Audio-Visual Person Recognition

By

Ahmad Ali

Spring-2022-MS-RIME 275850 SMME

Supervisor

Dr. Hasan Sajid

Department of Robotics and intelligent Machine Engineering

A thesis submitted in partial fulfillment of the requirements for the degree of Masters in Robotics and Intelligent Machine Engineering (MS RIME)

In

School of Mechanical and Manufacturing Engineering (SMME) ,

National University of Sciences and Technology (NUST),

Islamabad, Pakistan.

(June 2022)

Thesis Acceptance Certificate

Certified that final copy of MS/MPhil thesis entitled "**Audio-Visual Person Recognition**" written by **Ahmad Ali**, (Registration No **Spring-2022-MS-RIME 275850 SMME**), of School of Mechanical and Manufacturing Engineering (SMME) has been vetted by the undersigned, found complete in all respects as per NUST Statutes/Regulations, is free of plagiarism, errors and mistakes and is accepted as partial fulfillment for award of MS/M Phil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: _____

Name of Advisor: **Dr. Hasan Sajid**

Date: _____

Signature (HoD): _____

Date: _____

Signature (Dean/Principal): _____

Date: _____

Approval

It is certified that the contents and form of the thesis entitled "**Audio-Visual Person Recognition**" submitted by **Ahmad Ali** have been found satisfactory for the requirement of the degree.

Advisor: **Dr. Hasan Sajid**

Signature: _____

Date: _____

Committee Member 1: **Member 1**

Signature: _____

Date: _____

Committee Member 2: **Member 2**

Signature: _____

Date: _____

Committee Member 3: **Member 3**

Signature: _____

Date: _____

Dedication

This thesis is dedicated to all the deserving children who do not have access to quality education especially young girls.

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at Department of Robotics and intelligent Machine Engineeringat School of Mechanical and Manufacturing Engineering (SMME) or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at School of Mechanical and Manufacturing Engineering (SMME) or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Ahmad Ali**

Signature: _____

Acknowledgments

Glory be to Allah (S.W.A), the Creator of the Universe. Who only has the power to honour whom He please, and to abase whom He please. Verily no one can do anything without His will. From the day, I came to NUST till the day of my departure, He was the only one Who blessed me and opened ways for me, and showed me the path of success. Their is nothing which can payback for His bounties throughout my research period to complete it successfully.

Ahmad Ali

Contents

List of Figures

List of Tables

Abstract

Person authentication is a primary element to consider wherever privacy is necessary. Deep learning based authentication algorithms have a number of applications in the said field. Adding multiple modalities makes the system more robust. In this research a joint multi-modal audio-visual deep learning based method has been devised to authenticate a person based on their voice as well as face. This two-step verification process works by learning face-feature based embeddings as well as voice-feature based embeddings to serve two purposes: 1) if the face presented matches with an identity in a reference database and 2) if the voice matches any voice in the reference database. This strategy can help prevent important systems from impostor attempts using modalities that are commonly present and available in consumer devices.

Chapter 1

Introduction

The research presented in this dissertation explores the field of face-and-voice-based person recognition. The techniques explored as well as implementations done in this research are to develop a multi-modal audio-visual person recognition system that will allow recognition of the person in question, based on both the face features as well as the voice features of that person, given that the embedding-representation of voice and face features have a reference set

1.1 Background

The consumer devices present in today's day and age such as the modern-day smartphone, laptops, and computers widely support audio and visual input to the system. It is well known that the audio input to the system is commonly acquired from a microphone attached to the system and the most generic form of visual input is from a monocular camera. Concurrently, the voice and face features of the person are the most well-known and popular form of biometrics[23].

In today's day and age, biometrics-based person authentication has gained immense popularity and found its usage in various applications including but not limited to access control systems for commercial purposes such as digital access, domestic, and enterprise scenarios, as well as entrance verification systems. Moreover, recently there has been profound interest in usage of audio-visual biometric markers of a person in forensic scenarios. Naturally, voice features and face features are the two of the most effortlessly available biometric characteristics that accurately represent the identity of a

person. Consequently, the speaker recognition (SR) and face recognition (FR) systems are hot topics for researchers working on biometric representation of a person[15, 12]. The recent progress in the field of deep learning has allowed researchers to attain high performance in accomplishing these tasks. Various algorithms and architectures of deep neural networks along with a variety of loss functions have been under investigation to achieve successful results in FR and SR systems. This development has led to high-performance systems with high usability in commercial and forensic scenarios. The combination of these two systems promises to provide a higher reliability to the access control system as both representations when combined have a lower chance of being spoofed and have a boosted robustness against imposter attacks [16]. Moreover, by combining these two methodologies, there is a lower likelihood of false positives by providing lower FMR as compared to the audio-only and visual-only person recognition tasks [22].

1.1.1 Verification and Recognition Tasks

In the field of biometrics, verification refers to the tasks of verifying a person for their claimed identity. The identity of the person is claimed and therefore only requires that the person's biometric markers to be matched by the claimed identity. An illustration of a high-level person recognition system is shown in figure 1.1. For the person verification system to give a positive decision, the probe identity should match with the claimed identity.

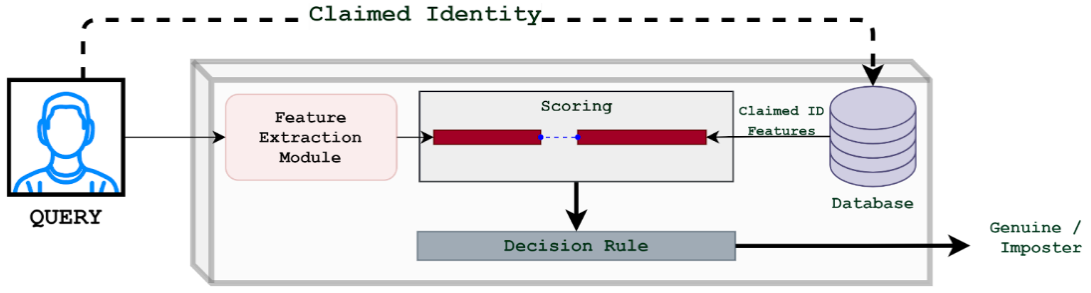

Figure 1.1: Person Verification System.

Whereas, in the recognition task, the objective is to match the probe identity with a set of identities already stored, often referred to as a "gallery set". Modern algorithms often use a 1xN dimensional vector which is used for scoring. The algorithms are developed

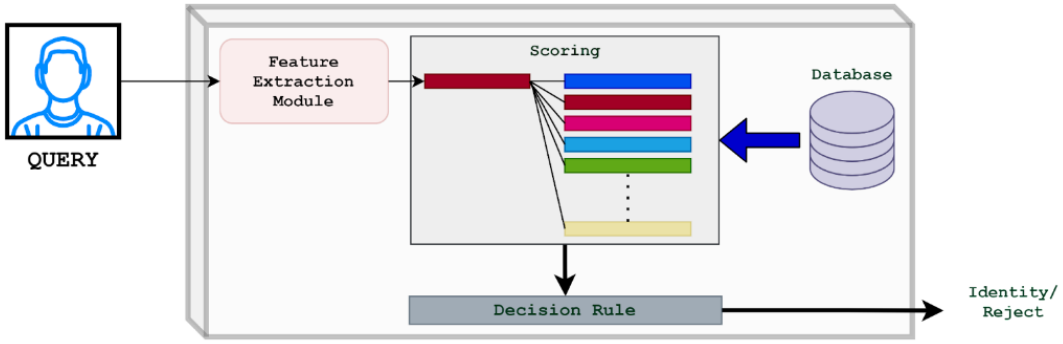

Figure 1.2: Person Recognition system

such that the said vector ("embedding") is representative of the identity of a person. This solves two problems: 1) Representation of a probe identity in a compact and mathematically viable form i.e., a vector. 2) Representation that is similar, using some metric, for the matching identity and dissimilar for the non-matching identities.

Figure 1.2 is a high-level illustration of modern recognition systems. The scoring system usually uses similarity or distance metrics such as Cosine similarity or Euclidean distance[15].

The scoring metric e.g., cosine similarity or Euclidean distance, is oftentimes used with combination of a threshold value that is defined by testing done on test set[1].

1.1.2 Face Recognition

Similar to the speaker recognition task described in section 1.1.2, the modern techniques of face recognition also involve learning face-level embeddings for identities. Modern techniques using deep convolutional neural networks (DCNNs) have allowed for high performance in embeddings-generation for facial recognition ("FR") task that approaches human-level performance.

[1]Test set refers to a sample of data that is disjoint from the data that is used for training. This set is typically not of the same distribution as the training data and due to the disjoint nature of the test set, results are comparable to the real world performance of the algorithm.

1.1.3 Speaker Recognition

Traditional probabilistic models, such as the Gaussian Mixture Model-Universal Background Model [4] and the i-vector [18], have given way to the more recent technique of deep speaker embedding learning as the preeminent modelling approach for the speaker identity. This shift occurred as a direct result of the rise in popularity of deep neural networks (DNN). When the amount of granularity of optimization is taken into account, mainstream deep speaker embedding learning may be broken down into two distinct categories: segment-level learning and frame-level learning.

Recent advancements in the field of deep neural networks (DNNs) have grabbed the lead in terms of modelling the attributes of an identity in terms of the algorithms used by biometric security systems. It has brought about a change away from the probabilistic modelling of speaker identities and into a new paradigm.

The deep speaker embedding models are categorized as follows:

1. Frame-level Speaker Embedding.
2. Segment-Level Speaker Embedding.

segment-level speaker embeddings are used for the whole utterance. These techniques were preliminarily text-dependent meaning that they usually rely on the speaker to utter specific word(s). While the frame-level speaker emebddings are more focused towards embedding the speech for a specific instance by using phonetics. These techniques, on the other hand, are inherently not text-dependent.

The DNN is trained in such a way to discriminate among speakers for each frame for frame-level speaker embeddings, and the embeddings are averaged for the whole segment to generate the necessary segment-level embeddings. In contrast, deep neural networks (DNNs) are utilised in the process of segment-level speaker embeddings in order to differentiate between the various speakers' individual voice segments. This creates a more accurate match between the optimization requirements of the training configuration and the deployment criteria.

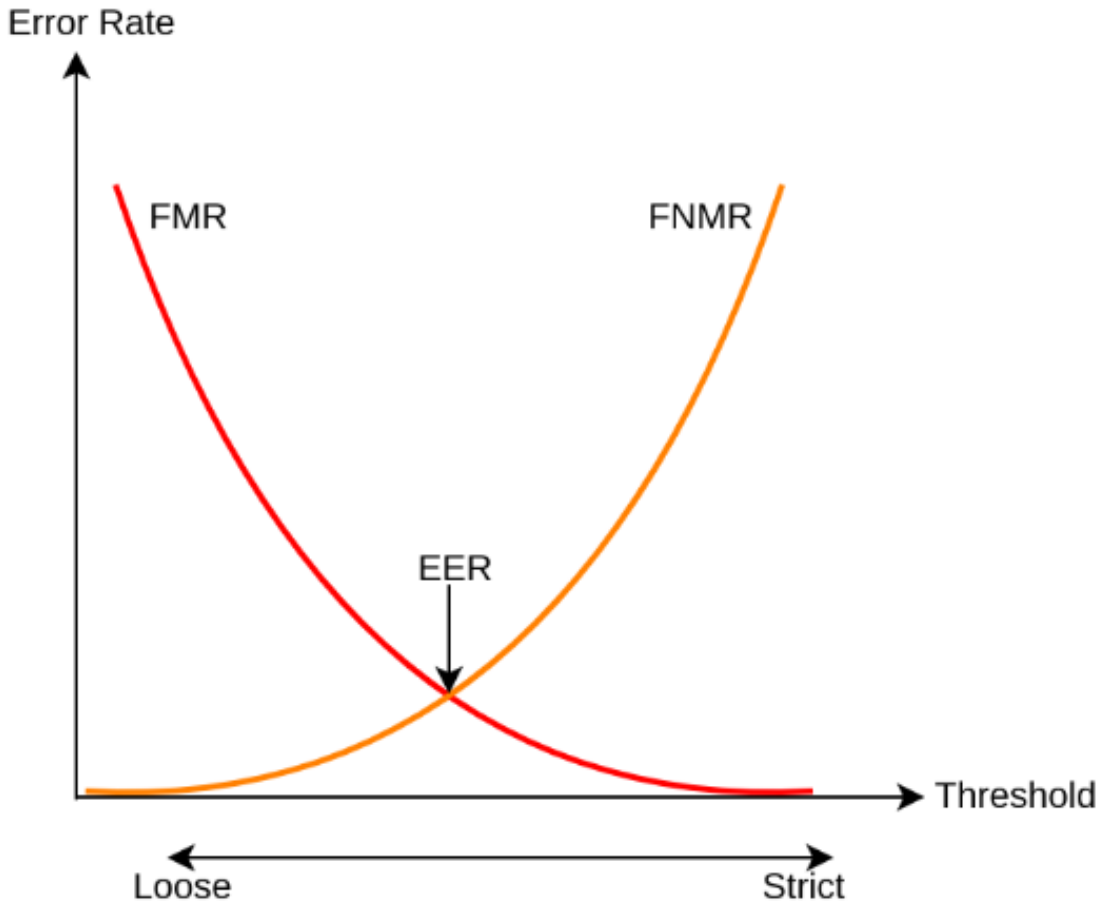

Figure 1.3: Person Recognition system

1.1.4 Performance Evaluation of Recognition systems

The recognition systems are often trained with a distance metric that they have to minimize. Thresholding on the scoring metric results in a certain False Matching error Rate (FMR) and False Non-Matching error Rate (FNMR). Sweeping threshold over the test set provides different values of FMR and FNMR. Equal Error Rate (EER) is the value where both FMR and FNMR cross. Performance of matching system is negatively proportional to EER.

As depicted in figure 1.3, the performance evaluation of a recognition system based on matching features is dependent on reduction of EER. Most modern day recognition system research work propose reduction in EER as a hallmark sign of gain in performance. This is actually true for the recognition systems where the decision is being made due

to a decision rule such as thresholding of scores. The scores are usually distance of similarity metrics aimed to determine the similarity of two feature embeddings based on the distance between them.

1.2 Motivation

The modern day consumer device, for example, a cell phone or a laptop computer comes with an integrated microphone and a camera. More than 67% of the world's population are smartphone users[25]. Recent days have witnessed as steady increase in abundance of multimedia data. While there has been a prevailing interest in analysis of multimedia data for many years, recently, due to abundance of video calling and multimedia communication services, the importance of this analysis as seen a increase. This has provoked advancements in multimedia processing such as speaker diarization and speaker identification. Recent advances in audio-visual person recognition datasets[11, 14] has allowed for accurately developing an algorithm for person recogntion in unconstrained evnironment. The evaluation of audio-visual person recognition challenge by NIST in 2019 stated that audio-visual fusion models provide over 85% improvement as compared to when only one modality of voice and face were used. Moreover, with the increase in commonality of smart devices such as a smartphone, tablets or laptops there is abundance of biometric markers, i.e. voice and face features, that are easily detectable by the camera and the micrphone built-in to the devices. Ease of biometric feature detection (face detection and voice detection) directly correlates to ease in data-processing of the said biometric features. Capturing data (keeping in view the ethical norms of data collection) are now easier than ever. Hence, the author argues that due to ease in data capturing of the voice and face features, and obtainment of better performance of the recognition algorithms due to using both of these modalities, there is a need for developing a solution that is accurately able to recognize a person using the audio-visual modalities.

1.3 Scope

This research is aimed for development of a multi-modal-input model which can transform both visual and audio features into a common embedding space E where contrastive

learning techniques are used for learning a metric, based on which recognition task can be performed. Essentially the purpose of the algorithm would be to learn to represent the voice and face features of a **single utterance, synchronized audio-video** of a person in the form of a $1 \times N$ dimensional vector (embedding). The embedding **e** would be so that distance between the embeddings of the voice-face features of the same person would be minimized and the distance between the embeddings of the voice-face features of the different persons would be maximized, using metric learning techniques.

So formally defining the scope of the research project, the algorithm has to minimize the distance function $D(e, e')$ for all instances where e & e' are embeddings belonging to the same identity j (a set of which is denoted as E_j) as denoted in equation (1.3.1), and maximize $D(e, e')$ where e & e' belong to different identities, as denoted in equation (1.3.2).

$$\min_{\forall e, e' \in E_j} D(e, e') \tag{1.3.1}$$

$$\max_{\forall e \notin E_j \ \& \ \forall e' \in E_j} D(e, e') \tag{1.3.2}$$

Chapter 2

Literature Review

Early audio-visual person recognition methods usually comprised of Hidden-Markov model. The techniques involved detection of face and processing on voice to combine the face and voice feature-vectors to develop a decision system [1, 2, 3]. Wu et al. provided early works in the fusion of audio and visual modalities for the purpose of person identification by presenting their work in usage of dynamic bayesian network for audio visual correlation[5]. The audio-visual fusion systems proved to be effective in the cases where sensitivity of the input system of one of the two modalities was detrimental for the performance of the algorithm. [6] provided a detailed analysis of fusion of modalities and usage of the audio-visual person recognition (AVPR) system for unconstrained test cases, and thus laying the path for research in the field of audio visual person recognition. Text-independent and text-dependent use cases were discussed. The use-cases of using Artificial Neural Networks apart from methods like GMM and HMM were discussed. Lip-movement was also studied for the purpose of its effect in the performance of audio visual person recognition systems.

Moreover, the work on i-vectors was gaining popularity due to their work in representation of audio speech data into low dimensional representation. These i-vectors were a great success of their time for determining the identity of a speaker, among other tasks [7].

Although there has been many advances in the field of person recognition and speaker recognition from images and speech data respectively, it is still a significant challenge under noisy and unconstrained situations. The CNN based face recognition methods paved way for methods that involved mapping the voice or face features to a vector

embedding (e) into an euclidean space where the distance between the embeddings is directly correspondent of the similarity of the speaker [8, 15, 10].

Similarly, DNN based embeddings gained popularity in mapping the speaker embeddings (x-vectors) to a feature space where the scoring between the embeddings in the lower dimensional feature space represented the similarity among speech. X-vectors were popular because large scale training data was leveraged better by the x-vector embeddings method as compared to the i-vectors method [17]. Moreover, work on DNN and CNN based speaker recognition continued to show improvements as compared to the acoustic i-vector method. Snyder et al. reiterated importance of triplet loss, which has proved itself for metric learning[9], for their work on usage of deep neural networks for text independent speaker verification using a modified version of the said loss function, although using tha PLDA backend which is the same back-end used for acoustic i-vector generation. This proved to improve the EER% in as compared to previous methods. As advancements in the DNN and CNN technologies grew, these advancements got incorporated in the field of speaker recogntion. Zeinali et al. introduced r-vectors by using ResNet architecture [19]. Later, use of Time Delay Neural Networks was emphasized for providing better results for speaker verification by Desplanques et al. The performance of the proposed ECAPA-TDNN architecture on the test sets of VoxCeleb datasets and in the VoxCeleb SR challenge held in 2019, proved to be much better than that of the most advanced TDNN-based systems proposed previously. [20].

As advances in the use of deep neural networks grew, the feild of audio-visual person recognition gained importance. Nagrani et al. made tremendous contributions in the field by introducing a large dataset for audio-visual person recognition [11]. Building upon their work, Shon et al. provided techniques for mid-fusion of voice and face vector embeddings and emphasized upon attention based techniques for state of the art reduction of EER% in the audio-visual person recognition system[16]. The audio-visual speaker recognition challenge by NIST boosted investigation in the novel field of multi-modal speaker recognition[21]. Chung et al. introduced even a larger datsaset of audio-visual speaker recognition and a baseline (VGGVox) for speech-based person recognition. This dataset was first of it's type to be comparable to the state of the art face recognition datasets [14]. In their investigation of audio-visual methods for speaker verification, Sari et al. began with conventional fusion methods for the purpose of learning joint audio-visual embeddings. Next, they proposed an innovative method

for managing cross-modal verification while the test was being conducted. They focused specifically on unimodal and concatenation-based AV fusion in their research. In light of the fact that these methods are incapable of performing cross-modal verification, the researchers developed a multi-view model that mapped audio and visual features into the same space by the use of shared classification technique [24].

2.1 Previous Research

Sari. et al [24] proposed a technique using a shared classification method to achieve better similarity between cross-modal featuers. Audio Embeddings and Visual Embeddings are passed on to a shared classifier. The outputs of the classifier are gathered from unimodal representation of speech embeddings and face emebddings as shown in equation 2.1.1 and equation 2.1.2. Weighted unimodal audio and visual losses are added and trained with arc-margin loss, as shown in equation 2.1.3. the depiction of the system developed by Sari et al. is given in figure 2.1. They achieved 2.0% EER by using their mid-fusion technique.

$$y_A = \mathcal{C}_{AV}\left(E_A\right) \tag{2.1.1}$$

$$y_V = \mathcal{C}_{AV}\left(E_V\right) \tag{2.1.2}$$

$$\mathcal{L}_{AV} = \lambda_A \mathcal{L}_A + \lambda_V \mathcal{L}_V \tag{2.1.3}$$

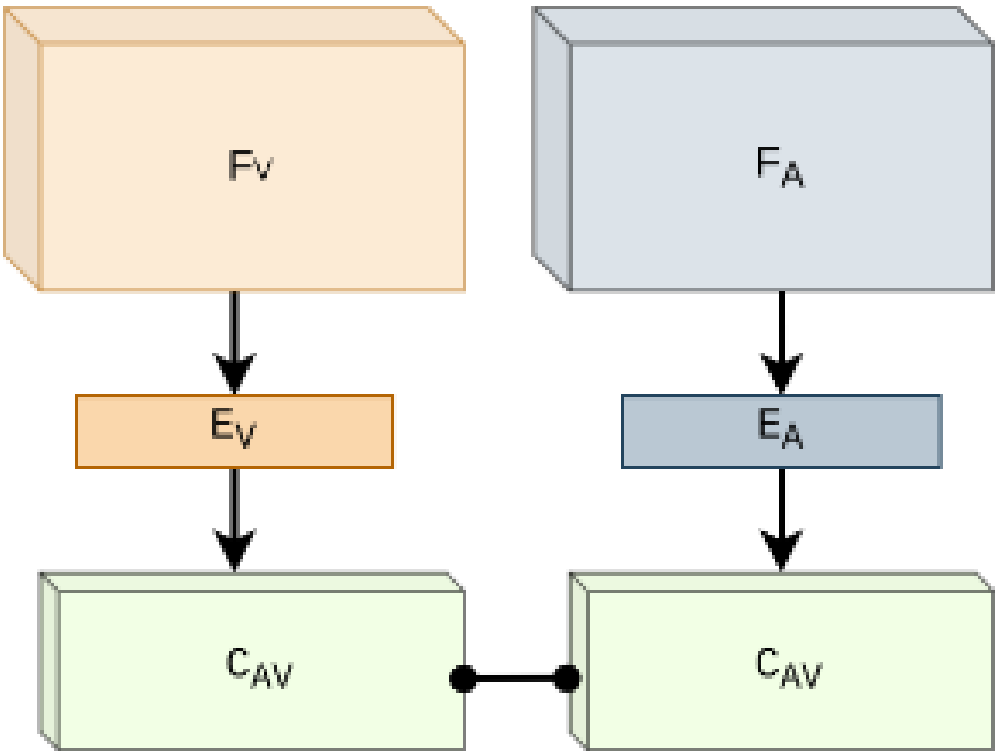

Figure 2.1: Person Recognition system

Chapter 3

Design and Methodology

3.1 Datasets

Following datasets have been used in the development of this research:

- VoxCeleb 2
- VGGFace 2 Dataset
- AGEDB 30
- Labeled Faces in the Wild

Table 3.1 depicts the characteristics of the VoxCeleb2 dataset. This dataset is first of it's kind to provide utterances and number of identities comparable to the modern day facial recognition datasets.

	dev	test
# of speakers	5,994	118
# of videos	145,569	4,911
# of utterances	1,092,009	36,237

Table 3.1: VoxCeleb2 Dataset Characteristics

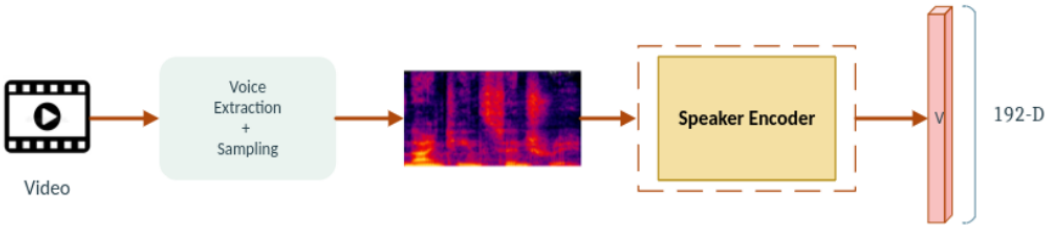

Figure 3.1: Audio Framework Overview

3.2 Audio Framework

Audio framework is the same as used in [20], This audio framework uses a novel architecture based on time delay neural networks. The time delay neural networks are coupled with Emphasized Channel Attention, Propagation and Aggregation techniques to provide superior representation of speech into a 1×194 dimensional vector.

Firstly, the video is passed on to the audio framework module, the audio framework then extracts the audio from the video and samples it at the desired sampling rate i.e. $16000KHz$. The network being used (ECAPA-TDNN) is trained on sampling frequency of $16000KHz$, so the audio input at any other sampling frequency is detrimental to performance of the system. This phenomena is also noted by MLOps experts commonly that misrepresentation of data which is called data-drift decreases the performance of the system to a great extent.

The audio is then passed through either filter banks or MFCC which allows it to be better represented for deep learning applications. This is a common technique also used in early works in the said field of audio processing introduced mainly by Das et al in 2008. Figure 3.1 depicts the overview of the audio framework.

Figure 3.2 provides detailed description of the speaker encoder. The speaker encoder used is a pretrained network of ECAPA-TDNN for speaker verification.

ECAPA-TDNN takes input of 80 dimensional input features, previously generated through filter banks or MFCC. The first layer in the network is a $Conv1D$ layer with a kernel size of 5 and dilation factor of 1. The output of the first layer are then passed on to the activation layer which is ReLU activation and then the output is batch normalized. The input of this layer is in the dimension of C where C is the number of channels for the $Conv1D$ layer.

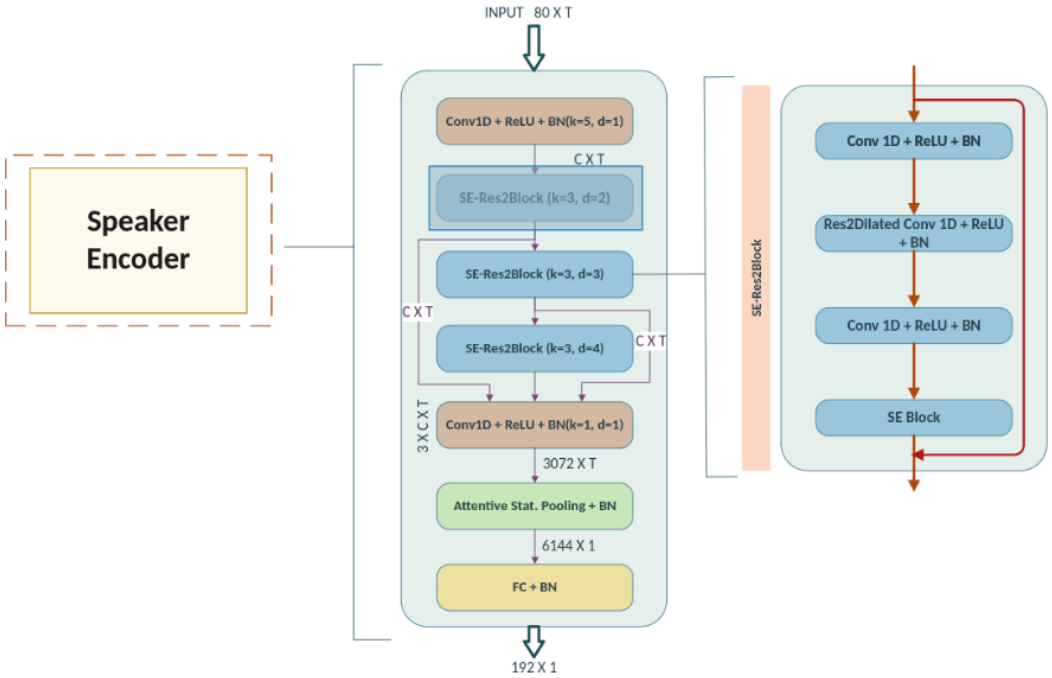

Figure 3.2: ECAPA-TDNN Based Speaker Encoder

The next module in the network is $SE-Res2Block$. The description of $SE-Res2Block$ is as follows. The $SE-Res2Block$ has a first layer of $Conv1D$ with ReLU activation and batch normalization. Then the output if this layer is passed to a $Res2DilatedConv1D$ block, also with ReLU activation and batch normalization. Then again a $Conv1D$ block similar to the previous, followed by a $SqueezeandExcitation$ block. The input of the $SE-Res2Block$ is added to the output of the block. This block has a kernel size of 3 and dilation factor of 2.

The network then has two more $SE-Res2Block$ blocks with dilation sizes of 3 and 4, respectively. This is so that the receptive field of the network increases as the depth of the network increases. Dilation increases the receptive field while keeping the number of parameters the same.

The output of the three $SE-Res2Block$ is concatenated, before being passed on to another $Conv1D$ layer. It then passes on to an $AttentiveStatisticalPooling$ layer before it is passed through a fully connected neural network layer with batch normalization.

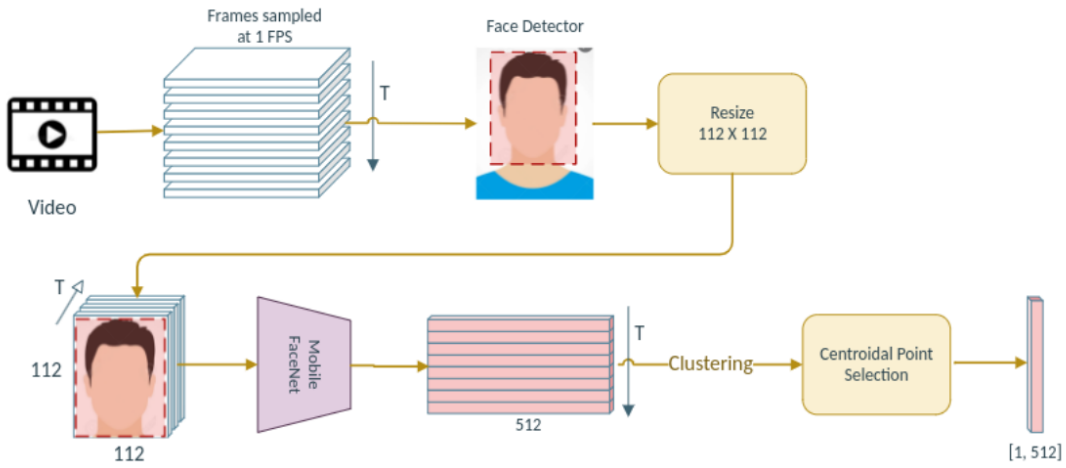

Figure 3.3: Overview of the Video Framework

3.3 Video Framework

A technique of temporal clustering and centroidal point selection was used in the video framework. This technique allows the author to obtain an embedding that is the most optimal embedding for that video. Figure 3.3 depicts the overview of the video framework.

When the video is provided to the video framework module, it is sampled at $1FPS$ and the frames are stacked. The stack of frames are presented to a face detection module which provides bounding box for the the temporally stacked frames. In the case where face is not detected, a zero frame is then added to the temporal stack. The stack of face crops is resized to 112×112. This stack of face crops is then provided to the face recognition module. The face recognition module consists of $MobileFaceNet$[13]. The recognition module then outputs $512 - D$ vectors for each time frame. This gives a $T \times 512$ matrix where each row is an embedding for each time step. These embeddings are then analysed in their latent space. A $centroidalpointselectionmodule$ is used to find the embedding closest to the centroid of the temporal cluster of face emeddings. So temporal clustering and centroidal point selection among that cluster is achieved.

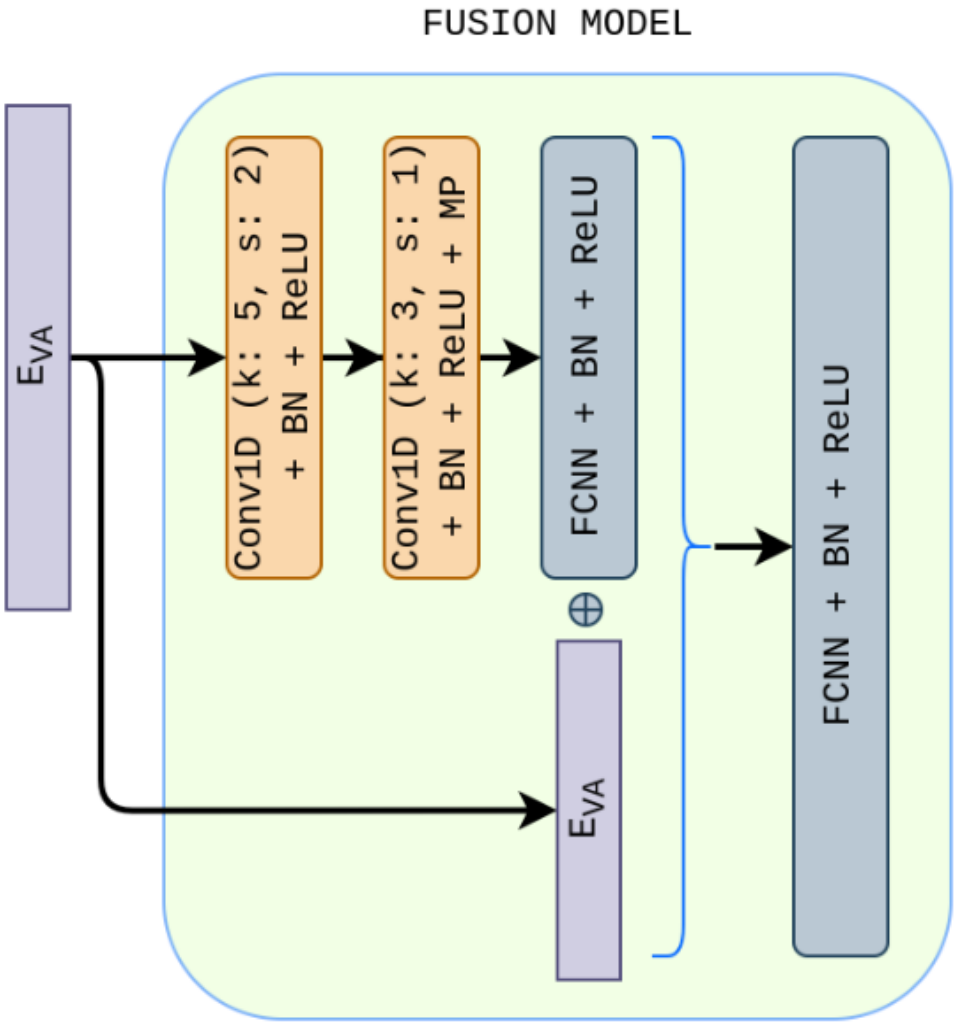

Figure 3.4: Fusion Network

3.4 Fusion Network

The fusion network consists of 4 layers. The first layer is a $Conv1D$ layer with kernel size of 5 and stride of 2. The next layer is another $Conv1D$ layer with kernel size of 3 and stride of 1. Both of these layers are following by ReLU and batch normalization. The depiction of the fusion network is shown in figure 3.4.

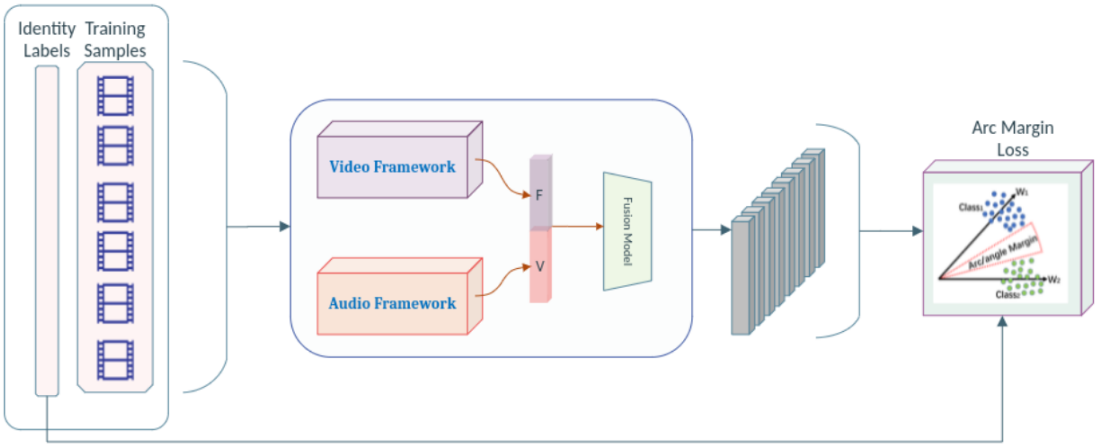

Figure 3.5: Training Routine

3.5 Training

During when the network is being trained, the system takes input a batch of videos along with their labels. The label is actually the label of the identity. The videos are passed to the $AudioFramework$ and $VideoFramework$ as described in the previous subsections. The output of the videos are concatenated and that forms a $704-D$ vector. that vector is then fed to the fusion model and the fusion model is trained with arc-margin loss. The arc margin loss is described in equation 3.5.1. The depiction of training routing is provided in figure 3.5.

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}}{e^{s\left(\cos\left(\theta_{y_i}+m\right)\right)}+\sum_{j=1,j\neq y_i}^{n}e^{s\cos\theta_j}} \tag{3.5.1}$$

$$s = \text{ Scale}$$

$$m = \text{ Margin}$$

3.6 Enrollment

During when the enrollment is required, the videos of the same identity are provided to the system. The embeddings for the whole batch of the videos is computed and the cluster of embeddings of that batch is provided to centroidal point selection mechanism described in previous sections. The centroidal point is the most optimal point to represent the person in those videos. The centroidal point along with the identity label is

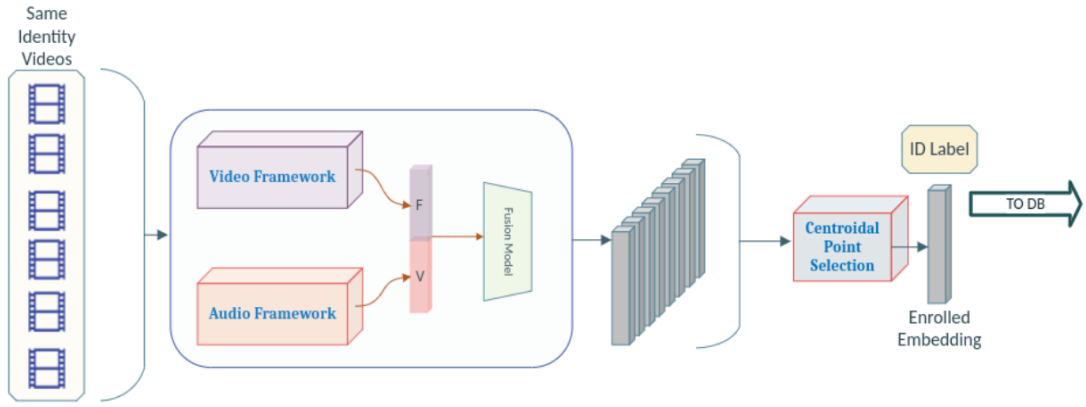

Figure 3.6: Enrollment Routine

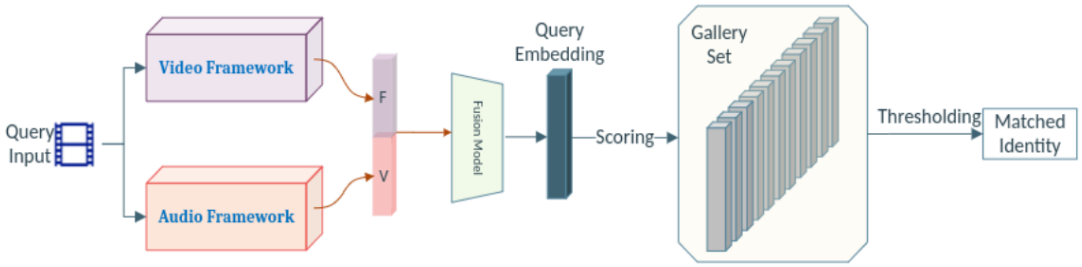

Figure 3.7: Recognition Routine

passed on to the database for storage. Figure 3.6 depicts the routine of the enrollment system.

3.7 Recognition

The input query video is passed on to the audio framework and the video framework. The audio framework compute the audio embeddings and the video framework computes the video embeddings. Both of the embeddings are combined and concatenated. The concatenated embedding is passed to the fusion framework which provides a fused embedding. This embedding is then stored with all the embeddings in the database. Decision rule is applied to the scores between the query embedding and the gallery embeddings. On the bases of the decision rule, which is actually a thresholding technique, the matched identity is the output of the system, in case there is an identity that fulfills the decision rule. Figure 3.7 represent the recognition routine.

3.8 Test Data Collection

Test data was collected for the purpose of evaluation of effect of vocabulary on the recognition system. The vocabulary of the collected data is as follows:

- **Common Greetings**
  1. Assalam o Alaikum
  2. Walaikum Assalam
  3. Good Morning
- **Nation's name**
  1. Pakistan
- **NATO phonetic alphabet**
  1. Bravo
  2. Charlie
  3. Foxtrot
  4. Uniform
- **Name of months**
  1. November
  2. January
  3. August

3.8.1 Data Collection App Overview

An app was developed to collect data from real world users. The home screen prompted users to enter their name (identity label) which proceeded them to phrase-list as shown in figure 3.8. The phrase list contained 11 phrases as described previously. There were some inherent checks within the app to support the scope of the project. The app checked if the face is too close or too far as shown in figure 3.9. More checks included face being aligned to the screen and presence of single face in the view as shown in figure 3.10. The app also enforced centering of the face. Only when all checks are passed, the user is allowed to proceed to recording as shown in figure 3.10.

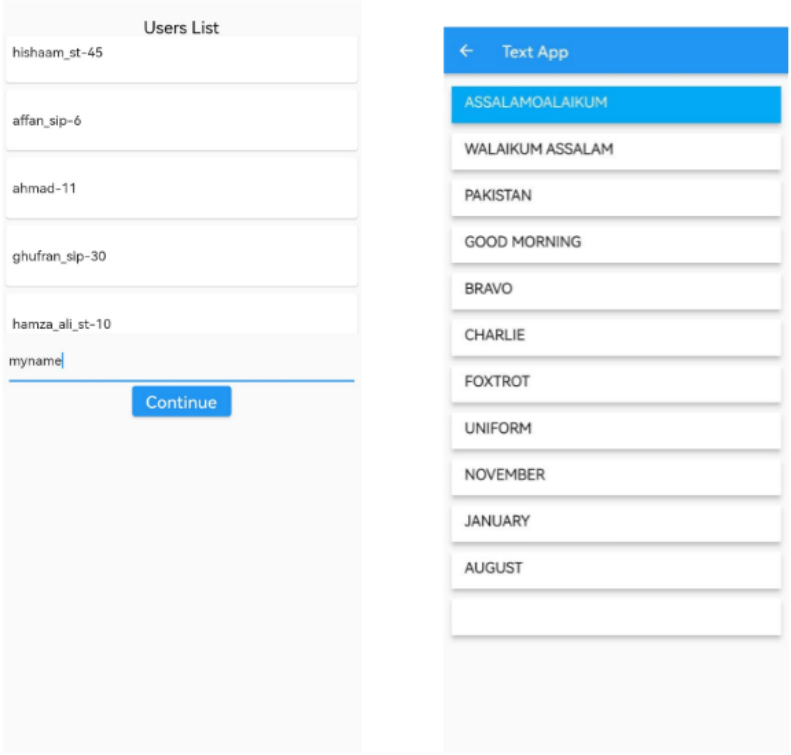

Figure 3.8: Home screen for user enrollment (left). Sentence list (right)

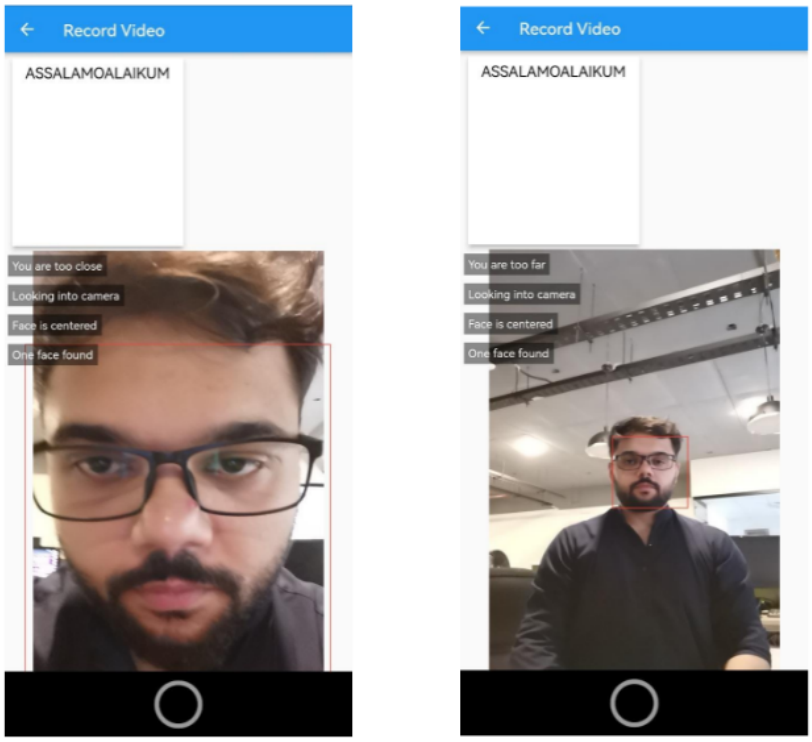

Figure 3.9: Face too close (left). Face too far (right).

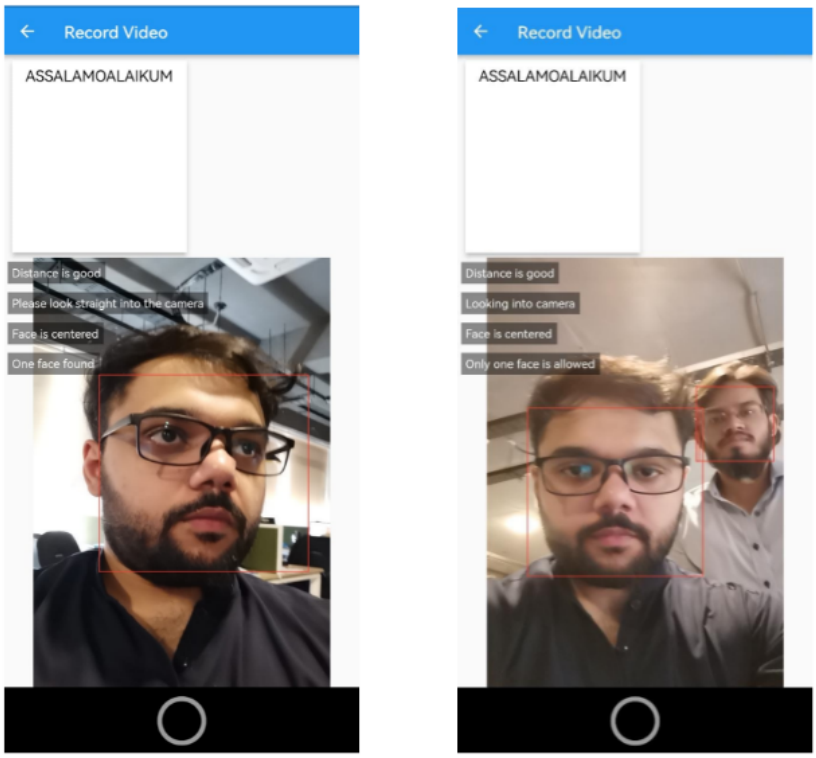

Figure 3.10: Looking away (left). Multiple faces (right).

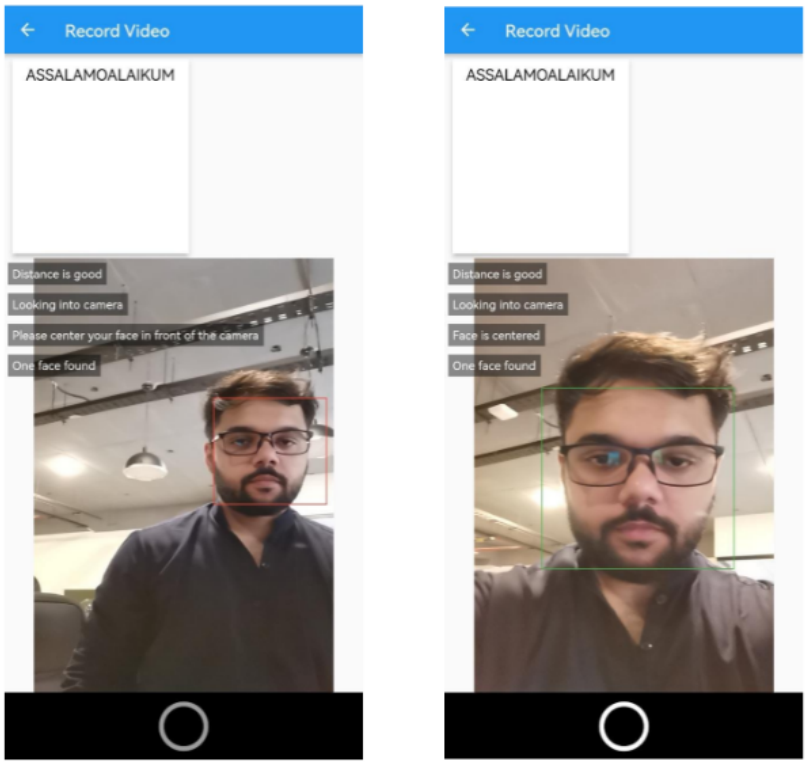

Figure 3.11: Face not centered(left). All checks passed (right).

3.8.2 Collected Data

The number of videos collected amounted to 275, for 25 users each. Each speaker spoke 11 words each for the data collection.

Chapter 4

Experiments and Results

4.1 Test Dataset

Test split of the VoxCeleb2 dataset was used for evaluation of the proposed algorithm. Characteristics of the test set of VoxCeleb2 dataset are briefed in table 4.1.

4.2 Model Selection

The model shown in section 3.3 was selected with the following in mind. It is infamous that the model's learning ability leans more towards one modality during the learning process. This is not the intent of the author. Which is why the author decided to concatenate the input layer into the 2nd last layer of the fusion model, so that even if the previous layers learn modality independent features, the output layer will still contain information that is directly coming from both the modalities.

The second reason was related to the learning ability of the model. $Conv1D$ layers are known to learn spatial information in one dimensional direction. Since the input of the model is in the form of the vector, so striding kernels over the vector would have superior

	test set
# of speakers	118
# of videos	4,911
# of utterances	36,237

Table 4.1: VoxCeleb2 Test Set Characteristics

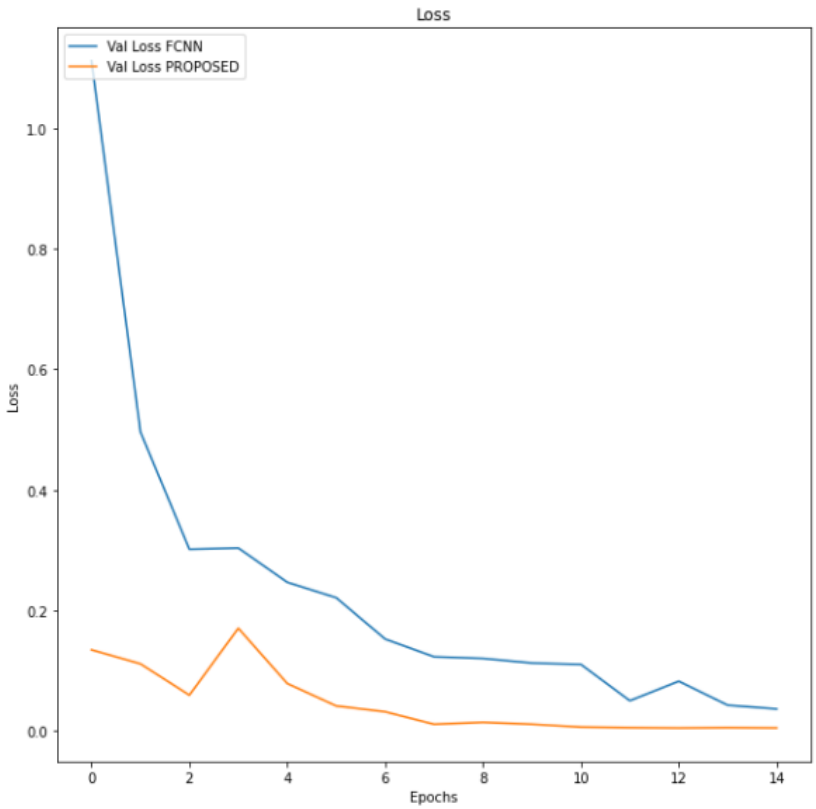

Figure 4.1: Model selection and comparison.

ability over the FCNN to learn the relationship between the voice and face embeddings. This is also depicted when the models are being trained, as shown in figure 4.1, the $Conv1D$ based model shows faster convergence over the FCNN-only based model.

4.3 Results

The positive and negative pairs were made from the dataset. The positive pairs consisted of a pair of embeddings belonging to the same identity, whereas, the negative pairs consisted of embeddings belonging to different identities. The distribution between the scores of imposter pairs (negative pairs) and genuine pairs (matching pairs) is shown in figure 4.2. In this figure, the overlap would depict that the scores of the matching and non-matching pairs are overlapping.

The FMR and FNMR curves are important factors of selecting a threshold value of the developed solution. Moreover, the curves visually depict the performance of the algorithm on the developed solution. Figure 4.3 illustrates the change in FMR and

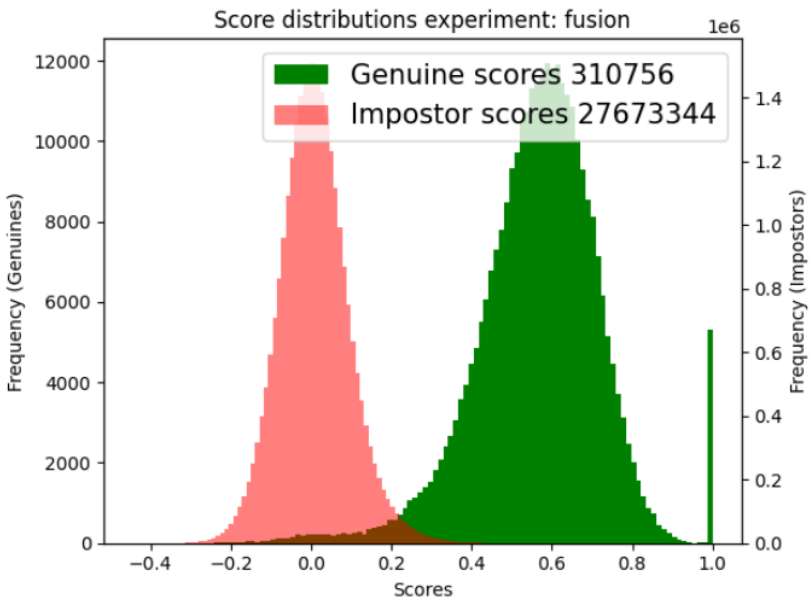

Figure 4.2: Score Distribution in the Test Dataset.

EER	0.020074089753085
Zero FMR Threshold	0.642737733764255
EER Threshold	0.202290341879147

Table 4.2: Performance Metrics of the Developed Solution.

FNMR for sweeping the value of the decision threshold.

Figure 4.4 illustrates the ROC curve of the developed solution. It can be seen that the Area Under The curve. The ROC curves illustrates the performance of the model at all thresholds. The developed system achieved Area Under the Curve value of 0.993.

Figure 4.5 and figure 4.6 illustrate the Detection Error Trade-off. The trends shown in these curves are a depiction of near to ideal classification between positive and negative pairs.

Table 4.2 provides numerical values for the system. The goal of the recognition systems are to mimize the EER%. The state of the art developed by [24] reported the same EER% on their mid-fusion model.

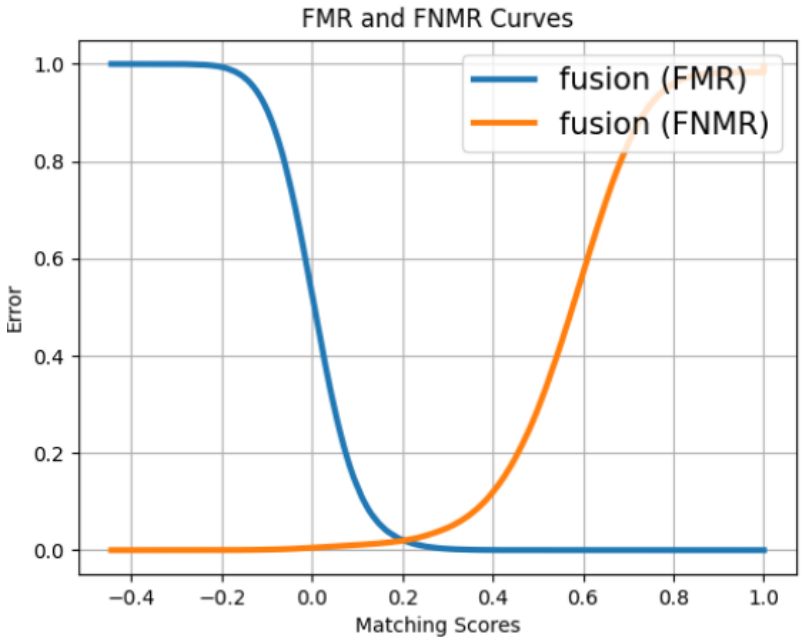

Figure 4.3: FNMR and FMR Curves of the Developed Solution on the Test Set.

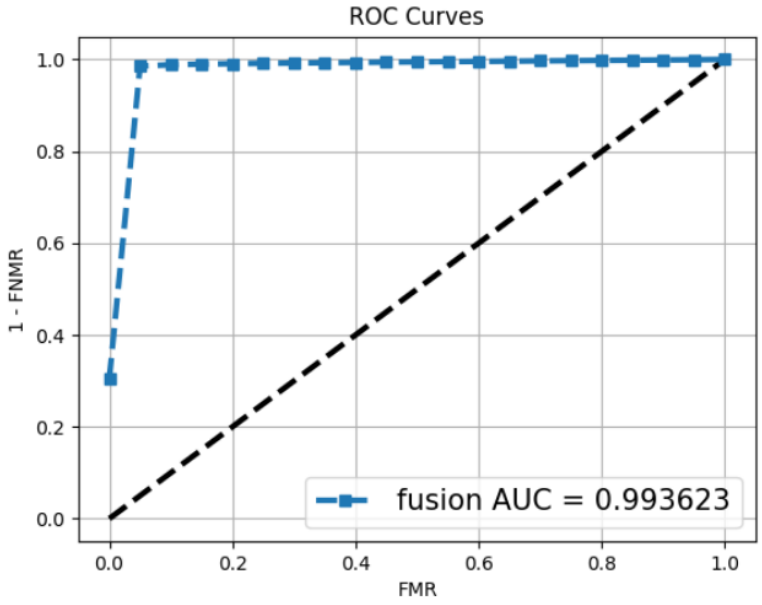

Figure 4.4: ROC Curve of the Developed Solution on the Test Set.

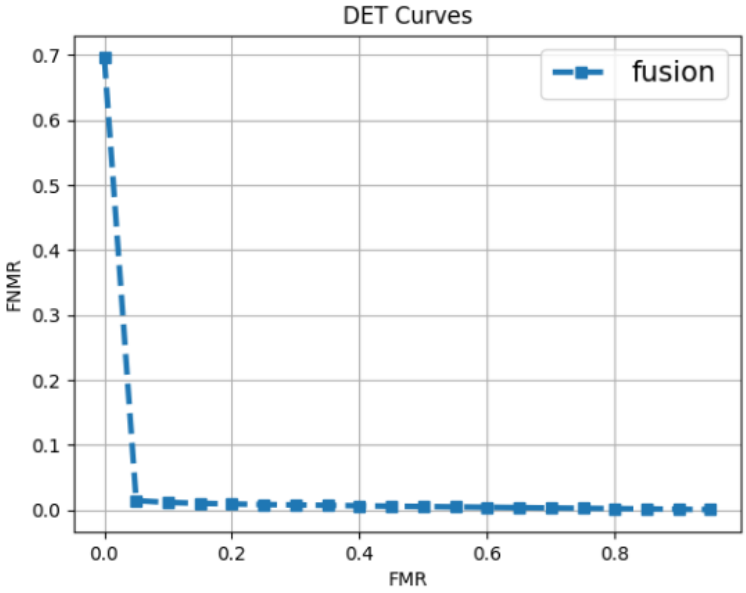

Figure 4.5: DET curve of the Developed Solution on the Test Set.

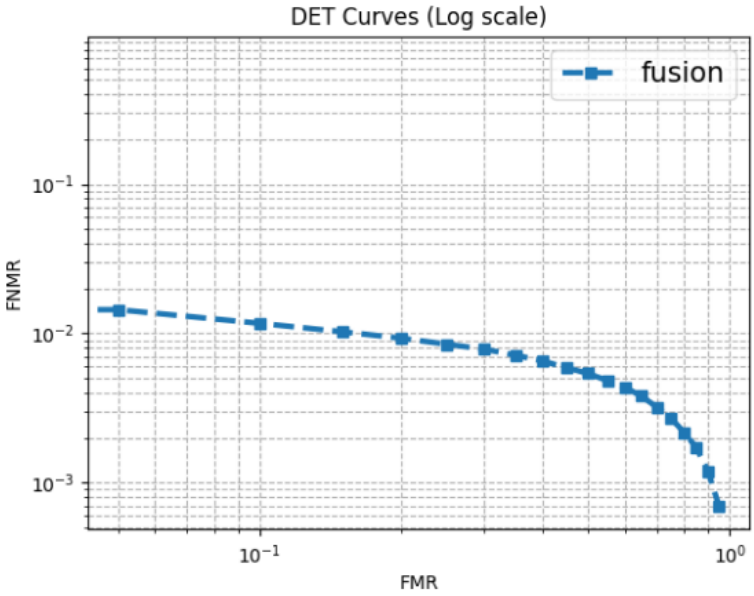

Figure 4.6: DET Curve in Log Scale of the Developed Solution on the Test Set.

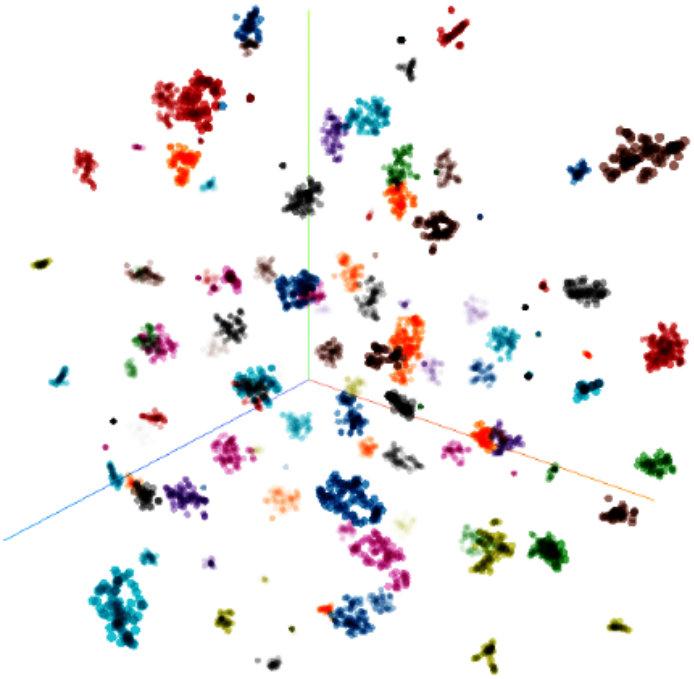

Figure 4.7: t-SNE Analysis of Sample from Test Set.

4.4 t-SNE Analysis of Sample from Test Set

A method for dimensionality reduction and visualization of high dimensional data points into a lower dimensional space is t-SNE[1]. figure 4.7 shows the grouping of the identities when analysed using t-SNE analysis.

4.5 Optimal Vocabulary Analysis

Optimal Vocabulary Analysis of the 11 words collected from 25 unique identities. Figure 4.8 Shows that the word "Walaikum Assalam" and "Pakistan" show the most optimal results. These words were the most optimal for 5 out of 25 users each.

[1] t-SNE analysis was done using projector.tensorflow.org. Accessed: 22-06-27

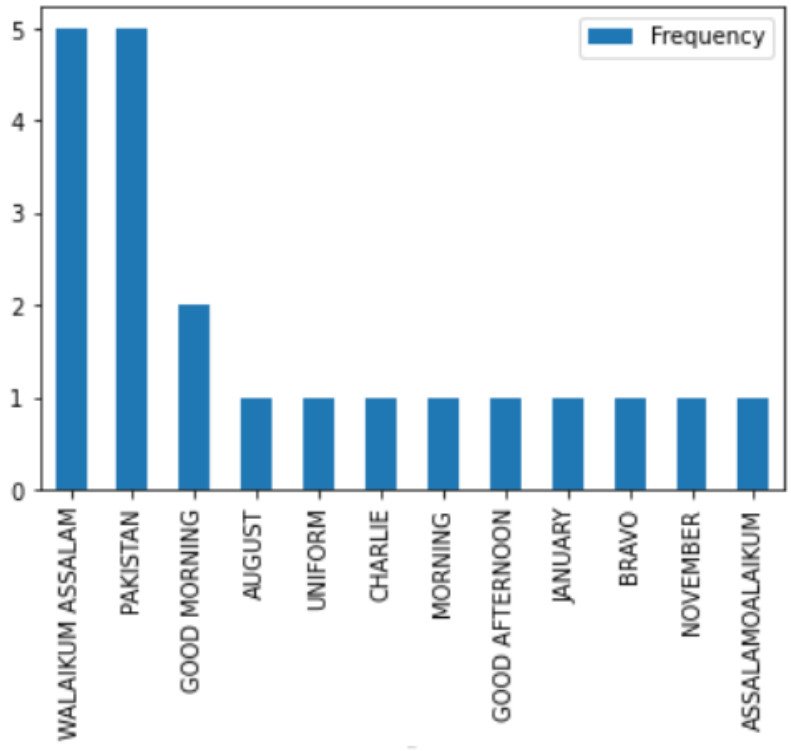

Figure 4.8: Optimal Vocabulary Analysis of the 11 words collected from 25 unique identities.

Chapter 5

Conclusion

Metric learning technique was used for the task of training a mid-fusion model. The metric learning loss that was used was Arc-Margin Loss. This loss was already used for training of the uni-modal voice and uni-modal face networks. The performance of the system comparable to the state of the art was achieved while exploration of FCNN and 1D ConvNet based models was performed. The 1D ConvNet showed better learning ability and quicker convergence as compared to the FCNN-only based network. Dimensionality reduction using t-SNE showed that when embeddings are plotted in lower dimensionality, seperations between the identities is visualized.

Vocabulary analysis was performed by collecting real world data from multiple users. The data collection was performed in an assisted way by providing a smartphone application user interface. Vocabulary analysis showed that long and complex words are more likely to be optimal for recognition.

Chapter 6

Future Work

VoxCeleb 1 and VoxCeleb 2, although first of their kind, do not still have the multilingual information of the data-points. These datasets contain English language utterances only. To study the effect of multilingual queries to the system, an analysis of using multilingual dataset needs to be performed.

Moreover, There is a need for determining effect of vocabulary on the recognition system. Although, due to inclusion of large vocabulary in the training data, one might presume the effect of vocabulary change might not be significant but findings in this research showed significance of such an analysis.

Collection of more real-world data to analyze the effects of difference of region and vocabulary of identities has large potential to scrutinize of such bio-metric systems.

References

[1] Souheil Ben-Yacoub et al. “Audio-Visual Person Verification”. In: vol. 1. IDIAP-RR 98-18. 1999, pp. 580–585. DOI: 10.1109/CVPR.1999.786997. URL: http://infoscience.epfl.ch/record/82501.

[2] C.C. Chibelushi, F. Deravi, and J.S.D. Mason. “A review of speech-based bimodal recognition”. In: *IEEE Transactions on Multimedia* 4.1 (2002), pp. 23–37. DOI: 10.1109/6046.985551.

[3] Tieyan Fu et al. “Audio-visual speaker identification using coupled hidden Markov models”. In: *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*. Vol. 3. 2003, pp. III–29. DOI: 10.1109/ICIP.2003.1247173.

[4] Rong Zheng, Shuwu Zhang, and Bo Xu. “Text-independent speaker identification using GMM-UBM and frame level likelihood normalization”. In: *2004 International Symposium on Chinese Spoken Language Processing*. 2004, pp. 289–292. DOI: 10.1109/CHINSL.2004.1409643.

[5] Zhiyong Wu, Lianhong Cai, and Helen Meng. “Multi-level Fusion of Audio and Visual Features for Speaker Identification”. en. In: *Advances in Biometrics*. Ed. by David Zhang and Anil K. Jain. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2005, pp. 493–499. ISBN: 978-3-540-31621-3. DOI: 10.1007/11608288_66.

[6] Petar S. Aleksic and Aggelos K. Katsaggelos. “Audio-Visual Biometrics”. In: *Proceedings of the IEEE* 94.11 (2006), pp. 2025–2044. DOI: 10.1109/JPROC.2006.886017.

[7] Najim Dehak et al. “Front-End Factor Analysis for Speaker Verification”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2011), pp. 788–798. DOI: 10.1109/TASL.2010.2064307.

[8] Yaniv Taigman et al. “DeepFace: Closing the Gap to Human-Level Performance in Face Verification”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1701–1708. DOI: 10.1109/CVPR.2014.220.

[9] Elad Hoffer and Nir Ailon. “Deep Metric Learning Using Triplet Network”. en. In: *Similarity-Based Pattern Recognition*. Ed. by Aasa Feragen, Marcello Pelillo, and Marco Loog. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 84–92. ISBN: 978-3-319-24261-3. DOI: 10.1007/978-3-319-24261-3_7.

[10] Florian Schroff, Dmitry Kalenichenko, and James Philbin. “FaceNet: A unified embedding for face recognition and clustering”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682.

[11] A. Nagrani, J. S. Chung, and A. Zisserman. “VoxCeleb: a large-scale speaker identification dataset”. In: *INTERSPEECH*. 2017.

[12] Chunlei Zhang and Kazuhito Koishida. “End-to-End Text-Independent Speaker Verification with Triplet Loss on Short Utterances”. In: *Proc. Interspeech 2017*. 2017, pp. 1487–1491. DOI: 10.21437/Interspeech.2017-1608.

[13] Sheng Chen et al. “MobileFaceNets: Efficient CNNs for Accurate Real-time Face Verification on Mobile Devices”. In: *CoRR* abs/1804.07573 (2018). arXiv: 1804.07573. URL: http://arxiv.org/abs/1804.07573.

[14] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. “VoxCeleb2: Deep Speaker Recognition”. In: *Interspeech 2018*. ISCA, Sept. 2018. DOI: 10.21437/interspeech.2018-1929. URL: https://doi.org/10.21437%2Finterspeech.2018-1929.

[15] Jiankang Deng, Jia Guo, and Stefanos Zafeiriou. “ArcFace: Additive Angular Margin Loss for Deep Face Recognition”. In: *CoRR* abs/1801.07698 (2018). arXiv: 1801.07698. URL: http://arxiv.org/abs/1801.07698.

[16] Suwon Shon, Tae-Hyun Oh, and James R. Glass. “Noise-tolerant Audio-visual Online Person Verification using an Attention-based Neural Network Fusion”. In: *CoRR* abs/1811.10813 (2018). arXiv: 1811.10813. URL: http://arxiv.org/abs/1811.10813.

[17] David Snyder et al. “X-Vectors: Robust DNN Embeddings for Speaker Recognition”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 5329–5333. DOI: 10.1109/ICASSP.2018.8461375.

[18] Shuai Wang et al. “Discriminative Neural Embedding Learning for Short-Duration Text-Independent Speaker Verification”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.11 (2019), pp. 1686–1696. DOI: 10.1109/TASLP.2019.2928128.

[19] Hossein Zeinali et al. *BUT System Description to VoxCeleb Speaker Recognition Challenge 2019*. 2019. arXiv: 1910.12592 [eess.AS].

[20] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification”. In: *Interspeech 2020*. ISCA, Oct. 2020. DOI: 10.21437/interspeech.2020-2650. URL: https://doi.org/10.21437%2Finterspeech.2020-2650.

[21] Seyed et al. “The 2019 NIST Audio-Visual Speaker Recognition Evaluation”. en. In: The Speaker and Language Recognition Workshop: Odyssey 2020, Tokyo, -1, 2020-05-18 2020. URL: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=929541.

[22] Ruijie Tao, Rohan Kumar Das, and Haizhou Li. “Audio-Visual Speaker Recognition with a Cross-Modal Discriminative Network”. In: *INTERSPEECH*. 2020.

[23] Yanmin Qian, Zhengyang Chen, and Shuai Wang. “Audio-Visual Deep Neural Network for Robust Person Verification”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 1079–1092. DOI: 10.1109/TASLP.2021.3057230.

[24] Leda Sari et al. “A Multi-View Approach to Audio-Visual Speaker Verification”. In: June 2021, pp. 6194–6198. DOI: 10.1109/ICASSP39728.2021.9414260.

[25] *Digital Around the World*. en-GB. URL: https://datareportal.com/global-digital-overview (visited on 06/19/2022).