Speaker ID on Apollo 11 corpus:
A Study using different Machine Learning Models

Serkan Tokgoz, Electrical Engineering, Erik Jonsson School of Engineering and Computer Science, The University of Texas at Dallas, , serkan.tokgoz@utdallas.edu

Thouseef Syed, Applied Cogntion & Neuroscience, Behavioural & Brain Sciences, The University of Texas at Dallas, thouseef.syed@utdallas.edu

Abstract— The main goal here is to match a voice sample from an unknown speaker to one of several labeled speaker models since speech is easily produced. For the feature extraction, Mel Frequency Cepstrum Coefficients will be used since it is one of the most common features used for speaker recognition. Before extracting the features, we will do pre-processing such as Voice Activity detection to ignore unvoiced parts of the speech. For classification and objective comparison, **K-Nearest Neighborhood (KNN), Convolutional Neural Network (CNN)** and **I-vectors/PLDA** results will be shared. The dataset used for the project is **FEARLESS STEPS** that consists of 10 hours of digitized recordings of the Apollo 11 Space Mission. These recordings were digitized by the **Centre of Robust Speech Systems (CRSS)** of The University of Texas at Dallas. It was typically used for speech activity detection, sentiment analysis and speaker recognition. In the research, there were a few challenges that were met using methods. Our main focus will be detecting speech parts of the speech signals and classifying the respective speakers in the given time frame.

Keywords—MFCC, KNN, CNN, i-Vector, PLDA

I. Introduction

Speaker identification (SI) techniques has been used in numerous commercial products over the last decades. In SI, the main purpose is to match a voice sample from an unknown speaker to one of the labeled speaker models. Figure 1 shows a basic structure of a speaker identification system. To be able accomplish this task, there are two operational phases, training (can be also termed as enrollment) and testing.

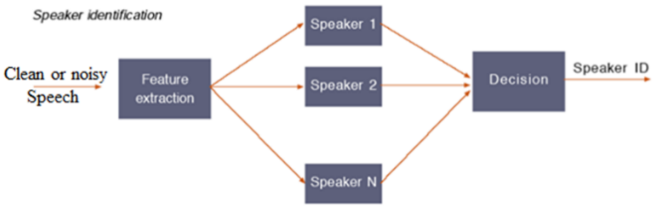

Figure 1 Basic structure of a speaker identification system

As shown in Figure 2, the signal separated to frames and applied Hamming window in both phases. Feature extraction and feature matching are the two key steps. In this work, we have extracted features with Mel-Frequency Cepstrum Coefficients (MFCC) because MFCC has accurate representation of the vocal tract, and a Voice Activity Detector (VAD) is implemented to extract features from the speech segments.

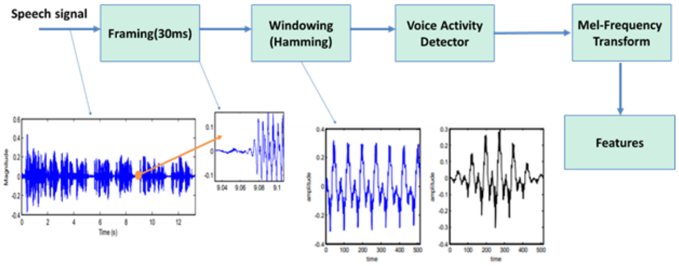

Figure 2 Block diagram of the speech signal processing stages.

We use Mel Frequency Cepstrum Coefficients for all speaker identification tasks to make a fair comparison since Mel scale is a scale that relates the perceived frequency of a tone to the actual measured frequency. It scales the frequency in order to match more closely what the human ear can hear.

$$M(f) = 1125\ln(1 + f/700) \tag{1}$$

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} & f(m-1) \le k \le f(m) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} & f(m) \le k \le f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{2}$$

Using equation (1), the upper and lower frequencies are converted to Mels depend on how many filter banks will be used for the system. After that we use equation (2) to convert these back to Hertz. The frequency resolution is required to put filters at the exact points calculated, thus those frequencies are rounded to the nearest FFT bin. To convert the frequencies to FFT bin numbers we need to know the FFT size and the sample rate. Figure 3 shows the steps for extracting the MFCC features from a speech frame.

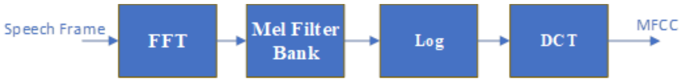

Figure 3 Block Diagram of the MFCC feature extraction

II. The Dataset

The Fearless Steps Corpus dedicated to Speaker Identification (SID) is essentially divided into two parts, one is called the Development set and the other Evaluation Set. It consists of around 183 speakers and 8394 utterances. They were typically communication between teams that played a vital role the Apollo-11 mission. The analysis of the data was done, and it was observed that there were many recordings that consisted of few utterances. Therefore, in order to tackle this challenge, refinement of the recording data was needed.

Figure 4 shows the histogram of the utterance counts per each speaker. As can ben seen from the figure, mosst of the data in the dataset includes a few utterances. Since it would be hard to classify with less data, we came up with a simple thresholding; if a speaker less than 6 utterances remove the recording before training the model. After removal of short audio data, 88 speakers were left in the set.

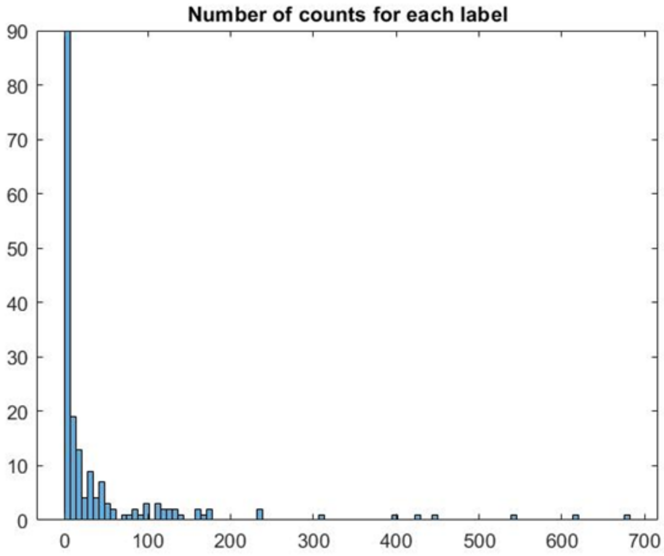

Figure 4 Histogram of the utterances for each speaker

A. *Models*

KNN: This is a classification technique that has knowledge of all the available cases and classifies new cases based on similarity parameter like the distance function. The features are used to train the classifier and the respective hyperparameters that emerge are the number of nearest neighbors, the distance to the nearest neighbor and therefore the weight of the distance metric. Figure 5 shows the block diagram of the training and testing phase for KNN classification.

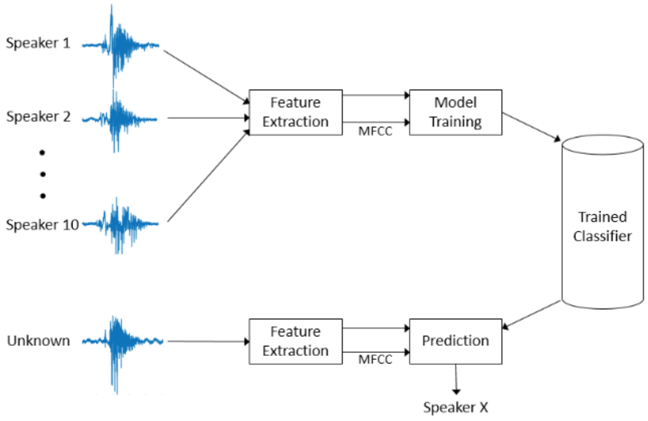

Figure 5 Block Diagram of KNN classification

CNN: The input to the network was 2 dimensional. This led to the extraction of MFCC features. The main objective of the convolutional layer is to extract high-level features. The first layer is known for extracting low-level features, hence with the aid of the additional layers, the network adapts to capture high-level features. However, on the other hand, every convolutional layer is cascaded to a max pooling layer that is responsible for extracting the dominant feature and eventually reduces the spatial size. Finally, it is fed to a fully connected layer where the speakers are classified

I-vector: A speech segment is represented by a low-dimensional "identity vector" (i-vector) extracted by Factor Analysis. The i-vector approach has become state-of-the-art in the speaker verification field. GMM supervector for speaker *s* at session *h*, $m_{s,h}$. The hidden variables $w_{s,h}$ + $N(0, I)$ in this case are called total factors. in this case are called total factors. + $N(0, I)$ in this case are called total factors. in this case are called total factors.

$$\mathbf{m}_{s,h} = \mathbf{m}_0 + \mathbf{T}\mathbf{w}_{s,h}. \tag{3}$$

III. Experiments

We were obsessed with the idea of exploring the dataset and its tremendous potential. Therefore, we assessed few machine learning models and found that for a start

A. *K-Nearest Neighbourhood*

For this method we used 88 different speakers for training. About **80%** of the data was used for **training** and the remaining **20%** was used for **testing.** Choosing the k value is important for KNN algorithm because it decides the distances between neighbors. In our experiments, optimal value for k was 7. If k is greater than 7, the system's performance was not improved a lot. Since, the increasing k value will have negative effect on the computational time, we preferred to keep the value less. We also used Euclidean distance between classes.

B. *Convolutional Neural Network*

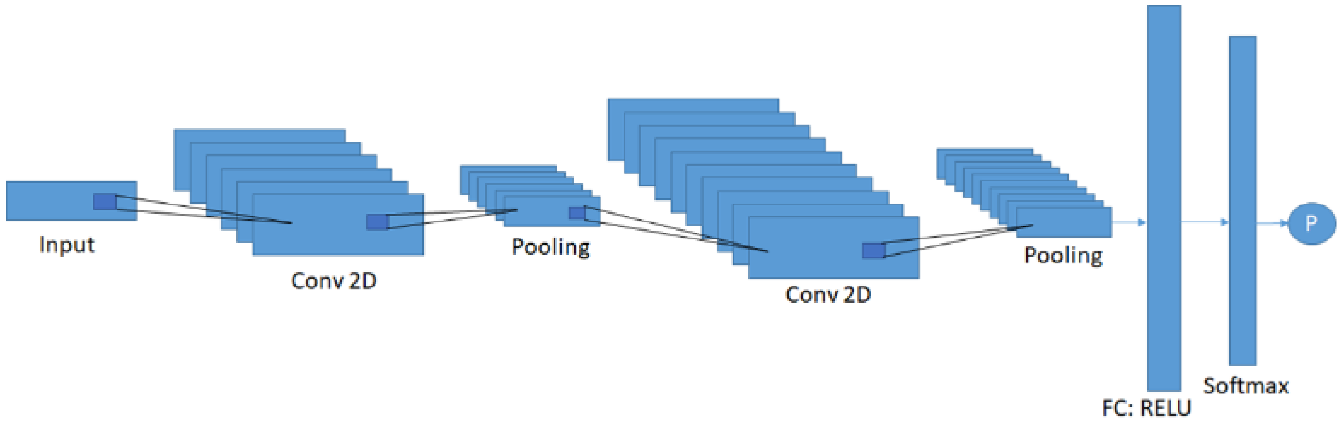

Figure 6 Speaker Identification System with Convolutional Neural Network

For the specifics, 2-layer CNN design shown in Figure 6. The size of input to the neural network is 39×N$_F$ where N$_F$ is the number of frames. Then, the input goes through a convolutional layer with 32 filters, kernel size = [1, 50], and RELU activation function to extract features from the data. A pooling layer of size [2, 2] was added and striding with 2 to reduce dimensions of output to *19× (N$_F$/2-1)*. After that it goes through another convolutional layer with 64 filters, kernel size = [1, 25], and RELU activation function to extract more features. This step is followed by a second pooling layer. Our model is implemented with TensorFlow machine learning library since it is one of the most common and powerful libraries in the field. We have also tried 3 convolutional layers, but the performance of the system was not improved noticeably.

C. *I-vectors*

The system was modeled using the state-of-the-art i-vector PLDA as a benchmark. We used open-sourced Kaldi to obtain our speaker models [5], and trained the Universal Background Model (UBM) using SRE-10 dataset. Due to the computational resources, the system is trained with less data. The PLDA model is trained with SRE-10 dataset, and adapted with i-vectors extracted from unlabeled train data [6]. Figure **7** describes the i-vector/PLDA based speaker identification system used for this task. The development set is then used as the training set and the evaluation set is used as the test set. The top 5 scores for the system are calculated to get the system performance.

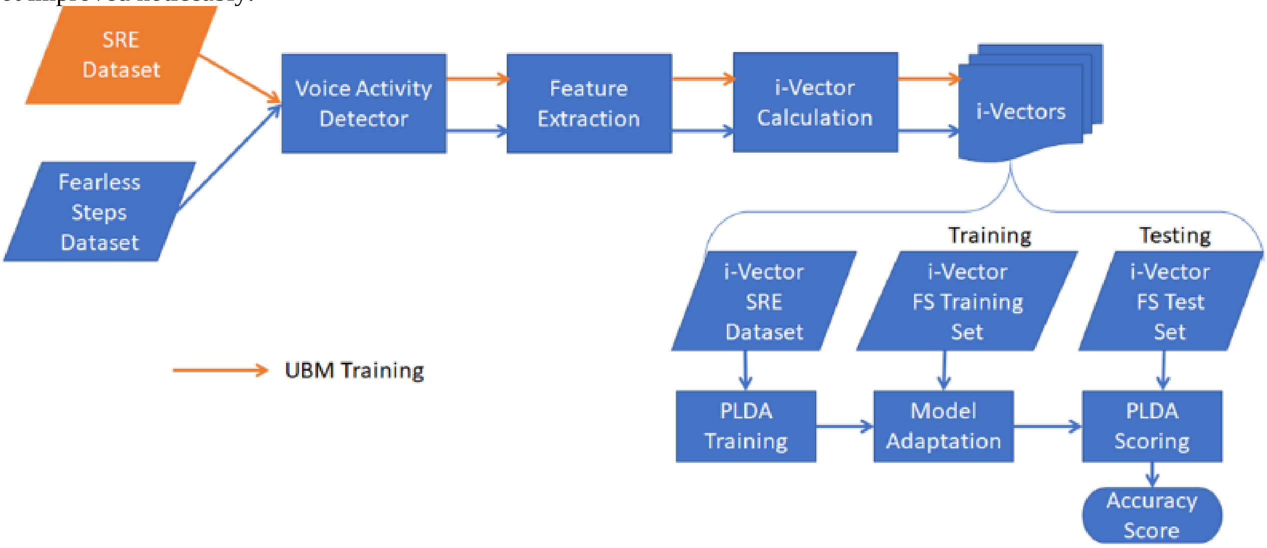

Figure 7 Block Diagram of Speaker Identification system with i-vector and PLDA

D. *Evaluation Metrics*

For the purpose of evaluation Top-5 accuracy was used as the metric. The 5 highest probability classes must match the expected output. In other words, the speaker that is identified, must match the predicted class of the speaker. The equation below depicts the Top-5 accuracy calculation.

$$Accuracy = \frac{\sum_{i \in S} N_{sys}(i)}{\sum_{i=1}^{M} N_{ref}(i)} \quad \text{for} \quad S = k \in [1, M] : N_{ref}(k) \subseteq N_{sys}(k)$$

where, $N_{ref}(i)$ represents speaker labels from ground truth for i^{th} segment, $N_{sys}(i)$ represents system predicted speaker labels for i^{th} segment, and M is the total number of frames. segment, and M is the total number of frames.

IV. The Results

The exploration of different machine learning models finally helped us achieve results. It was observed that CNN model performed much better than KNN and i-vectors. Since, CNN has intermediate layers like max pooling layers and a fully connected layer, the high dimensional features are extracted followed by classifying the targeted speaker. As far as KNN is concerned, it was observed that there existed an overlap between speakers because on the speaker space due to the similar features, therefore it misclassified the targeted speaker. Therefore, it decreased the overall accuracy. In contrast, for CNN, we believe that convolutional layers reduce spectral variations between speakers. Hence, it performs better for speaker identification. However, i-vectors performed poorly compared to CNN. On the other hand, it performs better compared to the FEARLESS steps competition's I-vector/ PLDA results. Since, we have pre-processed the data with less than 30 seconds or in other words the recordings that comprised at least 6 utterances or more certainly aided the model to learn and classify efficiently. Since, CNN has intermediate layers like max pooling layers and a fully connected layer, the high dimensional features are extracted followed by classifying the targeted speaker. Figure 8, demonstrates our Top-5 accuracy results.

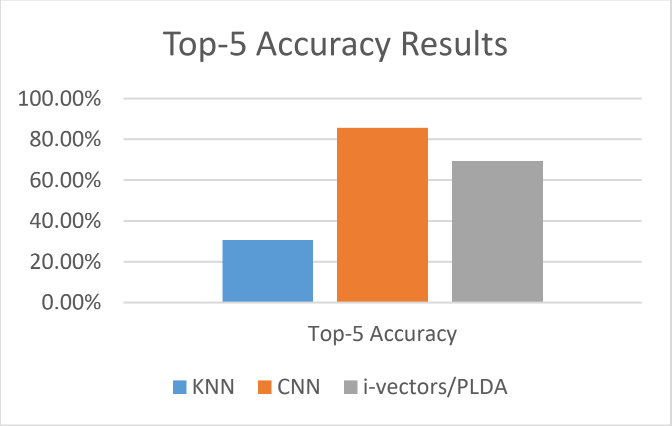

Figure 8 Accuracy results of the three systems

A. *Challenges*

The recordings pertaining to the mission posed a multitude of challenges. Each speaker had multiple utterances. Since, the mission needed to be orchestrated in a timely manner, there were speakers that needed to be switched on multiple occasions. The recordings had the following types of noise. Majority of the audio were degraded due the presence of transmission noise, channel noise, system noise, etc. As we mentioned earlier, the data was unbalanced as different speakers had varying durations. The microphone had also captured background conversations during the recording sessions. They were either engaged in long conversations or the remaining periods it was reported to be silent. Also, aging of the tape affected the quality of the audio. Another vital challenge was that computational resources were expensive and time consuming. Therefore, making a fair comparison between different models is quite challenging.

Conclusion

Speaker ID was performed on the Apollo-11 Corpus using different learning techniques like KNN, CNN and I-vectors. The ultimate purpose was to study the application of basic to complex methods implemented for Speaker ID. The data was challenging due to structure and quality of the audio. We proposed a refining technique that eliminates the recordings that had less utterances. This eventually aided the models to perform better. Finally, the best model was a CNN constructed with 2 convolutional and pooling layers. We were able to achieve Top-5 Accuracy with CNN and it outperformed both KNN and i-vectors. We were able to explore different toolkits and methods to perform Speaker ID.

References

[1] Hansen, J.H., Sangwan, A., Joglekar, A., Bulut, A.E., Kaushik, L., Yu, C. (2018) Fearless Steps: Apollo-11 Corpus Advancements for Speech Technologies from Earth to the Moon. Proc. Interspeech 2018, 2758-2762, DOI: 10.21437/Interspeech.2018-1942.

[2] Sangwan, Abhijeet & Kaushik, Lakshmish & yu, Chengzhu & Hansen, John & Oard, Douglas. (2013). 'Houston, We have a solution' : Using NASA Apollo Program to advance Speech and Language Processing Technology. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.

[3] J. H. L. Hansen and T. Hasan, "Speaker Recognition by Machines and Humans: A tutorial review," in *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 74-99, Nov. 2015.

[4] Yu, Chengzhu & Hansen, John & Oard, Douglas. (2014). 'Houston, We have a solution': A Case Study of the Analysis of Astronaut Speech during NASA Apollo 11 for Long-Term Speaker Modeling. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.

[5] Povey, Daniel, et al. "The Kaldi speech recognition toolkit." IEEE 2011 workshop on automatic speech recognition and understanding. No. CONF. IEEE Signal Processing Society, 2011.

[6] F. Bahmaninezhad and J. H. L. Hansen, "i-Vector/PLDA speaker recognition using support vectors with discriminant analysis," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, 2017, pp. 5410-5414.