Automatic Speech Recognition: Assignment

Presented by Elias Hossain

Table of Content

- Speaker Identification
- Speaker Diarization (non-overlapping region)
- Speech Diarization (overlapping region)
- Speech-to-gender Recognition

1. Speaker Identification

Fundamental Information

- **Speaker identification** is a task of identifying persons from their voices.
- It is known that a speaker’s voice contains personal traits of the speaker, given the unique pronunciation organs and speaking manner of the speaker, e.g. the unique vocal tract shape, larynx size, accent, and rhythm.
- Modern computational approaches are currently being utilized to measure voices of persons automatically and it is termed as **“Automatic Speech Recognition.”**
- It is used for the voice-based authentication of personal smart devices, such as cellular phones, vehicles, and laptops.
- Recently, deep neural network based approaches are placing top priority in the research community to achieve the task of identifying speech automatically.

Speaker Identification

Phase 1	*Loading audio samples*
Phase 2	*Extracting features*
Phase 3	*Modeling: ML/DL based network*
Phase 4	*Training the model*
Phase 5	*Testing the model*

Extracting Features: MFCC Technique

- The Mel frequency cepstral coefficient (MFCC) is commonly used to extract the features.
- MFCC, which maps the signal onto a non-linear Mel-Scale.
- Feature extraction helps feed input to ML algorithms as features are converted into numeric/vector form to make it more meaningful for the computer to understand.

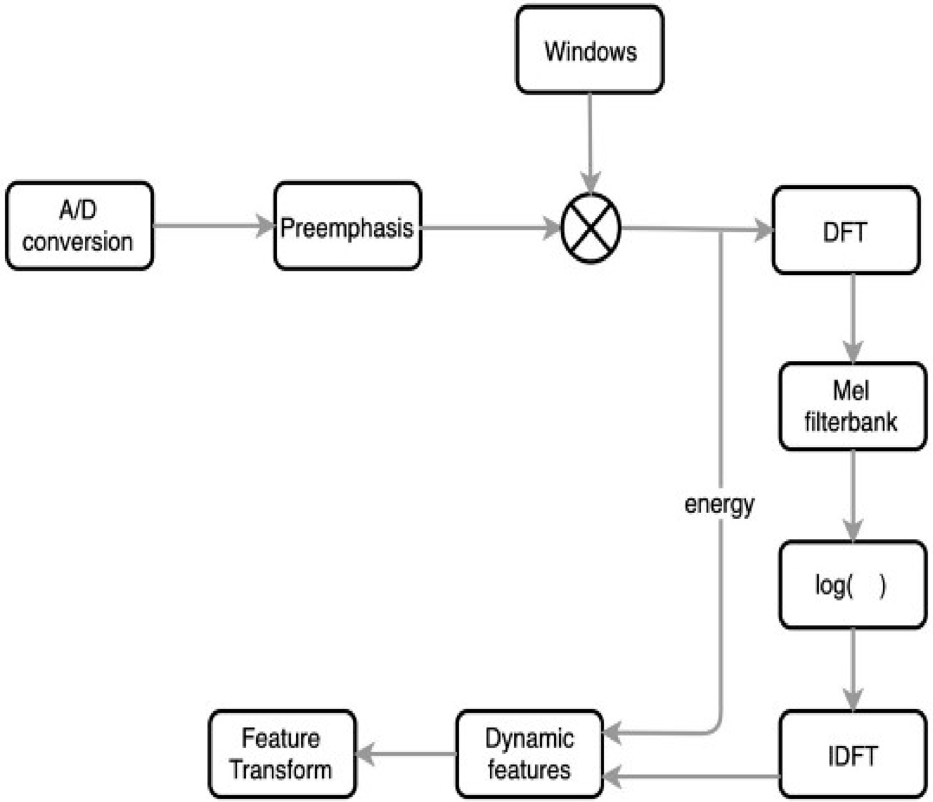

Figure 1: The road map of the MFCC technique

Selecting Model: One-Shot-Learning

- It is a special category of convolutional neural network called "Siamese neural networks **(SNNs).**"
- Assess the similarity and differences between the two images.
- One-shot learning aims to teach the model to set its own assumptions about their **similarities** based on the minimal number of visuals.
- Siamese neural networks are trained to **evaluate** the **distance** between features in two input images.
- Training an SNN for one-shot learning involves two stages: **verification** and **generalization.**

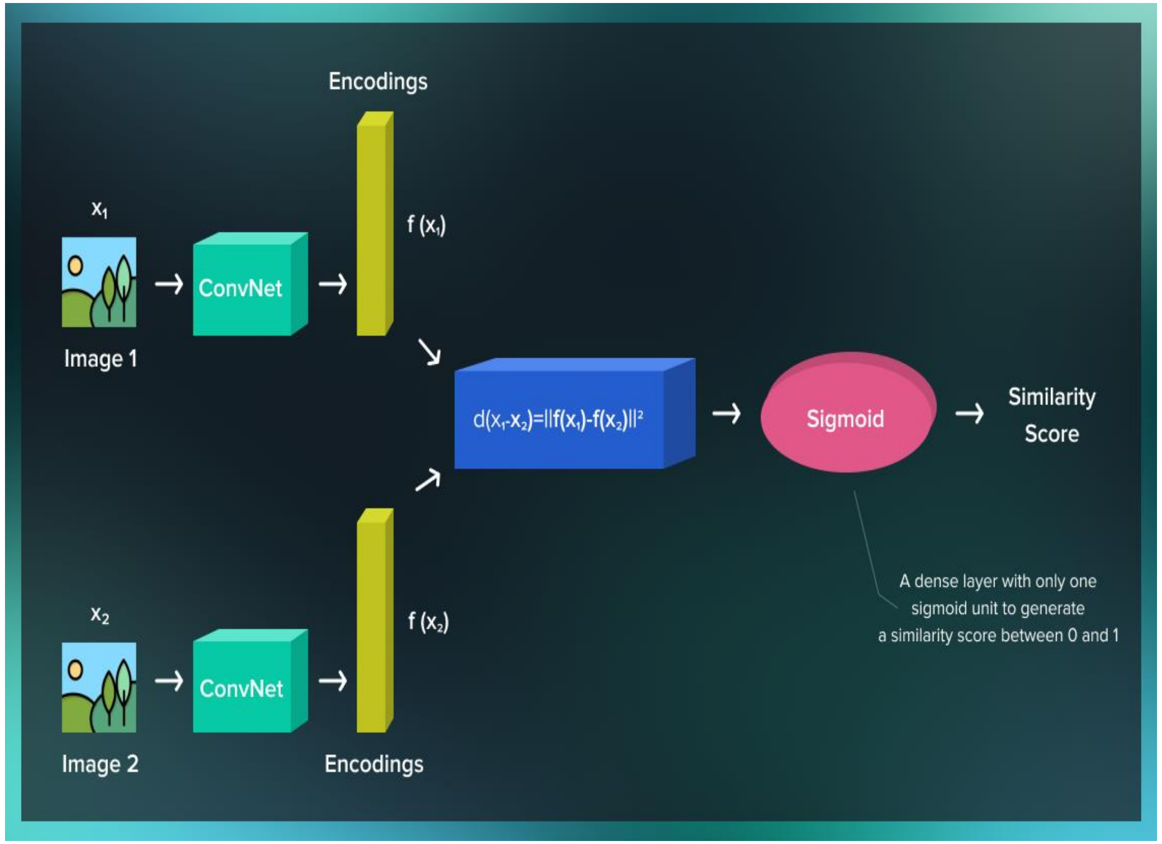
Encodings
x_1
Image 1
ConvNet
f (x_1)
$d(x_1\text{-}x_2)=||f(x_1)\text{-}f(x_2)||^2$
Sigmoid
Similarity Score
A dense layer with only one sigmoid unit to generate a similarity score between 0 and 1
f (x_2)
x_2
Image 2
ConvNet
Encodings

Figure 2: Architecture of the Siamese neural networks

2. Speaker Diarization (overlapping region)

Speaker Diarization: Working Flow

Speaker diarization systems consist of 3 main blocks:

- The voice activity detection module (VAD) [Hybrid energy based detector and model based decoder]
- The feature extraction module **[MFCC]**
- Clustering and Segmentation framing

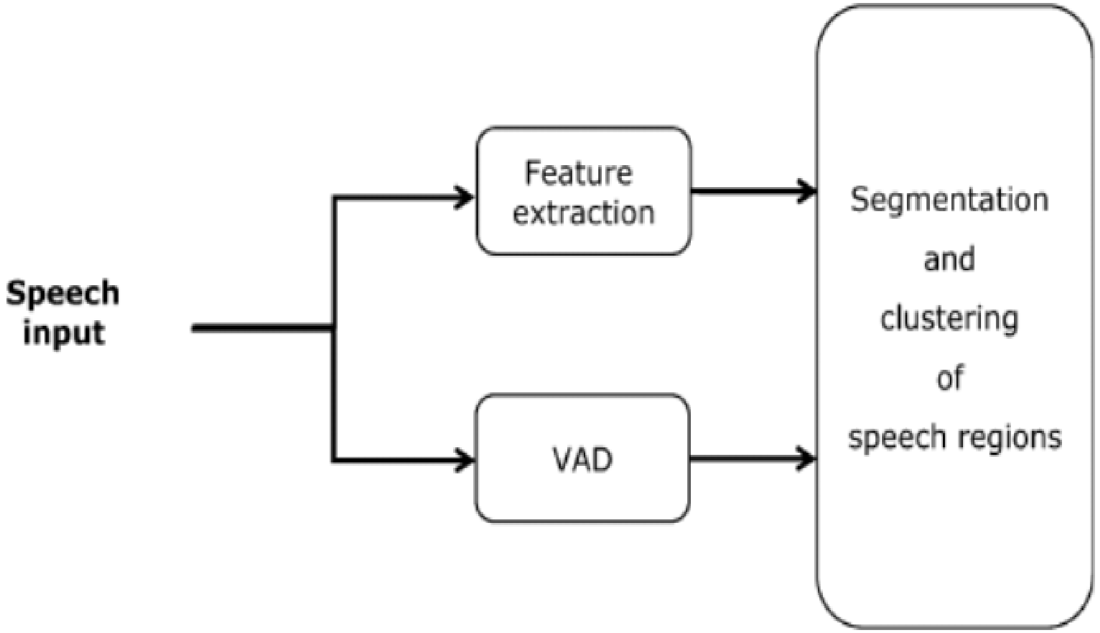

Figure 3: Simplified diagram of speaker diarization system

Speaker Diarization: Implementation

- **Speech Detection:** It is recommended to use the **pyannote.metrics library.**
- **Speech Segmentation:** This is achieved by segmenting the audio into windows with overlap. The size of the window determines the size of the segment. if the window size is 2 seconds, and set an overlap of 0.5 seconds, first window would be : **(start = 0.0s , stop = 2.0s),** next window will be: **(start = 0.5s, stop = 2.5s)** … and so on until full audio is covered.
- **Embedding Extraction:** We need to find MFCC (Mel Frequency Cepstral Coefficient) of the audio segment. The SciPy library of python has a separate module for finding MFCCs. In the next step, we need to apply the **LSTM based network** which takes in the MFCCs and outputs a vector representation (embedding) which is called a d-vector.
- **Clustering:** Clustering is an Unsupervised machine learning method which tries to create clusters (or groups) of data in an n-dimensional space. However, it is suggested to use **Spectral Clustering algorithm.**

Speaker Diarization: Overlapping Region

The following approaches are applied in a research paper to detect the overlapping region before clustering toward improving the performance of the speaker diarization system.

✓ Assigning speaker labels in overlap regions according to the labels of the neighboring segments.

✓ In addition, the use of **cross correlation features** with **MFCC’s** reduces the performance gap due to overlaps, so that there is little gain from removing overlapped regions before clustering.

✓ Another way is to deal with the overleaping region is to **pre-process** the overlapped speech signal with a **source separation algorithm.**

✓ Spectral autocorrelation peak valley ratio (SAPVR) approaches also used by many researchers to solve the underlined problem.

✓ Mel-warped cepstral coefficients **(MFCC's)** methods are currently being applied by the research community.

Speaker Diarization: Non-overlapping Region

Nowadays **Resemblers** are considered for many voice recognition tasks and the following features distinguish them from all others:

- **Resemblyzer** allows us to derive a high-level representation of a voice through a deep learning model. However, it is considered as voice encoder.
- It is a python package to analyze and compare voices with deep learning.
- Resemblyzer can be used for speaker verification, diarization, fake speech detection, and more.
- Given an audio file of speech, it creates a summary vector of 256 values that summarizes the characteristics of the voice spoken.

3. Speech-to-gender Recognition

Gender Recognition: Working Flow

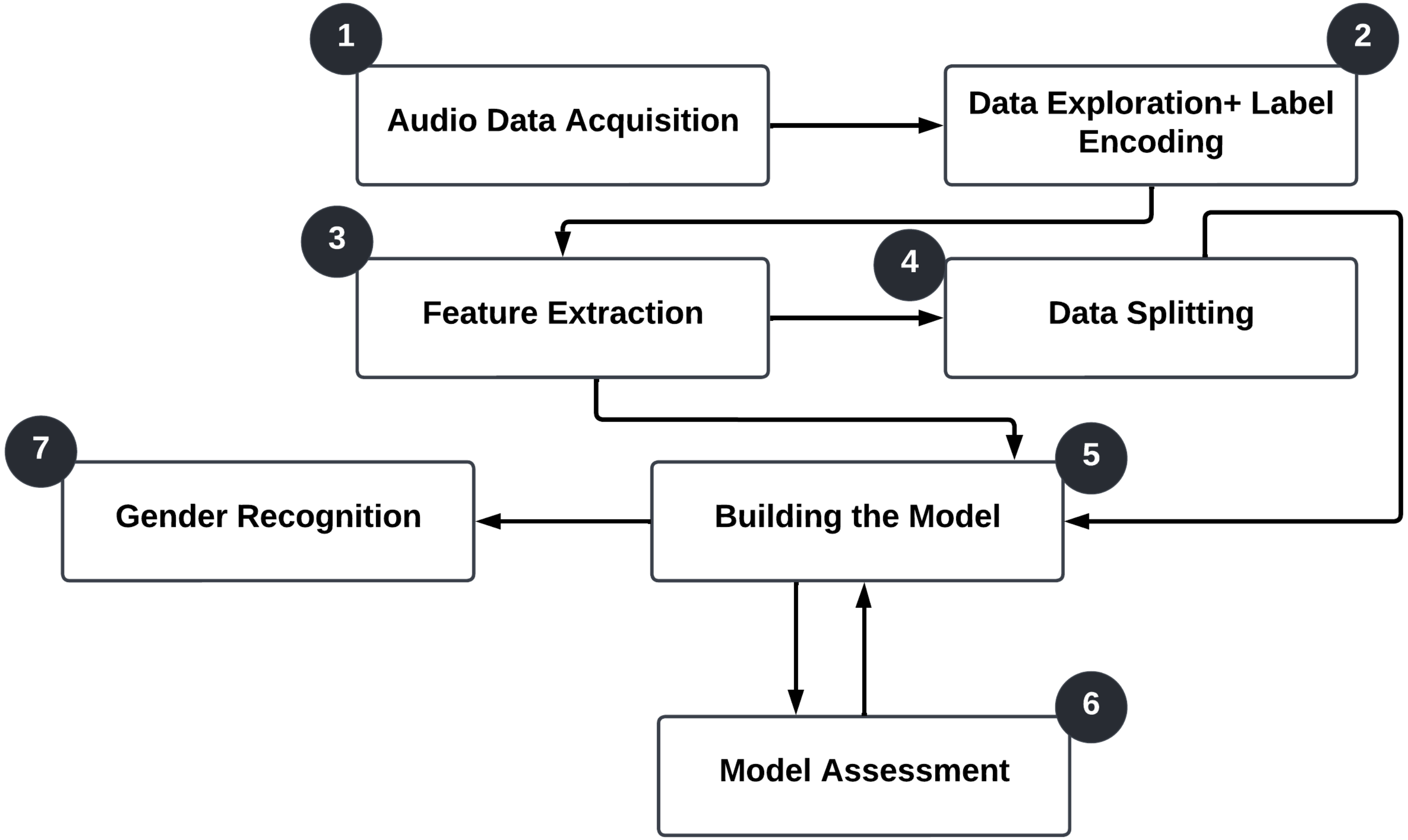

Figure 4: Overall steps of identifying gender based on voices

THANK YOU