"Determination of Time Offset between Audio and Video in Online Repositories (YouTube)"

B.Tech. Project End-Term Evaluation

Vikram Voleti
09EE3501

Motivation

- To measure time offset between audio signal and video signal in a video
- To eliminate redundant videos in a search

Bilabial Consonants

- *‘p,’ ‘b,’ ‘m’* in the International Phonetic Alphabet
- Require the lips to close
- Serve as reference points to check for synchronicity

Detection of bilabial consonants in audio:

Mel Frequency Cepstral Coefficients (MFCC)

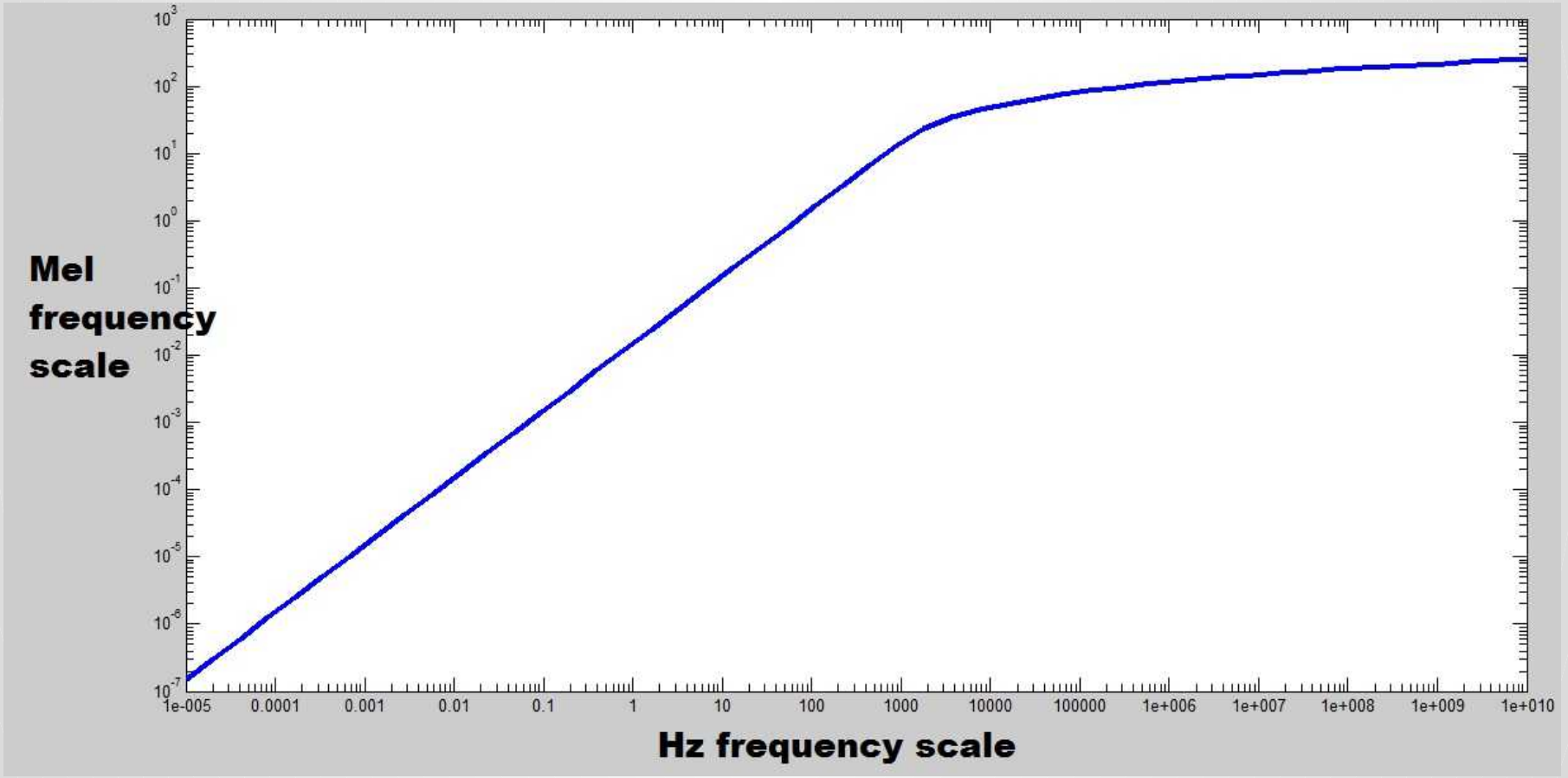
- Derived from the audio signal using digital signal processing operations
- Well-suited to parameterize human speech
  - Reduce noise
  - Mapping is on the Mel scale

Gaussian Mixture Models (GMM)

- Parametric probability density function
- Weighted sum of Gaussian component densities

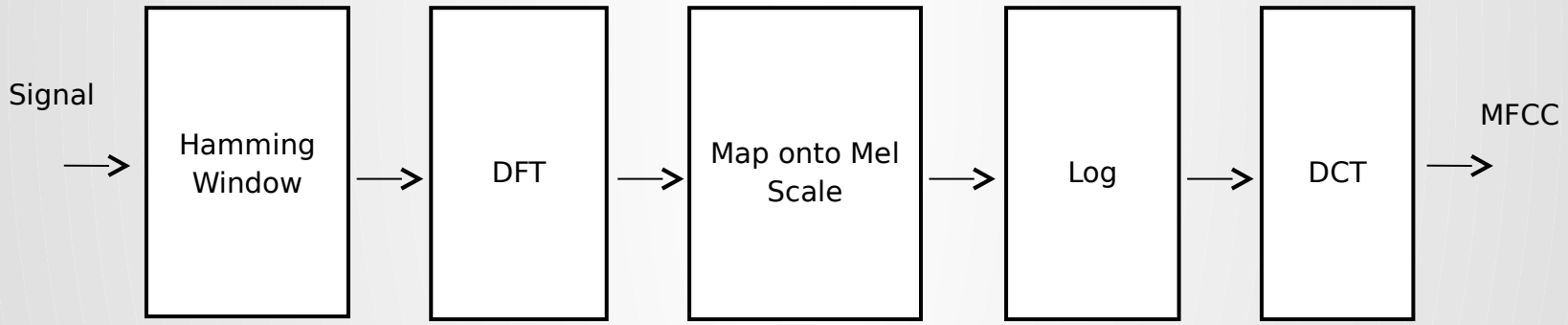
Detection of bilabial consonants in audio:

Relation between the Hertz and the Mel scales



Detection of bilabial consonants in audio:

Mel Frequency Cepstral Coefficients (MFCC)



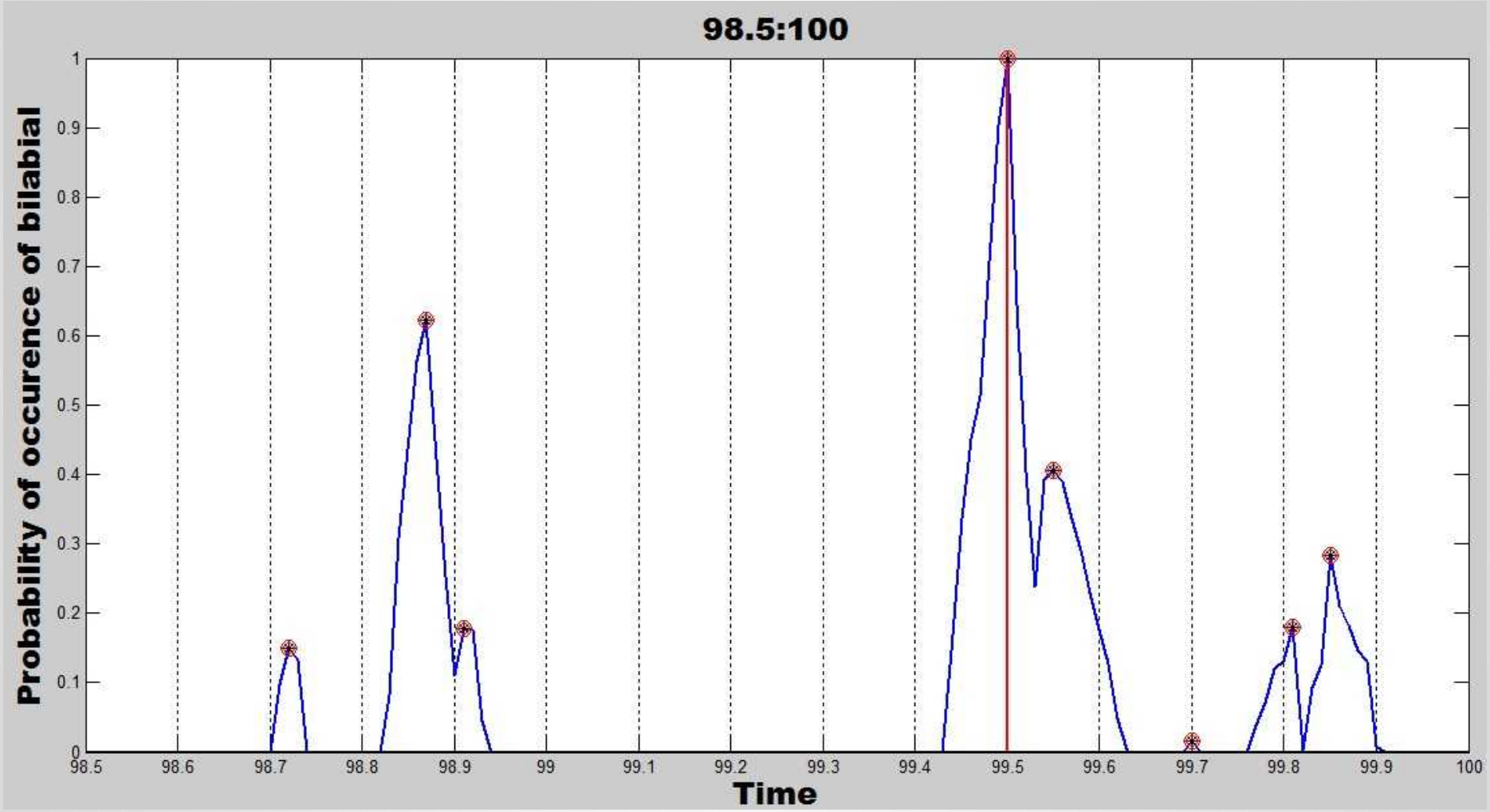
Detection of bilabial consonants in audio: Training

- Calculate MFCC's for audio signal
- Manually identify windows of bilabial sounds in audio signal
- Train GMM1 using MFCC's from bilabial sounds
- Train GMM2 using MFCC's from non-bilabial sounds
- Training: Using an Expectation Maximization (EM) algorithm to fit Gaussian curves into the feature vectors

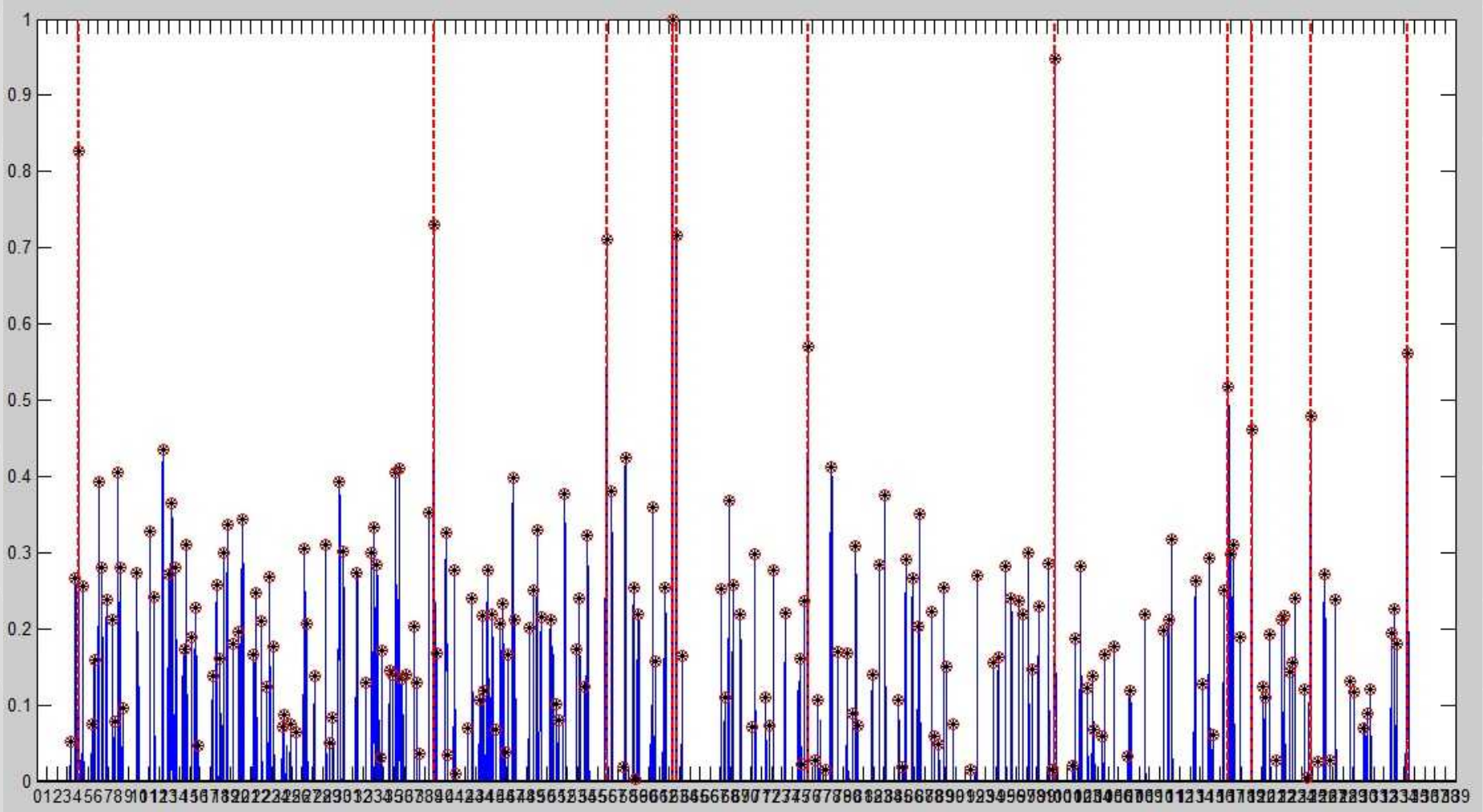
Detection of bilabial consonants in audio: Testing

- Calculate MFCC’s for test case
- Calculate probability of MFCC features belonging to GMM1 and to GMM2
- Subtract the two probabilities and find peaks in the graph of the difference
- Maintain a minimum threshold, and note down time instances of peaks

Detection of bilabial consonants in audio: Testing



Detection of bilabial consonants in audio: Testing



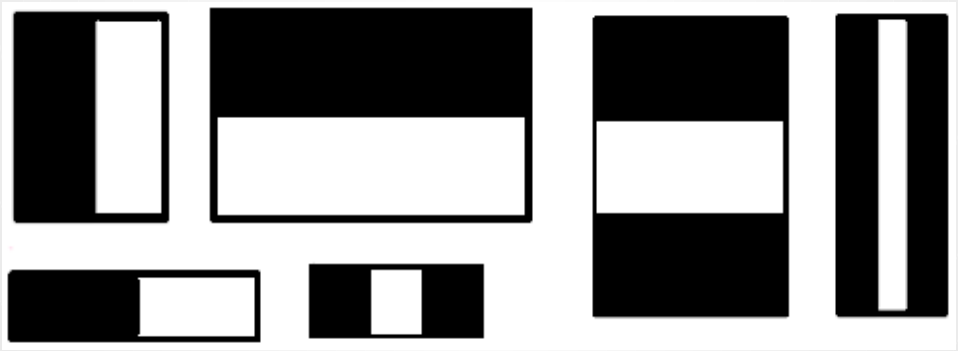
Detection of bilabials in video

- Detect Face using Viola-Jones Algorithm
- Detect Mouth in Face using Viola-Jones Algorithm
- Filter Mouth with suitable colour range to make binary image of Lip region
- Choose a vertical column 3 pixels wide, and record changes in pixel values down the column
- Changes in pixel values indicate lip closure or non-closure

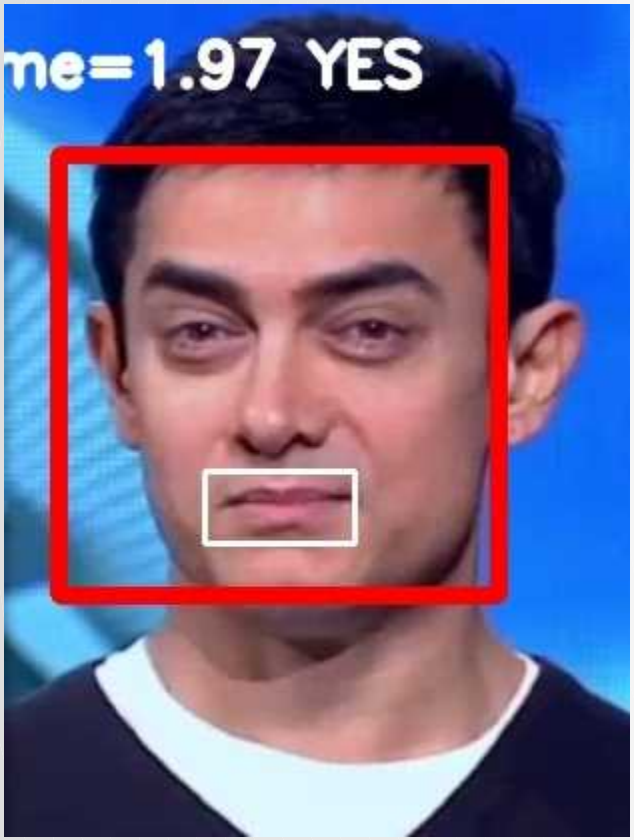
Detection of bilabials in video

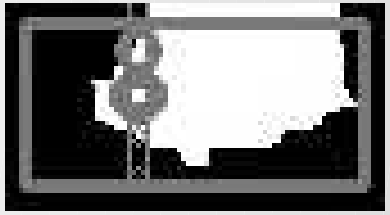
Viola-Jones algorithm to detect Face and Mouth

1. Detect Haar Features
2. Check if Haar feature is present
3. Use AdaBoost Algorithm

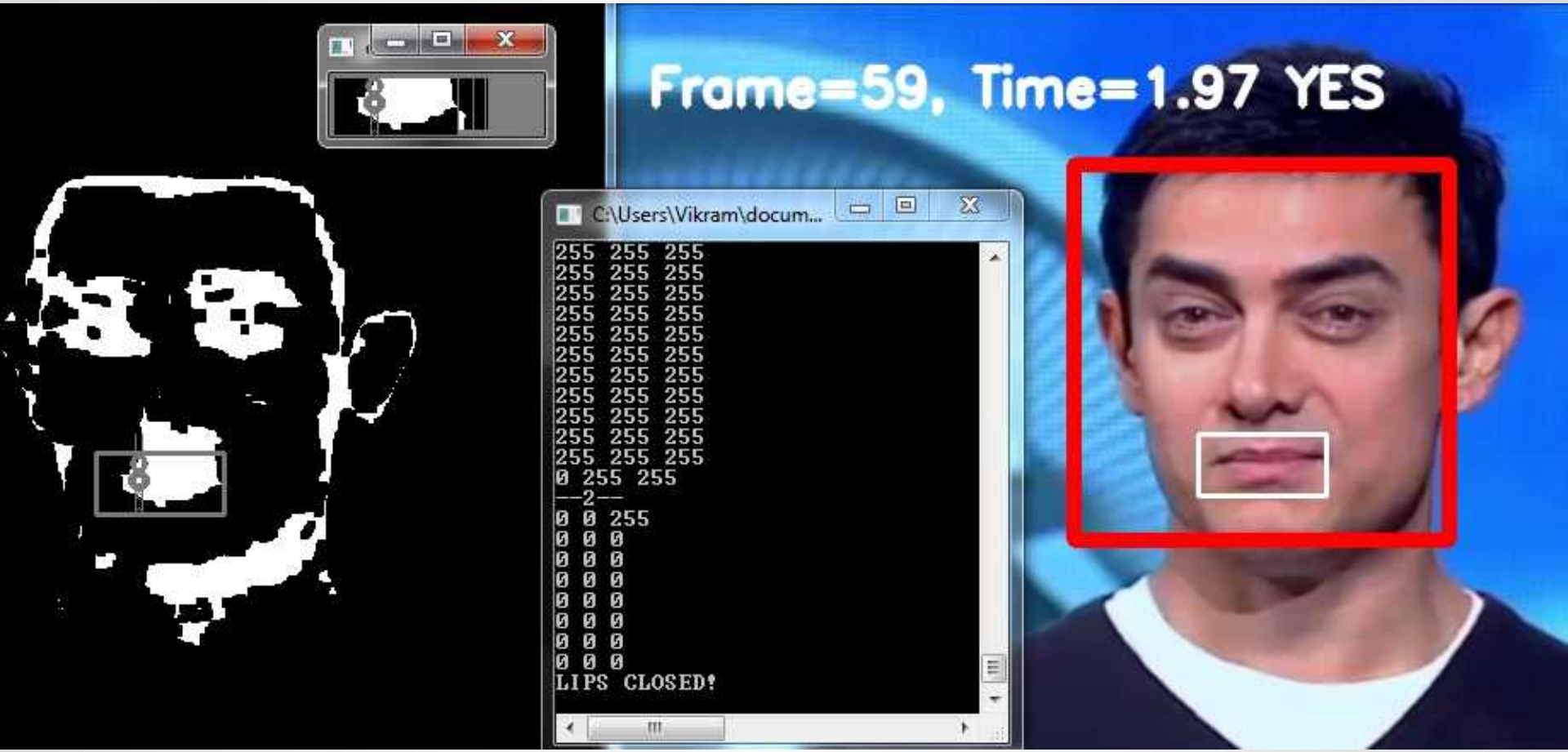


Detection of bilabials in video

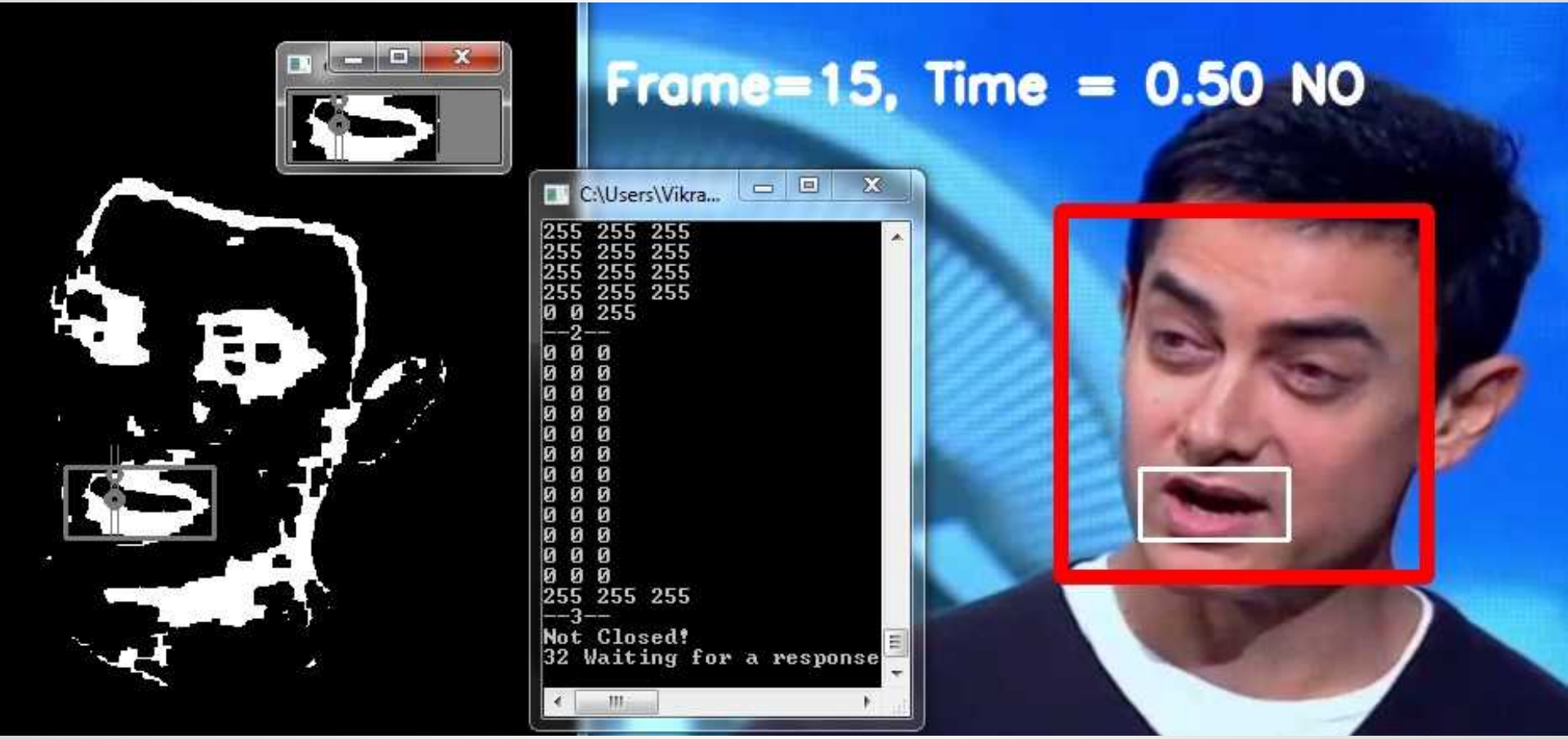




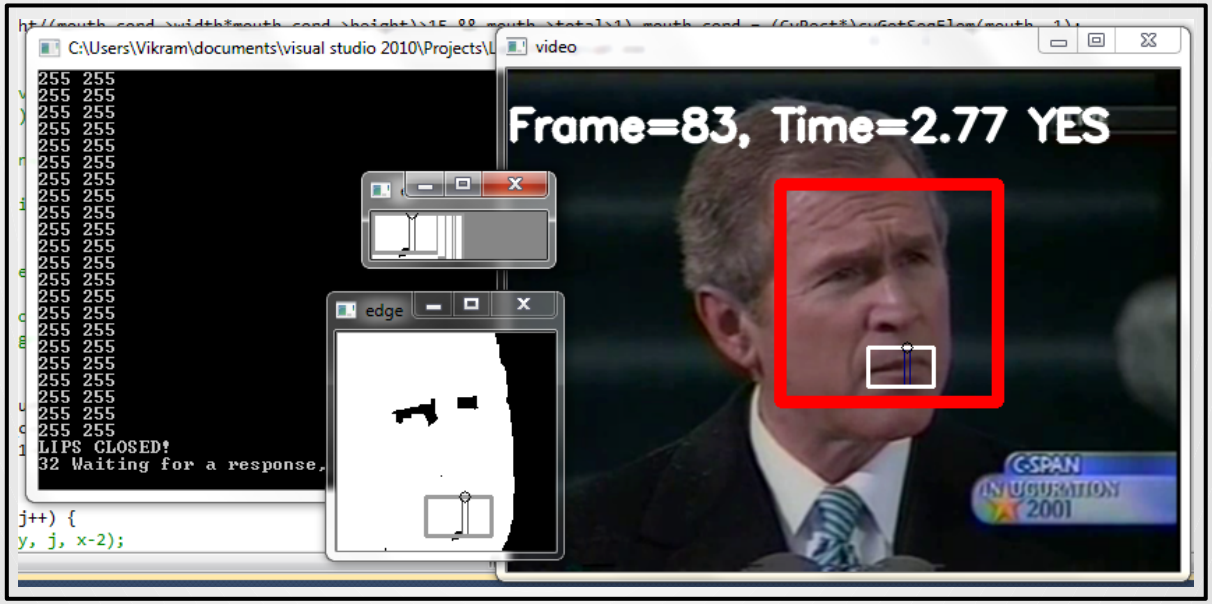
Detection of bilabials in video



Detection of bilabials in video



Detection of bilabials in video



Pro's & Con's:

Detection of bilabials in audio

- Successfully uses MFCCs to compute a probability with good success rate.
- Is very flexible; parameters of the GMM can be easily changed to better fit the model

- This method can only provide a probabilistic determination of bilabials, rather than a classification.
- There can be many correct sets of parameters for the GMM to fit the model.
- Requires a lot of training data.

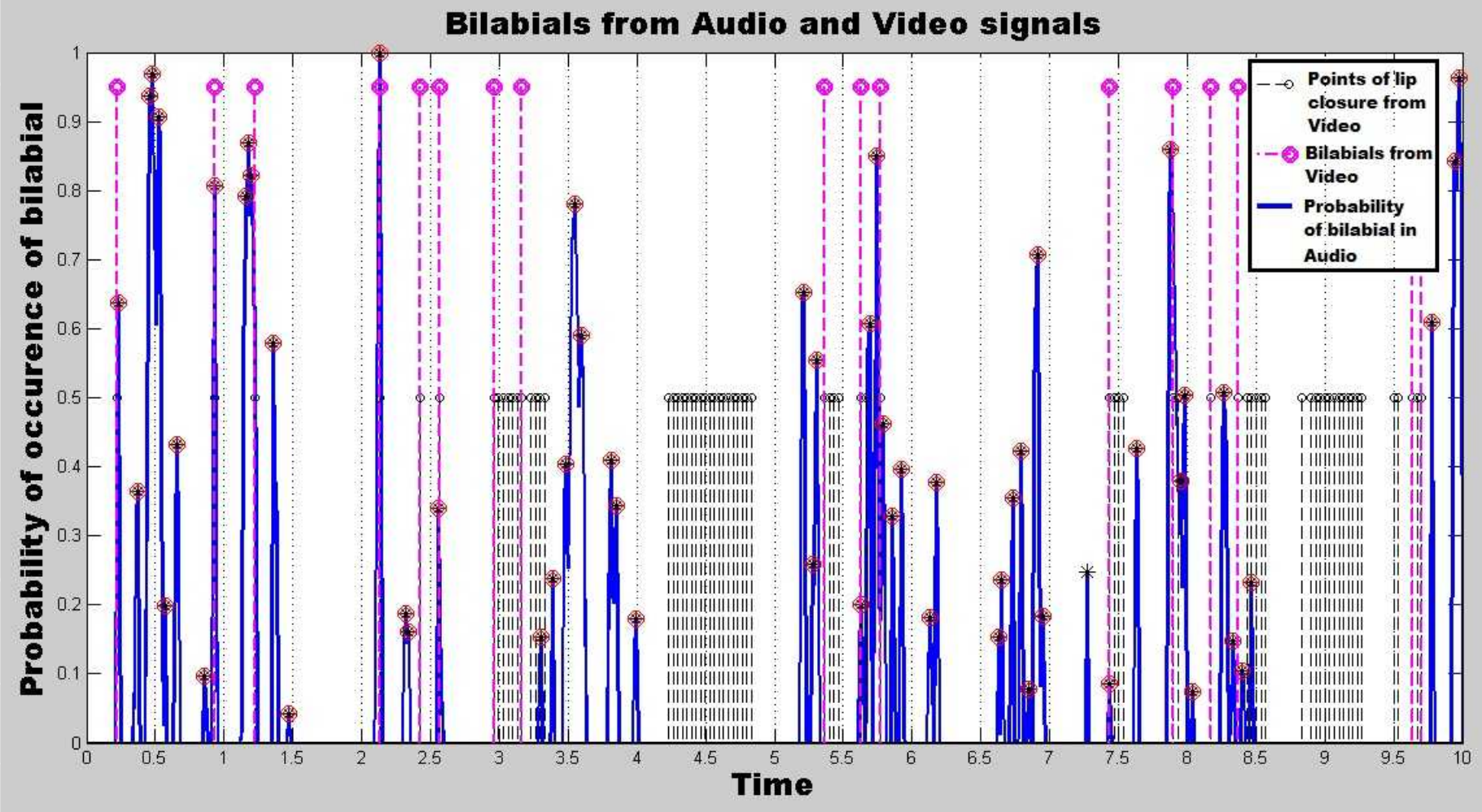
Pro's & Con's:

Detection of bilabials in video

- Checking for lip-closure in video is more error-free than checking for bilabials in audio.
- This method can work for videos with low quality, as are very often found in online repositories such as YouTube.
- No training required.
- Can be easily implemented in OpenCV rather than MATLAB for faster results.

- Cannot detect face and lip regions if head is tilted beyond a certain angle (only in rare cases).
- Colour range chosen to detect lips may need to be dynamically varied depending on the gender and ethnicity of the speaker
- Frame rate of the video affects the accuracy of the time at which lip closure is detected.

Detection of bilabials in video & audio



Vowel Detection

- The speech sample is zero-centred and normalized.
- The Power Spectrum of the normalized zero-centred speech sample is determined using the Yule-Walker method.
- The frequencies where the first three peaks occur in the Power Spectrum are noted as the Formants of the speech signal.
- The three frequencies, combined with their respective weights, are compared to pre-determined formant frequency values for the different vowel sounds.
- If the error between the formant frequencies of the speech sample and any one of the vowels is less than a threshold value, then it is claimed that the speech sample in consideration contains a vowel.
- This method was judged redundant after experimentation.

Vowel Detection

Vowel Symbol	Example Word	F1 (Hz)	F2 (Hz)	F3 (Hz)
/ow/	Coat	570	840	2410
/oo/	Loot	300	870	2240
/u/	Put	440	1020	2240
/a/	Hot	730	1090	2440
/uh/	Cut	520	1190	2390
/e(r)/	Bird	490	1350	1690
/ae/	Bat	660	1720	2410
/e/	Bet	530	1840	2480
/i/	Bit	390	1990	2550
/iy/	Beat	270	2290	3010

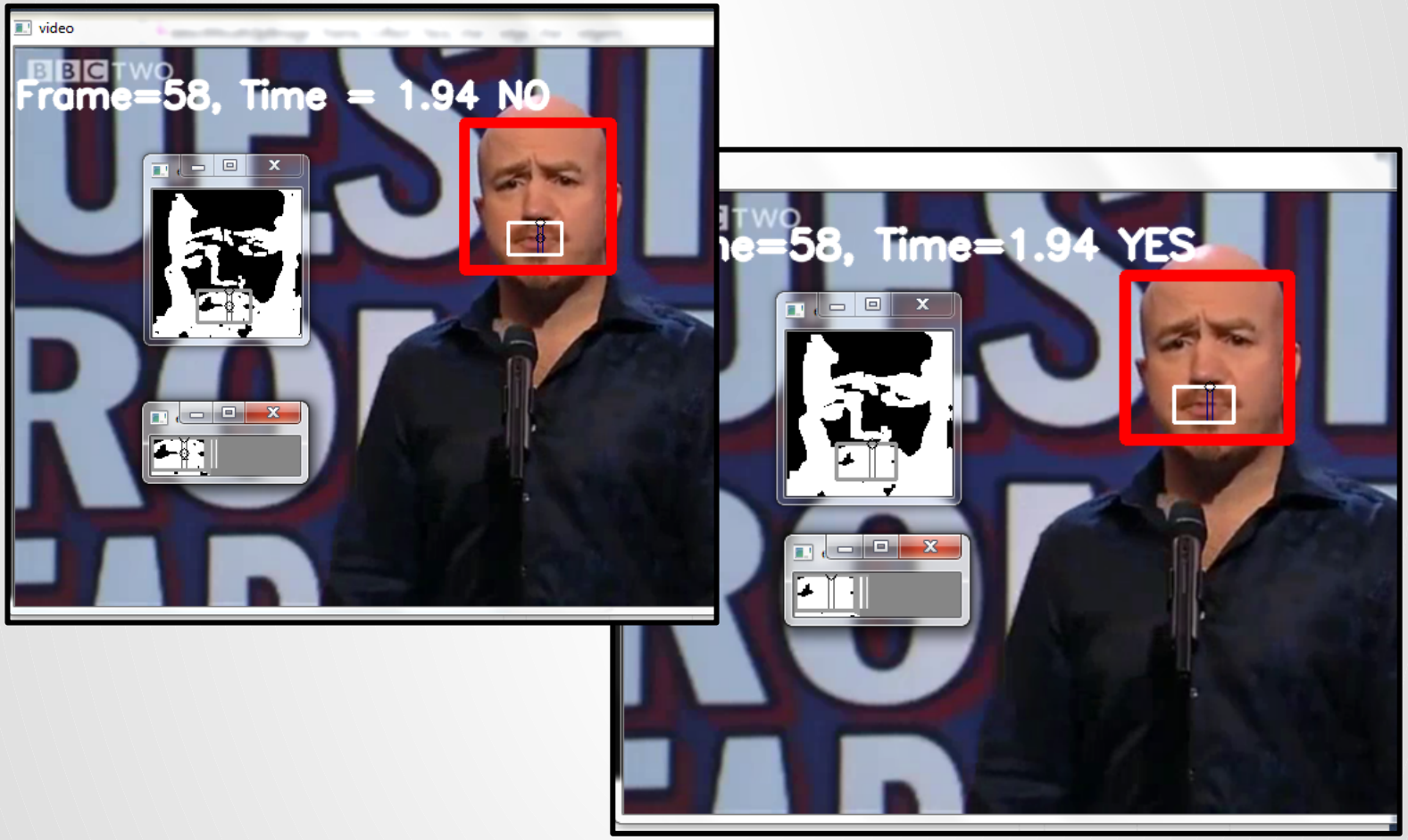
Errors in Lip Closure Detection in Video

- **1. Error due to Viola – Jones Algorithm used to detect face and mouth**

1) Limitation to head tilt
2) Experimental threshold values of Haar Classifiers

Errors in Lip Closure Detection in Video

- **2. Error due to Colour Filtering**



Errors in Lip Closure Detection in Video

- **3. Error due to Frame Rate**

Frame rate of a typical video is limited to 29fps.

This implies a gap of around 35msecs between two frames.

Thus, lip closure is detected with an error margin of +35msecs to -35msecs.

Errors in Bilabial Detection in Audio

- **1. Error due to MFCC**

  Only an approximate model of human auditory response

- **2. Error in Training Data**

  1) inclusion of non-bilabial points into the GMM,

  2) exclusion of bilabial points from the GMM.

Errors in Bilabial Detection in Audio

- **3. Error in the Gaussian Mixture Model**

1) Number of Iterations if the Expectation Maximisation algorithm

- Increase in number of iterations can lead to a more accurate model, but would take more time to train

2) Over-fitting

- ‘ma’ occurs more frequently than ‘pa’ or ‘ba’
- More number of training points within a small area could over-work the training algorithm

3) Number of Gaussians in the Mixture

- More the number of Gaussians, more accurate is the fit, but the training is slower

Conclusion & Future Work

- Time offset can be detected for audio samples of highest probability of bilabial consonant, and a lip closure upto an error margin of 35msecs

- Better training data for GMM

- Inclusion of other algorithmic checks to reduce errors in lip detection, such as Scale-invariant feature transform (or SIFT), and Principal Component Analysis

- Making of a synthetic video as a test case for the program

Thank you.