Summer Internship Report

on

AUTOMATIC SPEAKER VERIFICATION

&

AUTOMATIC SPEECH RECOGNITION

by

Sonu Kumar Bhagat

Under the Supervision of

Dr. Syed Shahnawazuddin

Department: Electronics and Communication Engineering

National Institute of Technology Patna

Declaration:

I, Sonu Kumar Bhagat (B210066EC) from National Institute of Technology Sikkim declare that the work is done mainly for the summer internship at National Institute of Technology Patna from 10th June 2023 to 20th July 2023 on the **topic automatic speaker verification (ASV)** and **automatic speech recognition (ASR) system**.

Date.....

Acknowledgement:

We would like to give sincere thanks and gratitude to our esteemed supervisor Dr. Syed Shahnawazuddin for providing his valuable guidance and encouragement to learn new things and enhancing our knowledge in the field of speaker verification and speech recognition. His kind cooperation and suggestions throughout the course of this research guided us with an impetus to work and successfully completed the project. We would also like to express sincere gratitude to HOD, Dr. Bharat Gupta, Department of Electronics and Communication Engineering. National Institute of Technology Patna for all the facilities provided in the Institute.

Objective:

Objective of this internship is to learn about Kaldi toolkit, ASR and ASV system.

- An objective for this position should emphasize the skills you already possess in the area and your interest in learning more.
- Utilizing internships is a great way to build your resume and develop skills that can be emphasized in your resume for future jobs. When you are applying for a Training Internship, make sure to highlight any special skills or talents that can make you stand apart from the rest of the applicants so that you have an improved chance of landing the position.

.

Abstract:

Automatic speaker verification is the authentication of individuals by doing analysis on speech utterances. Speaker verification falls into pattern matching problem. Different approaches of training and adaptation techniques are studied in order to improve the recognition accuracy. Furthermore, as data size is a very important point in order to achieve enough recognition accuracy, the role of it, is also studied. Automatic Speech Recognition comes the closest to allowing real conversation between people and machine intelligence and though it still has a long way to go before reaching an apex of development Speech recognition and verification technology has experienced significant advancements in recent years, revolutionizing the way we interact with machines and improving the overall user experience. Automatic speech recognition (ASR) by machine has been a field of research for more than 60 years.

Content:

Chapter-1

1 Automatic Speaker Verification (ASV):

1.1 Introduction:

Automatic Speaker Verification (ASV) is a technology that aims to authenticate or verify the identity of a speaker based on their voice or speech patterns. It is a subfield of biometrics that uses speech-related characteristics to establish the identity of an individual. ASV systems analyze various features of a person's speech, such as pitch, rhythm, spectral patterns, and phonetic content, to create a unique voiceprint or speaker model. During enrollment, a person's voice is recorded and processed to generate a reference model. Subsequently, during verification, the system compares the speech of an individual with the reference model to determine whether the claimed identity matches the speaker. Various techniques can be used for comparison, such as pattern matching algorithms, statistical modeling, or machine learning algorithms.

1.2 Acoustic feature Representation:

In order to build an ASV system model architecture, MFCC is most frequently used acoustic feature.

1.2.1 MFCC:

MFCC stands for **Mel Frequency Cepstral Coefficients**. It is a feature extraction technique widely used in speech and audio signal processing, particularly in automatic speech recognition (ASR) systems.

This also includes mapping the normal frequency representation to a representation that is based on how humans perceive tones at different frequencies. A mel is a measure of the perceived frequency of a tone.

The relationship between mels and actual frequency is approximate linear below 1 kHz and logarithmic above.

The **mel scale** is derived by an experiment where test persons were exposed to a reference tone at 1 kHz and then asked to increase the frequency until they perceived certain tones.

Calculations of MFCC:

1. **Pre-emphasis**: A pre-emphasis filter is applied to the audio signal to amplify higher frequencies. It emphasizes the high-frequency components and reduces the influence of low-frequency components.

2. **Framing**: The pre-emphasized signal is divided into small frames typically ranging from 20 to 40 milliseconds. Overlapping frames are commonly used to capture temporal information.

3. **Windowing**: A window function, such as the Hamming window, is applied to each frame to reduce spectral leakage caused by discontinuities at frame boundaries.

4. **Fast Fourier Transform (FFT)**: The windowed frame is transformed from the time domain to the frequency domain using the FFT algorithm. This process provides a spectrum representation of the frame.

5. **Mel Filtering**: The Mel filter-bank is applied to the power spectrum obtained from the previous step. The Mel scale is a perceptual scale that maps frequency to the perceived pitch of sound. The Mel filter-bank consists of a series of triangular filters spaced uniformly on the Mel scale.

6. **Logarithm**: The output of the Mel filter-bank is converted to the logarithmic scale to mimic the human perception of loudness. This is typically done using the natural logarithm.

7. **Discrete Cosine Transform (DCT)**: The DCT is applied to the log filter bank energies to decorrelate the coefficients and extract the most significant information. The resulting coefficients are called Mel Frequency Cepstral Coefficients.

8. **Cepstral Mean Normalization (CMN):** CMN is an optional step where the mean of the MFCC coefficients is subtracted from each frame to remove channel-dependent variations.

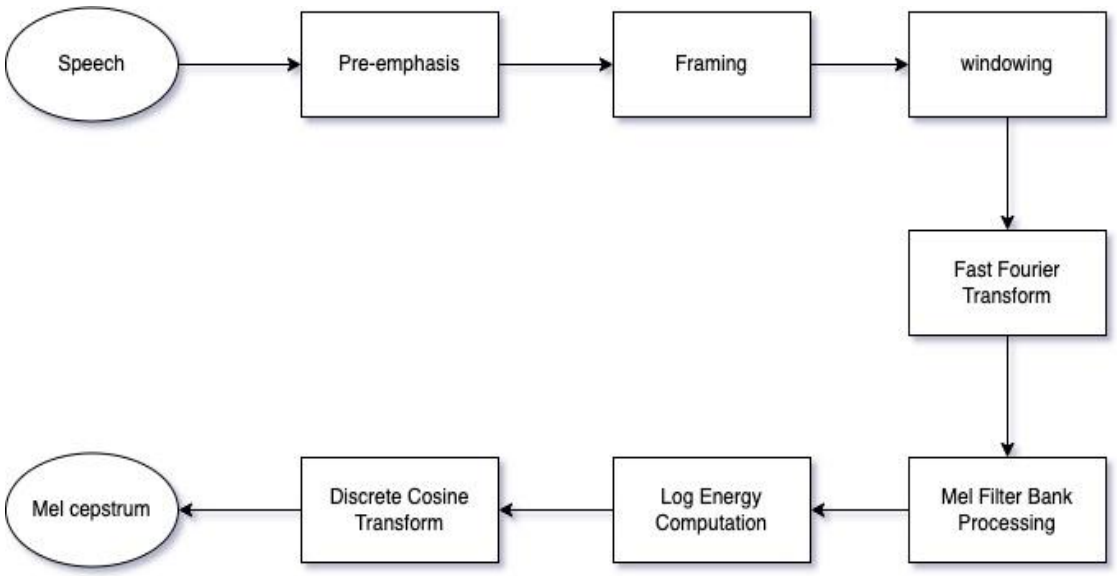

Fig 1.1: Block diagram of MFCC

1.3 X-vector in ASV:

X-vectors are a type of speaker embedding that are used in Automatic Speaker Verification (ASV). They are extracted from a **Deep Neural Network (DNN)** that has been trained on a large corpus of speech data. X-vectors are typically 128-dimensional vectors that represent the speaker's vocal characteristics.

Some of the benefits of using x-vectors in ASV:

Robustness to Noise:

X-vectors are more robust to noise than other types of speaker embeddings. This makes them well-suited for applications where the speech signal may be noisy, such as in a call center or in a noisy environment.

Discriminative Power:

X-vectors have high discriminative power, meaning that they can better distinguish between different speakers. This makes them well-suited for applications where it is important to accurately identify the speaker, such as in voice authentication or fraud detection.

Scalability:

X-vectors are scalable, meaning that they can be used to train large ASV systems. This makes them well-suited for applications where there are many speakers, such as in a call center or in a large organization.

Overall, X-vectors are a powerful tool for ASV. They are robust to noise, have high discriminative power, and are scalable. As a result, they are the state-of-the-art in ASV and are used in a variety of applications.

1.4 DNN:

DNN stands for **Deep Neural Network**. In deep-learning networks, each layer of nodes trains on a distinct set of features. based on the previous layers output. The further you advance into the neural net, the more complex the features your nodes can recognize, since they aggregate and recombine features from the previous layer. This is known as feature hierarchy, and it is a hierarchy of increasing complexity and abstraction. It makes deep-learning networks capable of handling very large, high-dimensional data sets with billions of parameters that pass through nonlinear functions. Above all, these nets are capable of discovering latent structures within unlabeled, unstructured data, which is the vast majority of data in the world. Another word for unstructured data is raw media; i.e. pictures, texts, video and audio recordings.

1.5 LDA:

In ASV, LDA is commonly used as a feature transformation technique to improve the discrimination between different speakers. It aims to reduce the dimensionality of the feature space while maximizing the separation between speakers' voice characteristics.

By applying LDA, ASV systems can reduce the influence of irrelevant or redundant features while enhancing the discriminatory power of relevant speaker-specific information. This transformation helps to improve the accuracy and robustness of speaker verification systems, making them more effective in various applications such as access control, authentication, and forensic speaker analysis.

1.6 Experimental Evaluation:

1.6.1 Equal Error Rate (EER):

In ASV, the Equal Error Rate (EER) is a measure of the system's accuracy. It is the point at which the false rejection rate (FRR) is equal to the false acceptance rate (FAR). FRR is the probability that the system will incorrectly reject a genuine speaker, while FAR is the probability that the system will incorrectly accept a non-speaker.

A low EER indicates a high-accuracy ASV system. An EER of 0% means that the system will never make a mistake, while an EER of 100% means that the system will always make a mistake.

Here are some of the factors that can affect the EER of an ASV system:

- The quality of the speech signal.
- The length of the speech signal.
- The type of ASV algorithm used.
- The amount of training data used.
- The noise level in the environment.

The **ASV_Tamil dataset** was used for the speaker verification system and the kaldi toolkit was used for the same. Equal error rate was used as evaluation parameter. The baseline system for ASV gave an EER of 12.09%.

In order to reduce the EER; several configuration changes as well as data augmentation through speed perturbation was performed.

1) Changing the configuration file of same ASV Model:

We had changed the value of number of cepstral coefficient as well as LDA dimension in our TDNN model to check kits effectiveness over baseline system.

ASV System	**lda_dim**	**num_cep**	**tdnn_dim**	**EER (%)**
Baseline	150	30	512	12.09
	75	30	512	12.22
	250	30	512	12.28
	150	35	512	-----
	150	25	256	12.48
	150	**25**	**512**	**10.36**

Table: 1.1

lda_dim: Dimension of Linear Data Analysis

num_cep: No. of Cepstral Coefficients

tdnn_dim: Dimension of Time Delay Neural Network

EER (%): Equal Error Rate

2) Speed Perturbation Technique:

In order to increase the dataset volume, 3_ways speed perturbation was implemented to generate the Tamil data synthetically.

The data -increment helped in further reducing the EER.

Speed Perturbation factor	EER (%)
1	**10.08**
0.9	**15.05**
1.1	**12.59**
0.9, 1, 1.1 (3-ways speed perturbation)	**6.314**

Table: 1.2

1.6.2 Conclusion of ASV Experiment:

The best EER ASV system build over LDA dimension 150 with the cepstral coefficient 25 and TDNN dimension 512.Changing the configuration of ASV system like changing the dimension of the LDA and TDNN, altering cepstral coefficient lead us to reduction in EER by gaining a relative improvement of 14.30.

After that we employed 3_ways_perturbation....._Which further reduced the EER to 6.314% with a relative improvement of approximate **48%**.

Chapter-2

2 Automatic Speech Recognition(ASR):

2.1 Introduction:

Automatic Speech Recognition or ASR, as it's known in short, is the technology that allows human beings to use their voices to speak with a computer interface in a way that, in its most sophisticated variations, resembles normal human conversation. The statistical approach to automatic speech recognition aims at modeling the stochastic relation between a speech signal and the spoken word sequence with the objective of minimizing the expected error rate of a classifier. Automatic speech recognition (ASR) has become a widespread and convenient mode of human-machine interaction, but it is still not sufficiently reliable when used under highly noisy or reverberant conditions.

The process of ASR involves multiple steps:

Acoustic Signal Processing: The first step is to capture and process the acoustic signal, which is the audio input containing the spoken speech.

Feature Extraction: The next step is to extract acoustic features from the pre-processed audio signal. Commonly used features include Mel-frequency cepstral coefficients (MFCCs), which capture the spectral characteristics of the speech signal.

Acoustic Modeling: In this step, statistical models, often based on Hidden Markov Models (HMMs) or deep neural networks (DNNs), are used to model the relationship between the extracted acoustic features and the corresponding linguistic units, such as phonemes, words, or sub-word units.

Language Modeling: Language modeling involves capturing the statistical regularities and dependencies within a given language.

Decoding: The decoding process uses the acoustic and language models to determine the most likely sequence of words or linguistic units that match the observed acoustic features.

Post-processing: Once the decoding is complete, post-processing steps may be applied to enhance the output.

Evaluation and Refinement: ASR systems are typically evaluated using objective metrics such as Word Error Rate (WER).

2.2 X-vector in ASR:

The x-vector system is an evolution of i-vectors originally developed for the task of speaker verification. X-vectors are a type of speaker embedding that are used in ASR. They are extracted from a deep neural network (DNN) that has been trained on a large corpus of speech data. They have been shown to improve the performance of ASR systems in a variety of ways. For example, they can be used to improve the accuracy of speaker verification, which is the task of determining whether two utterances were spoken by the same person.

In an ASR system, x-vectors can be used for a variety of tasks, including:

Speaker verification: This is the task of determining whether two utterances were spoken by the same person. X-vectors can be used to improve the accuracy of speaker verification by providing a robust representation of the speaker's voice.

Speaker diarization: This is the task of identifying the speakers in a multi-speaker recording. X-vectors can be used to improve the accuracy of 3 different utterances.

Acoustic model adaptation: This is the process of improving the accuracy of an acoustic model by adapting it to the specific speaker or environment in which it is being used. X-vectors can be used to improve the accuracy of acoustic model adaptation by providing a way to represent the speaker's voice in a way that is independent of the acoustic environment.

2.3 Experimental Evaluation:

2.3.1 Word Error Rate (WER):

Word Error Rate (WER) is a measure of the accuracy of an ASR system. It is calculated as the number of words that are incorrectly recognized divided by the total number of words in the reference transcript.

WER is a widely used metric for comparing the performance of different ASR systems. It is also used to track the progress of an ASR system as it is being developed or improved.

A lower WER indicates a higher accuracy ASR system. A WER of 0% means that the system will never make a mistake, while a WER of 100% means that the system will always make a mistake.

WER is calculated as follows:

WER = (Number of words incorrectly recognized + Number of words deleted + Number of words inserted) / Total number of words

For example, if an ASR system incorrectly recognizes 10 words, deletes 5 words, and inserts 2 words in a transcript of 100 words, then the WER would be 17%.

The **ASR_DEMO dataset** was used for the speech recognition system and the kaldi toolkit was used for the same. Word error rate was used as evaluation parameter.

The baseline system for ASR gave an WER of 60.09%.

In order to reduce the WER, several configuration changes as well as data augmentation through speed perturbation was performed.

After running Perturb_data_dir_speed_3way.sh file:

WER = 54.52%

When no. of data becomes five times the original data:

WER = 50.89%

2.3.2 Conclusion of ASR Experiment:

The WER of Baseline ASR System is 60.09%.

After that we employed **Perturb_data_dir_speed_3way.sh** Which further reduced the WER to 50.89% with a relative improvement of approximate 16%.

3 Conclusion:

We have learnt to train the ASR and ASV model using Kaldi Toolkit.

Both ASR and ASV systems are based on the same underlying principles of speech processing and machine learning. However, there are some key differences between the two technologies. ASR systems typically require a large amount of training data, while ASV systems can be trained on much smaller datasets. ASR systems are becoming increasingly accurate and are being used in a wide variety of applications. ASV systems are also becoming more accurate and are being used in applications such as security access control and fraud detection.

Both ASR and ASV systems are based on the same underlying principles of speech processing and machine learning.

4 References:

[1] http://kaldi.sourceforge.net

[2] https://www.sciencedirect.com/topics/engineering/automatic-speech-recognition

[3] https://ieeexplore.ieee.org/document/8461375

[4] https://www.danielpovey.com/files/2018_icassp_xvectors.pdf

[5] https://www.knowledgehut.com/blog/data-science/linear-discriminant-analysis-for-machine-learning

[6] https://eprints.whiterose.ac.uk/102623/7/Zeiler_submitted%20(1).pdf