Active Speaker Detection

Kashaf Khan
Computer Systems Engineering
NED University of Engineering and Technology
Karachi, Pakistan
kashafsarwarkhan@gmail.com

Hafsa Zafar
Computer Systems Engineering
NED University of Engineering and Technology
Karachi, Pakistan
zhafsa729@gmail.com

Aleesha Ahmed
Computer Systems Engineering
NED University of Engineering and Technology
Karachi, Pakistan
aleeshaahmed30@gmail.com

Abstract

Active speaker detection is a crucial task in various applications such as video conferencing, surveillance, and multimedia content analysis. This research paper presents an innovative approach to active speaker detection by combining audio detection with advanced preprocessing techniques and convolutional neural network (CNN) models. The proposed methodology also integrates speaker diarization using Gaussian Mixture Models (GMM) and video model implementation utilizing the Retina-Face library for facial area detection.

The research begins with an exploration of audio preprocessing methods that enhance the audio data quality and extract relevant features for speaker detection. A CNN-based audio model is then implemented to classify audio segments into active and non-active speaker categories. This model is trained on a diverse dataset of audio recordings to achieve accurate speaker detection.

To further enhance the accuracy of active speaker detection, a speaker diarization model based on GMM is introduced. This model segments audio streams into distinct speaker segments, allowing for a more refined analysis of the active speaker.

Additionally, the paper presents the integration of visual cues by utilizing the Retina-Face library for facial area detection in video frames. This visual information is fused with the audio-based active speaker detection to improve the overall detection accuracy, considering both audio and visual context.

Experimental results demonstrate the effectiveness of the proposed approach in accurately detecting active speakers in various scenarios. The combined use of audio detection, preprocessing, CNN audio models, GMM-based speaker diarization, and Retina-Face-based video analysis yields promising results, showcasing the potential of multi-modal active speaker detection.

In conclusion, this research contributes to the advancement of active speaker detection methodologies by incorporating audio and visual cues, preprocessing techniques, and advanced machine learning models. The proposed approach has the potential to enhance the performance of active speaker detection systems across different applications.

Keywords— CNN, GMM, Retina-Face, multimodal detection, speaker diarization

I. INTRODUCTION

In the modern era of digital communication and multimedia content consumption, the effective identification and tracking of active speakers in various scenarios have gained paramount importance. Active Speaker Detection (ASD) is a fundamental task that finds applications in video conferencing, remote collaboration, surveillance systems, content analysis, and more. The ability to accurately discern the active speaker among multiple participants or entities in an audio-visual stream holds the potential to enhance communication quality, facilitate content analysis, and improve user experience.

ASD plays a crucial role in enabling efficient and natural communication in various contexts. In video conferencing platforms, for instance, identifying the active speaker aids in focusing the visual display on the individual who is currently speaking, allowing participants to better understand the ongoing conversation. Similarly, in surveillance scenarios, ASD can be leveraged to prioritize monitoring of specific individuals or regions where speech is detected, contributing to efficient resource allocation and threat assessment.

The complexity of ASD arises from the multi-modal nature of audio-visual data. In a typical scenario, audio cues provide essential information regarding speech activity, while visual cues such as facial expressions and lip movements can further refine the detection process. Achieving accurate ASD necessitates the integration of diverse technologies spanning audio signal processing, machine learning, computer vision, and data fusion techniques.

This research paper presents a comprehensive exploration of ASD methodologies that leverage advanced techniques from audio processing, machine learning, and computer vision domains. The objective is to contribute to the enhancement of existing ASD systems by addressing challenges such as noise interference, speaker overlap, and robustness across different environments. The proposed approach seeks to exploit the synergies between audio and visual information, enabling a more precise identification of active speakers.

The structure of this paper is as follows: Section II provides an overview of related work and the current state-of-the-art in ASD techniques. Section III elaborates on the proposed methodology, encompassing audio preprocessing, convolutional neural network (CNN) audio models, Gaussian Mixture Models (GMM) for speaker diarization, and the utilization of the Retina-Face library for video-based facial area detection. Section IV presents experimental results and

discusses the findings, highlighting the effectiveness of the proposed approach. Lastly, Section V concludes the paper and outlines avenues for future research in the field of ASD.

In essence, this research endeavors to contribute to the advancement of ASD systems that can adapt to diverse scenarios, enhance communication experiences, and facilitate efficient content analysis through the integration of audio and visual cues.

II. RELATED WORK

Modeling audiovisual data from a single speaker is the main focus of current active speaker detection techniques. This approach may work well in situations with a single speaker, but it hinders reliable detection when trying to figure out which of several potential speakers is speaking.[1]

For the first time, Adekunle Akinrinmade, Emmanuel Adetiba, Joke A. Badejo employed the standard deviation of the mouth region from frame to frame for the prediction of active speakers in a unique concept for active speaker recognition in digital videos. They proposed a technique to determine as the speaker's lips open and shut, revealing the speaker's mouth's inside contents. The standard deviations of color histograms of the mouth region can be used to identify active speakers in movies. [2]

A concept of three features—audio, video, and social—that are based on participant gaze—whose direction is considered to be the same as the orientation of the head—are used to estimate speakers in a meeting situation. The outcomes might be enhanced by additional social behavior study. Francisco Madrigal, Frédéric Lerasle,Lionel Pibre, Isabelle Ferrané proposed a speaker detection framework incorporating audiovisual elements from the conference environment. CNN processes visual cues by analyzing motion and raw pixels (RGB pictures). [3]

A speaker diarization technique to enable speech recognition on multi-speaker audio recordings is used. By combining the most recent advancements in neural approaches, this research makes a significant contribution to the field by offering a survey work that will help the field move closer to an effective speaker diarization.

RetinaFace, a reliable single-stage face detector, uses joint extra-supervised and self-supervised multi-task learning to achieve pixel-wise face localization on various scales of faces. The technique of Retina Face for face identification, is put forth to address the difficult problem of concurrent dense localization and alignment of faces of arbitrary scales in images. Retina Face also produces far more accurate results in detecting faces when paired with cutting-edge techniques.

1.

III. METHODOLOGY

We use a simple, straightforward approach. We assess the probability of each model to implement the multimodal approach to achieve better results. The multimodal technique uses two audio models and one visual model to detect active speakers in the scene. We then use score based fusion of audio and visual models to determine a single active speaker in the scene. [4]

A. Audio Extraction from Video

Frame by frame extraction of audio for 1 second around the current frame is performed using moviepy library. Each extracted audio is then written to an audio file in the wav format to be further processed by both the audio models.

B. Audio Preprocessing

Audio dataset consists of features youtube identifiers, start_time_stamps, end_time_stamps and output speech having labels clean_speech, no_speech, speech_with_music, speech_with_noise. Our dataset consists of some duplicate and missing values that may give an incorrect view of the overall statistics of data. Duplicate data values are dropped and there were many null or missing values that were filled using numerical methods like mean and mode. .The count of missing values (null or NaN) in each column of a Data Frame named df is determined via sum (). The amount of missing values for each column is displayed in a series that is returned.

Missing timestamp feature values are filled with arithmetic mean to remove the null values. Missing values of youtube identifier feature and speech are filled using mode in order to remove the null values. A data frame with the name df uses the function value_counts() to count the instances of unique values in a certain column. It gives back a series that displays the counts of every distinct value in the given column.

Youtube videos are downloaded using pytube library. All the unique youtube identifiers in our dataset having different time stamps are attached with the URL of youtube and are redirected to that link. The audios are then filtered and are streamed from those YouTube videos. Once the audios are extracted they are saved in a separate folder named ‘audios’ in the form of mp4.

After this step, audio feature extraction is done using mfcc.

C. Audio Classification Model

For the model, CNN is implemented. The layers of the Keras Sequential API are used to define the model architecture.Conv1D With 32 filters and a kernel size of 3, represents a 1D convolutional layer. Rectified Linear Unit, or ReLU, is the activation function that is used. The shape of a single training example is represented by the input_shape setting of (7138, 1).MaxPooling1D layer uses a pool size of 2 layers to perform max-pooling. The output of the preceding layer is flattened into a 1D array by this layer. Fully connected and dense layer has 64 units and ReLU activation. Layers.Dense has four units (for four classes) and a softmax activation function that transforms the output into probabilities. The training data (X_train and y_train) are used to train the model. Training is done with a batch size of 32 across 5 epochs. The weights of the model are modified throughout training in order to reduce the desired loss. Finally, the accuracy of the model is evaluated.

D. Speaker Diarization Model

To identify whether or not each frame contains voice activity, we calculate the RMS (Root Mean Square) energy of the audio frames and compare it with a predefined threshold. We use a boolean array to represent speech activity for each frame returned which determines whether a person is audible or not.

We then calculate the ideal number of clusters for GMM-based speaker diarization and Bayesian Information Criterion (BIC) scores for various cluster counts.We then load the audio data, extract the MFCC features, and then train a GMM using the optimal clusters determined by plotting the BIC score graph . Agglomerative Clustering is also carried out on the probabilities of segments belonging to various clusters. As a result we get the cluster assignments and the segment likelihoods.We use the cluster assignments from the GMM-based clustering to create a binary array where each frame is associated with a specific cluster label by assigning cluster labels to individual frames.We build a Data Frame that represents speaker diarization by processing the clustering findings. Based on the results of the clustering, determine speaker change points, separates the audio into speaker segments, and generates a Data Frame containing the name of the audio file, the start and end times, and the speaker label.[5]

E. Video Model

Retina-Face library is used for face detection and for extracting features. It is a deep learning-based face identification and alignment library that specializes in recognizing faces in images with different sizes. It is intended to perform effectively with high-resolution images, making it appropriate for applications such as facial recognition.It takes frames of the video as input and performs pixel-wise face localization on various scales of the faces. After analyzing, it returns facial area coordinates and some landmarks (eyes, nose and mouth) with a score.[6]

When the face and the facial features are detected, separate matrices of landmarks of every face in each frame are formed. The difference between the two matrices of subsequent frames are calculated and the resulting matrix is compared with the threshold. If the resulting matrix is greater than the threshold the speaker is active. If the resulting matrix is less than the threshold then the speaker is non-active.[7]

F. Audio Visual Fusion

The audio and visual models are fused on the basis of score of audio classification model and visual model. For every face identified in the scene, the score of that person is fused with the score of the ‘clean speech’ label predicted by the audio model. The fusion technique used is simple addition technique,

G. Active Speaker Detection

An active speaker is detected by taking the maximum fused score and comparing it with a threshold of 1.3. The facial coordinates of the speaker with the maximum score and value > threshold are then found and the camera is focused on the active speaker.

.

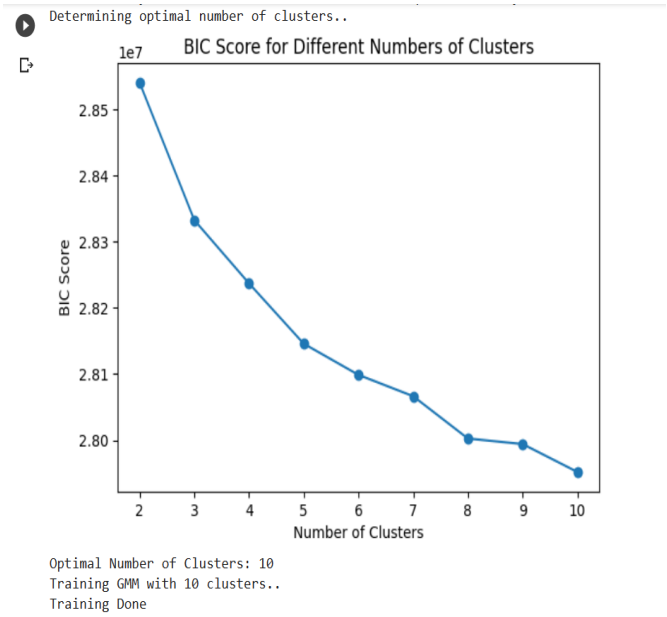
IV. EXPERIMENTAL RESULTS

In this section, we provide an empirical analysis of the proposed method.

The dataset provided consisted of around 7000 entries in the csv file. From the data sample features are extracted, which are then used to train the CNN model. The training statistics of the CNN model are given below:

```
Epoch 1/5
125/125 [==============================] - 18s 134ms/step - loss: 1.7741 - accuracy: 0.2592
Epoch 2/5
125/125 [==============================] - 18s 145ms/step - loss: 1.3375 - accuracy: 0.3433
Epoch 3/5
125/125 [==============================] - 17s 138ms/step - loss: 1.1687 - accuracy: 0.4787
Epoch 4/5
125/125 [==============================] - 17s 135ms/step - loss: 0.8764 - accuracy: 0.7065
Epoch 5/5
125/125 [==============================] - 18s 148ms/step - loss: 0.5345 - accuracy: 0.8710
0.8550000190734863
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 7136, 32)          128

 max_pooling1d (MaxPooling1D  (None, 3568, 32)         0
 )

 flatten (Flatten)           (None, 114176)            0

 dense (Dense)               (None, 64)                7307328

 dense_1 (Dense)             (None, 4)                 260

=================================================================
Total params: 7,307,716
Trainable params: 7,307,716
Non-trainable params: 0
```

The CNN model is trained with 85% accuracy. The audio is then given to the speaker change model which determines number of clusters by plotting the BIC score graph and trains the GMM.



For the video model, the retina face library detects and extracts faces.

The three models are then used to finally detect an active speaker based on score based fusion and thresholding. We finally get the facial coordinates of the active speaker and crop the frame to focus on the person currently speaking.

V. CONCLUSION

A. Future Enhancements

- Subtitle support can be provided while detecting an active speaker. This helps to understand what the speaker is speaking thus providing more ease.
- The active speaker detection system can be applied to real-time systems like live video conferencing where the speaker is detected in live videos.
- Noise-Resistant models can also be implemented for accurate detection in cluttered environments.
- Privacy measures can also be taken in order to handle the data securely.
- User-friendly real time interfaces can also be integrated for easy visualization.
- Deep learning architecture can also be applied like RNNs, CNNs for enhanced detection of the speaker.
- Use of contrast learning approach and positional encoding for better detection. [8]

REFERENCES

[1] Alcázar, Juan León, et al. "Active speakers in context." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

[2] Akinrinmade, Adekunle & Adetiba, Emmanuel & Badejo, Joke & Oshin, Oluwadamilola. (2023). An Active Speaker Detection Method in Videos using Standard Deviations of Color Histogram. 1-6. 10.1109/SEB-SDG57117.2023.10124488.

[3] F. Madrigal, F. Lerasle, L. Pibre and I. Ferrané, "Audio-Video detection of the active speaker in meetings," 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 2536-2543, doi: 10.1109/ICPR48806.2021.9412681.

[4] Alcázar, Juan León, et al. "Maas: Multi-modal assignation for active speaker detection." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.

[5] Tae Jin Parka,∗ , Naoyuki Kandab,∗ , Dimitrios Dimitriadisb,∗ , Kyu J. Hanc,∗ , Shinji Watanabed,∗ , Shrikanth Narayanan "A Review of Speaker Diarization: Recent Advances with Deep Learning " (2021)

[6] Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I. and Zafeiriou, S., 2019. Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*.

[7] Fan, Yue, et al. "Cn-celeb: a challenging chinese speaker recognition dataset." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[8] Wuerkaixi, Abudukelimu, et al. "Rethinking audio-visual synchronization for active speaker detection." *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2022.