Speaker Identification with Convolutional Neural Networks

Ashwij Kumbla
Robotics Engineering Dept
Worcester Polytechnic Institute
Worcester, MA
akumbla@wpi.edu

Shivaram Srikanth
Robotics Engineering Dept
Worcester Polytechnic Institute
Worcester, MA
ssrikanth@wpi.edu

Irakli Grigolia
Computer Science Dept
Worcester Polytechnic Institute
Worcester, MA
igrigolia@wpi.edu

***Abstract*—In today's world, there is a need for automatic recognition systems for a person based on different behavioral or physical characteristics. Speaker identification is a powerful, and inexpensive bio-metric technique. It is a classification task that aims to identify a subject from a given time-series sequential data. In this paper, we discuss the Convolutional Neural Network (CNN) approach for speaker identification. Several models were trained to investigate the correlation between the number of speakers and the amount of data needed to get the best prediction accuracy. Further, we investigate what minimal length of the voice message we need to obtain a reliable identification score.**

***Index Terms*—Convolutional Neural Network, speaker identification, MFCC, TSNE, HMM, CNN**

I. Introduction

In today's world, most bio-metric authentication systems are restricted to finger print scanning and retina scanning. Though the human voice is a metric to identify people due to the uniqueness of physiology of the glottis and the vocal chord, it is not widely used in the commercial space for identification. Building such a system, promises very useful user experience advancements for a wide range of applications. One such advancement could be a voice-based personal signature. But to achieve that, we need to have good speaker identification and verification systems in the first place.

The purpose of automated speaker recognition is to extract, classify, and recognize information about a speaker's identity such that he or she may be identified by his or her voice.Text-dependent and text-independent approaches can be used in the systems. Text-dependent systems need the speaker to say a certain phrase, whereas text-independent methods capture the qualities of the voice regardless of the text said.

This research focuses on building such a text independent system for speaker identification using CNN and further addresses some of the complexities involved in creating such systems.

II. BACKGROUND

Automatic speech recognition has been a subject of major interest for more than half a century. With the introduction of digital computing and signal processing, this topic became even more interesting for different people. Speech recognition has a wide range of applications, including voice-controlled apps with full speech-to-text software, automation of operator-assisted services, and voice recognition assistance for people with disabilities. In voice recognition, there are various approaches: Pattern recognition, acoustic-phonetic, machine learning, and neural network approaches are all viable choices. These modern approaches have evolved from decades of research on techniques as mentioned in [2] and [8], like the Hidden Markov Model (HMM) and Gaussian Mixture Models (GMM). Speech recognition has a high potential to play a significant role in human-computer interaction soon. A good speech recognition system must determine not just characteristics that are present in the input pattern at one moment in time, but also features that change over time. Following are of methods that we researched to create such a system.

A. Hidden Markov Model

During the research phase, different possible approaches and methods were considered, one of them was the Hidden Markov Model (HMM). HMMs are a type of probabilistic graphical model that allows us to predict a series of unknown (hidden) variables given a collection of observed variables. An HMM may be thought of as a Bayes Net that has been unrolled across time, with observations made at various time steps being used to forecast the optimum hidden state sequence. Decades ago, the key focus of a variety of useful acoustic modeling algorithms was HMM. The model's success was due to its analytic abilities in the speech phenomena, as well as its accuracy in real speech recognition systems. Another essential aspect of HMM was its convergent and reliable parameter training approach. In representing spoken utterances, a non-stationary sequence of feature vectors was utilized. To statistically evaluate a voice sequence, it was split into stationary sections.

But HMM has its cons. Here are a few of the key disadvantages:

- The presumption that subsequent observations are unrelated to one another. Subsequent observations are rarely independent of one another.

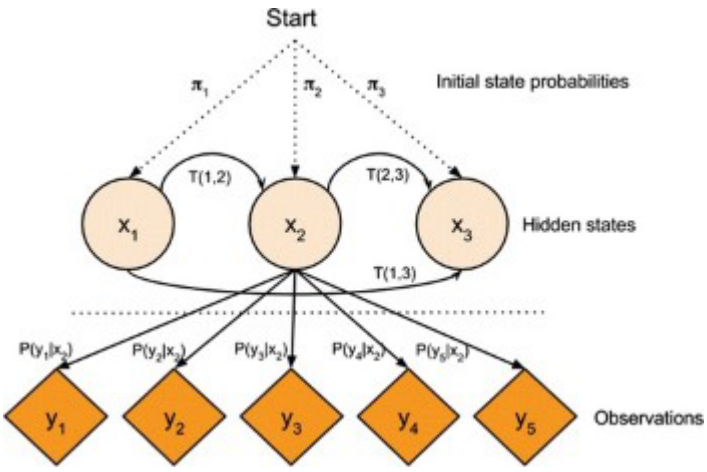

Fig. 1. Hidden Markov model

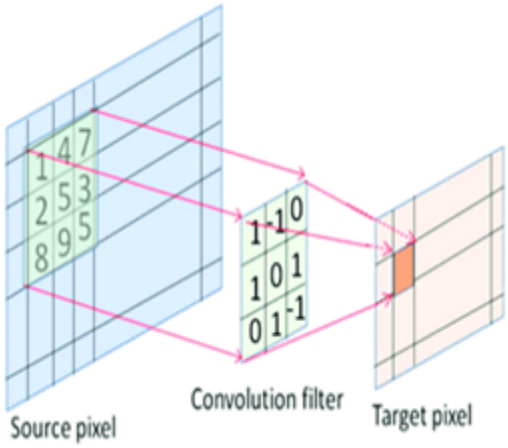

Fig. 2. Convolutional Layer

- Markov property asserts that the probability of existing in a particular state at time t solely depends on the state at time t-1. This is the basis for modeling. This is not always the case with speech sounds, where dependencies might span many stages.
- Observation frames of constant length. This constraint limits the options for feature extraction.

B. Convolutional Neural Network

A Convolutional Neural Network (CNN) is a Deep Learning system that can take an input image and assign importance to distinct features of the image, as well as differences between them. The amount of pre-processing required by a CNN is much less than that required by other classification techniques. CNN's can learn these filters/characteristics with adequate training, whereas simple techniques need hand-engineering of filters. Due to the reduced number of parameters involved and the reusability of weights, the architecture performs superior fitting to picture data sets. In other words, the network may be trained to better recognize the image's complexity. CNN's have 3 main types of layers: Convolutional Layer, Pooling layer, fully connected layer.

The convolutional layer as shown in figure 2 is the most important component of a CNN since it is where most of the processing takes place. It requires input data, a filter, and a feature map, among other things. Let's pretend the input is a color picture, which is made up of a 3D matrix of pixels. This implies the input will have three dimensions: height, width, and depth, which match the RGB color space of a picture. A feature detector, also known as a kernel or a filter, will traverse over the image's receptive fields, checking for the presence of the feature. Convolution is the term for this procedure.

The pooling layer as shown in figure 3 is a dimensionality reduction technique that reduces the number of factors in the input. The pooling process sweeps a filter across the whole input, like the convolutional layer, however, this filter does not contain any weights. Instead, the kernel uses an aggregation function to populate the output array from the values in the receptive field.

In partly linked layers, the pixel values of the input pictures are not directly connected to the output layer. Each node in the output layer, on the other hand, links directly to a node in the preceding layer in the fully-connected layer. This layer performs classification tasks based on the characteristics retrieved by the preceding layers and their various filters. While convolutional and pooling layers often utilize ReLU functions to categorize inputs, FC layers typically use a SoftMax activation function to provide a probability from 0 to 1.

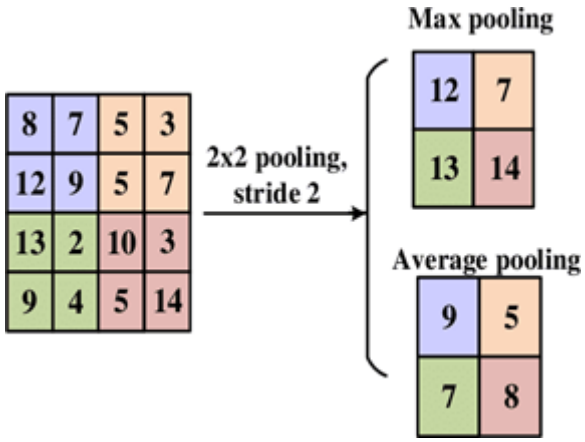

Fig. 3. Pooling layer

III. METHODOLOGY

A. Data collection

For this study, audio recordings of 30 healthy individuals with no symptoms of hoarse voice were collected at a sample rate of 22KHz. Each individual spoke for 10-12 minutes based on the pace. The subjects read the phonetically rich English literature and the resulting raw audio clips were saved in WAV format and further pre-processing was done.

B. Data pre-processing and augmentation

The collected audio clips were initially split using silence detection to remove the silent chunks of the audio. Finally, these audio clips were cut into audio samples of different lengths (2 seconds, 5 seconds, 10 seconds, 20 seconds). An example of the signal wave image is shown in Figure 4.

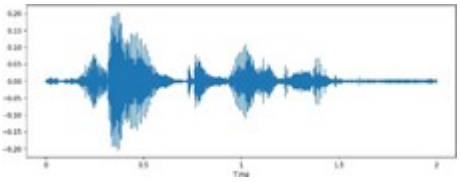

Fig. 4. Raw signal wave

C. Mel-frequency Cepstrum Coefficient

The presence of noise in the audio stream deems raw audio signals unfit for training the model. From the research done in [3] and [4] extracting features from the audio signal and applying them as input to the basic model produces significantly better results than utilizing the raw audio signal as input. The most extensively used method for extracting characteristics from an audio source is MFCC which is displayed in Figure 5. These MFCC features from each audio sample are fed into the speech recognition model as images.

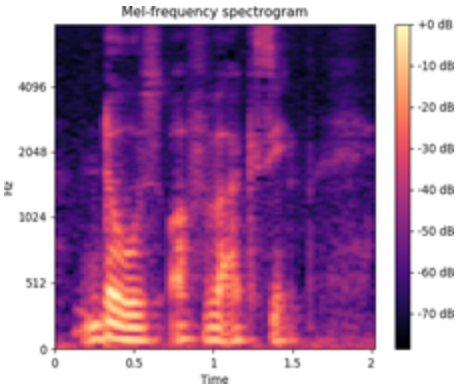

Fig. 5. Mel-frequency Spectrogram

The steps to extract the MFCC from a raw audio are as follows:

Pre-emphasis

Pre-emphasis involves improving the energy of the sound at higher frequencies. Human speech generally does not have as much energy at high frequency as it does at low frequency. This step is done to compensate for this issue.

Framing and Windowing

The process of segmenting the digital sample of the signal into segments of about 20 seconds. This step is carried out to simplify the huge amount of information contained within an audio signal. On average a person speaks three words per second with 4 phones and each phone will have three states resulting in 36 states per second or 28ms per state which is close to our 25ms window.A Hamming window is used to extract the features by grouping close frequencies together.

Fast Fourier Transform

Human speech is a convolution of the glottal pulse and vocal tract impulse response. FFT is performed to convert them into the frequency space.

Mel Filter Bank Processing

There is a difference between how a machine and human perceive high frequencies. Human auditory perception has lesser resolution at higher frequencies hence the frequencies are plotted in the mel map to show actual frequency versus the frequency perceived by humans. The equation for conversion to the mel map is given by following equation:

$$mel(f) = 1127ln(1 + f/700)$$

Inverse Fourier Transform

This step converts the log of mel into the time domain. The result is a vector of coefficients for every sample in the audio signal.

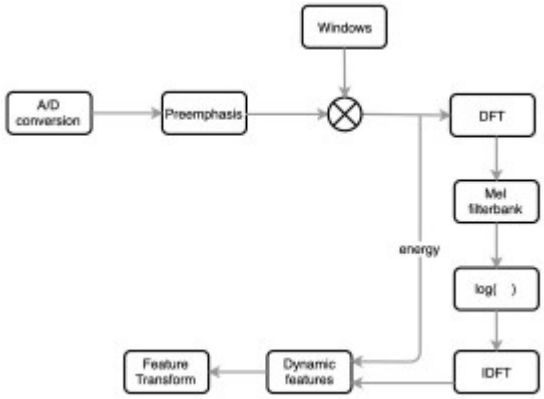

Fig. 6. Procedure to extract MFCC

D. CNN architecture

After experimenting with several architectures for the problem, a CNN architecture was chosen which took as input, Mel frequency coefficient of size 128X90X1 extracted from the audio clips during preprocessing. The complete CNN architecture has been summarized in table 1. The regularization term was set to 0.03 and batch normalization was carried out after every convolution to avoid over-fitting. To further prevent overfitting, we included a dropout layer between the dense layers with a dropout rate of 0.25. Throughout, the ReLU activation function was used except for the output layer in which SoftMax function was used to get the probabilities. In addition, speaker similarity was measured using standardized Euclidean Distances (ED) between pairs of speakers. This gave an intuition of how mfcc features embeds the variation of voices.

TABLE I
CNN ARCHITECTURE

Layer	Filter Dimension	Filter Number	Stride length	Data Size
Conv 1	3X3	16	1X1	128X90X16
Maxpool 1	-	-	2X2	64X45X16
Conv 2	3X3	32	1X1	64X45X32
Maxpool 2	-	-	2X2	32X22X32
Conv 3	3X3	64	1X1	32X22X64
Maxpool 3	-	-	2X2	16X11X64
FC 4	1024	-	-	1024
FC 5	2XUnique speaker	-	-	2XUnique speaker
FC6	Unique speaker	-	-	Unique speaker

IV. RESULTS

For this experiment, we trained 4 different models, namely model_5, model_10, model_20, model_30 where the numerical value represents the number of subjects the model has been trained on. All these models were tested on a single subject whose data is trained in every above-mentioned model and will be addressed as the primary subject henceforth. We collected 30 recordings of 20 seconds duration of the primary subject to evaluate these models. Further, we split these 20 seconds data into 2,5,10 duration samples to evaluate the minimum length of audio for which the model gives the best prediction.

From table 2 it is observed that with the increase in the duration of the data, the model prediction gets better. For all 4 models, the prediction is higher for 20-second data when compared to 2-second data. The performance has noticeably risen from 5 seconds to 10 seconds. This trend is true for all the models. Though there is an increase in accuracy with the increase in the length of the audio during testing, the rate of improvement in accuracy tends to wane with increasing length i.e the increase in accuracy from 5 seconds to 10 seconds audio samples is much greater than that between 10 seconds and 20 second samples. Since audio samples of longer duration can be hard to collect, a trade off between accuracy and data must be achieved for optimality. In this study, it is found to be a duration of 10 seconds.

TABLE II
MODEL ACCURACY COMPARISON

	Duration(second)			
	2	5	10	20
Model_5	36.6	60	86.6	96.6
Model_10	26	66.6	83.3	90
Model_20	23.3	43.3	63.3	80
Model_30	16.6	46.6	53.3	73.3

The other trend seen is that the model performance decreases with an increase in the number of subjects. With the same amount of data being used to train all the models, increasing the number of subjects in the model makes it hard for the model to generalize.

Inspired by the above results, we created 4 versions of Model_10 from the 30 subjects. From the trend we see in table 2, we expected the accuracy of all 4 versions to be similar. But the results deviated from our intuition. To better understand this, MFCC of 15 subjects reading the same text was plotted using the t-SNE technique which is shown in figure 6.

We found that MFCCs of subjects with similar accents were closely spaced in comparison with those with different accents. This can be clearly seen in figure 6 where we find 2 clusters represented by black bounding circles of people with different accents. Furthermore, within groups, female subjects were noticeably distant from male ones as seen in red bounding circles. We found out that out of 4 versions of the model_10 the one with higher accuracy was trained on subjects belonging to different groups.

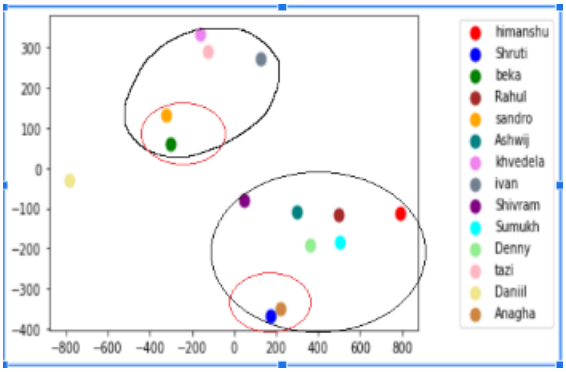

Fig. 7. MFCC plots of different speakers

V. CONCLUSION

For a speaker identification system for a smaller population limited amount of data is sufficient for a good prediction. But the accuracy reduces when the population increases. Further, it was observed that for a model to generalize on subjects with similar accents, a larger data set is required. In addition, it was found that an optimal speaker identification system needs an audio file of length anywhere between 5 to 10 seconds to get better accuracy.

VI. FUTURE WORK

This work can be extended on a larger data set of diverse population. The accuracies of all the models can be further improved with more data. A study can be conducted to find the phrase which can get more reliable identification.

REFERENCES

[1] Bai Z, Zhang XL, "Speaker recognition based on deep learning: An overview," 2021 Neural Networks. 2021 Mar 17

[2] P. Kumar, N. Jakhanwal and M. Chandra, "Text Dependent Speaker Identification in Noisy Environment," 2011 International Conference on Devices and Communications (ICDeCom), 2011, pp. 1-4, doi: 10.1109/ICDECOM.2011.5738533.

[3] R. Jahangir et al., "Text-Independent Speaker Identification Through Feature Fusion and Deep Neural Network," in IEEE Access, vol. 8, pp. 32187-32202, 2020, doi: 10.1109/ACCESS.2020.2973541.

[4] Kekre DH, Kulkarni V "Speaker identification using power distribution in frequency spectrum," Technopath, Journal of Science, Engineering Technology Management. 2010 Jan;2(1).

[5] Campbell WM et al., "Support vector machines for speaker and language recognition," Computer Speech Language. 2006 Apr 1;20(2-3):210- 29.

[6] Preti A et al., "Confidence measure based unsupervised target model adaptation for speaker verification," 2007 Eighth Annual Conference of the International Speech Communication Association 2007.

[7] Jahangir R et al., "Speaker Identification through artificial intelligence techniques: A comprehensive review and research challenges" Expert Systems with Applications. 2021Jun 1;171:114591.

[8] Salehghaffari H. Speaker verification using convolutional neural networks. arXiv preprint arXiv:1803.05427. 2018 Mar 14.

[9] D. Snyder et al., "X-Vectors: Robust DNN Embeddings for Speaker Recognition," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 5329-5333, doi: 10.1109/ICASSP.2018.8461375.

[10] C. Li, et al. "Deep speaker: an end-to-end neural speaker embedding system," arXiv preprint arXiv:1705.02304, 2017.

[11] Sarthak Yadav, and Atul Rai., "Learning discriminative features for speaker identification and verification," Interspeech, pp.2237 - 2241, 2018.

[12] A. Nguyen Nang, N. Quang Thanh, and Y. Liu, "Deep CNNs with selfattention for speaker identification," IEEE Access, vol. 7, pp. 85327-85337, 2019.

[13] R.Singh "The Voice Signal and Its Information Content-2," In: Profiling Humans from their Voice. Springer, Singapore, 2019.

[14] I. Goodfellow, Y. Bengio, and A, Courville, "Deep learning," MIT press,2016.

[15] A. Pai, et al, "Characterization of errors in deep learning-based brain MRI segmentation," In: Deep Learning for Medical Image Analysis. Academic Press. Elsivier, pp. 223-242, 2017 .