictor of “ brightness” of a sound

* + Spectral roll off: This feature is used to differentiate between the harmonic sound (below roll off) and the

noise sound (above roll off). It is known as the energy spectrum under a specific percentage that is defined by

the used (85% by default).

* + Chroma: representation for audio where the spectrum is divided onto 12 bins representing the 12 distinct

semitones (or chroma) of the musical octave.

* + Mel Frequency Cepstral Coefficients (MFCC). MFCCs are widely used features for speech recognition. The Mel-frequency scale represents subjective or perceived pitch as its construction is based on pairwise comparisons of sinusoidal tones.

We use the feature set and partition it to training, validation and testing sets using **“from sklearn.model\_selection import train\_test\_split”.**

We use **“from sklearn.preprocessing import StandardScaler”** to normalize the data.

The Feature Scaling and Normalization process is to transform features to a similar scale/range so that no feature is dominating others in the ML Algorithm and thus improving the accuracy of the model.