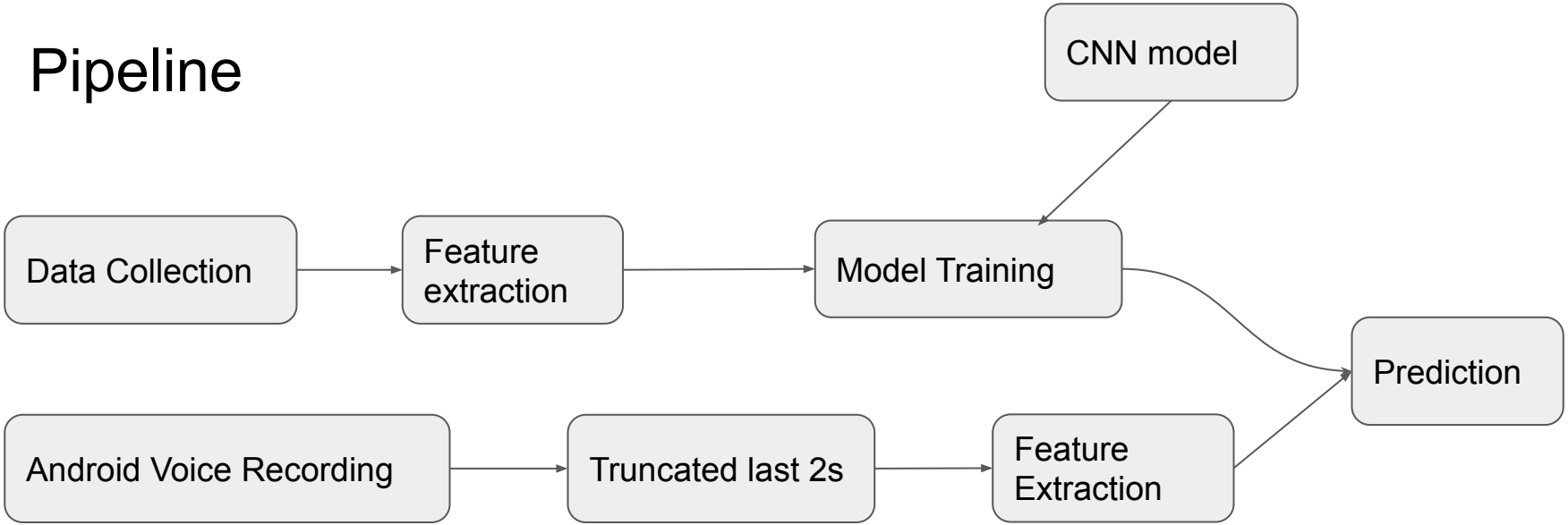
Voice Classification App

Zaikun Xu
ETH PhD Interview

Pipeline


CNN model
Data Collection
Feature extraction
Model Training
Prediction
Android Voice Recording
Truncated last 2s
Feature Extraction

Data Collection

1. Data Download
    a. Download youtube videos as wav files for singing and speaking categories
    b. Record silence data with some background noises using smart phone
    c. To make the dataset diverse enough, chinese, english and german languages with men as well as woman voces are included

2. Split into Training and Validation, Testing
    a. with limited desktop computational power, 3257 audio clips are used for training and 337 clips for validation 284 clips for testing (clips belong to one audio will be in the same training or validation or test)

3. Audio Clip generation
    a. Cut each big wav files into 3-4 second chunk During training, validation or testing , sample 2s from the audio chunk

Feature Extraction

1. For training, extract mfcc features with librosa library, num of mfccs = 40

2. On Android, since librosa is not available, use third-party implementation with same parameters as training

3. Pad features if necessary with fix length for training (input dim = 1280)

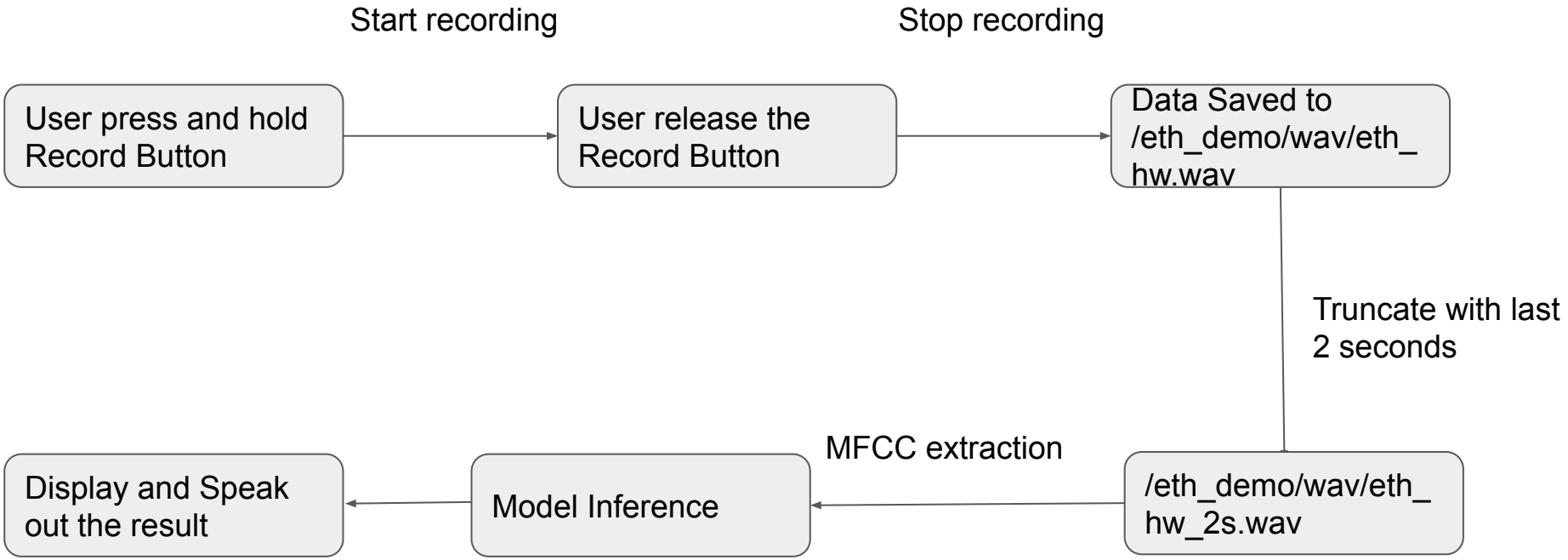
Model Training and Transformation

1. CNN
    a. Train a simple cnn with 3 conv2d layer + 2 fully connected layer + softmax , loss with cross-entropy and adam as optimizer

    b. Validation/Test acc is 78%/80% (improved with more data collected)

    c. Transformation model as tflite and put into asset folder of android app

Android User Interaction Logics



Future Work

1. Collect more data and train with a deeper network
2. Extract more features and fuse with mfcc for training
3. Try out other machine learning approaches
4. Improve UI