Signal Processing Cup-2022

Subodha Charles, Darukeesan Pakiyarajah, Muhunthan Shandirasegaran, Heethanjan Kanagalingam
Ragavan Ravichandran, Prarththanan Sothyrajah, Thenukan Pathmanathan, Nirajkanth Ravichandran
Nerththiga Neminathan, Vithurabiman Sethuran, Mayooran Thavendra
Electronic and Telecommunication department, University of Moratuwa, Colombo district, Sri Lanka

I. Introduction

Digital Multimedia is becoming common today and the possibility of manipulating digital multimedia objects is becoming very common. For instance, fake synthetic speech audio tracks can be generated through a wide variety of available methods and algorithms. These range from simple cut-and-paste techniques, to complex neural networks. The goal is to design and develop a system for synthetic speech attribution.Given an audio recording representing a synthetically generated speech track, our approach should detect to which method among a list of candidate ones has been used to synthesize the speech.

II. Approach

A. Feature extraction

Speech recognition domain has been vastly evolved in recent times. Our team identified multiple features which is capable of recognising the distinct algorithms used to construct the above audios.

1. Mel- frequency Cepstral Coefficients (MFCC)
2. Characteristic Spectograms
3. Short term and long term (STLT) feature
4. Bi-coherence
5. Fundamental frequency of speech signal
6. Spectral envelope
7. Excitation signal
8. Linear prediction Cepstral Coefficients
9. Perceptual Linear predictive analysis
10. 1st and 2nd order differential coefficients of cepstrum
11. Relative Spectral Analysis filters (RASTA)
12. Gammatone Frequency Cepstral Coefficients (GFCC)
13. Pitch
14, Zero Crossing Rate (ZCR)

B. Feature selection

Meaningful features which corresponds to the required problem are selected by finding the testing accuracy using VGG16 and Resnet50 (convolutional neural network models) for a certain feature extraction applied to the whole data set individually. Extracting algorithm for most of the above mentioned features are already existed in famous libraries such as (Librosa,matpoltlib,scipy and spafe). Some of the extractors were constructed referring to the mathematical guidance mentioned in the corresponding research papers which introduced or enhanced the feature. MFCC and Characteristic spectograms portrayed significant accuracy among all.

Identify applicable funding agency here. If none, delete this.

C. Feature concatenation

The feature array of the above two features were interpreted into grey scale images and re-scaled to a common size using area interpolation method in openCV. The images were weighted and linear blended.

$$g(x) = (1-k)f_0(x) + kf_1(x)$$

g(x) : output image
fi(x): input image for feature i
k : scale

The final image was obtained by this pipeline was fed to VGG16 and classification is observed.

III. Extracted Features

A. MFCC

The main point to understand about speech is that the sounds generated by a human are filtered by the shape of the vocal tract including tongue, teeth etc. This shape determines what sound comes out. If we can determine the shape accurately, this should give us an accurate representation of the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope.

Mel Frequency Cepstral Coefficents (MFCCs) are a feature widely used in automatic speech and speaker recognition. They were introduced by Davis and Mermelstein in the 1980's, and have been state-of-the-art ever since. Prior to the introduction of MFCCs, Linear Prediction Coefficients (LPCs) and Linear Prediction Cepstral Coefficients (LPCCs) and were the main feature type for automatic speech recognition (ASR), especially with Hidden Markov Models (HMM) classifiers.

B. Characteristic Spectogram

The characteristics spectrogram is extracted by linearly superimposing several short time spectrogram of a given audio sample at the same frequency for a certain time period

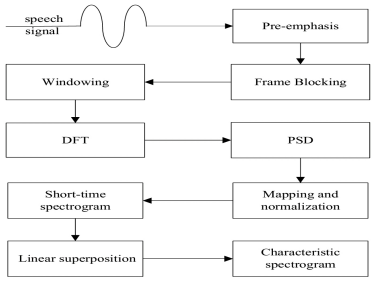

Fig. 1. Characteristic Spectogram

IV. Implementation pipeline

The implementation pipeline is visualized below.

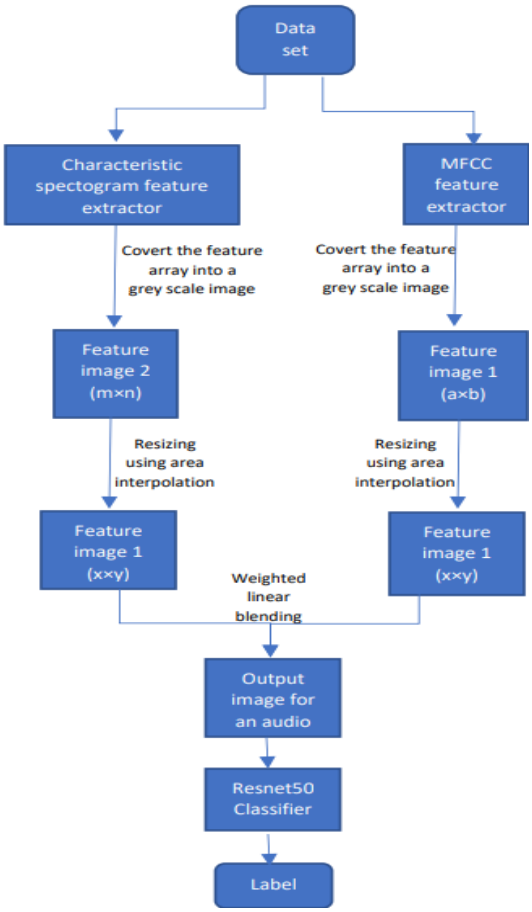

Fig. 2. Implementation pipeline

V. Results

TABLE I
Results Summary of the approaches

Feature Name	Model	Accuracy
LPC	Custom neural network	62
0.75*MFCC + 0.25*LPC	Custom neural network	59
STLT feature	Custom neural network	20
Energy plot(Amp vs time)	VGG16	73
MFCC image	VGG19	87.5
Characteristic Spectogram	Resnet50	57.8
MFCC	Resnet50	79.3
MFCC + Ch. Specto.	Resnet50	57.8

The approaches to the features are as follows;

- *LPC*-1D feature vector(1,13) was obtained for each audio and trained on 5000 training dataset
- *LPC + MFCC*-Linearly combined LPC(1,13) and MFCC(1,13) vectors for each audio and trained on 5000 training dataset
- *STLT*-1D feature vector(1,16) was obtained for each audio and trained on 5000 training dataset
- *Energy plot (Amplitude vs time) of audio*-Energy plot of each audio file was plotted using matplotlib library. Those 6000 images were used to train VGG16 model using transfer learning approach. A global max pooling layer, dropout (0.2) and a dense layer with softmax activation were added to the last layer of VGG16 and trained while freeing the other layers. Final model gave 43
- *MFCC image*-MFCC images of 6000 training audio samples were used for training. VGG19 model was used as a feature extractor here. 1D feature vector with the length 25088 was obtained from VGG19 and used to train a linearSVC model. It resulted in 77
- *MFCC + Characteristic Spectogram*- Unfortunately, the combination of the above two gave a poor accuracy. Finally, we planned to stick to the approach of using MFCC as 1-D array as the only feature. The feature vector was fed into a customized tensorflow CNN for classification.

References

[1] C.Borrelli, P.Bestagini, F.Antonacci, A.Sarti, and S.Tubaro, "Synthetic speech detection through short-term and long-term prediction traces", EURASIP Journal on Information Security, 2021.

[2] H.S.Kim, Linear predictive coding is all-pole resonance modeling, unpublished. Available: https://ccrma.stanford.edu/ hskim08/lpc/lpc.pdf

[3] J.Lyons (2013, March 12), "Mel frequency cepstral coefficient (MFCC) tutorial"[Online]. Available: http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/ .

[4] A. Mahmood (2019, April20), "Audio Classification with Pre-trained VGG-19 (Keras)"[Online]. Available : https://towardsdatascience.com/audio-classification-with-pre-trained-vgg-19-keras-bca55c2a0efe

[5] S.Amiriparian, *et al*. "Towards cross-modal pre-training and learning tempo-spatial characteristics for audio recognition with convolutional and recurrent neural networks" EURASIP Journal on Audio,Speech,and Music Processing,2020