An Active Speaker Detection Method in Videos using Standard Deviations of Color Histogram

Adekunle Akinrinmade (adekunleakinrinmade@gmail.com)
Covenant University https://orcid.org/0000-0002-2016-0832
Emmanuel Adetiba
Covenant University College of Engineering
Joke A. Badejo
Covenant University COE: Covenant University College of Engineering

Research

Keywords: Active Speaker Detection, Color Histograms, Standard Deviations

Posted Date: July 21st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1848123/v1

Abstract

Active Speaker Detection (ASD) refers to the process of predicting who amongst a number of speakers whose faces appear on screen is speaking (if any) at any given time within the duration of a video. This paper proposes a novel method for determining active speakers in videos based on the standard deviations of Color Histograms (CHs) of the mouth region from frame-to-frame. The reasoning behind this is that the lips of an active speaker will open and close exposing and concealing the inner contents of the mouth such as the vocal cavity, teeth and tongue at fairly regular intervals in the process which are of different colors. Therefore, if the mouth region can be accurately localized and the changes in the color activities in that region analyzed during speaking such information can be used to detect if a person is actively speaking or not. The lips of a non-speaker are usually closed and at rest, so the CHs for such mouth region are expected to be fairly constant and as such the standard deviations should be low. If an experimentally determined threshold could be set, it can draw the line between active and non-active speakers. In this work, 53 videos available online from Channels TV news, one of Nigeria's most popular TV stations were used to create 250 video clips totaling 3.6 hours, each ranging from between 15 seconds to 1 minute in such a way that the faces of two speakers were always simultaneously visible in any order in the duration of each video clip. The active speakers in each second of the video clips were manually labeled and used to evaluate the performance of the proposed methodology which achieved a prediction accuracy of up to 99.19%.

1. Introduction

Active speaker detection (ASD) refers to the process of recognizing amongst a group of people who are visible in a video the one(s) speaking (if any) at any point in time. ASD seeks to classify if a given face at a given time in a video is speaking or not [1] [2]. Active speaker determination proves useful in a number of tasks such as human-computer/human-robot interactions [3], where in a field of view of multiple speakers, a robot needs to know who is talking in order to turn its head or fix it's gaze in that direction to visually pay attention and better manage conversations with the different speakers [4], audio-visual diarization (auto annotation of descriptions in video scenes) [5], allowing deaf audience to better appreciate movies [6], video conferencing systems to allow zooming in on the current speaker [7], a necessary step in the auto curation of audio samples from videos where the face image of the subjects are known [8], speaker naming where in addition to detecting the active speaker, the identity is also made known [6], speech enhancement, video re-targeting for meetings [1] and is a basic prerequisite for artificial cognitive systems in the acquisition of language in social settings [9]. Research in active speaker detection from videos is faced with challenges such as presence of multiple people leading to variability of possible speakers in a video, poor resolution [10], visibility of speaker in video (speakers who are off screen) [11], faces turned at inconvenient in-plane angles to the recording camera, recordings from YouTube are from varying demographics, have different illumination settings and faces are occluded in some cases. Talk-like mouth movements such as facial expressions, yawning, eating or chewing,

laughing, smiling, nodding, lip licking or squeezing, sighs, groans, grunts, humming and coughing poses challenges to the task [1, 2] because these actions can be confused with talking.

In times past Canonical Correlation Analysis (CCA) was used to determine how well the audio and visual streams of a video correlate to determine the faces talking in videos, for example, [12] employed a 1-norm based CCA in the localization of pixels of active speakers, [13] applied a non-linear CCA to seek the most correlated mouth region with hand-crafted audio features. In [14], MFCC features [15] were extracted from the audio stream while Spatio-temporal visual features were used for the identification of moving parts in the scene. CCA was applied to seek canonical audio and visual sub-spaces maximizing the correlation of these two features and the highly synchronized regions were regarded as the dominant source of sound in the image. Eigen-face features of faces in videos had also been correlated with MFCC features from voice using CCA to detect talking faces in videos [16]. In more recent times, researchers have relied more on deep networks (usually Convolutional Neural Networks (CNNs) [17, 18] and Long Short-Term Memory (LSTMs)) for extraction of embeddings and the use of such in the localization of active speakers in videos, these networks seek to find how well the audio stream information synchronizes with the corresponding visual counterpart and predicts the active speaker as the talking face with the highest correlation between both information [19–22]. These networks have also been extended to speaker naming where in addition to detecting the active speaker the identity is also provided, for instance [23] used an LSTM network for the learning of shared weights between audio and visual modalities. CNNs are also used in learning the optimal fusion functions for audio and facial cues. However, there are no hard-and-fast rules in this research area as researchers have tried other unique approaches and have used varying metrics for the evaluation of their algorithms, for example the mean Average Precision (mAP) [1, 24], Lc performance that provides evaluation from an energy perspective [25], F-Score [26], ratio of correctly predicted samples to total number of test samples [27] and area under Receiver Operating Characteristic (auROC) curve [1, 10]. Some researchers have used only facial cues [3, 4, 9, 28, 29], others have used just audio cues for example [5] and others have used a combination of both cues [1, 2]. Some researchers in addition to facial cues have used head movements, hand gestures and prosody [4, 28], yet others such as [3, 30–33] in order to determine active speakers, rely on the use of an array of multiple microphones and cameras because such setup provides directional and spatial information respectively, the problem with such methods apart from the extra overhead is that in most real-life scenarios such as YouTube videos they are not applicable.

2. Related Works

The researchers in [25] experimented with 10 videos incorporating two features for their unsupervised detection of active speakers, these were the low-rank matrix decomposition of the background of audio/visual information and a kernel of sparse matrix that captured the correlation of foreground components between audio/visual information. The image frames represented the visual features while audio spectrogram magnitudes were used for the audio features. These features were decomposed into sparse and low-rank components. The sparse matrices from the audio and visual features were then mapped to a kernel space. The derived matrix with non-zero elements indicated the position of the pixels

corresponding to the active speaker. In [4], a 21 minutes audio-visual corpus consisting of 4 speakers in fixed positions talking one at a time was created and annotated. The head and facial landmark (lips, eyes, eyebrows and jaw) coordinates were then tracked and the average rate of change, mean and standard deviations of these coordinates were used to predict the active speaker. Models specific to subjects can be built from cross-modal supervision from videos and used in an audio-visual combined training. An active speaker classifier which was video-based employing a concatenation of Histogram of Optical Flow (HoF), Histogram of Oriented Gradients (HoG) and Motion Boundary Histogram (MBH) features trained using directional audio was used in the training of a video-based subject-specific active speaker detection system in a new dataset. These video classifiers learnt online were subsequently used in the supervision of the training of subject-specific voice models achieving near-perfect active speaker detection in [34]. Speaker naming in [6] was realized by an attention-based architecture which accepted a video as input and extracts face images cropped to 160 x 160 pixels fed into a face network that extracted a 512 dimensional face embeddings. It also extracted audio spectrograms of size 257 x 200 x 1 fed to networks also generating a 512-dimensional voice embeddings. The facial and voice embeddings were then concatenated and fed to an attention module that predicted the identity of the active speaker in the video. To identify and label speakers in videos, [27] applied deep fusion of the face and voice modalities, the VGG architecture was used to extract facial features while the Mel-Frequency Cepstrum Coefficients (MFCC) features [15] were extracted from the audio stream. The MFCC feature was fed into a two-layered LSTM. One feature from the layered LSTM network was concatenated with the feature from the VGG network and fed to another CNN which generated the Face Attention feature that was finally fused via a Factorized Bilinear Model (FBM) with the second feature extracted by the layered LSTM network and fed to a fully connected network to identify the active speaker. Researchers in [1] approached the active speaker detection step as a joint mapping of facial and audio signals where a function having three input parameters; the track of face image thumbnails, frequency domain representation of waveform and the weights to be trained is decomposed into two networks built from scratch which were jointly trained. One network was for the audio while the other for the face, these two networks were fused by a third network. The design in [2] comprised three separate modules capable of independently using either of the face, voice or audio-visual modalities for active speaker detection. The first module made use of the audio segment to generate narrow-band spectrograms serving as input to the VGGVox network [35]. The corresponding face image frames served as input to the second module which was an extended version of ResNet50 model [36]. The third module drawing inspiration from the brain's superior colliculus for multi-sensory combination was used to fuse the features from the previous two modules to determine active speakers in videos. Voice Activity detection (VAD) in the audio stream of a video could be used to weakly supervise a video-based active speaker detection by training a classifier with the faces that appeared in the frames during VAD, this training though learned on one video was improved from generic models to person-specific models and adapted to speakers in a new video for active speaker detection [26]. Using two embedding networks; one for audio and the other for visual the researchers in [10] proposed two methods of fusing face images with optical flow features. The first method of fusing these two modalities was by stacking them as input to the MobileNet network while in the second method, each served as input to its respective MobileNet architecture and the outputs from the two MobileNet model

were concatenated. Any of the output of both methods could be concatenated with the output from the audio embedding network which finally served as input into the prediction network that produced a "yes' or "no" answer to the question of whether a speaker was active or not. The researchers in [28] followed a similar approach to those in [4] using facial landmarks, specifically the lips and head movement, they assumed and verified that one-second prior to speaking, the speaker often need to articulate himself in preparation and this prosody can be visually captured by tracking voice activity in the audio stream. Although the activities detected in these portions of speech were less accurate, when fused with the features in the active speech region they improved results.

The experiment carried out in [29] was to design a model capable of giving two distinct outputs- active or non-active. This was done using only visual cues referred to as facial Action Units (AUs) such as jaw, lips, cheek, eye movements and an 8-state Hidden Markov Model (HMM) spatio-temporal modeling. Their model was a multivariate Gaussian distribution trained and tested using the [37] dataset on a 51 dimensional feature vector obtained by concatenation of 17 AU raw features, 17 first and 17 second order AU differences. The design for active speaker detection by [24] comprised of two CNN architecture at the front-end. The first was used to obtain the 512 dimensional audio features from a 20-frame and 13-MFCC input [15] while the second obtained the 512 dimensional visual features from a 3D 5-image frame input. Both features were fed into two different LSTM networks and the outputs of the LSTMs were concatenated and eventually fed to a linear classifier that determines if a speaker was active or not. These researchers were part of the participants at the ActivityNet Challenge 2019 - Task B Active Speaker Detection (AVA) using the AVA ActiveSpeaker dataset. The design in [9] made use of only visual cues to determine active speakers in videos weakly supervised by the audio stream for automatic labelling of the image frames. Once the face image frames were labeled through stochastic optimization, features were extracted using a CNN which were classified experimenting with a non-temporal (Perceptron) and a temporal (LSTM model). The output of each method is a probability distribution over the two possible outcomes—actively speaking or non-active. Probabilities greater than 0.5 threshold depicted speaking activity. Faces could also be detected in the frames of a video and speech activity detection performed on the audio counterpart of video to remove face-frames at portions of speech inactivity. Contiguous face-frames are then grouped together and then split into 2 seconds segments along with their corresponding audio segments [38]. These smaller segments were fed into a network called SyncNet [39] which performs synchronization between the audio and visual input and predicts how well they correlate thus detecting which segments of the video speakers were active or otherwise. Some researchers in order to minimize false alarms in active speaker detection made use of a hybrid method to check the correlation between the visual and audio streams, for instance [40] used a variant of SyncNet [39] and also an audio-visual speech enhancement network [41] that isolated the target speaker's speech from the sound mixture. The active speaker was only accepted if the prediction from both models agreed.

3. Methodology

3.1 Problem Definition

During speaking activity, the lips open and close intermittently revealing inner content of the mouth which are of varying colors, that is, color changes happen at the mouth region in this process. The degree of color changes can be captured using standard deviation of color histograms (CHs) of the mouth region. For an active speaker, this standard deviation is expected to be considerably higher compared with that for a non-active speaker whose lips are at rest with minimal color changes. We aim to solve the problem of determining the threshold value, λ, capable of best discriminating between active and non-active speakers using standard deviations of CHs. Given two inputs, S1 and S2, where S1 and S2 are the standard deviations of the CHs from frame-to-frame of the mouth region for speaker 1 and speaker 2 respectively. The task is to maximize the function in Eq. (1) by experimentally determining λ, where A is the accuracy of prediction, considering all the videos to be tested.

$$\mathrm{A} = f(\mathrm{S1}, \mathrm{S2}, \lambda) \tag{1}$$

3.2 Proposed Algorithm

The standard deviations are computed over the set of features, C1, C2, C3… CN, which represents the CHs of the mouth region from frame-to-frame for a 1-second consecutive frames of a video, where N represents the frame rate of the video. Let a 1-secomd contiguous segment of a video contain N frames of faces, F_1, F_2, F_3, … F_N with corresponding mouth regions, M_1, M_2, M_3, … M_N of corresponding color histograms C_1, C_2, C_3, … C_N for a certain interval of a certain bin as shown in Fig. 1. Then a face is considered to be talking if Eq. (2) is fulfilled,

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_i - \mu)^2} \geq \lambda \tag{2}$$

Where λ is an experimentally determined threshold and μ is the mean color histogram of the mouth region for a certain speaker.

If Eq. (2) is true and assuming the start time of the video is 0, then the corresponding audio sampling points for the interval of the frames in Fig. 1 can be obtained from the sampling points in Eq. (3) using the relationship between the sampling frequency and frame rate of the video.

$$F1*\frac{Fs}{Fr}, F1*\frac{Fs}{Fr}, \ldots, FN*\frac{Fs}{Fr} \tag{3}$$

Where FR is the image frame rate, FS is the audio sampling frequency of the video, F1 is the start frame number and FN is the last frame number. The task in this work, therefore, is by selecting a certain interval of bin for the computation of CHs of the mouth region, calculating their standard deviations within 1-second segments in the video and comparing with an experimentally derived threshold, λ, which best discriminates between active and non-active speakers, determine the active speakers in the videos.

A histogram is a mathematical function m_i, counting the number of observations falling into each categories or bins according to Eq. (4).

$$n = \sum_{i=1}^{k} m_i \quad (4)$$

Where n is the total number of observations and k is the number of bins

A color histogram gives the picture of how the colors in an image are distributed. The RGB color space used in this work consist of 3 sub-color spaces; R (red), G (green) and B (blue), each space having 2^8 or 256 representations in the range 0-255. If we divide the color range into a number of intervals or bins, then count the number of color pixels that falls within each group across all color space, we obtain the CH for those specific intervals or bins. In this regard, Eq. (4) transforms to Eq. (5) so that for each interval of bin, the observation is a sum of the three color components of the mouth region. A CH has no information about the position of the colors in the image rather it focuses on the amount of different colors that falls within certain intervals within the distribution of the image.

$$n = \sum_{i=1}^{k} m_i = \sum_{R=1}^{k} m_R + \sum_{G=1}^{k} m_G + \sum_{B=1}^{k} m_B \quad (5)$$

The reasoning in this research is that in the process of talking, the lips opens and closes revealing/concealing mouth parts such as the vocal cavity, teeth and tongue which are all of different colors at fairly regular intervals. This implies there is intermittent change in the color component of the mouth region from one video frame to another as a person speaks. A non-speaker's lips are usually closed and the color of the lips remains virtually the same from one video frame to another as such the deviations in color properties are expected to be near zero and lower than for an active speaker. Therefore, if the mouth region can be accurately localized in consecutive frames and we can compute the CH of the mouth region for each frame using certain bins then compute the standard deviation of the CH of this set of frames, there exist certain bin intervals of the CH that gives optimal discriminative information between an active and non-active speaker. If a threshold, λ, for the standard deviations of CHs can be experimentally determined, then the proposed method can be used to predict an active speaker for speakers having computations higher than the threshold. Specifically, the standard deviations of CH bins are computed for all the frames that appear within each second for all the seconds in the video to determine the active speaker(s) in these instances. The proposed model is described in Algorithm 1.

Algorithm 1: Determination of active speakers using standard deviation of color histograms

Input: $\mathbf{M_{ij}}$ #Set of mouth regions of speakers

Output: $\mathbf{S_j}$ #Speaker status (**Active**|**non-Active**)

1. numberOfFrames **=** total count of mouth regions
2. numberOfSpeakers = 2
3. **for** i = 1 to numberOfFrames **do**
4. **for** j = 1 to numberOfSpeakers **do**
5. Locate mouth region $\mathbf{M_{ij}}$
6. Compute Color Histogram of mouth region $\mathbf{MCH_{ij}}$
7. **end**
8. **end**
9. **for** j = 1 to numberOfSpeakers **do**
10. $\mathbf{S_j}$ = StdDev($\mathbf{MCH_{ij}}$), $\forall$ i ϵ {1, 2, 3, … numberOfFrames}
11. **if** $\mathbf{S_j} \geq \lambda$
12. $\mathbf{S_j}$ = **Active**
13. **else**
14. $\mathbf{S_j}$ = **Non-Active**
15. **end**
16. **end**

The algorithm accepts the set of mouth regions for speakers in 1-second segments of videos, computes the CHs of the set of mouth regions and finds the standard deviations for the different speakers in the segment. By comparing the computed standard deviations with a threshold, λ, it decides if the respective speakers are active or non-active.

3.2.1 Determining the Cut-Off Threshold for Active Speakers

To determine the threshold, λ which effectively discriminates between active and non-active speakers, 20% of the 250 video clips were randomly selected. The selected videos were processed to localize the mouth regions in the segments of videos were applicable as described in section 3.4 and shown in the flow chart of Fig. 2. The standard deviations of CHs for all mouth regions in the segments were

computed. This set of standard deviations contains values for both active and non-active speakers. Therefore, the task ahead is to cluster these standard deviations into two classes.

However, since the set of values contain outliers due to poor mouth localization as a result of the poor resolution of some of the videos, these outliers need to be removed first before clustering as they can significantly affect the result of clustering [42]. The 25th and 75th percentile of the set of standard deviations were computed. The positive difference between the two values yielded the interquartile range (IQR) value. The IQR was used to compute the upper outer, upper inner, lower inner, and lower outer fences as described in [43]. These fences were used to eliminate the outliers in the set.

K-means clustering was finally applied to separate the standard deviations into two classes and the centroids of the two classes were obtained. The lower centroid represented the average standard deviations for the non-active speakers while the higher centroid represented the average standard deviation for the active speakers. The cut-off between the two centroids was computed as the mid-point between the two values, it was computed to be 103.78. The flow chart in Fig. 2 is a summary of the processes described in the computation of the cut-off standard deviation, λ, which discriminates between active and non-active speakers.

Interval 1 of bin 2 were chosen for the computation of CHs. Any of the interval of bin 2 is expected to be the most sensitive to color changes at the mouth region. This is because the mouth region is represented by a fixed number of pixels (32 x 64), therefore, as the number of bin increases, so does the number of intervals. This makes the available number of pixels in each interval drops making them less sensitive to color changes.

3.3 Dataset

A dataset is necessary to evaluate the performance of the proposed ASD algorithm. The dataset was created using YouTube videos available online from Channels TV news, a popular Nigeria news channel. It contains the host of the station interviewing different guests like politicians, activists, professionals in different spheres of life, performing artistes and so on. 53 of such videos were used to create 250 video clips each ranging between 15 seconds to 1 minute. These videos were clipped in such a way that two speakers' faces appeared simultaneously in any part between the start and end of each video. Although, the speakers were at most two, the camera randomly zooms in on any of the speakers in any order in random parts of each video. There are parts of the video clips where the speakers spoke one at a time, interrupted each other in the course of the discussions or were both silent. The total length of the videos was about three and a half hours.

3.3.1 Benchmark Dataset

Since the videos in this work were obtained from the wild for which evaluation datasets do not exist [11], a benchmark dataset consisting of manual labels of the seconds in the videos for which the speakers were active was created for the purpose of evaluating the performance of the proposed method for ASD.

Each speaker in the video clips have their individual labels. The rows in the table of labels as shown in Fig. 3, represents the videos while the columns represents the seconds in each video where the respective speakers were active. The columns where the speakers are active were labeled with the position in seconds in the video while the silent portions were omitted. Each second of all the video clips where applicable was manually labeled to reflect the speaker(s) who spoke in those instances. In cases of doubts, the MATLAB2017b sound function was used to replay the segment in question to ascertain the ground truth speaker(s) in such segments. The online sites for the 53 videos used and the portions of the videos that were clipped together with the labels for all the seconds of all 250 videos will be made publicly available.

3.4 Face and Mouth Detection

Face detection was achieved using the Viola-Jones [44] algorithm available in MATLAB 2017b image processing toolbox. The MergeThreshold parameter of this function determines how strict the algorithm detect faces. Higher values detect faces more accurately and over looks poor quality images but takes longer time while lower values can detect faces even in poor quality videos/images, are faster but more prone to errors. The only extra contribution in this section was to interchangeably use both high (50) and low (1) values of this parameter in such a way as to get the best of the algorithm. The higher value was only used to detect the first appearance of face after which the coordinates of detection are noted and subsequent faces in frames were searched only within the neighborhood of the first occurrence using the lower parameter until no further faces were detected in the region. This setup speeds up face detection because only the smaller parameter was used most of the time and the entire frame need not be searched during this period. Also the Viola-Jones algorithm is not robust to faces tilted beyond certain angles, so whenever a face could no longer be detected in the region of search, that region was gradually rotated in steps while applying the algorithm in hopes of detecting faces that could otherwise have been missed out. Mouth region detection was also done using Viola-Jones algorithm implemented using MATLAB 2017b image processing toolbox function. Most of the times this algorithm confuses the eyes for the mouth, therefore in this work, the mouth region was only searched within the lower 45% part of the face. This overcame the first problem, yet sometimes the algorithm returns more than one region for the mouth like the chin or parts of the lips. For scenarios like this, the algorithm was placed in a loop where the MergeThreshold parameter was gradually incremented until one output was obtained by the algorithm which turns out to be the most accurate mouth region.

4. Experiment

The proposed method depicted in Fig. 4 was applied to all 250 video clips in the dataset. The detected faces and corresponding mouth regions were cropped and resized to 80x80 and 32x64 pixels respectively for uniformity. Contiguous frames of faces were then separated into groups.

Each group was sub-divided into 1-second segments and in each segment the mouth region of the speakers that appeared analyzed by computing their CHs using interval 1 of bin 2 and the computed

threshold, λ = 103.78 that was experimentally derived to effectively discriminate between active and non-active speakers as illustrated in Fig. 5 which shows a typical standard deviation computed for an active speaker to be 147.43 while that computed for a typical non-active speaker was 12.55. It is observed that the experimentally derived cut-off, λ = 103.78, effectively discriminates between these two values. The standard deviations of the CHs for the different speakers were computed and speaker(s) having values above threshold (λ) were deemed to be the active speaker(s).

The speakers predicted by the algorithm were compared with the labels in the dataset for all the seconds in the 250 videos to see how well the proposed method worked. The main evaluation parameter used in this work was the ASD Accuracy defined as the ratio of the total number of correct predictions to the total number of predictions in percentage considering all the 250 video clips analyzed as seen in Eq. (6). Asides this, other parametric terms were also coined, one such parameter was the ASD Effectiveness, Eq. (7), defined as the ratio of the total number of correct predictions to the total number of active speaker samples present in the dataset in percentage for the 250 video clips, this gives a picture of how much of the active speaker's voice samples the algorithm was comfortable to predict. The other parameter was the ASD Confidence, which is an indication of the extent to which the prediction made can be trusted, defined as the total number of videos where all the predictions were 100% correct to the total number of videos tested in percentage, Eq. (8).

$$ASDAccuracy = \frac{Total\ number\ of\ correct\ predictions}{Total\ number\ of\ predictions} * 100\% \quad (6)$$

$$ASDEffectiveness = \frac{Total\ number\ of\ correct\ predictions}{Total\ number\ of\ active\ speaker\ instances} * 100\% \quad (7)$$

$$ASDConfidence = \frac{Total\ number\ of\ videos\ where\ predictions\ were\ 100\%\ correct}{Total\ number\ of\ videos\ tested} * 100\% \quad (8)$$

4.1 Experimental Results

The results of the experiment carried out for ASD are captured in Table 1 and Fig. 6. The table show the overall results of the 250 videos processed for the various performance evaluation metrics, 99.19% of the predicted speakers were accurate. Figure 6 shows the result of Table 1 in bar chart form, it gives more granularity by showing the percentages of videos tested as they contributed to the overall accuracy.

Table 1
Active speaker detection performance

Performance Metrics	λ = 103.78
Effectiveness (%)	45.67
Confidence (%)	92.97
Accuracy (%)	99.19

5. Discussion

Effectiveness in this work refers to the percentage of the total talk-time of speakers the algorithm was able to make accurate predictions while the accuracy refers to the percentage of the predictions made that were correct. A 3rd parameter that indicates the extent to which the predictions can be trusted is the confidence parameter, the higher this value, the more reliable the prediction is. Table 1 shows that predictions were only made for about 45.67% of the cases the speakers were active, however, 99.19% of the predictions made were accurate. Finally, in 92.97% (232 of 250) of the videos tested, the predictions were 100% accurate. Figure 6 shows the distribution of accuracies of ASD considering the number of videos tested to show granularity as against the overall result. The chart shows that 96.4567% of the videos tested had prediction accuracies in the range 91–100%, none of the videos tested was below 50% accuracy. 1.9685% of the videos tested fell within the accuracy range of 81–90% and less than 2% of the videos tested fell within accuracy range 51–80%. In the few missed cases, the algorithm was confused with mouth movement activities such as lip licking, squeezing, mouth opening in anticipation to talk without speaking and chewing, these can be addressed by combining the visual cues in this work with audio cues in future work.

The result of our proposed algorithm for ASD on our dataset was compared with a similar work [28], which made use of variations in lips and head coordinates for ASD. In both cases, the evaluation metric was the same, that is, the percentage of the number of video segments where the active speakers were correctly detected. The comparison is shown in Table 2. They tested using 63 video segments and their best result correctly predicted 35 of the segments, this yielded an effective accuracy of 55.56%, their method did not make a prediction in 5 of the segments. Excluding these 5 cases, their result achieved an accuracy of 60.34%. In terms of the effectiveness of prediction, our proposed algorithm achieved 45.67% accuracy but in terms of accuracy of predictions, we achieved 99.19%. These results show that our proposed algorithm using standard deviations of CHs of the mouth region is more effective than tracking variations lips and head coordinates for ASD.

Table 2
Active speaker detection performance comparison

Authors	ASD Effectiveness Accuracy (%)	ASD Prediction Accuracy (%)
[45]	55.56	60.34
The proposed method	45.67	99.19

It is worthy of note that our proposed algorithm would rather not make a prediction than make a wrong decision. As shown in Table 2, this was why our algorithm makes prediction in only about 45.67% of the total cases but the predictions in those cases were 99.19% accurate, unlike in [28] where predictions were made in about 92% of the total cases but in those cases only 60.34% of the predictions were accurate.

It is also worthwhile to note that while the video in [28] was created using a high quality camera for the recording of effectively 3 individuals whose positions were fixed and only 63 video segments were tested,

our work made use of online videos in the wild. A considerable number of these video segments were of poor resolution quality and the position of the two individuals in the video segments can vary. The videos tested in our work cut across 45 unique individuals and we tested over 7,000 segments of videos.

6. Conclusion

A novel concept for ASD in digital videos was proposed. To the best of our knowledge, this is the first time only standard deviations of CHs of the mouth region from frame-to-frame is being used for prediction of active speakers. Experiments carried out using interval 1 of bin 2 of CHs on 250 video clips from 53 videos on YouTube obtained a prediction accuracy up to 99.19% of the times predictions were made using only visual cues (CHs of the mouth region). The result was benchmarked with a similar work and performed better with greater confidence. This shows CHs are excellent features for active speaker prediction because during speaking, the various content of the mouth which are of different colors are intermittently revealed and concealed resulting in changes in color activities at the mouth region. The dataset used for the experiment along with the speaker labels will be made publicly available. Future work will explore fusion of visual and audio cues to further improve the obtained results.

Abbreviations

ASD
Active Speaker Detection
CH
Color Histogram
CCA
Canonical Correlation Analysis
CNN
Convolutional Neural Network
LSTM
Long Short-Term Memory
mAP
mean Average Precision
auROC
area under Receiver Operating Characteristic
HoF
Histogram of Optical Flow
HoG
Histogram of Optical Gradients
MBH
Motion Boundary Histogram
MFCC

Mel-Frequency Cepstrum Coefficients
FBM
Factorized Bilinear Model
VAD
Voice Activity Detection
AU
Action unit
HMM
Hidden Markov Model
AVA
ActivityNet Challenge 2019 - Task B Active Speaker Detection
IQR
InterQuartile Range

Declarations

Availability of data and materials

Dataset: https://drive.google.com/drive/u/1/folders/1dIasEM4LVmwslWiqh4466jMND47HbZTR

Please contact any of the authors for access.

Competing interests

The authors declare that they have no competing interests.

Funding

This research is fully sponsored by Covenant University Centre for Research, Innovation and Development (CUCRID), Covenant University, Ota, Nigeria.

Authors' contributions

Akinrinmade provided conceptualization, investigation, formal analysis, methodology and write-up; Prof. Adetiba and Dr. Badejo provided supervision, visualization, corrections, and validation. The authors have read and approved the final manuscript.

Acknowledgements

This research is fully sponsored by Covenant University Centre for Research, Innovation and Development (CUCRID), Covenant University, Ota, Nigeria.

References

1. J. Roth, S. Chaudhuri, O. Klejch, R. Marvin, A. Gallagher, L. Kaver, *et al.*, "Supplementary Material: AVA-ActiveSpeaker: An Audio-Visual Dataset for Active Speaker Detection," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 3718–3722
2. G. Assunção, N. Gonçalves, P. Menezes, "Bio-Inspired Modality Fusion for Active Speaker Detection," *arXiv preprint arXiv:2003.00063*, 2020
3. K. Stefanov, J. Beskow, G. Salvi, "Vision-based active speaker detection in multiparty interaction," in *Grounding Language Understanding GLU2017 August 25, 2017, KTH Royal Institute of Technology, Stockholm, Sweden*, 2017
4. F. Haider, N. Campbell, S. Luz, "Active speaker detection in human machine multiparty dialogue using visual prosody information," in 2016 *IEEE global conference on signal and information processing (GlobalSIP)*, 2016, pp. 1207–1211
5. X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, O. Vinyals, Speaker diarization: A review of recent research. IEEE Trans. Audio Speech Lang. Process. **20**, 356–370 (2012)
6. J. Pyo, J. Lee, Y. Park, T.-C. Bui, S.K. Cha, "An Attention-Based Speaker Naming Method for Online Adaptation in Non-Fixed Scenarios," *arXiv preprint arXiv:1912.00649*, 2019
7. P. Besson, V. Popovici, J.-M. Vesin, J.-P. Thiran, M. Kunt, Extraction of audio features specific to speech production for multimodal speaker detection. IEEE Trans. Multimedia **10**, 63–73 (2007)
8. J.S. Chung, A. Nagrani, A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018
9. K. Stefanov, J. Beskow, G. Salvi, "Self-Supervised Vision-Based Detection of the Active Speaker as Support for Socially Aware Language Acquisition". IEEE Trans. Cogn. Dev. Syst. **12**, 250–259 (2019)
10. C. Huang, K. Koishida, "Improved Active Speaker Detection Based on Optical Flow," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 950–951
11. K. Hoover, S. Chaudhuri, C. Pantofaru, I. Sturdy, M. Slaney, "Using audio-visual information to understand speaker activity: Tracking active speakers on and off screen," in 2018 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6558–6562
12. E. Kidron, Y.Y. Schechner, M. Elad, "Pixels that sound," in 2005 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 2005, pp. 88–95
13. K. Li, J. Ye, K.A. Hua, "What's making that sound?," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 147–156
14. H. Izadinia, I. Saleemi, M. Shah, "Multimodal analysis for identification and segmentation of moving-sounding objects. IEEE Trans. Multimedia **15**, 378–390 (2012)
15. A.A. Akinrinmade, E. Adetiba, J.A. Badejo, A.A. Atayero, "Creation of a Nigerian Voice Corpus for Indigenous Speaker Recognition," in *Journal of Physics: Conference Series*, 2019, p. 032011
16. D. Li, C. Taskiran, N. Dimitrova, W. Wang, M. Li, I. Sethi, "Cross-modal analysis of audio-visual programs for speaker detection," in 2005 *IEEE 7th Workshop on Multimedia Signal Processing*, 2005,

pp. 1–4

17. C. Lawal, A. Akinrinmade, J. Badejo, "Face-based Gender recognition Analysis for Nigerians Using CNN," in *Journal of Physics: Conference Series*, 2019, p. 032014
18. J.A. Badejo, E. Adetiba, A. Akinrinmade, M.B. Akanle, "Medical image classification with hand-designed or machine-designed texture descriptors: a performance evaluation," in *International Conference on Bioinformatics and Biomedical Engineering*, 2018, pp. 266–275
19. R. Arandjelovic, A. Zisserman, "Look, listen and learn," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 609–617
20. R. Arandjelovic, A. Zisserman, "Objects that sound," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 435–451
21. A. Owens, A.A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 631–648
22. A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, I. So Kweon, "Learning to localize sound source in visual scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4358–4366
23. J. Ren, Y. Hu, Y.-W. Tai, C. Wang, L. Xu, W. Sun, *et al.*, "Look, listen and learn-a multimodal lstm for speaker identification," *arXiv preprint arXiv:1602.04364*, 2016
24. J.S. Chung, "Naver at ActivityNet Challenge 2019–Task B Active Speaker Detection (AVA)," *arXiv preprint arXiv:1906.10555*, 2019
25. J. Pu, Y. Panagakis, M. Pantic, "Active Speaker Detection and Localization in Videos using Low-Rank and Kernelized Sparsity" IEEE Signal Processing Letters, 2020
26. P. Chakravarty, T. Tuytelaars, "Cross-modal supervision for learning active speaker detection in video," in *European Conference on Computer Vision*, 2016, pp. 285–301
27. X. Liu, J. Geng, H. Ling, Y. Cheung, "Attention guided deep audio-face fusion for efficient speaker naming. Pattern Recogn. **88**, 557–568 (2019)
28. F. Haider, S. Luz, C. Vogel, N. Campbell, "Improving Response Time of Active Speaker Detection Using Visual Prosody Information Prior to Articulation," in *INTERSPEECH*, 2018, pp. 1736–1740
29. K. Stefanov, A. Sugimoto, J. Beskow, "Look who's talking: visual identification of the active speaker in multi-party human-robot interaction," in *Proceedings of the 2nd Workshop on Advancements in Social Signal Processing for Multimodal Interaction*, 2016, pp. 22–27
30. J. Cech, R. Mittal, A. Deleforge, J. Sanchez-Riera, X. Alameda-Pineda, R. Horaud, "Active-speaker detection and localization with microphones and cameras embedded into a robotic head," in *2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2013, pp. 203–210
31. I.D. Gebru, X. Alameda-Pineda, R. Horaud, F. Forbes, "Audio-visual speaker localization via weighted clustering," in *2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2014, pp. 1–6

32. P. Chakravarty, S. Mirzaei, T. Tuytelaars, H. Van hamme, "Who's Speaking? Audio-Supervised Classification of Active Speakers in Video," in *Proceedings of the* 2015 *ACM on International Conference on Multimodal Interaction*, 2015, pp. 87–90
33. I.D. Gebru, S. Ba, G. Evangelidis, R. Horaud, "Tracking the active speaker based on a joint audio-visual observation model," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015, pp. 15–21
34. P. Chakravarty, J. Zegers, T. Tuytelaars, H. Van hamme, "Active speaker detection with audio-visual co-training," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, 2016, pp. 312–316
35. A. Nagrani, J.S. Chung, A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017
36. K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778
37. K. Stefanov, J. Beskow, "A multi-party multi-modal dataset for focus of visual attention in human-human and human-robot interaction," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016, pp. 4440–4444
38. R. Ahmad, S. Zubair, H. Alquhayz, "Speech Enhancement for Multimodal Speaker Diarization System". IEEE Access. **8**, 126671–126680 (2020)
39. J.S. Chung, A. Zisserman, "Out of time: automated lip sync in the wild," in *Asian conference on computer vision*, 2016, pp. 251–263
40. J.S. Chung, J. Huh, A. Nagrani, T. Afouras, A. Zisserman, "Spot the conversation: speaker diarisation in the wild," *arXiv preprint arXiv:*2007.*01216*, 2020
41. T. Afouras, J.S. Chung, A. Zisserman, "The conversation: Deep audio-visual speech enhancement," *arXiv preprint arXiv:1804.04121*, 2018
42. S. Gupta, R. Kumar, K. Lu, B. Moseley, S. Vassilvitskii, "Local search methods for k-means with outliers," *Proceedings of the VLDB Endowment*, vol. 10, pp. 757–768, 2017
43. H. Vinutha, B. Poornima, B. Sagar, "Detection of outliers using interquartile range technique from intrusion dataset" Information and Decision Sciences, ed: Springer, 2018, pp. 511–518
44. P. Viola, M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the* 2001 *IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, 2001, pp. I-I
45. O.M. Parkhi, A. Vedaldi, A. Zisserman, "Deep face recognition" 2015

Figures

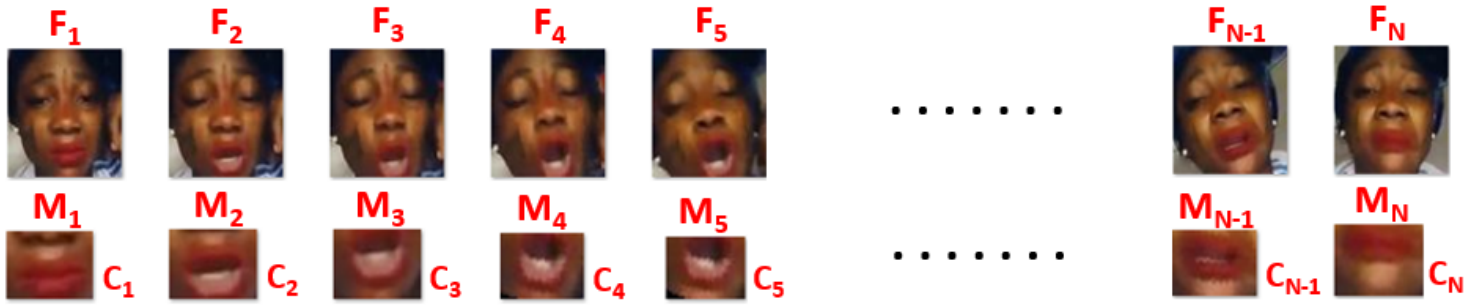

Figure 1

Identifying speaking subjects from contiguous frames of mouth region

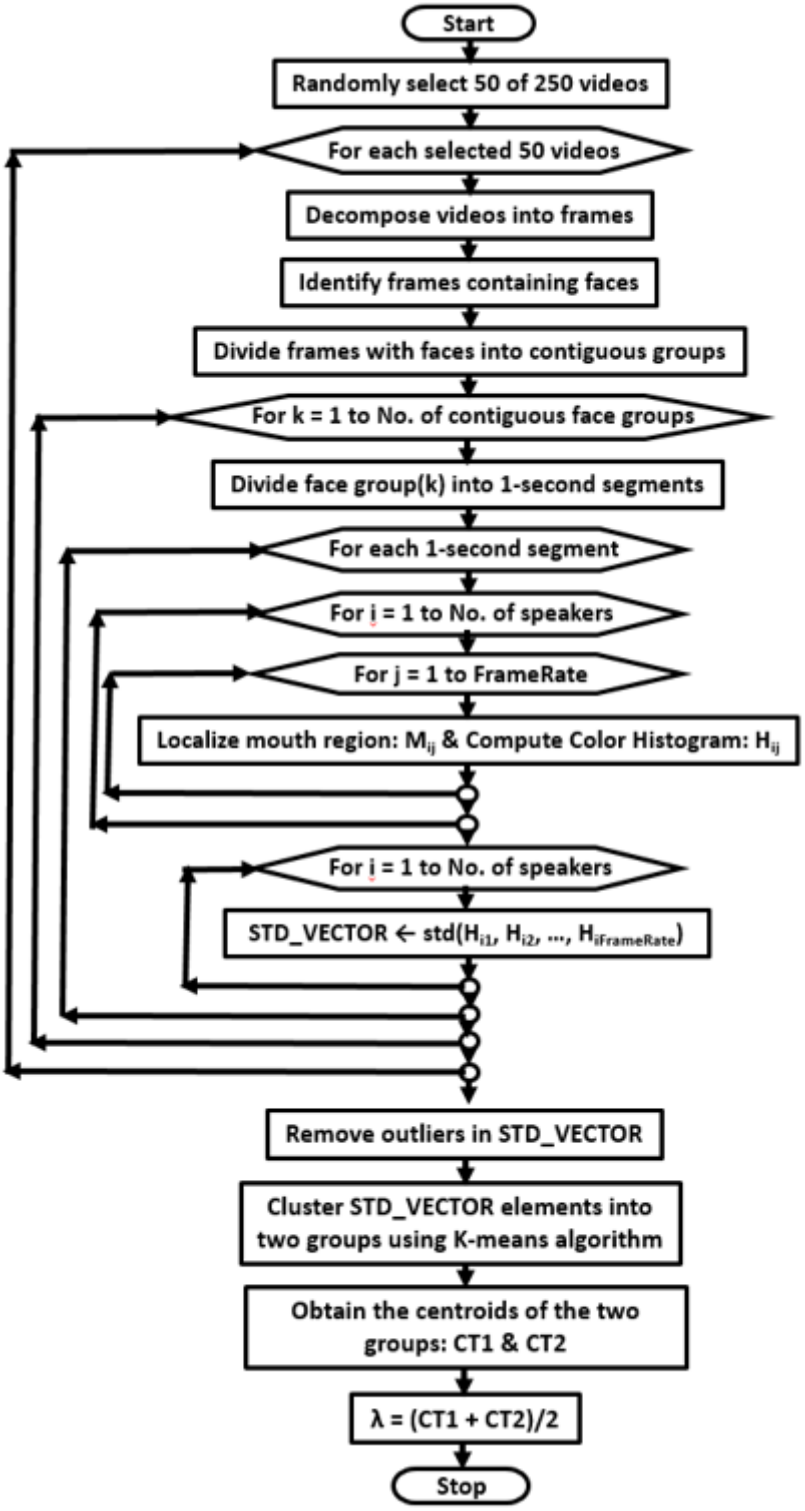

Figure 2

Determination of Active Speaker Cut-Off Threshold

```
1: 1 2 3 4 5 6 7 8
2:
3: 1 2 3 4 5 6 7
4: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
5: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29
6: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
7: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
8:   2 3 4 5 6 7 8 9 10 11 12
9:       4 5 6 7 8 9 10 11 12 13 14
10:1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
11:    3 4 5 6 7 8 9 10 11 12 13 14
12:1 2                                        40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
13:1 2                           26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
14:1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
15:1 2 3 4                    23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42
```

(a) Sample snapshot of audio labels for speaker 1

```
1:         9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
2: 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
3:        8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
4:                           28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53
5:                              32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56
6:                       22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49
7:
8:1             13 14 15 16 17 18 19 20
9:  3              16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37
10:                  18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44
11:                16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
12:   4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38
13:  3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
14:   4 5 6                          29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53
15:    5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
```

(b) Sample snapshot of audio labels for speaker 2

Figure 3

Sample snapshot of voice labels for the speakers in a video

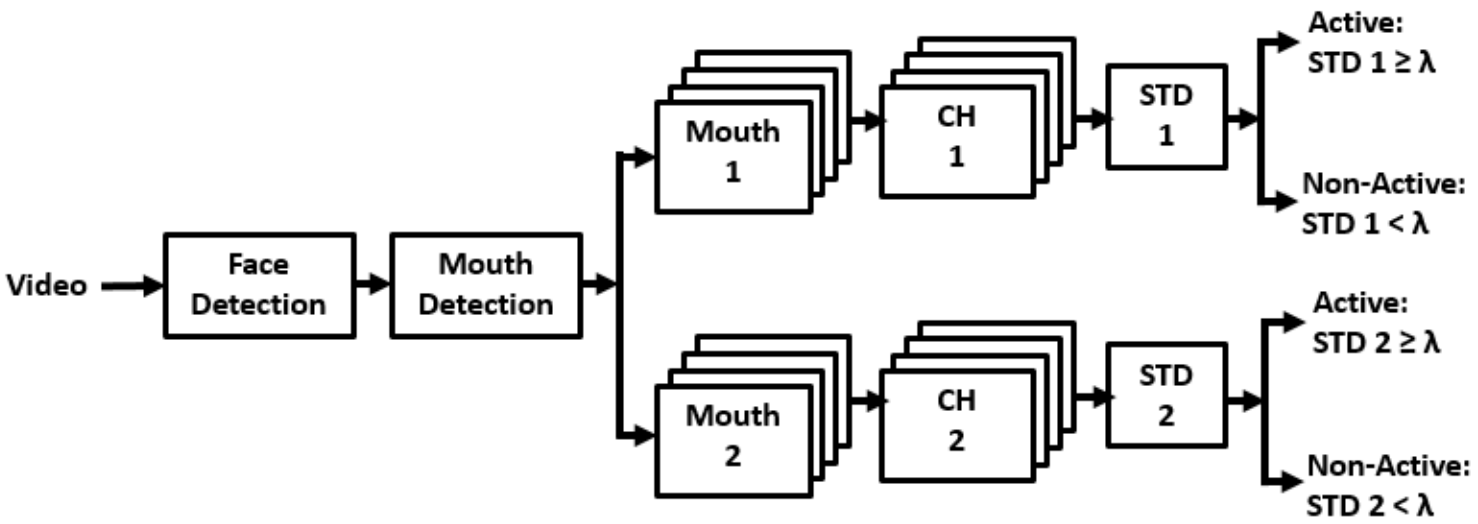

Figure 4

Overview of the proposed ASD pipeline

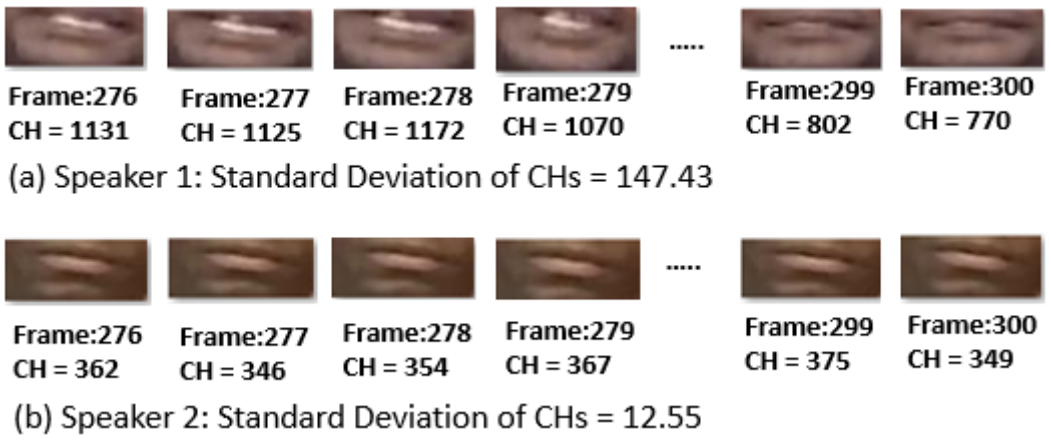

Figure 5

Samples of contiguous frames of mouth region for the 12th second of the 1st video with CH values for interval 1 of bin 2. The standard deviation of (a), the active speaker is significantly greater than the standard deviation of (b), the non-active speaker

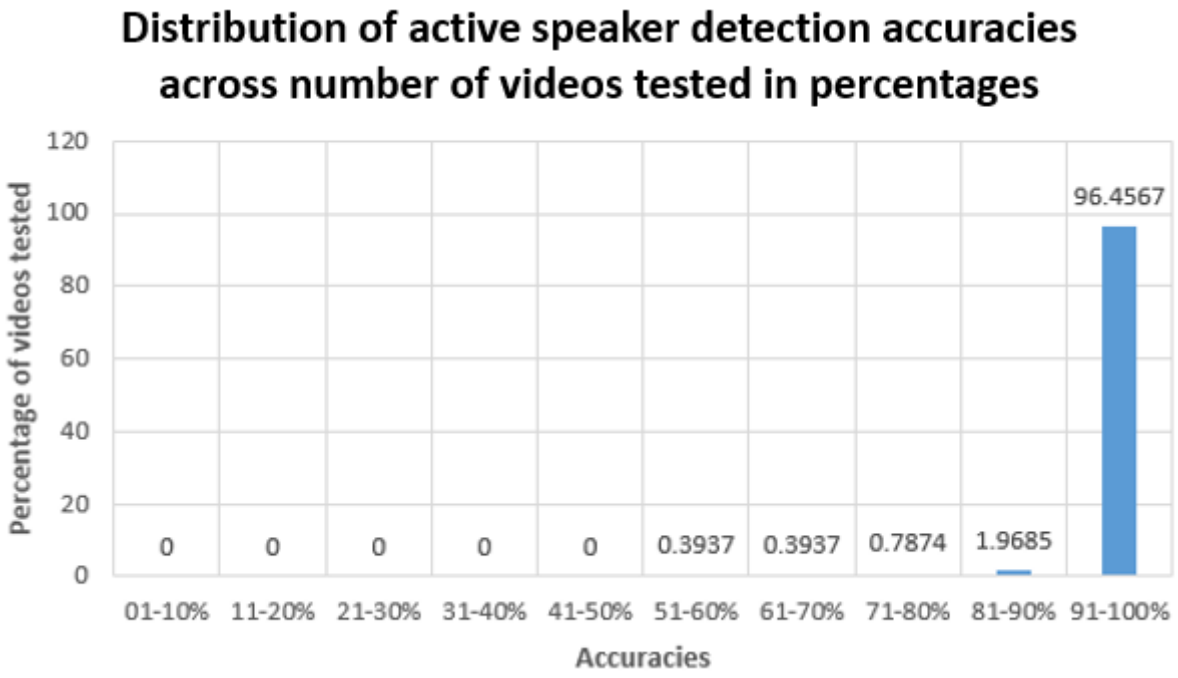

Figure 6

Distribution of Active speaker detection accuracies across number of videos tested in percentages