# Towards Speaker Detection using FaceAPI Facial Movements in Human-Machine Multiparty Dialogue

FASIH HAIDER


KTH Computer Science
and Communication

Master of Science Thesis
Stockholm, Sweden 2013

# Towards Speaker Detection using FaceAPI Facial Movements in Human-Machine Multiparty Dialogue

F A S I H  H A I D E R

DT218X, Master's Thesis in Speech Communication (30 ECTS credits)
Master Progr. in Project Management and Operational Development 60 cr
Supervisor at CSC was Samer Al Moubayed
Examiner was Björn Granström

TRITA-CSC-E 2013:001
ISRN-KTH/CSC/E--13/001--SE
ISSN-1653-5715

Royal Institute of Technology
*School of Computer Science and Communication*

**KTH** CSC
SE-100 44 Stockholm, Sweden

URL: www.kth.se/csc

**Abstract**

In multiparty multimodal dialogue setup, where the robot is set to interact with multiple people, a main requirement for the robot is to recognize the user speaking to it. This would allow the robot to pay attention (visually) to the person the robot is listening to (for example looking by the gaze and head pose to the speaker), and to organize the dialogue structure with multiple people. Knowing the speaker from a set of persons in the field-of-view of the robot is a research problem that is usually addressed by analyzing the facial dynamics of persons (the person that is moving his lips and looking towards the robot is probably the person speaking to the robot).This thesis investigates the use of lip and head movements for the purpose of speaker and speech/silence detection in the context of human-machine multiparty dialogue. The use of speaker and voice activity detection systems in human-machine multiparty dialogue is to help the machine in detecting **who** and **when** someone is speaking out of a set of persons in the field-of-view of the camera. To begin with, a video of four speakers (S1, S2, S3 and S4) speaking in a task free dialogue with a fifth speaker (S5) through video conferencing is audio-visually recorded. After that each speaker present in the video is annotated with segments of speech, silence, smile and laughter. Then the real-time FaceAPI face tracking commercial software is applied to each of the four speakers in the video to track the facial markers such as head and lip movements. At the end, three classification techniques namely Mahalanobis distance, naïve Bayes classifier and neural network classifier are applied to facial data (lip and head movements) to detect speech/silence and speaker.

In this thesis, three types of training methods are used to estimate the training models of speech/silence for every speaker. The first one is speaker dependent method, in which the training model contains the facial data of testing person. The second one is speaker independent method, where the training model does not contain the facial data of testing person. It means that if the test person is S1 then the training model may contain the facial data of S2, S3 or S4. The third one is hybrid method, where the training model is estimated using the facial data of all the speakers and testing is performed on one of the speaker.

The results of speaker dependent and hybrid methods show that the neural network classifier provides the best results. In the speaker dependent method, the accuracies of neural network classifier for speaker and speech/silence detection are 97.43% and 98.73% respectively. However, in the hybrid method, the accuracy of neural network classifier for speech/silence detection is 96.22%. The results of speaker independent method shows that the naïve Bayes classifier provides the best results with an optimal accuracy of 67.57% for speech/silence detection.

**Sammanfattning**
**Gentemot Talaren Detektering med FaceAPI Facial rörelser i Människa-Maskin Multiparty Dialog**

I fleraparter med fleramodala dialoginställningar, där roboten är inställd på att interagera med flera personer. Det är en viktig förutsättning för roboten att känna igen att användaren talar till den. Detta skulle göra det möjligt för roboten att uppmärksamma (visuellt) den person roboten lyssnar till (till exempel genom att titta i blicken och på huvudet för att känna igen talaren) och att organisera dialogens struktur med flera personer. Talaren från en upp sättning av personer i roboten synfält är ett forskningsproblem som vanligtvis riktar sig till att analysera dynamiken i ansiktsuttryck för personer (den person som rör på sina läppar och riktar blicken mot roboten är förmodligen den person som talar till roboten). Denna avhandling undersöker användningen av läpp och huvudrörelser i syfte av att upptäcka högtalare och tal/tystnad i samband med människa-maskin flerpartisystem dialog. Användningen av högtalare och röstaktivitetsdetekteringssystem i människa-maskin flerpartisystem dialog är att hjälpa maskinen att upptäcka vem och när någon talar i kamerans synfält. Till att börja med, en video av fyra högtalare (S1, S2, S3 och S4) talar i en uppgift utan dialog med en femte högtalare (S5) genom videokonferenser blir ljud-visuellt inspelat. Sedan tillämpas realtid FaceAPI tracking kommersiell programvara på vardera fyra högtalarna i videon, för att spåra ansiktets markörer som huvud-och läpprörelser. I slutet finns tre klassificeringstekniker nämligen Mahalanobis distans, naiva Bayes klassificeraren och neuralanätverk klassificerare, som tillämpas på ansiktet (läpp och huvudrörelser) för att upptäcka tal/tystnad och talare.

I denna avhandling har tre typer av träningsmetoder använts för att uppskatta utbildningsmodellerna för tal/tystnad för varje talare. Den första är en talarberoende metod, där utbildningsmodellen innehåller uppgifter om ansiktsdrag från testpersonen. Den andra är en talaroberoende metod, där träningsmodellen inte innehåller ansiktsdrag från testpersonen. Det innebär att om testpersonen är S1 kan utbildningsmodellen innehålla data om ansiktsdrag från S2, S3 eller S4. Den tredje är en hybrid metod, där utbildningsmodellen beräknas utifrån data från alla talares ansiktsdrag men tester utförs på en av talarna.

Resultaten av talarberoende och hybridmetoderna visar att den neurala nätverksklassificeraren ger bästa resultat. Utifrån data från alla talares ansiktsdrag är, noggrannheten på neurala nätverk klassificerare för talare och tal/tystnad upptäckt är 97,43% och 98,73% respektive. I hybridmetoden, är däremot noggrannheten hos neurala nätverksklassificeraren för tal/tystnad detektering 96,22%. Resultaten av talaroberoende metod visar att den naïve Bayes klassificerare ger de bästa resultaten med en optimal noggrannhet på 67,57% för tal/tystnad detektering.

# Acknowledgment

I would like to thank my supervisor Samer Al Moubayed in the Department of Speech, Music and Hearing (TMH) at The Royal Institute of Technology (KTH) Stockholm Sweden for giving me the opportunity to write the thesis under his supervision. I would also like to thank all of the staff members at TMH-KTH, especially those who participated in the recording of the video.

Finally, I would like to thank my family and friends for putting up with me during my years of studies.

# Contents

# Chapter 1

# Introduction

Building a conversational system that is able to carry out dialogue with multiple humans (a multiparty dialogue system) brings about several complexities compared to a dyadic human-machine dialogue. An essential requirement for such a system to operate is the need to know when and which person is speaking to the system. This ability allows the system to coordinate the dialogue to accommodate the different interlocutors. This becomes even more demanding if the dialogue system is an embodied system – communicating with the subjects using a human-like face and/or body, such as the KTH Furhat robot head [6], shown in Fig: 1.1, that is specifically designed for multimodal multiparty dialogue. In multimodal multiparty interactions, the system then needs to generate multimodal output depending on who the speaker is and in turn whom the system is talking to: this includes, for example, the system targeting its attention to the speaker in real-time (orienting the eyes and head towards the speaker). There are various examples of the multiparty dialogue, where robots need to pay attention (visually) to the participants of a dialogue (for example in smart kiosks [11, 12, 5], or working as a receptionist).

Figure 1.1: Furhat in multiparty dialogue

In this thesis, a solution for identification of a speaker using the lip and head movements in a human-machine multiparty dialogue is presented that helps controlling the dialogue management and multimodal output of the

robot in multiparty settings. The flow chart of proposed approach is shown in Fig: 1.2.

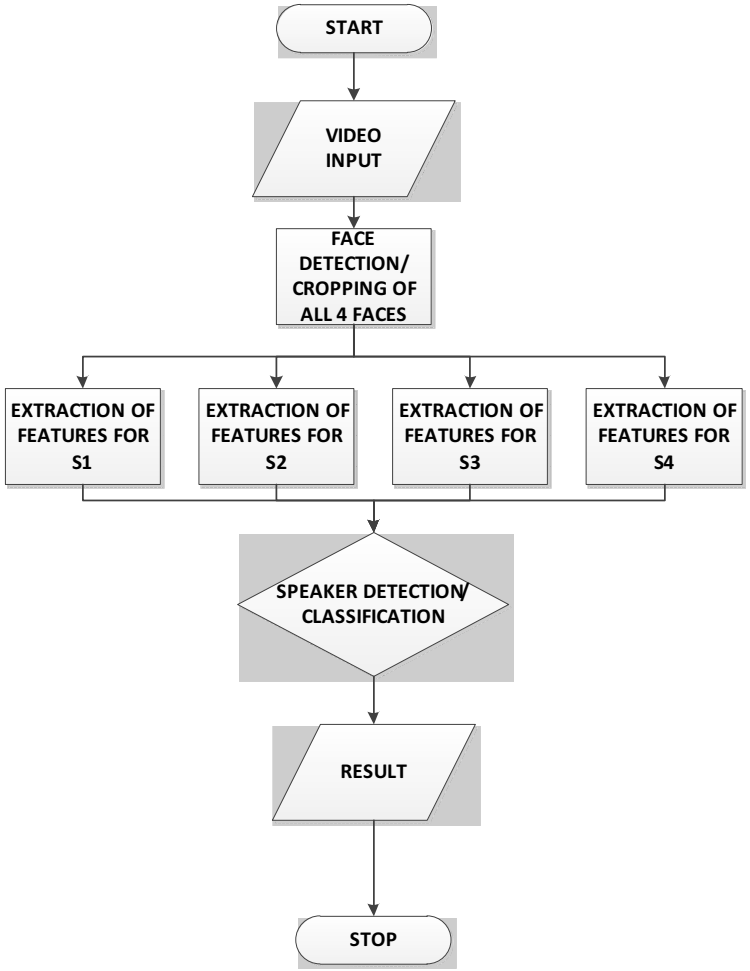



Figure 1.2: Speaker detection/ Voice activity detection flow chart

## 1.1 Thesis Overview

This thesis is organised as follow. In chapter 2, related work is discussed. Chapter 3 explains the recording environment, recording settings and restrictions imposed on the participants of the video. Later, it sheds light on the annotation of the video as segments of speech, silence, smile and laughter. It also exhibits the total number of frames present in each class e.g. speech and silence. The discussion continues while describing the cropping of faces, feature extraction through faceAPI, selected features. Chapter 4 discribed the classification methods (e.g. naïve Bayes classifier).

Chapter 5 addresses the results and discussions of the speaker detection and voice activity detection systems. It constitutes explanation of the speaker dependent, speaker independent and hybrid methods. The conclusion and future work are described in chapter 6 and 7 respectively.

# Chapter 2

# Related Work

Human speech is a bimodal signal. In noisy environment, speech intelligibility improves in the presence of the speaker's face. The visual information such as lip movements of a speaker helps in speech comprehension [17]. In this chapter, voice activity detection and speaker detection systems based on audio-visual or solely visual modality of speech are discussed.

## 2.1 Voice Activity Detection

The Voice Activity Detection (VAD) or speech activity detection systems are used to discriminate between speech and non-speech segments in an audio/video signal. A typical VAD system mainly consists of three steps as shown in Fig: 2.1. In the feature extraction step, suitable discriminative speech features are extracted (such as height and width of lips [14] or optical flow vector of mouth region [2]). In the decision making step, the decision (speech or non-speech) is made based on some thresholds. Several decision making methods have been proposed in the literature such as Euclidean distance [10], support vector machine [9] and genetic algorithm [13]. VAD algorithms, Those work on frame by frame basis, generally need a decision smoothing algorithm such as hang-over algorithm to improve the robustness against noise.

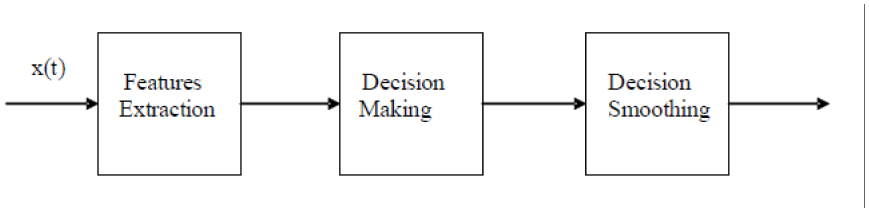



Figure 2.1: Voice activity detection

The major difficulty in the detection of voice is encountered in situations where signal-to-noise ratio is very low. It may be impossible to discriminate between speech and noise by using audio features when parts of the speech utterances are masked by noise. The use of facial features (such as the features extracted from the mouth region) to support a VAD system can improve

its performance in noisy environments since it does not only rely on acoustic features. Several studies have already targeted the integration of facial features into audio VADs. In [8], a system based on audio and visual features is proposed. The audio and visual features have been fused by using two methods. The first one (feature fusion method) is to combine the audio and visual features for a single classifier. The second one (decision fusion method) is to classify based on audio and visual features separately and then fuse the results of both audio and visual classifiers. The results show that the performance of the feature fusion method is typically better than the decision fusion method.

In [16], an automatic lip-reading system is proposed, using the coherence between the audio signal and lip movements. The lip movements assist the VAD system to classify speech parts or particularly silence parts. In [3], two VAD methods have been proposed by Aubrey et al. to exploit the bi-modality of speech. The first one is based on the appearance parameters of the speaker's lips that are extracted from an active appearance model. The second one is based on the retinal filtering to obtain the essential parameters of lips region. These two methods are shown to have high accuracies for silence detection.

## 2.2 Speaker Detection

In multiparty dialogue, the problem of voice activity detection (VAD) is naturally extended to a problem of speaker detection. Speaker Detection (SD) systems are used to detect a speaker among multiple subjects from the incoming audio or video signal at every point in time as shown in Fig: 2.2. The task of speaker detection becomes even less accurate and increasingly complex if to depend merely on the audio signal, since the detection of "speech" is not sufficient enough, but techniques such as speaker identification or localisation need to be in place. However, similarly to VAD systems, visual information such as lip movements might provide a feasible and relatively more practically solution. In [4], the combination of both audio and visual information is used for speaker detection purpose.

The application of SD systems in human-machine multiparty dialogue is to help the machine to identify a speaker in a dialogue. These are used to control the multimodal outputs (gaze, head pose) of a robot. The SD systems are also used to zoom in the speaker in a video conference.

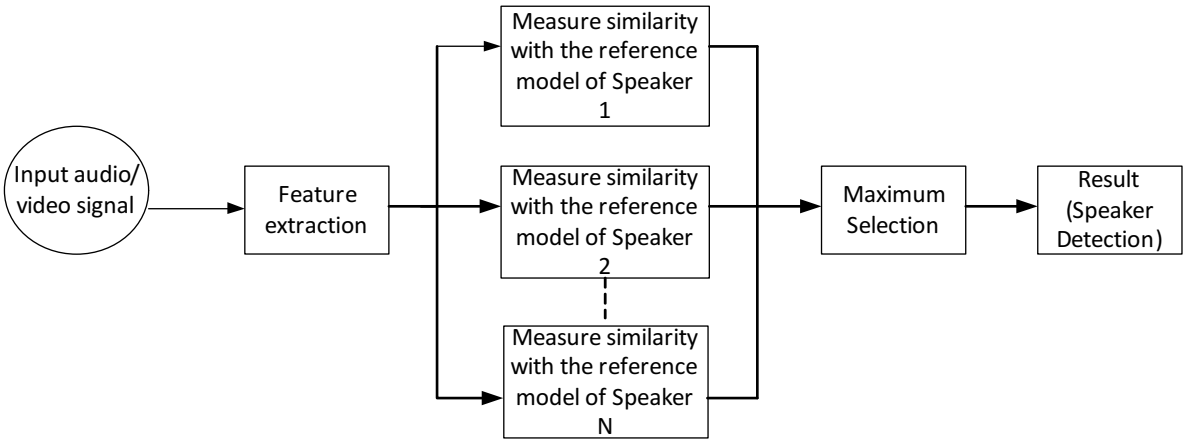



Figure 2.2: Speaker detection system

# Chapter 3

# Data Collection and Annotation

A video recording of human-machine multiparty dialogue is required for the proposed approach. The objective is to target the lip and head movements for speaker and specch/silence detection. Moreover different classification techniques (Mahalanobis, naïve Bayes classifier and neural network classifier) are applied to detect speech/silence and speaker and are compared with a baseline that relies only on the derivative.

## 3.1 Recording of Data

A video of four speakers speaking in a task free dialogue with a fifth speaker through video conferencing is audio-visually recorded. To simulate a multiparty human-machine dialogue, the fifth person communicated with the four speakers through video conferencing and the video is recorded from the perspective of the remote person (simulating the view point of the machine) as shown in Fig: 3.1. One restriction was imposed on the participants, which is not to talk to each other but rather address the remote speaker in their dialogue. However, they could smile, move their heads and may not look into the camera during the dialogue.

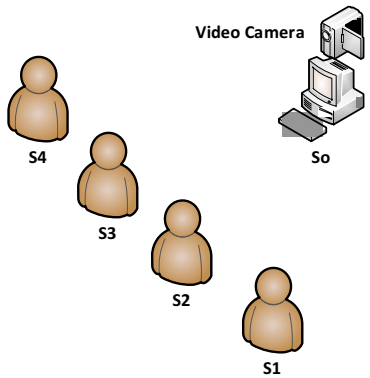



Figure 3.1: Recording setting (There are four speakers (S1, S2, S3 and S4) in front of the camera and computer. All four speaker are speaking in a free dialogue with a 5th speaker through video conferencing.)

The video is recorded using a high definition JVC video camera and the duration of the resulted video is 21 minutes with a frame rate of 25 frames/second. Two snapshots from the recorded video are shown in Fig: 3.2. The audio signal is also recoded with a stereo microphone as this corpus may be used for future studies. However, in this study, it is not important.

The recording session consists of two parts. In part one, the subjects (in front of the camera) ask questions from the fifth subject one by one through video conferencing and the fifth subject answers them. In second part, the fifth subject asks the questions from the participants of the video and s/he answers him. Some sample questions are as follows:

- Where is the nearest train station?
- How can I reach to the football ground?
- Where can I find a place for lunch at the KTH?

Figure 3.2: Two snapshots from the recorded video. In upper figure, all subjects are facing the camera and in lower figure, subject number 4 is not facing the camera.

## 3.2 Annotation of Data

The annotation of the video is performed by ELAN annotation software[7] as shown in Fig: 3.3. All speakers are annotated with segments of speech, silence, smile and laughter by the author. As a result, each speaker provided around 70 speech segments on average. For the annotation, a “speech” frame is defined as when any subject is talking while other subjects are silent. However, a “silence” frame is defined as when subjects are not smiling, laughing or speaking.

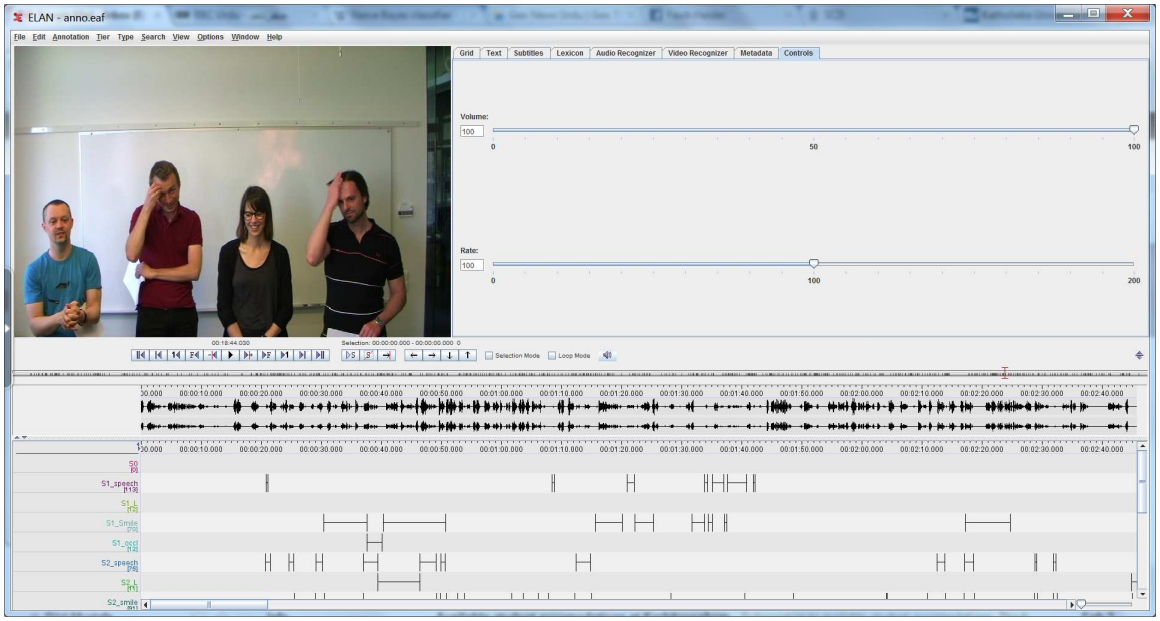



Figure 3.3: Video annotation by ELAN (A snapshot showing the annotation interface on the recorded video.)

In Table.3.2, the number of frames present in different classes are displayed. This study focuses on the speech and silence frames. However, the smile and laughter frames are not used at this stage.

| | Speech | Silence | Laughter | Smile |
|---|---|---|---|---|
| S1 | 1590 | 19291 | 607 | 7364 |
| S2 | 808 | 13743 | 1050 | 13575 |
| S3 | 515 | 19845 | 926 | 8677 |
| S4 | 519 | 21379 | 675 | 7016 |

Table 3.1: Number of frames present in each class for every speaker

## 3.3 Features Extraction

In this section, the strategy for extraction of the visual features is briefly described. A commercially availale software namely faceAPI [1] is used to track the facial markers of every speaker. FaceAPI is capable of estimating the head pose and the location of the lips, jaw, eyebrows, and the eyes in real-time and to a high accuracy. The limitation however is that, at the moment, it only works on one face at a time. Since there are four subjects present in our case, the images of each person are automatically cropped by manually selecting the subject face area as shown in Fig: 3.4. An example frame with overlaid markers is shown in Fig: 3.4

The features and their Identification (ID) numbers are shown in Fig: 3.5. Features used in this study are the lips inner height, outer height and width calculated by the face landmark ID numbers 101, 104, 202, 206, 200 and 204 and the head rotation along x, y and z axis as shown in Fig: 3.5 and Fig: 3.6.

Figure 3.4: Tracking of FaceAPI on the cropped images. In upper figure, a frame of the dialogue video showing the different speakers separated manually into different parallel video files and in lower figure, a frame of the dialogue video showing the faceAPI tracking on the different speakers.

[1] http://www.seeingmachines.com/product/faceapi/

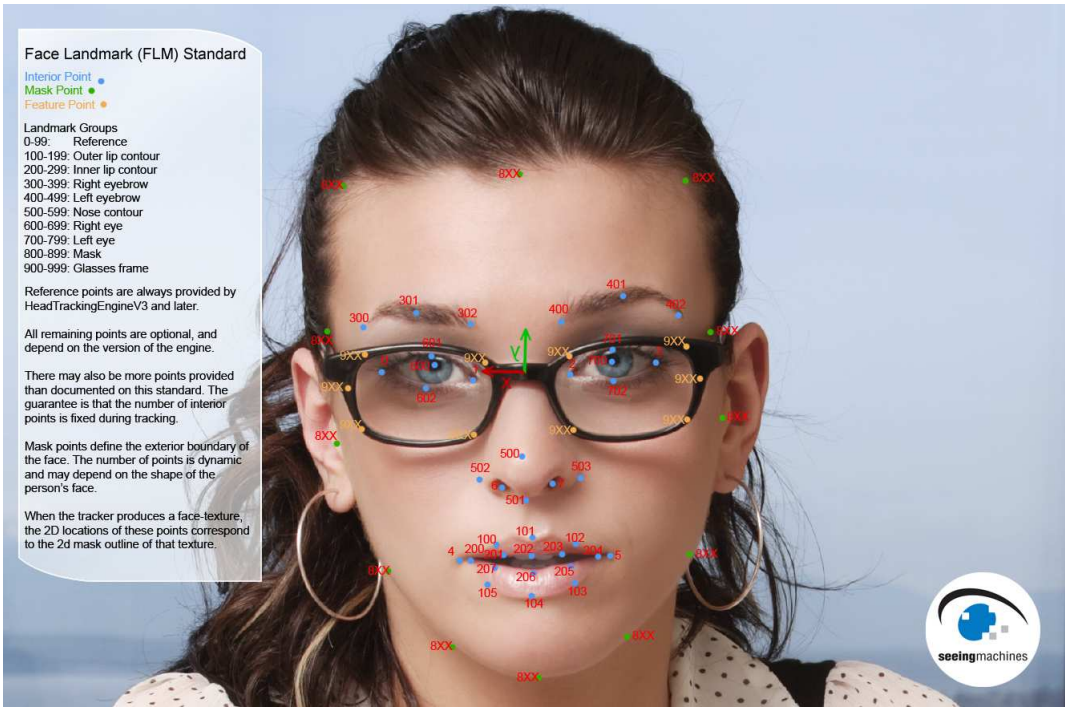



Figure 3.5: Face landmarks(Identification numbers of face landmarks (tracked by faceAPI)) [1]

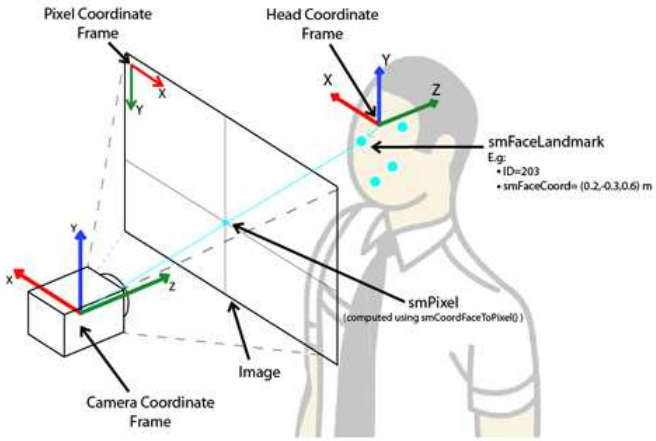



Figure 3.6: The face trakingAPI coordinate frames [1]

All the speech and silence data are concatenated. For classification purpose, different window-sizes are applied on the features (head and lip movements).

# Chapter 4

# Classification Methods

## 4.1 Baseline

For the baseline, we use the temporal derivative of tracked features. This is done by calculating the difference between every two consecutive frames and averaged over the size of the window. This derivative value represents the average rate of movement of each feature under the time window it is calculated over. Using the average derivative value, the speaker is detected by selecting the subject with highest derivative value. This means that the subject who moves the lips most is considered to be the speaker. Here, hundred frames of speech and silence are considered as training data. For the detection of speech/silence using this baseline, the average derivative value for speech and silence is calculated using the training data (hundred frames of speech and silence), and then compared to the testing data. If the difference between the mean of speech derivative value and testing data is less than the difference between the mean of silence derivative value and testing data then the frame is classified as speech frame.

## 4.2 Mahalanobis Distance

P. C. Mahalanobis introduced a distance measure in 1936 called Mahalanobis distance, to identify and analyze the different patterns by correlating the variables. It measures similarity of unknown data set to a know data set. For example, $x = (x_1, x_2, x_3, .....x_N)^T$ and $y = (y_1, y_2, y_3, .....y_N)^T$are two random vector and have same distribution with covariance matrix $S$. Then, Mahalanobis distance is defined as follows:

$$d(\overrightarrow{x}, \overrightarrow{y}) = \sqrt{(\overrightarrow{x} - \overrightarrow{y})S^{-1}(\overrightarrow{x} - \overrightarrow{y})} \tag{4.1}$$

In training phase, the distribution of the different lip and head movements is calculated using the for speech and non-speech segments for each speaker. During testing, the distance between the feature and the distribution is calculated, and the distribution that provides the shortest distance corresponds to the distribution (speech/non-speech) that feature belongs to. Speaker detection is performed by selecting the speaker with minimum average Mahalanobis distance using the speech distributions of all speakers as shown in Fig: 4.1.

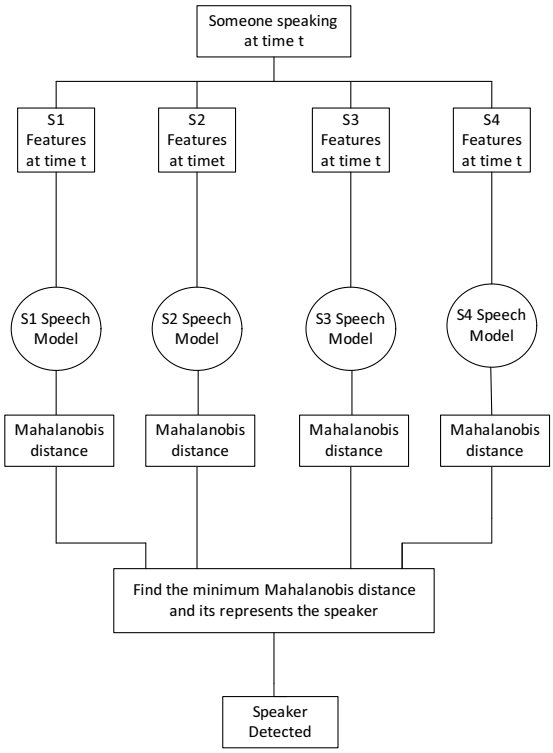



Figure 4.1: Speaker detection system block diagram (Mahalanobis distance)

For speech activity detection, done on each speaker separately, the Mahalanobis distance is calculated between the feature and the *speech* distribution calculated over the training data, and the *silence* distribution. The one that provides the shorter distance is the distribution the feature belongs to.

## 4.3 Naïve Bayes Classifier

The naïve Bayes classifier requires a small amount of data set for training to estimate the means and variances of data sets for classification purpose. The naïve Bayes classifier is computationally efficient as compared to Mahalanobis distance, where the whole covariance matrix needs to be calculated.

In naïve Bayes classifier, a probability model (in our case, it is normal distribution) is assumed to determine the posterior probability for the classification purpose. One main constraint of statistical models is that they perform well only when the underlying suppositions are fulfilled. The performance of naïve Bayes classifier depends to a large extent on different suppositions or clauses under which the models are developed. A good information of both data properties and model aptitudes are needed before the models can be effectively applied. The naïve Bayes classifier is used to calculate the posterior probability as shown in Eq:4.2. The $P(\omega_j)$ represents the prior probability of group j. The $f(x|\omega_j)$ represents the probability density function, then Bayes rule says:

$$P(\omega_j|x) = \frac{f(x|\omega_j)P(\omega_j)}{f(x)} \tag{4.2}$$

$f(x)$ is the probability density function and represented by the following formula.

$$f(x) = \Sigma_{j=1}^{M} f(x|\omega_j) P(\omega_j) \tag{4.3}$$

$P(\omega_j|x)$ denotes the posterior probability. $M$ is the total number of observed classes like in our case it is 4 for SD approach and 2 for VAD approach.

$$P(\omega_k|x) = \max_{i = 1, 2, ...M} P(\omega_i|x) \tag{4.4}$$

lip movements are assumed as normal process and naïve Bayes classifier is applied. During training, the classifier is trained using 70% of the available data to estimate the mean and variance of the normal distribution and 30% is used for testing.

## 4.4 Neural Network Classifier

The neural networks have appeared as a key method for classification problem. The latest research activities in neural classification have found that the neural networks are a promising substitute to different conventional classification techniques. The neural networks have ability to adjust themselves without any explicit knowledge of underlying model aptitudes. The neural networks provide direct estimates of posterior probability and importance of this capability is summarized by Richard and Lippmann.

"*Interpretation of network outputs as Bayesian probabilities allows outputs from multiple networks to be combined for higher level decision making, simplifies creation of rejection thresholds, makes it possible to compensate for difference between pattern class probabilities in training and test data, allows output to be used to minimize alternative risk functions, and suggests alternative measures of network performance.*"[15]

The single-layer feed-forward neural network is applied on the lip and head features for speaker and speech/silence detection and the results are discussed in chapter 3. The MATLAB neural network tool box is used for simulation of network. Neural network is trained, using 60% of the data and validated with 20% of the data. 20% is used for testing.

# Chapter 5

# Results and Discussion

## 5.1 Voice Activity Detection

In this section, the results of VAD system using the speaker dependent, speaker independent and hybrid methods are shown and discussed. In VAD approach, we are 50% sure about the class (speech/silence) of every frame. For example, if the speech frames for subject number one are 1590 then the same numbers of silence frames are present in the data (speech and silence).

### 5.1.1 Speaker Dependent Method

In the speaker dependent method, data from each speaker is used as part of the training and testing, which means that all subjects have been already seen in the training and only models trained on that speakers are used in the testing. The speaker dependent method hence requires an identification of the speaker to run his own associated models. For example, if the training is performed using the facial data of S1 then the testing will be performed using the facial data of S1 as shown in Fig: 5.1.

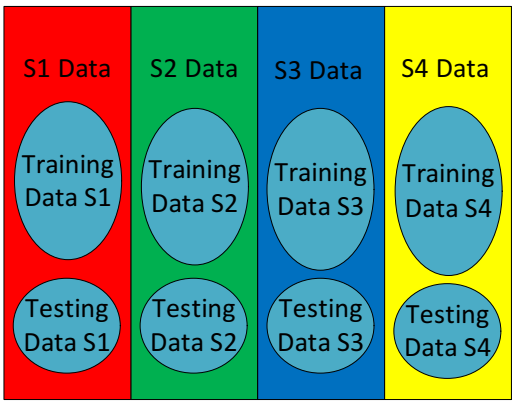



Figure 5.1: Speaker dependent method (A plot showing the principal separation between the testing and training data for the speaker dependent method.)

The results of baseline, Mahalanobis distance and naïve Bayes classifier using the combination of head and lip movements are shown in Fig: 5.2.

The naïve Bayes classifier has the best results with an optimal accutacy of 94.23%. The Mahalanobis distance and baseline provide the optimal accuracies of 67.01% and 49.92% respectively.

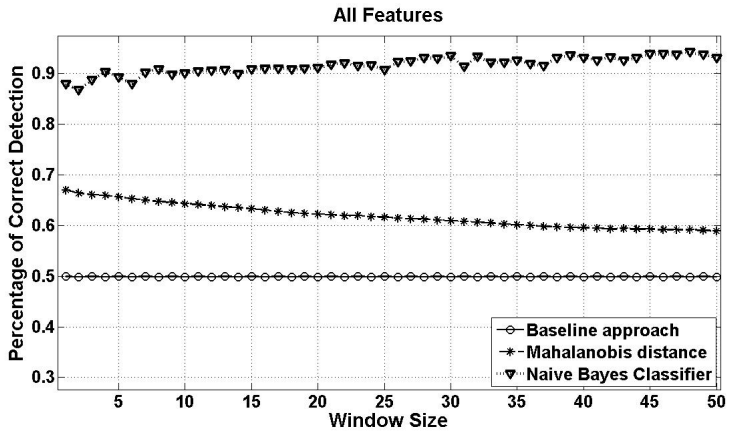



Figure 5.2: Results of speaker dependent VAD (lip and head movements as features)

The results of baseline, Mahalanobis distance and naïve Bayes classifier using the lip movements are displayed in Fig: 5.3. The naïve Bayes classifier has the best results with an optimal accuracy of 76.95%. The Mahalanobis distance and baseline approach provide the optimal accuracies of 42.49% and 49.95% respectively.

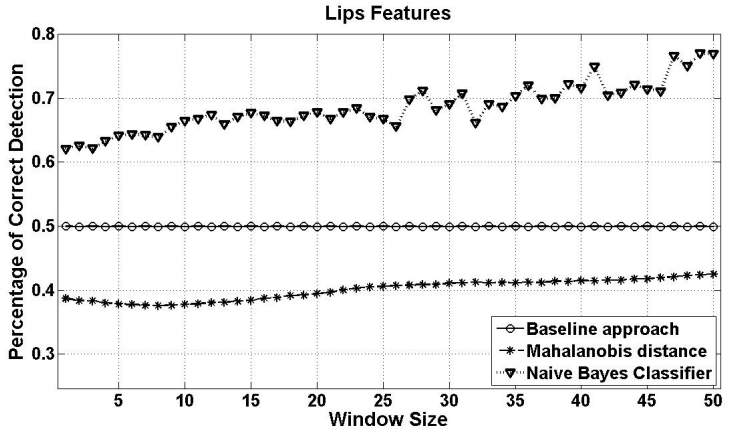



Figure 5.3: Results of speaker dependent VAD (lip movements as features)

The results of baseline, Mahalanobis distance and naïve Bayes classifier using the head movements are displayed in Fig: 5.4. The naïve Bayes classifier has the bests results with an optimal accuracy of 93.56%. The Mahalanobis distance and baseline approach provide the optimal accuracies of 67.75% and 49.92% respectively.

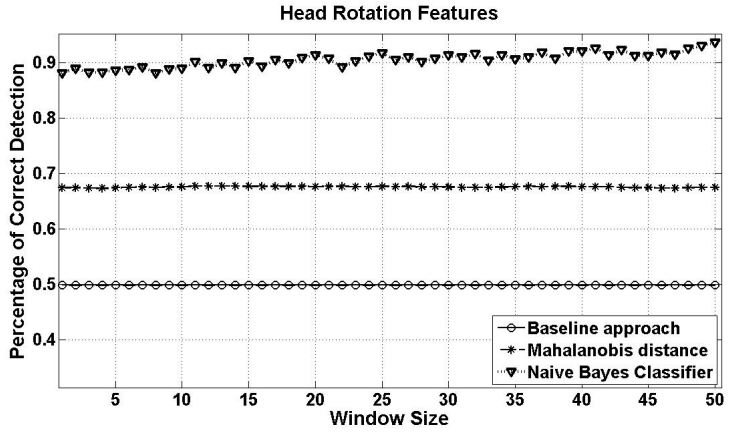



Figure 5.4: Results of speaker dependent VAD (head movements as features)

The results of neural network classifier using the combination of lip and head movements are displayed in Fig: 5.5. The neural network classifier provides an optimal accuracy of 98.73%.

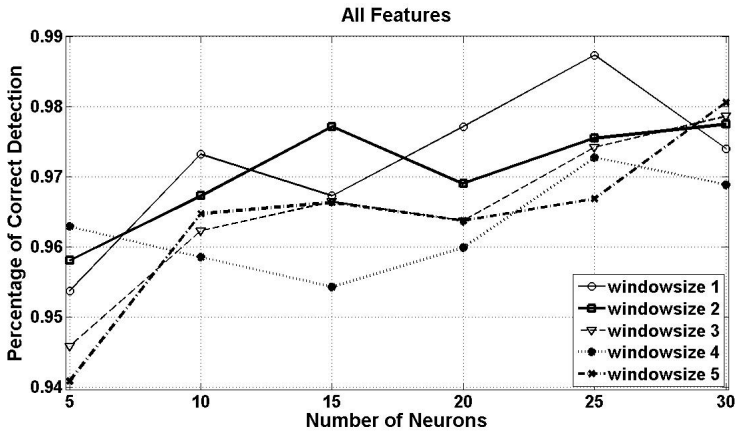



Figure 5.5: Results of speaker dependent VAD using neural network (lip and head movements as features)

The results of neural network classifier using the lip movements are displayed in Fig: 5.6. The neural network classifer provides an optimal accuracy of 78.31%.

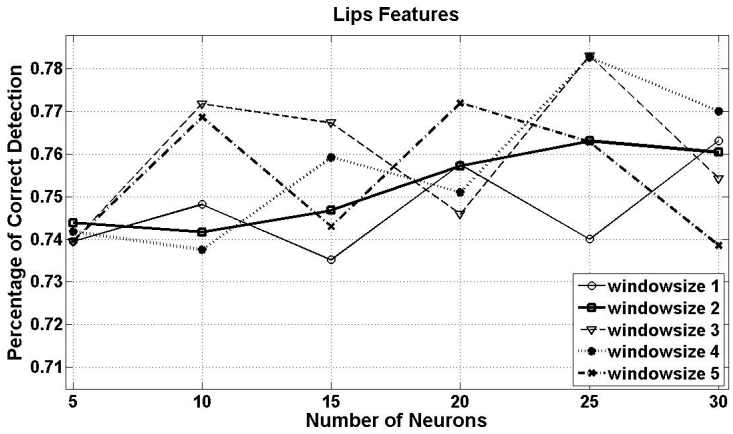
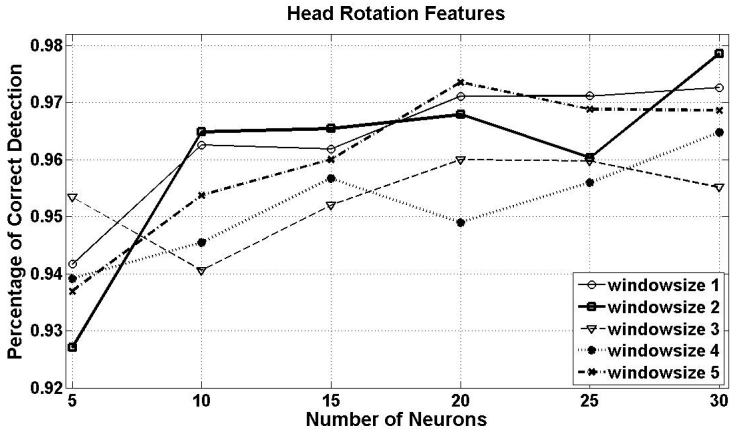



Figure 5.6: Results of speaker dependent VAD using neural network (lip movements as features)

The results of neural network classifier using the head movements are displayed in Fig: 5.7.The neural network classifer provides an optimal accuracy of 97.87%.




X-axis represents the number of neurons in the hidden layer of neural network and y-axis represents the percentage of correct speech/non-speech detection.

Figure 5.7: Results of speaker dependent-VAD using neural network (head movements as features)

The neural network classifier provides the best result using the combination of lip and head movements with an optimal accuracy of 98.73% as shown in Table: 5.1. However, the results of the head movements are better than lip movements.

| Features | Baseline Approach | Mahalanobis Distance | naïve Bayes Classifier | Neural Network Classifier |
|---|---|---|---|---|
| lip | 49.95% | 42.49% | 76.95% | *78.31%* |
| Head rotation | 49.92% | 67.75% | 93.56% | *97.87%* |
| All | 49.92% | 67.01% | 94.23% | *98.73%* |

Table 5.1: Results of speaker dependent VAD

The results of naïve Bayes classifier are better than Mahalanobis distance and baseline. However, Mahalanobis distance provides better results than baseline using the head movements. The combination of head and lip movements improves the performance of the VAD system as compared to lip and head movements. However, in Mahalanobis distance, the combination of head and lip movements causes a slightly decrease in the accuracy of VAD system as compared to head movements.

In speaker dependent method, training and testing are performed on the same speaker. The results of both neural network and naïve Bayes classifiers are promising. Moreover, the results of head movements are better than lip movements. Fig: *5.3* depicts that the performance of naïve Bayes classifier increases with an increase in window size. It is due to the fact that more frames are used for calculating means and variances of the training and testing data thus resulting in more accurate performance.

### 5.1.2 Speaker Independent Method

In Speaker Independent (SI) method, training and testing data does not belong to same speaker as shown in Fig: 5.1.2. It means that the testing is not performed using the speaker's own training model.

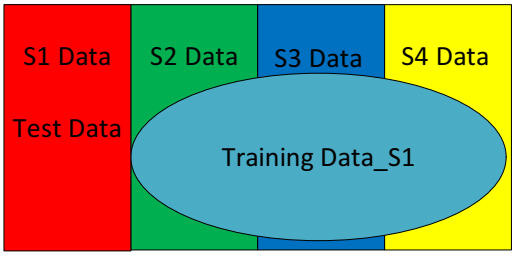



Figure 5.8: SI-VAD (A plot showing the principal separation between the testing and training data for the speaker independent method.)

The results of baseline, Mahalanobis distance and naïve Bayes classifier using the combination of head and lip movements are shown in Fig: 5.9. The naïve Bayes classifier provides the best result with an optimal accuracy of 67.57%. The Mahalanobis distance and baseline provide the optimal accuracies of 60.50% and 49.88% respectively.

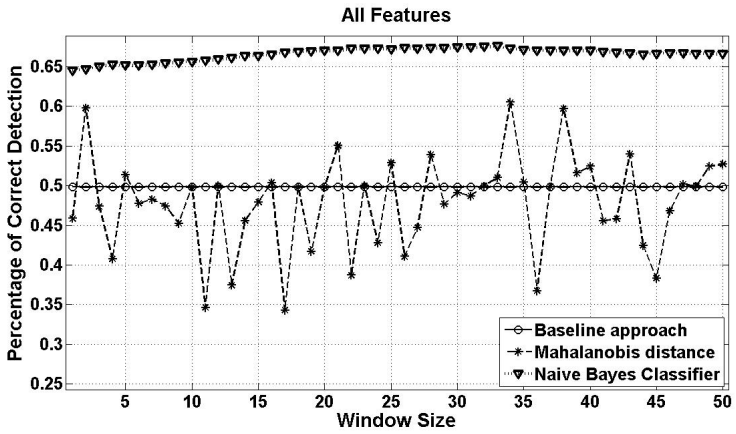



Figure 5.9: Results of SI-VAD (lip and head movements as features)

The results of baseline, Mahalanobis distance and naïve Bayes classifier using the lip movements are displayed in Fig: 5.10. The naïve Bayes classifier has the best result with an optimal accuracy of 64.56%. The Mahalanobis distance and baseline provide the optimal accuracies of 52.43% and 49.95% respectively.

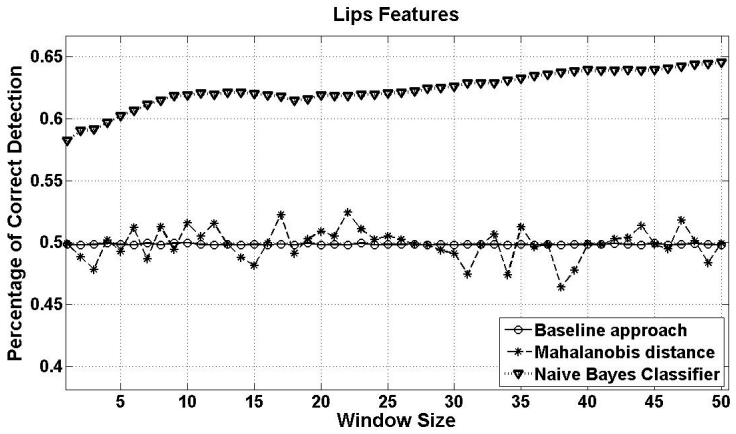



Figure 5.10: Results of SI-VAD (lip movements as features)

The results of baseline, Mahalanobis distance and naïve Bayes classifier using the head movements are displayed in Fig:5.11. The naïve Bayes classifier provides the best result with an optimal accuracy of 62.71%. The Mahalanobis distance and baseline approach provide the optimal accuracies of 63.90% and 49.89% respectively.

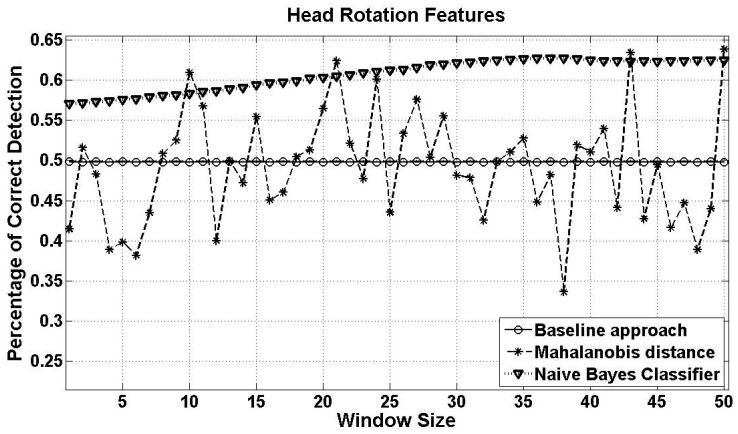



Figure 5.11: Results of SI-VAD (head movements as features)

The results of neural network classifier using the combination of lip and head movements are displayed in Fig:5.12. The neural network classifier provides an optimal accuracy of 57.61%.

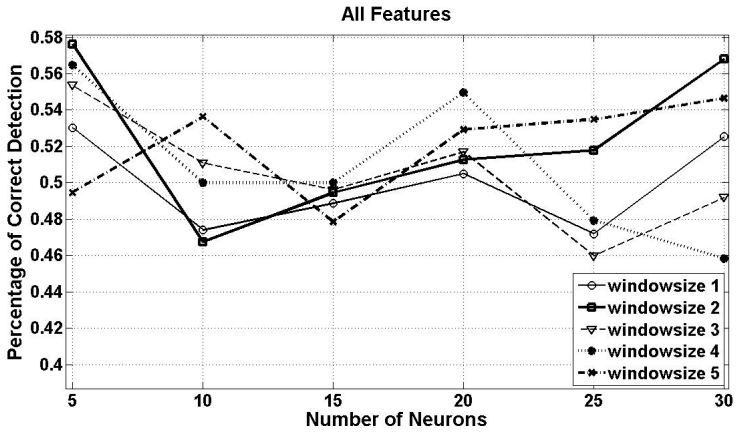



Figure 5.12: Results of SI-VAD using neural network (lip and head movements as features)

The results of neural network classifier using the lip movements are displayed in Fig:5.13. The neural network classifier provides an optimal accuracy of 52.17%.

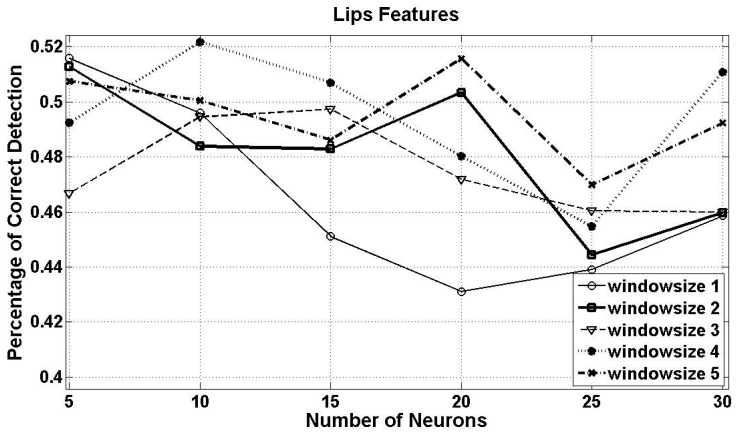



Figure 5.13: Results of SI-VAD using neural network (lip movements as features)

The results of neural network classifier using the head movements are displayed in Fig:5.14 . The neural network classifier provides an optimal accuracy of 58.36%.

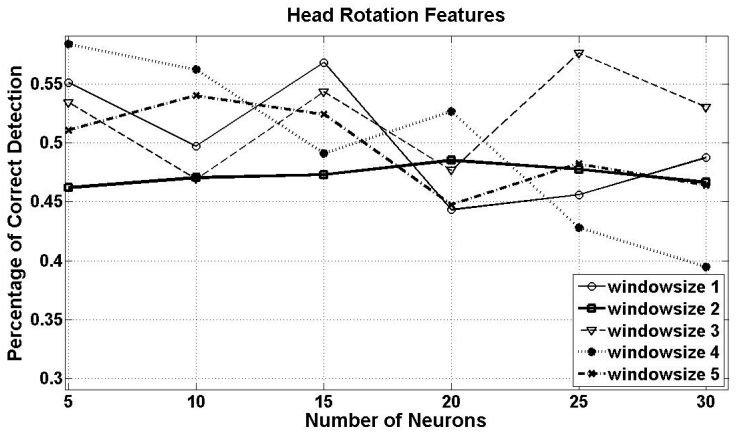



Figure 5.14: Results of SI-VAD using neural network (head movements as features)

The naïve Bayes classifier provides the best result using the combination of lip and head movements with an optimal accuracy of 67.57% as shown in Table: 5.2. However, the results of the lip movements are better than head movements for naïve Bayes classifier.

| Features | Baseline Approach | Mahalanobis Distance | naïve Bayes Classifier | Neural Network Classifier |
|---|---|---|---|---|
| lip | 49.86% | 52.43% | *64.56%* | 52.17% |
| Head rotation | 49.85% | *63.9%* | 62.71% | 58.36% |
| All | 49.98 % | 60.5% | *67.57%* | 57.61% |

Table 5.2: Results of SI-VAD

The Mahalanobis distance and naïve Bayes classifier provide results using the head movements with the optimal accuracies of 63.90% and 62.71% respectively. However, Mahalanobis distance provides better results than neural network classifier. Moreover, in Mahalanobis distance and neural network classifier, the combination of head and lip movements causes a slightly decrease in the accuracy of VAD system as compared to head movements.

In this approach, the training model does not contain the facial data of test person, that's why it might have been difficult for the neural network to deal with the unexperienced situations. The naïve Bayes classifier works on the normal distribution assumption that's why it might have been less affected by the training data as compared to the neural network classifier. The head and lip movements are extremely different for every person resulting in the poor performance of the classifiers. In this case, the training model contains the facial data of three different subjects. The neural network performance might be increased by introducing more speakers in the training model. It is due to the fact that the training model probably contains some similar behavior of lip and head movement from another speaker as compared to the test speaker.

### 5.1.3 Hybrid Method

In Hybrid (H) method, the training model contains the facial data of all the four subjects as shown in Fig: 5.15.

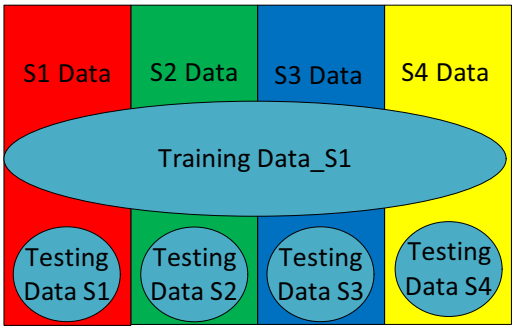



Figure 5.15: H-VAD (A plot showing the principal separation between the testing and training data for the hybrid method.)

The results of baseline, Mahalanobis distance and naïve Bayes classifier using the combination of head and lip movements are shown in Fig: 5.16. The naïve Bayes classifier provides the best result with an optimal accuracy of 74.29%. The Mahalanobis distance and baseline provide the optimal accuracies of 57.11% and 49.86% respectively.

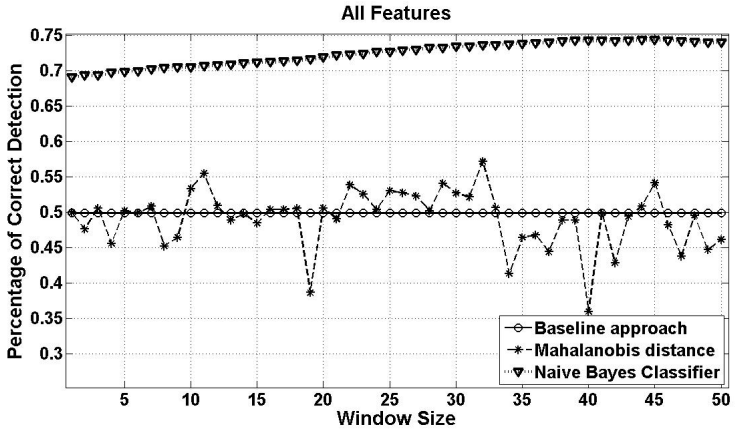



Figure 5.16: Results of H-VAD (lip and head movements as features)

The results of baseline, Mahalanobis distance and naïve Bayes classifier using the lip movements are shown in Fig: 5.17. The naïve Bayes classifier provides the best result with an optimal accuracy of 66.61%. The Mahalanobis distance and baseline provide the optimal accuracies of 55.61% and 49.86% respectively.

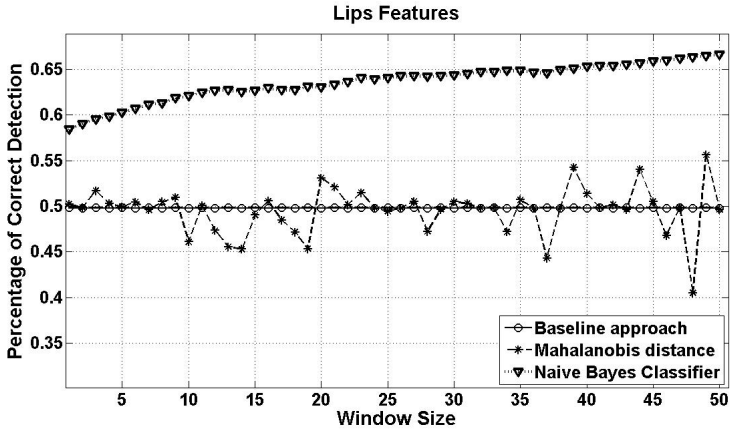



Figure 5.17: Results of H-VAD (lip movements as features)

The results of baseline, Mahalanobis distance and naïve Bayes classifier using the head movements are shown in Fig: 5.18. The naïve Bayes classifier provides the best result with an optimal accuracy of 70.78%. The Mahalanobis distance and baseline provide the optimal accuracies of 61.10% and 49.86% respectively.

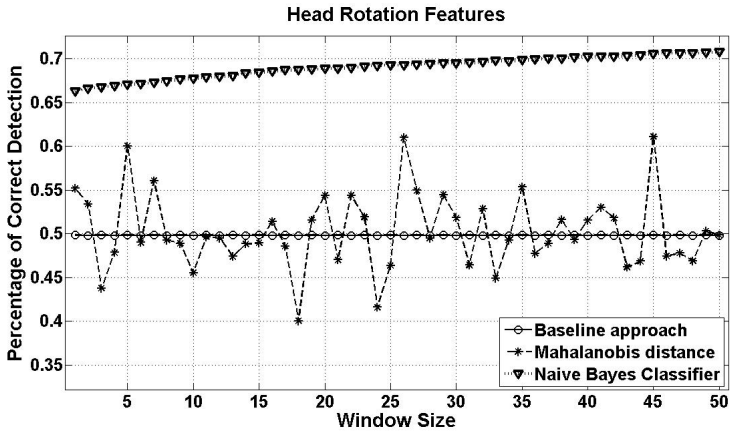



Figure 5.18: Results of H-VAD (head movements as features)

The results of neural network classifier using the combination of lip and head movements are displayed in Fig: 5.19. The neural network classifier provides an optimal accuracy of 96.22%.

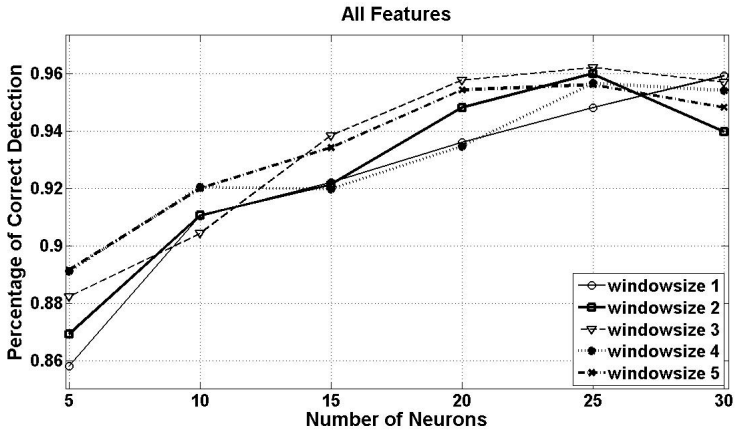



Figure 5.19: Results of H-VAD using neural network (lip and head movements features)

The results of neural network classifier using the lip movements are displayed in Fig: 5.20. The neural network classifier provides an optimal accuracy of 70.85%.

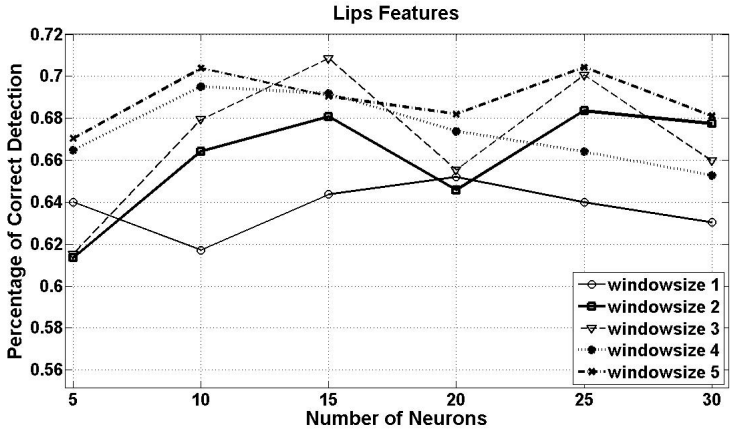
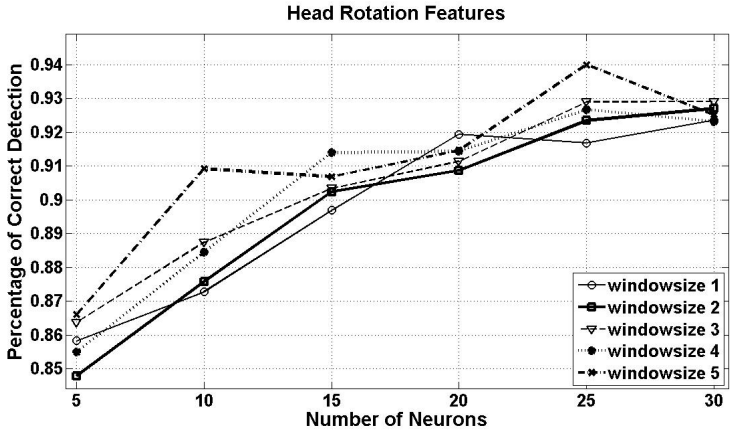



Figure 5.20: Results of H-VAD using neural network (lip movements as features)

The results of neural network classifier using the head movements are displayed in Fig: 5.21. The neural network classifier provides an optimal accuracy of 92.91%.




.

Figure 5.21: Results of H-VAD using neural network (head movements as features)

The neural network classifier provides the best result using the combination of lip and head movements with an optimal accuracy of 96.22% as shown in Table: 5.3. However, the results of the head movements are better than lip movements. Moreover, in naïve Bayes classifier and neural network classifier, the combination of head and lip movements causes an increase in the accuracy of VAD system.

| Features | Baseline Approach | Mahalanobis Distance | naïve Bayes Classifier | Neural Network Classifier |
|---|---|---|---|---|
| lip | 49.86% | 55.61% | 66.61% | *70.85%* |
| Head rotation | 49.86% | 61.1% | 70.78% | *92.91%* |
| All | 49.86% | 57.11% | 74.29% | *96.22%* |

Table 5.3: Results of H-VAD

In this approach, training model contains the facial data of all the four speakers. The head movements results are better than the lip movements due to the fact that the speakers close their lip during speech. The results of the hybrid methods are better than the speaker independent method. Here, we can see the possibility of more accurate detection in the case of speaker independent method. It is due to the fact that some of the training data belong to the test speaker. If our training data contains the lip and head movements of the various people then probably we can find the similar relationship among the multiple speakers and the test subject.

## 5.2 Speaker Detection (SD)

### 5.2.1 Speaker Dependent Method

In this approach, training and testing data of speech frames belong to the same speaker but testing data is not a part of training data as shown in Fig: 5.2.1.

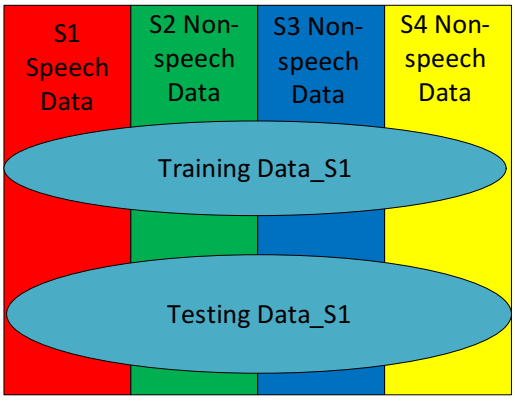



Figure 5.22: Speaker detection (A plot showing the principal separation between the testing and training data for the speaker dependent framework. In this case, the subject S1 is speaking and the other subjects (S2, S3 and S4) are silent.)

The results of baseline, Mahalanobis distance and naïve Bayes classifier using the combination of head and lip movements are shown in Fig: 5.23. The naïve Bayes classifier has the best results with an optimal accutacy of

76.87%. The Mahalanobis distance and baseline provide the optimal accuracies of 45.14% and 25.69% respectively.

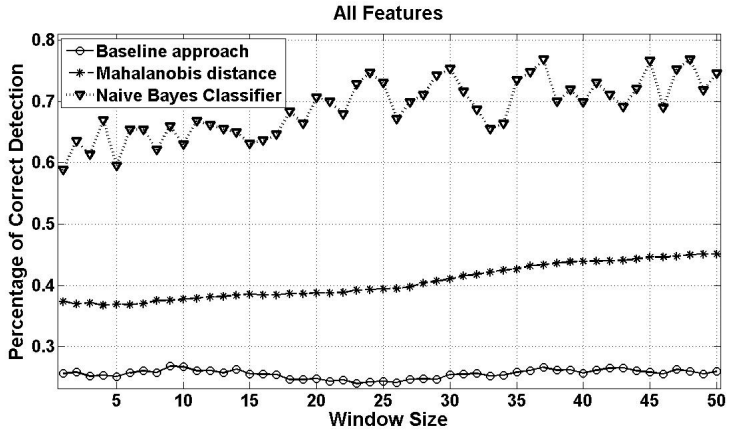



Figure 5.23: Results of SD (lip and head movements as features)

The results of baseline, Mahalanobis distance and naïve Bayes classifier using the lip movements are shown in Fig: 5.24. The naïve Bayes classifier has the best results with an optimal accutacy of 58.28%. The Mahalanobis distance and baseline provide the optimal accuracies of 29.68% and 27.11% respectively.

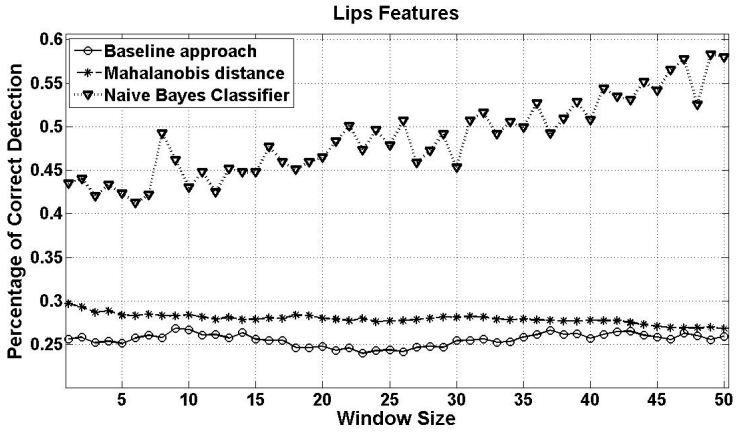



Figure 5.24: Results of SD (lip movements as features)

The results of baseline, Mahalanobis distance and naïve Bayes classifier using the head movements are shown in Fig: 5.25. The naïve Bayes classifier has the best results with an optimal accutacy of 72.60%. The Mahalanobis distance and baseline provide the optimal accuracies of 51.93% and 25.41% respectively.

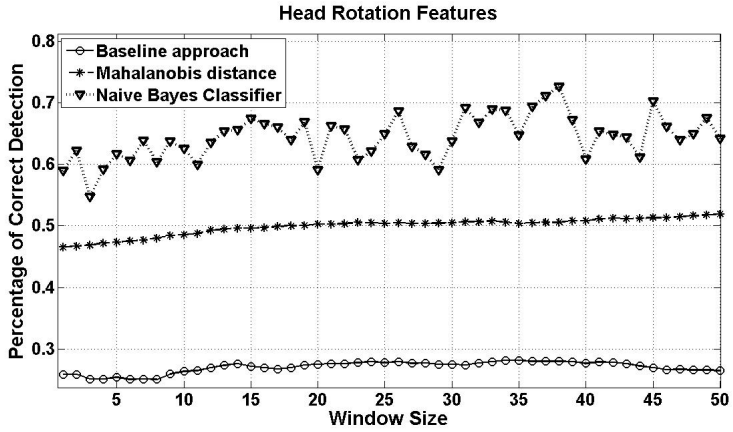



Figure 5.25: Results of SD (head movements as features)

The results of neural network classifier using the combination of lip and head movements are displayed in Fig: 5.26. The neural network classifier provides an optimal accuracy of 97.43%.

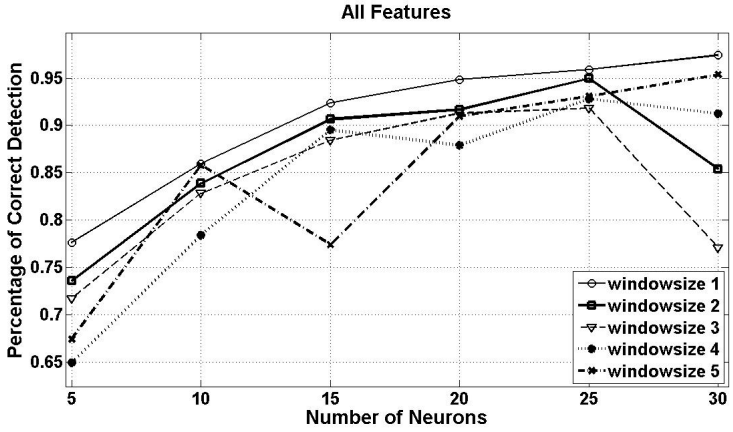



Figure 5.26: Results of SD using neural network (lip and head movements as features)

The results of neural network classifier using the lip movements are displayed in Fig: 5.27. The neural network classifier provides an optimal accuracy of 71.29%.

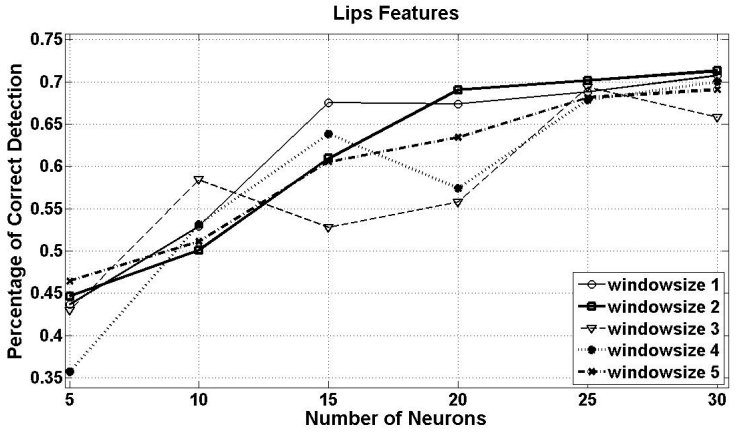



Figure 5.27: Results of SD using neural network (lip movements as features)

The results of neural network classifier using the head movements are displayed in Fig: 5.28. The neural network classifier provides an optimal accuracy of 97.13%.

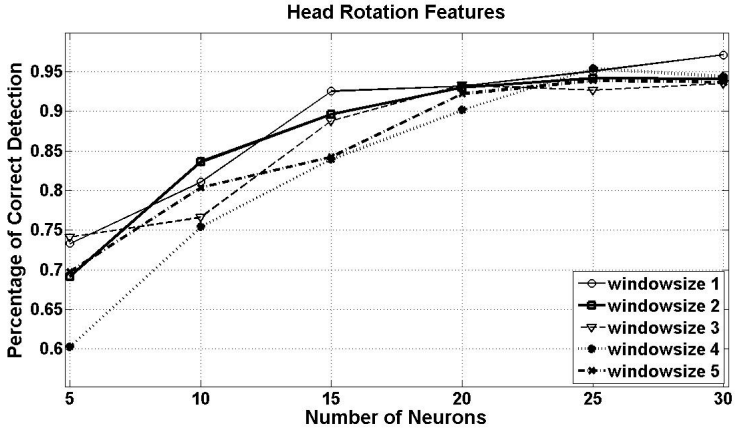



Figure 5.28: Results of SD using neural network (head movements as features)

The neural network classifier provides the best result using the combination of lip and head movements with an optimal accuracy of 97.43% as shown in Table: 5.4. However, the results of the head movements are better than lip movements. Moreover, in naïve Bayes classifier and neural network classifier, the combination of head and lip movements causes an increase in the accuracy of SD system.

| Features | Baseline Approach | Mahalanobis Distance | naïve Bayes Classifier | Neural Network Classifier |
|---|---|---|---|---|
| lip | 26.85% | 29.68% | 58.28% | *71.29%* |
| Head rotation | 28.16% | 51.93% | 72.6% | *97.13%* |
| All | 26.85% | 45.14% | 76.87% | *97.43%* |

Table 5.4: Results of speaker detection

The results of the neural network and naive Bayes classifiers are promising. The results of head movements are better than lip movements. It is due to the fact that the speakers also close their lip during the speech and we have selected those frames as speech frames in the training data. It might have been difficult for the Mahalanobis distance approach and naïve Bayes classifier to distinguish between the lips closed during speech and lips closed during silence. The mean and variance values of speech and non speech frames also indicate that there are no significant difference between the lip movements during speech and silence. Fig: 5.24 shows that the performance of naïve Bayes classifier increases with an increase in window size. It is due to the fact that more frames are used for calculating means and variances of the training and testing data thus resulting in more accurate performance.

# Chapter 6

# Conclusion

This paper proposes speaker detection and voice activity detection systems using the lip and head movements in the context of human-machine multiparty dialogue. The naïve Bayes and neural network classifiers provide promising results. The results show that the performance of naïve Bayes classifier is less affected in the speaker independent method as compared to the neural network classifier. It also raises the point that movements of lip and head may vary from speaker to speaker.

In the speaker dependent and hybrid methods, the results of the neural network classifier provides the best results with the optimal accuracies of 98.73% and 96.22% repectively for the speech/silence detection. However, in speaker independent method, the results of naïve Bayes classifier provides best results with an optimal accuracy of 67.57%. Moreover, in speaker detection, the speaker dependent method provides an optimal accuracy of 97.43% using the neural network classifier.

The head movements are also found to be working for the speaker detection and voice activity detection methods. The classification by fusion of head and lip movements provides better results than that of lip and head movements separately. In speaker dependent and hybrid methods, the results of head movements are better than lip movements. However, in speaker independent method, the results of lip movements are better than head movements.

# Chapter 7

# Future Work

Previous work done on facial voice activity detection has usually been using full face high quality videos with almost constant environment conditions such as lighting and illumination. In this work we tried to push the boundary and really try to test the approaches using a more ecological multiparty multimodal setup, simulating a task free multiparty dialogue with far field microphone and one video camera. To do this, we decided to use a proprietary real-time face tracker (FaceAPI). Although FaceAPI provides relatively accurate tracking, it is however difficult to get insights on how quantitatively good it is, and whether its performance can be increased by providing context dependent training data. A future work we envision is to try out state-of-the-art research methods for lip tracking.

Results from research also show that other facial parameter might help regulate multiparty interaction, signals such as gaze direction, head orientation, and eyebrows movements might be useful parameters to include in future studies.

In this work we also avoided using facial data when subjects were smiling and laughing, due to the difficulty to visually distinguish a smile or laughter from speech. A possible future work can be the detection of smile and laughter in voice activity detection systems. In this study, commercially available software namely FaceAPI is used to extract the features. FaceAPI works for only one face at a time thus making the speaker detection system unrealisable for real time applications and so algorithms are required for multiple faces tracking out of a video, to make the system realisable for real time applications.

# Bibliography


[1] http://www.seeingmachines.com/product/faceapi/.

[2] Y.A. Aubrey, A.J.; Hicks and J.A. Chambers. Visual voice activity detection with optical flow. vol.4:pages 463 – 472, 2010.

[3] Andrew Aubrey; Bertrand Rivet; Yulia Hicks; Laurent Girin; Jonathon Chambers and Christian Jutten. Two novel visual voice activity detectors based on appearance models and retinal filtering. *15th European Signal Processing Conference (EUSIPCO 2007), Poznan, Poland*, 2007.

[4] J.M.; Pavlovic V.; Pentland A Choudhury, T.; Rehg. Boosting and structure learning in dynamic bayesian networks for audio-visual speaker detection. vol.3:pages. 789–794, 2002.

[5] A. D. Christian and B. L. Avery. Digital smart kiosk project. *ACM SIGCHI 98, Los Angeles, CA, April 18-23*, pages 155–162, 1998.

[6] Al Moubayed S.; Beskow J.; Skantze G. and Granström B. Furhat: A back-projected human-like robot head for multiparty human-machine interaction. *In Esposito, A. et al. (Eds.), cognitive behavioural systems. Lecture notes in computer science. Springer.*, 2012.

[7] Brugman H. and Russel A. Annotating multi-media / multimodal resources with elan. *In proceedings of LREC. (Lisbon, Portugal), 2065-2068.*, 2004.

[8] Shin ichi Takeuchi; Takashi Hashiba; Satoshi Tamura and Satoru Hayamizu. Voice activity detection based on fusion of audio and visual information. 2009.

[9] Ramírez J; Yélamos P; Górriz J and J.C M Segura. Svm-based speech endpoint detection using contextual speech features. *IEE Electronics Letters*, vol. 42, 7., 2006.

[10] Górriz JM; Ramírez J; Segura JC and Puntonet CG. An effective cluster-based model for robust speech detection and speech recognition in noisy environments. *Journal of the Acoustical Society of America*, vol. 120, 1:pages 470–481, 2006.

[11] K. Waters; J. M. Rehg; M. Loughlin; S. B. Kang and D. Terzopoulos. Visual sensing of humans for active public interfaces. *Computer Vision for HumanMachine Interaction*, Cambridge University Press:pages 83–96, 1998.

[12] J. M. Rehg; M. Loughlin and K. Waters. Vision for a smart kiosk. *Computer Vision and Pattern Recognition*, pages 690–696, 1997.

[13] Estevez P.A.; Becerra-Yoma N.; Boric N. and Ramirez J.A. Genetic programming based voice activity detection. *Electronics Letters*, vol. 41, 20:pages 1141–1142, 2005.

[14] Takami Yoshida; Kazuhiro Nakadai and Hiroshi G. Okuno. An improvement in audio-visual voice activity detection for automatic speech recognition.

[15] M. D. Richard and R. Lippmann. Neural network classifiers estimate bayesian a posteriori probabilities. vol. 3:pages 461–483, 1991.

[16] David Sodoyer; Bertrand Rivet; Laurent Girin; Jean-Luc Schwart and Christian Jutte. An analysis of visual speech information applied to voice activity detection. 2006.

[17] W. Sumby and I. Pollack. Visual contributions to speech intelligibility in noise. *Acoust. Soc. Am.*, vol. 26 Issue 2:pages 212–215, 1954.

TRITA-CSC-E 2013:001
ISRN-KTH/CSC/E--13/001-SE
ISSN-1653-5715

www.kth.se