Audio-Visual Isolated Digit Recognition for Whispered Speech

Xing Fan, Carlos Busso and John H.L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering & Computer Science
Department of Electrical Engineering
University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.

EUSIPCO 2011

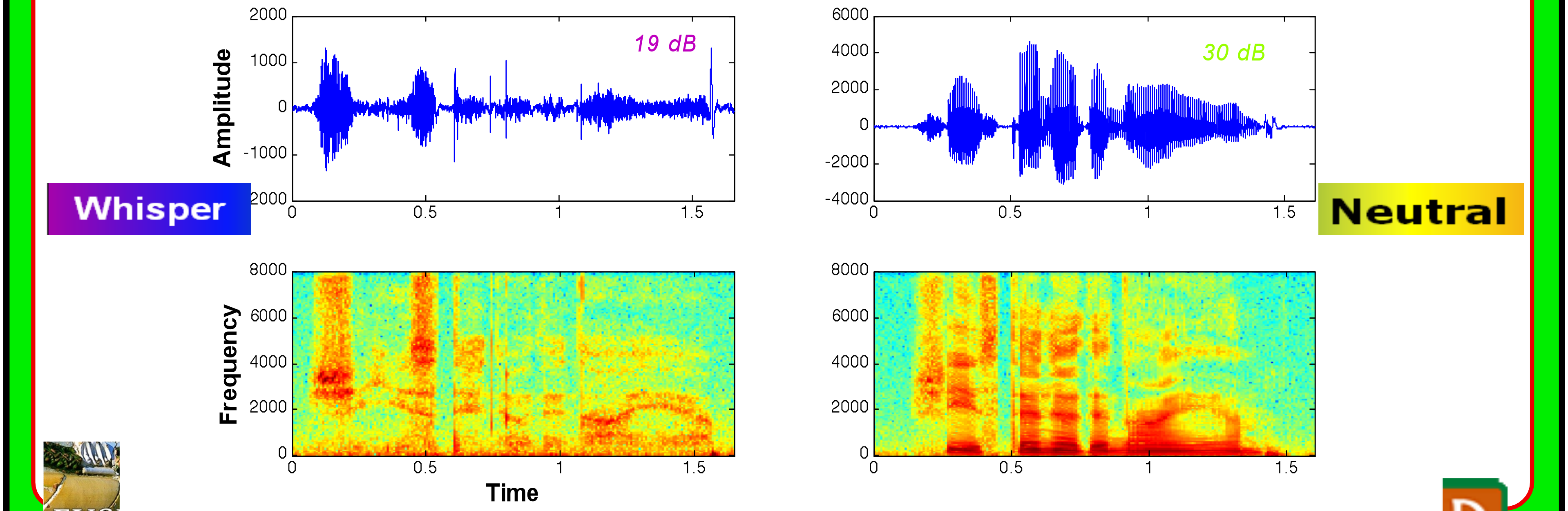
August 29 - September 2, 2011 Barcelona, Spain

Difference between Whispered and Neutral Speech

- Absence of periodic excitation
- Formant shifting in F1, F2
- Low volume and longer duration

Acoustic Domain

Significant degradation in performance for ASR system trained with neutral speech



Corpus Setup

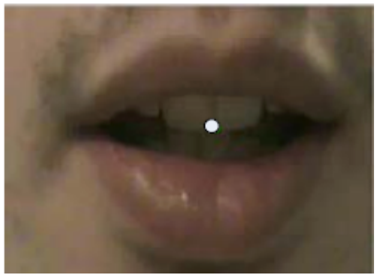
WHISPER-NEUTRAL Audio-Lip DATA

- 1 male native American English speaking subject
- 300 digits numbers from 0-9 are randomly ordered and read by the subject in both whispered and neutral mode.
- Both audio and visual information are recorded by a camera.
- The video is of the size 720*576, captures in color at a rate of 25 frames/sec
- The audio is collected with the video at a sample rate of 44.1 kHz with microphone in the camera ~70cm from the subject, and down-sampled to 16 KHz

Feature Extraction

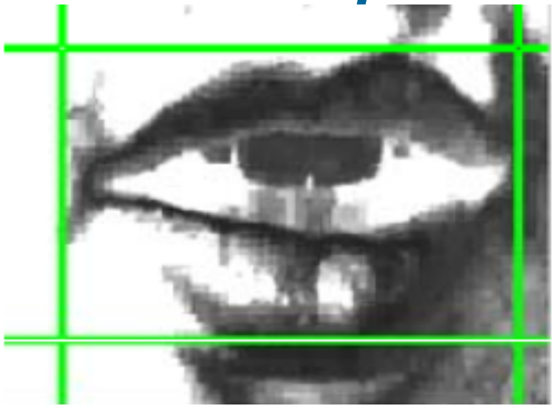
Pre-processing (extraction of ROI)

Step1 : Using gravity center detection to find the center of the lips



Step2 : Using RGB color vector to detect the lip boundary

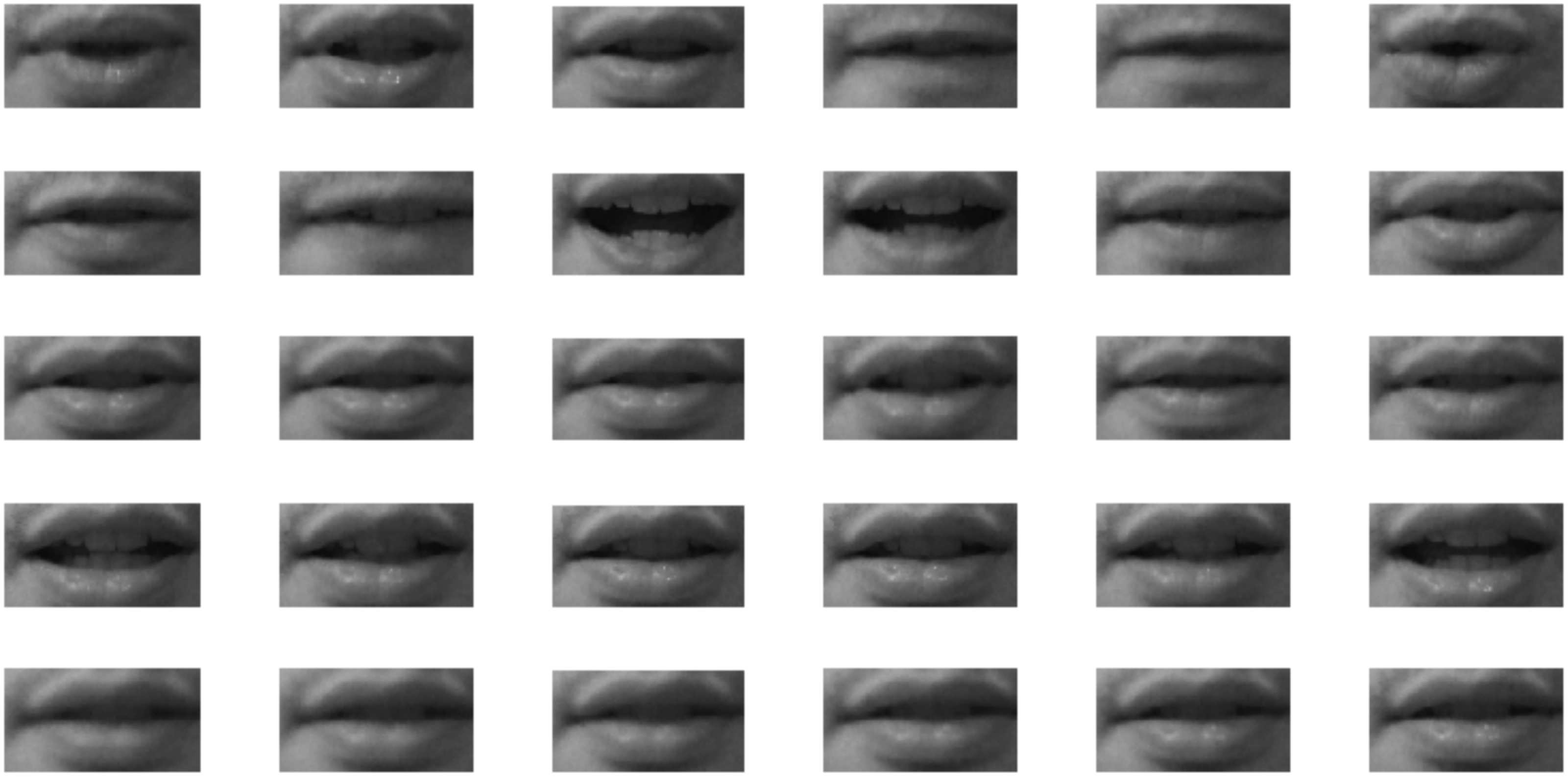
- Use developing lip sample to train a Lip color GMM
- For each input image, each pixel will test the GMM, the boundary will be determined by the output score matrix



EUSIPCO 2011

Feature Extraction

Eigenlips: A total number of 400 lip samples are employed for calculating the eigenlips



Feature Extraction

Steps for calculating the eigenlips

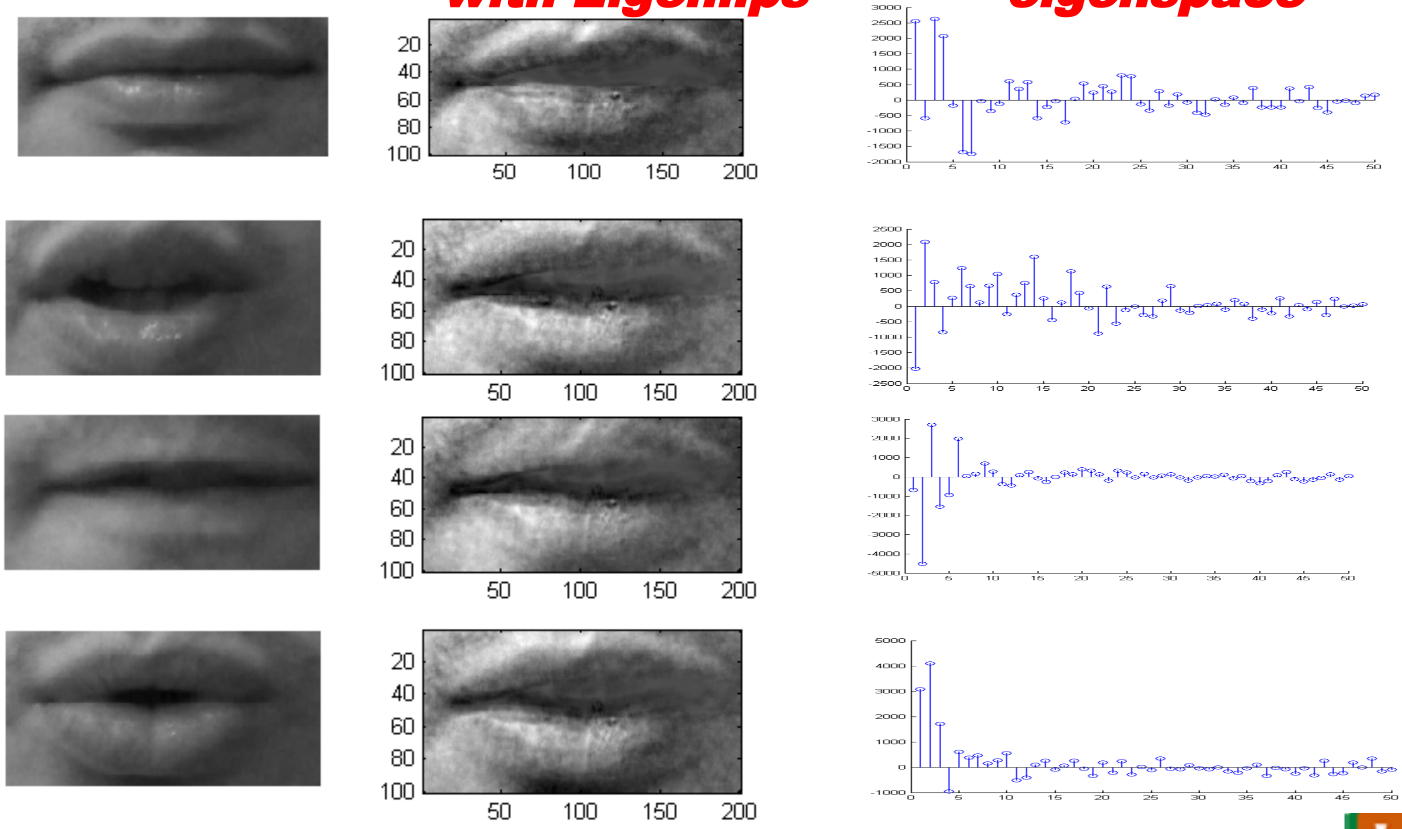
- Calculate the mean lip image structure
- Normalizing of each image
- Calculate the covariance matrix C
- Obtain the eigenvectors and eigenvalues of C
- Choose and sort the eigenvalues and find the corresponding eigenvectors

→ *Eigenvectors with top 15 largest eigenvalues are chosen as eigenlips*

Feature Extraction

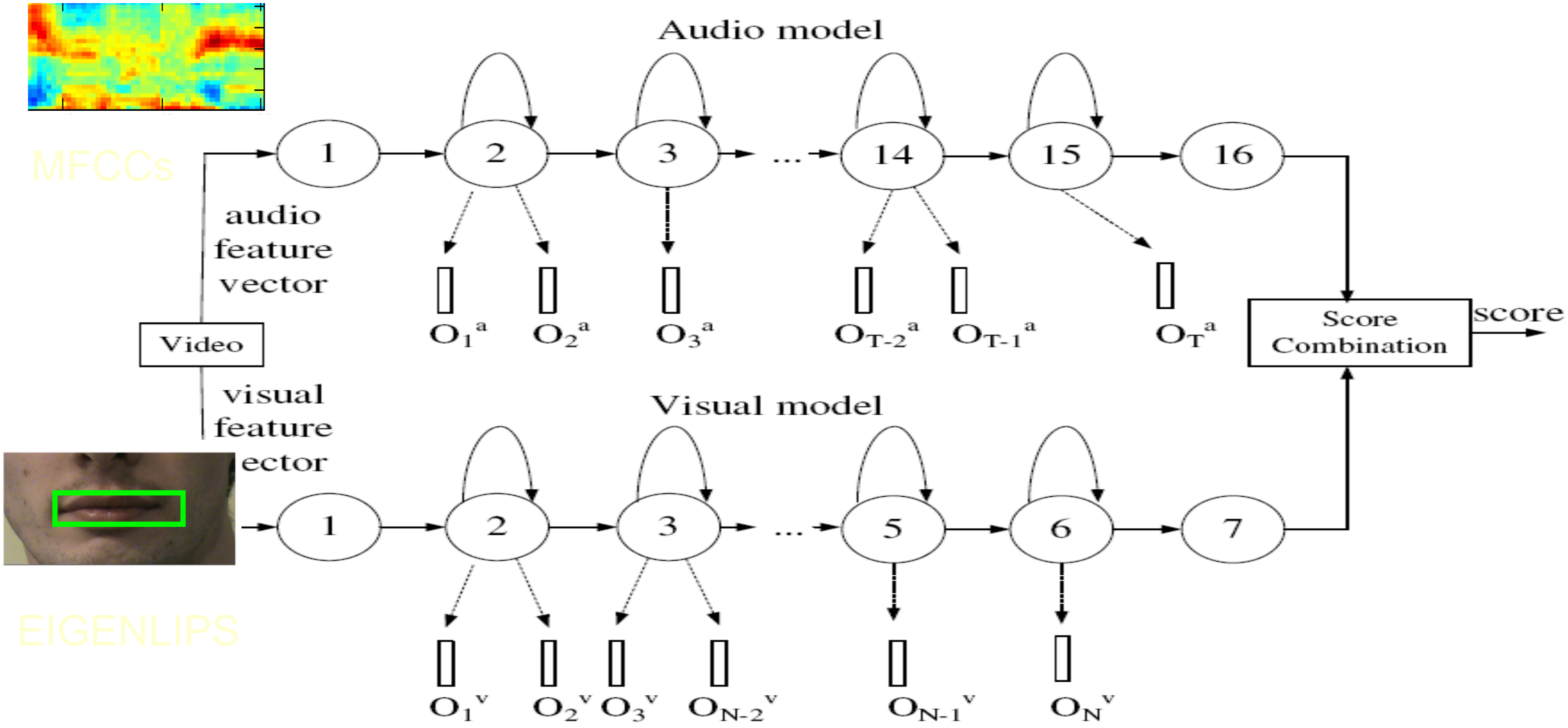
Mean Lip

Original Lips | **Lip Reconstruction with Eigenlips** | **Projection on to the eigenspace**



System Description

HMM Structure used in this study



System Description

Audio and Video HMMs are trained independent in this study.

Audio:

- Training feature: MFCC_0_D_A (39d)
- State num: 16 states (2 non-emitting state)
- Mixture number: 1

Video:

- Training feature: eigenlips_D (30d)
- State num: 7 states (2 non-emitting state)
- Mixture number: 1

System Description

Decoding:

- Synchrony is constrained on the boundary of each word

- Scores are log-linearly combined

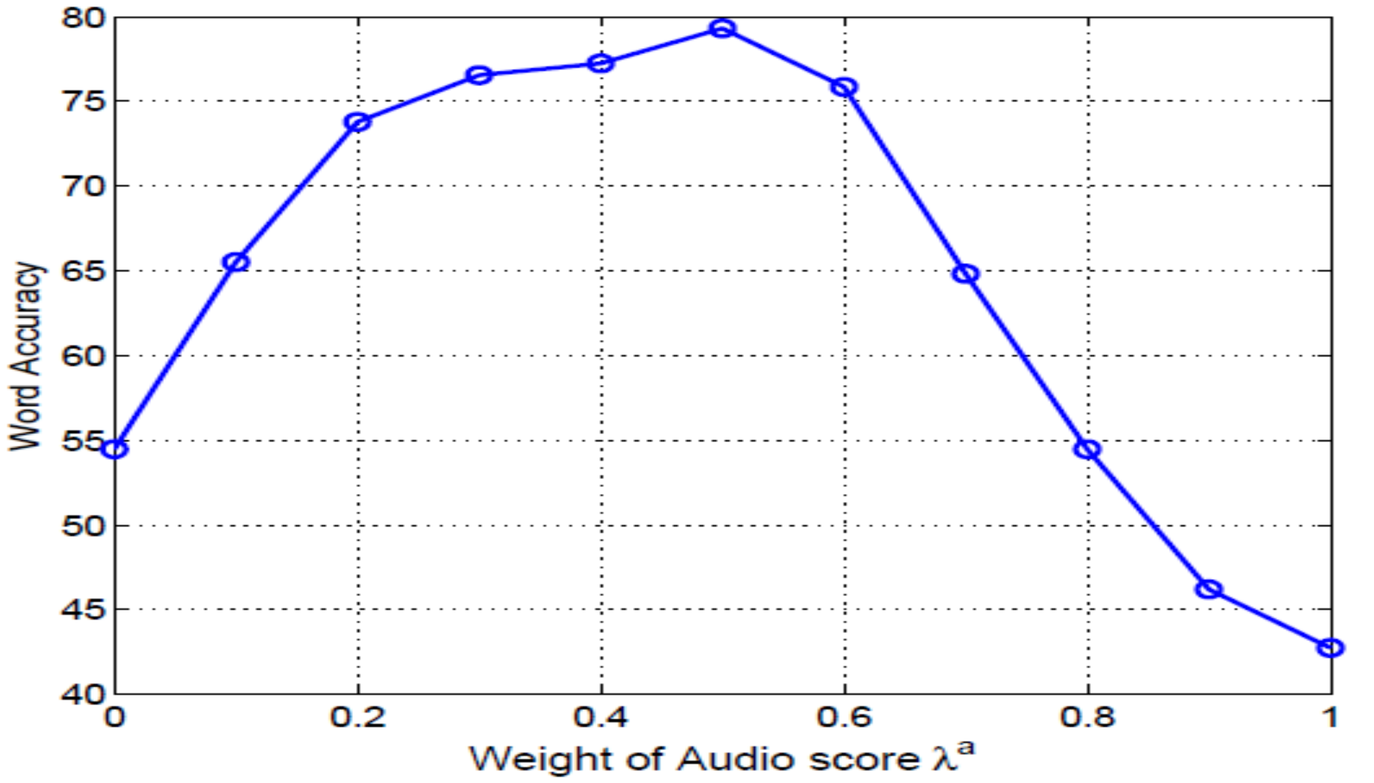
$$P(O|M) = \{\sum_{X} a^{a}_{x(0)x(1)} \prod_{t=1}^{T} b^{a}_{x(t)}(o_t) a^{a}_{x(t)x(t+1)}\}^{\lambda^{a}} +$$

$$\{\sum_{X} a^{v}_{x(0)x(1)} \prod_{t=1}^{T} b^{v}_{x(t)}(o_t) a^{v}_{x(t)x(t+1)}\}^{\lambda^{v}}$$

$$\lambda^{a} + \lambda^{b} = 1.$$

Experimental Results

Adjusting the value of the weight of the audio score versus lip score, we have word accuracy: (training data only contains neutral sp



EUSIPCO 2011

Experimental Results

Overall Word accuracy:

Stream	*training*	*test*	*Word Accuracy(%)*
Audio data	neutral	neutral	98.7
Audio data	whisper	whisper	83.3
Audio data	neutral	whisper	42.7
Video data	neutral	neutral	70.7
Video data	whisper	whisper	68.0
Video data	neutral	whisper	54.7
combined	neutral	whisper	79.7

-56%

+38%

- Audio based system achieves good baseline 98.7%
- Testing with whisper audio → significant ASR performance loss 56%
- Combine Audio-Visual improved performance to 79.7%

Conclusion

- A small digit corpus is developed for an exploratory study of audio-visual speech recognition for whispered speech.
- An eigenlip based feature extraction method is applied for visual data
- Multistream framework is built using audio and video stream HMMs
- Significant improvement in word accuracy is presented using this multi-stream model system