IOSR Journal of Mobile Computing & Application (IOSR-JMCA)
e- ISSN: 2394-0050, P-ISSN: 2394-0042.Volume 11, Issue 3 (May. – June. 2024), PP 10-14
www.iosrjournals.org

Encoder-Only Transformer Architecture Forspeaker Modeling And Identification Using Malayalam Speech Corpus

Manju G
(Department Of Computer Science, Govt. College, Ambalapuzha, University Of Kerala, India

Abstract:
Background*: Human biometrics encompasses various types of data used for identification and authentication, including fingerprints, iris patterns, voiceprints etc. Each type has unique characteristics suitable for different security levels. Voiceprints or speaker recognition analyses unique voice features and determines a speaker's identity using the acoustic features of their voice. Voice prints can enhance user experience and security in mobile applications. Malayalam is s a low resource south Indian language in terms of availability of speech corpora. Transformers in machine learning have shown promising results in natural language processing tasks, including speech recognition. This opens up new possibilities for improving speaker identification and voice recognition systems ultimately contributing to advancements in fields such as security, forensics, and human-computer interaction.*
Materials and Methods*: Using a publicly available Malayalam speech dataset, this paper examines the application of encoder-only transformers for identifying speaker. The model is trained and tested fordifferent number of encoder layers.*
Results: *It has been shown that the model with one encoder layer has the highest accuracy for the dataset used. As the number of layers increases to two the accuracy of training and validation decreases slightly.*
Conclusion: *Adding too many layers can increase the risk of overfitting, especially if the model capacity exceeds what the dataset can support. The experiment shows tha, to identify a speaker, a single encoder layer is sufficient for the given data set.*
Key Word*: Speaker Identification, Malayalam Language, transformers, Encoder-only model.*

Date of Submission: 04-06-2024 Date of Acceptance: 14-06-2024

I. Introduction

Human biometric includes various types of data that can be used for identification and authentication purposes. This can include fingerprints, iris patterns, facial recognition, voiceprints, and even DNA. Each type of biometric data has its own unique characteristics and can be used in different scenarios depending on the level of security required. Voiceprints, also known as speaker recognition, are a type of biometric data that analyses the unique characteristics of an individual's voice, such as pitch, tone, and pronunciation. Speaker Identification (SI) is a process of extracting the identity of a speaker by using the acoustic features of the given utterance. Accurate speaker identification systems can have a wide range of applications. They can be used in security systems, such as access control and surveillance, to verify the identity of individuals based on their voice. Speaker identification can also be useful in forensic investigations, where it can help in identifying suspects or analysing audio evidence. Accurate voice recognition systems can enhance human-computer interaction by allowing devices and applications to respond to specific individuals based on their voice commands[1].

Speech allows for the direct transmission of information through the speech signal, which includes not only words but also tone, pitch, and other vocal cues that convey meaning and emotion. Pitch refers to the perceived frequency of a person's voice, which can vary from low to high. Intonation refers to the rise and fall of pitch patterns within speech, providing additional cues for speaker recognition. Spectral information refers to the unique pattern of frequencies present in a person's speech signal. It plays a crucial role in speech perception and can provide valuable insights into a speaker's identity, emotional state, and even health conditions. By analysing spectral information, we can uncover intricate details about the speaker, making it an essential component of speech analysis and recognition systems.

Speech formants are another important aspect of spectral information in speech. Analysing formants can provide valuable insights into a speaker's articulation patterns and vocal characteristics, making it an essential tool for speech analysis and recognition. Spectral information captures the unique distribution of

energy across different frequencies, further enhancing the accuracy of speaker identification systems[2]. This paper examines the application of encoder –only transformers for identifying speakers over publicly available Malayalam speech dataset. Malayalam is a regional language spoken by the people of Kerala, a southern state of India and is a low resource south Indian language in terms of availability of speech-to-text corpora

II. Literature Review

Statistical modelingtechniques like Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) to model the distribution of feature vectors for each speaker. Speaker identification using GMMs and HMMs involves training statistical models to represent the distribution of feature vectors for each speaker. GMMs are powerful statistical models that can capture the unique characteristics of each speaker's features. They analyse the statistical properties of the feature vectors, such as mean and variance, to create a representation of the speaker's voice. By comparing these representations with the test speaker's features, the system can accurately determine the most likely speaker identity. HMMs are particularly useful for capturing the temporal dynamics in the speech signal. They model the transitions between different states of the speech signal, allowing the system to understand the sequence and timing of speech sounds. This information helps in accurately determining the speaker's identity by considering not just the static features, but also the dynamic aspects of their speech[3].

Deep learning modelingusesneural networks, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs), to create more robust and accurate speaker models.. These models automatically learn hierarchical representations of input data, capturing complex patterns and dependencies in speech signals. This makes them more robust and accurate in identifying speakers, especially in challenging conditions such as noisy environments or speech variations. Deep learning models can handle large amounts of data and be trained on massive datasets. This enables them to generalize well to unseen speakers and improve overall performance[4,5,6].

Recent advancements in deep learning and other hardware techniques have greatly impacted the field of automatic speaker identification (SI)[7, 8]. This paper describes an encoder- only transformer model or Speaker recognition.

III. Methodology

Dataset

For the purposes of developing speech technologies for Malayalam, particularly text-to-speech, the IMaSC Malayalam speech corpus has been made publicly available by ICFOSS. The corpus contains 34,473 text-audio pairs of Malayalam sentences spoken by 8 speakers, totalling approximately 50 hours of audio.The diverse set of speakers in the corpus enables researchers to analyze and understand speech patterns in Malayalam. The IMaSC Malayalam speech corpus was developed through a rigorous process of data collection and annotation. Native Malayalam speakers were recorded while reading a range of sentences. The audio recordings were then aligned with the corresponding text, resulting in a valuable resource for training and evaluating speech technologies in Malayalam.

Feature extraction

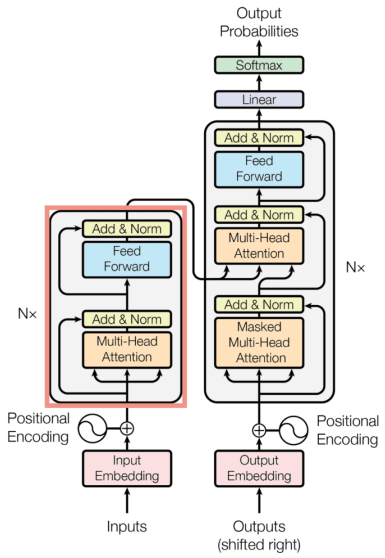
There are several techniques used for speech signal feature extraction, including Mel-frequency cepstral coefficients (MFCCs) , Perceptual Linear Prediction, Wavelet extraction andlinear predictive coding (LPC). These techniques extract relevant information from the speech signal, enabling accurate speaker identification and recognition. Here MFCC is used to extract speech features. MFCC addresses taking into account the non-linear nature of human perception of sound. By converting the speech signal into a logarithmic Mel scale, MFCC captures the frequency bands that are most relevant to human hearing. This allows for more accurate and robust feature extraction, making MFCC a valuable tool in speech processing applications. The logarithmic Mel scale is a perceptually based frequency scale that approximates the way humans perceive sound. It takes into account the non-linear nature of human hearing, where our perception of pitch is not directly proportional to the physical frequency of the sound. By converting the speech signal into the Mel scale, MFCC focuses on the frequency bands that are most important for human hearing, allowing for more accurate and meaningful feature extraction. This makes MFCC a valuable tool in speech processing applications, as it aligns with how our auditory system processes and interprets sound[9,10,11].

Speaker Modelling

This experiment uses the encoder-only transformer model to identify speakers. A transformer model is a type of neural network architecture that has been widely used in natural language processing. The paper ‘Attention Is All You Need’ introduces and an architecture called Transformers. It consists of multiple layers of self-attention and feed-forward neural networks, allowing the model to weigh the importance of different

parts of the input sequence dynamically. An encoder-only model for speaker recognition refers to a transformer model that only consists of an encoder component. Encoders are designed to learn embedding that can be used for various predictive modeling tasks such as classification. Unlike traditional transformer models that have both an encoder and a decoder, encoder-only models focus solely on encoding the input sequence of acoustic features to extract relevant speaker-specific information[12].

Figure 1: Encoder Architecture from the paper "Attention Is All You Need,"



Input Layer defines the shape of the input data and is feeded with MFCC features of the audio data. It is followed by a stack Transformer encoder layers. Each of these layers consists of the multi-head attention and feed-forward network. Multi-Head Self-Attention enables the model to focus on different parts of the input sequence simultaneously, capturing various types of relationships and dependencies. Feed-Forward Neural Networksare applied to each position in the sequence independently, enabling the model to learn complex transformations of the input data. LayerNormalization and Dropout stabilize and regularize the training process, helping the model to generalize better and prevent over fitting. After the transformer layers, the model includes two dense layers with ReLU activation and dropout for further transformation and regularization. Output Layer consists of **a** dense layer with Softmax activation to produce the probability distribution over the classes for classification.

Training And Testing

The data set is properly labelled and is split into training and testing sets. The acoustic features extracted are standardized and reshaped to input data for the encoder-only transformer model. And the model is trained for different batch size and number of layers for 20 epochs. The following figures shows the training and validation loss after each epoch along with the sparse categorical accuracy during testing and training.

Figure 2. Encoder layers-2, Batch size -1

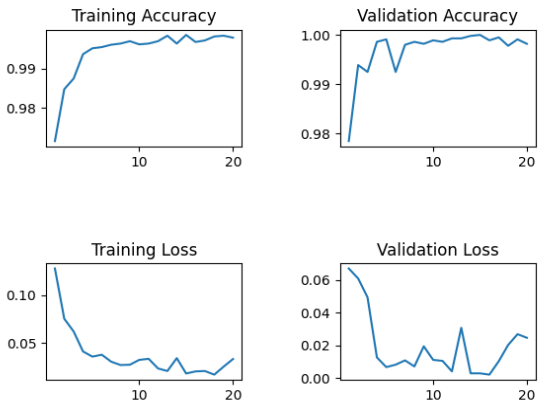


Figure3. Encoder layers-2, Batch size- 2

Figure 3. Encoder layers-1, Batch size- 1

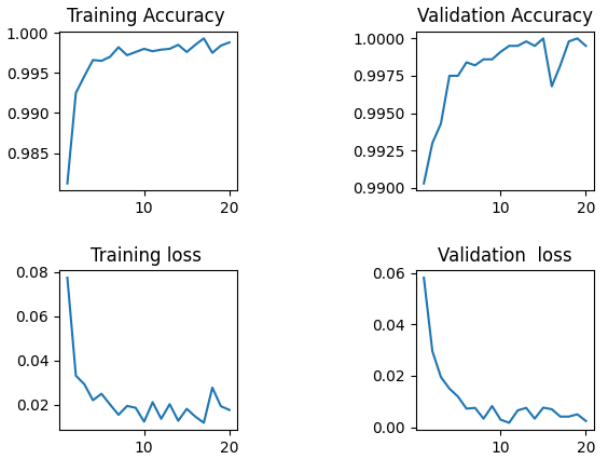


Figure 4. Encoder layers-1, Batch size- 2

IV. Discussion

The highest level of accuracy is achieved by the model with one encoder layer. As the number of layers increases to two the accuracy of training and validation decreases slightly for a batch size of 1. The model shows around 57% accuracy when the batch size is increased to 2. However, when the number of layers increases to two, the accuracy of both training and validation decreases significantly. The model with one layer outperforms the model with two layers in speaker recognition accuracy. The impact of batch size on the model accuracy is evident in the results. While the model with one layer maintains relatively high accuracy with a batch size of 1 than batch size of 2 and the model with two layers also experiences high accuracy with a batch size of 1 than batch size of 2.

V. Conclusion

Increasing the number of layers in the model can lead to a more complex network that captures more intricate patterns and features in audio data. Adding too many layers can increase the risk of over fitting, especially if the model capacity exceeds what the dataset can support. The experiment shows that, to identify a speaker, a single encoder layer is sufficient for the given data set. It is important to note that this experiment only explored the impact of layer depth and batch size on speaker recognition accuracy. Future research could consider other factors such as the type of audio data, the size of the dataset, and the use of different machine learning algorithms or architectures. Additionally, investigating the optimal combination of layer depth, batch size, and other hyper parameters could further enhance the accuracy of speaker recognition models. Having a sufficient amount of speech corpora in Malayalam is crucial for accurate speaker identification and voice recognition systems. It enables the development of robust algorithms and models that can effectively analyse and identify individual speakers based on their unique vocal characteristics, contributing to advancements in various fields such as security, forensics, and human-computer interaction.

References

[1]. Marcos Faundez-Zanuy,Enric Monte-Moreno,Escolauniversitariapolitècnica De Mataró, State-Of-The-Art In Speaker Recognition,Talp Research Center (Bercelona)Avda. Puig I Cadafalch 101-111,08303 Mataro (Barcelona) Spain

[2]. A Review On Speech Recognition Technique, International Journal Of Computer Applications, Vol 10, No.3, Pp. 16-24. 2010.

[3]. Adwa Abakarim, Abdenbiabenaou,Comparative Study To Realize An Automatic Speaker Recognition System,International Journal Of Electrical And Computer Engineering (Ijece),Vol. 12, No. 1, Pp. 376~382,2022

[4]. Xu Xiang, Shuai Wang, Houjun Huang, Yanminqian, Kai Yu,Margin Matters: Towards More Discriminative Deep Neural Network Embeddings For Speaker Recognition,2019

[5]. R. Sharma, D. Govind,J. Mishra, A. K. Dubey, K. T. Deepak, S. R. M. Prasanna,Milestones In Speaker Recognition,Artifcial Intelligence Review 57:58,2024

[6]. Rashid Jahangir , Ying Wahteh , Henry Friday Nweke , Ghulammujtaba ,Mohammed Ali Al-Garadi , Ihsan Ali ,Elsevier, Expert Systems With Applications,Volume 171, 1 June 2021.

[7]. Brown Et Al, Speakerbox: Few-Shot Learning For Speaker Identification With Transformers. Journal Of Open Source Software, 8(83), 5132, 2023.

[8]. Kavyamanohar,Gokul G. Menon, Ashishabraham, Rajeevrajan,A. R. Jayan ,3automatic Recognition Of Continuous Malayalam,Speech Using Pretrainedmultilingual, Transformers, International Conference On Intelligent Systems For Communication, Iotand,Security (Iciscois),2023.

[9]. P. V. Janse Et Al., "A Comparative Study Between Mfcc And Dwt Feature Extraction Technique," International Journal Of Engineering Research And Technology, Vol. 3, No. 1, Pp. 3124-3127, 2014.

[10]. A. Winursito, R. Hidayat, A. Bejo, And M. N. Y. Utomo, "Feature Data Reduction Of Mfcc Using Pca And Svd In Speech Recognition System," 2018 International Conference On Smart Computing And Electronic Enterprise (Icscee), Pp. 1-6, 2018

[11]. Y. Wang And B. Lawlor, "Speaker Recognition Based On Mfcc And Bp Neural Networks," 2017 28th Irish Signals And Systems,Conference (Issc) , Pp. 1-4,2017

[12]. Ashishvaswani Et.Al, Attention Is All You Need,Google, 2023.