Audio-Visual Integration in Multimodal Communication by CHEN, RAO

Presented by
Tuneesh k Lella

Agenda

- Introduction
- Bimodality of human speech
- Lip Reading
- Speech-driven face animation
- Lip Synchronization
- Lip Tracking
- Audio-to Visual Mapping
- Bimodal person verification
- Conclusions

Introduction

- Traditional Information Processing techniques focus on one media type-text or audio or video
- Interaction between audio and video is the most interesting
- Audio-Visual integration aids in
  - Automatic Lip reading
  - Lip synchronization
  - Joint audio-video coding
  - Bimodal person authentication

2. Bimodality of Human Speech

- McGurk Effect demonstrates bimodality of speech perception

Audio +	Visual →	Perceived
ba	ga	da
pa	ga	ta
ma	ga	na

- Reverse McGurk Effect also exists
- Speech production is also bimodal

Viseme

- Basic unit of mouth movements (like phoneme for speech)
- Many-to-one mapping between phonemes and Visemes
- Viseme groups obtained by analyzing confusions in stimulus response matrices

Viseme Groups for English Consonants

1	**f, v**
2	**th, dh**
3	**s, z**
4	**sh, zh**
5	**p, b, m**
6	**w**
7	**r**
8	**g, k, n, t, d, y**
9	**l**

Viseme contd..

- Subject asked to identify syllables visually
 (C-V-C words)
- Viseme groups are identified as those clusters of phonemes in which at least 75% of all responses occur within the cluster
- Fisher's observations
  - Viseme groupings for Initial and final consonants differed
  - Confusions between consonants in a viseme class could be directional

3. Lip Reading (Speech Reading)

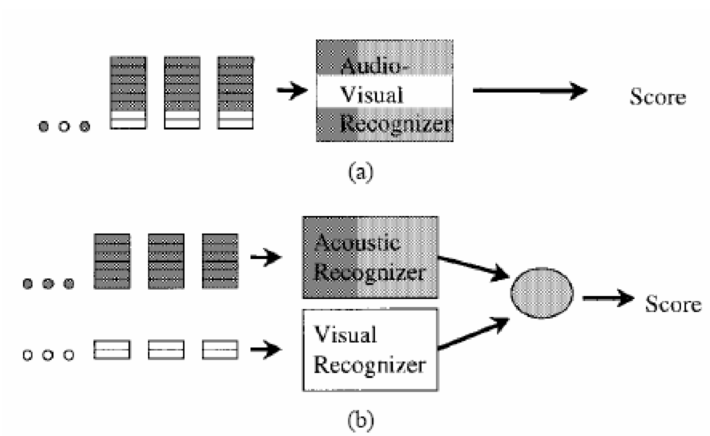
- Human Lip Reading
- Automated Lip Reading

Human Lip Reading

- Infers the meaning of spoken sentences by looking at the configurations and motion of visible articulators of speech
- Useful in situations like cocktail party
- Recognition of audio-visual cues degrades less rapidly than acoustic cues alone
- Lip reading performance affected by
  - Viewing conditions
  - Coarticulation (Berger)

Human Lip Reading contd..

- Frame rate importance with impaired listeners was studied by Frowein et. al.
  - 15Hz frame rate is necessary for speech understanding
- Effects of frame rates on Isolated viseme recognition were observed by williams et.al.
  - At different frame rates, viseme groupings were different
  - Minimum frame-rate for continuous speech greater than 5 Hz
- Early and Late Integration



Automated Lip Reading (ALR)

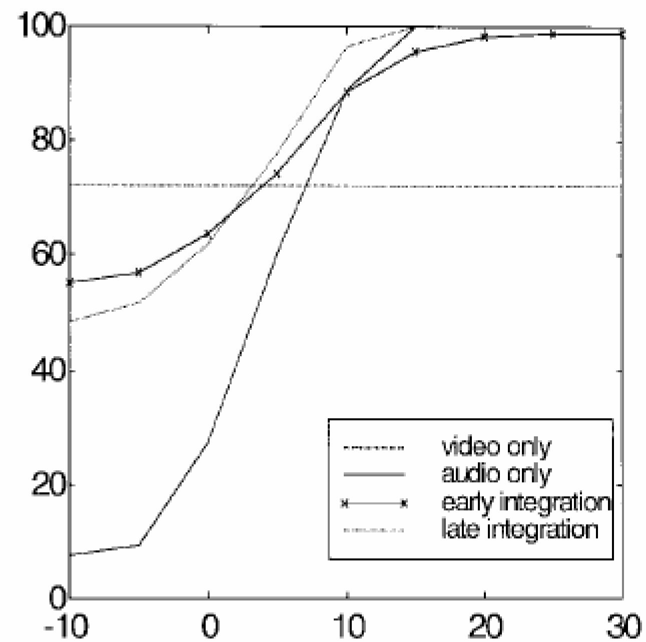
- No clear consensus on optimal audio-visual recognizer
- Petajan developed one of the 1st audio-visual recognition systems
  - Binary mouth images are analyzed to derive the mouth open area, the perimeter, the height, and the width
  - Audio and visual speech recognizers in serial fashion
- Dynamics of visual feature set also useful for speech recognition (Goldschen; Mase & Pentland)
- Physical dimensions of mouth can provide good recognition performance (Finn & Montgomery)

ALR Contd..

- Yuhas et.al. used neural networks for the fusion
  - Pixel values of mouth are fed to multilayer network directly
  - Estimated acoustic spectrum combined with true spectrum
- Stork et.al. used time delayed neural networks (TDNN)
  - coarticulation was considered
  - Early and late integrations were used
  - Late integration was better and could replicate McGurk effect
- Many other researchers also found that audio-visual recognizers clearly dominates either audio or visual recognizers used alone

Experiment

- Experiments done with isolated word recognizer using audio visual data (zero to nine)
  - 4 HMMs- one each for visual information, acoustic information, early integration, late integration
  - Integrations performed worse than visual-only info in high noise environments



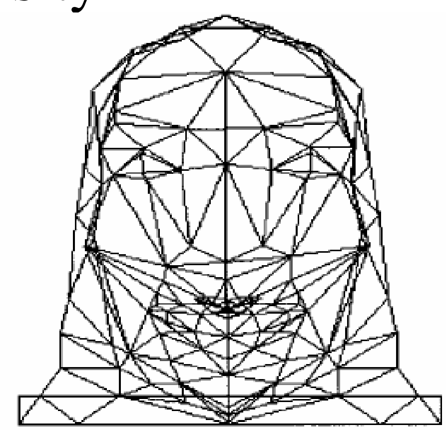
Results of joint audio-visual speech recognition.

4. Speech-Driven Face Animation

- Visual speech from auditory speech
- Two approaches to generate talking-head images are
  - Flipbook method
  - Wireframe model (2-D or 3-D approach)
- Flipbook method
  - a number of mouth images of a person, called key frames, are captured and stored
  - according to the speech signal, the corresponding mouth images are “flipped” one by one to the display to form animation.
  - Less computationally intensive, requires more data

Speech-Driven Face Animation

- Wireframe model
  - Composed of a large number of triangular patches
  - Vertices can be manipulated to synthesize new expressions (FACS)
  - Must be combined with lighting models that specify how to map shape and position of wireframe into intensity
  - Texture is necessary for More realism
  - Computationally intensive, flexible, less data required



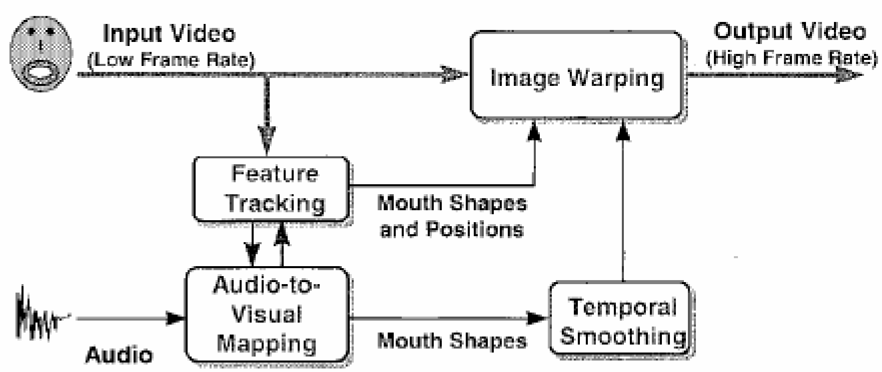
The wireframe model "Candide."

How to make Talking-heads *"say"* sentences?

- Morishima et.al. used 3-D wireframe model to synthesize lip motion
  - Lip parameters form a 8-D vector and extracted from text or speech
  - In speech input, LPC Cepstra are vector quantized and centroids of corresponding lip-feature vectors were computed, used for classifying the input speech
- HMM based technique was used by Rabiner & Juang
- Some others used TDNN based approaches

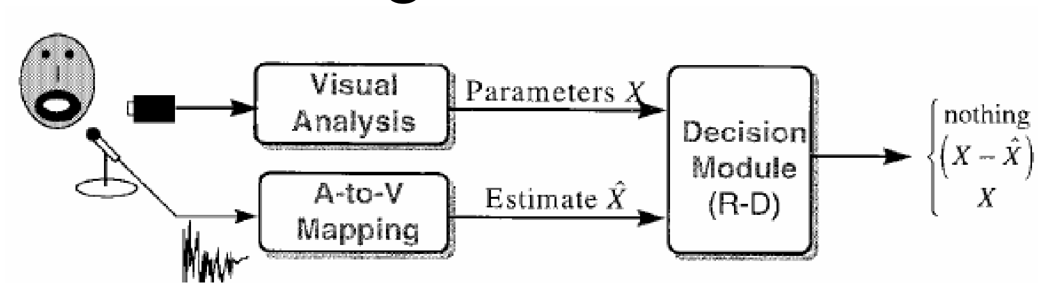
5. Lip Synchronization

- One of the most important issues in video telephony & conferencing
- What to do if Frame rate is not adequate for lip sync perception?
  - Warp the acoustic signal to make it sound synchronized with the person's mouth movement
    e.g.-dubbing in movie production
  - Time-warp the video



Lip Synchronization contd..

- Transmission also affects lip sync
- Delay more for video than audio
  - Solve this by warping the mouth image of speaker to be in sync with the audio
  - We can embed speech interpolation into video codec
- Can be useful in dubbing of foreign movies, cartoon animation etc.
- Cross modal predictive coding



6. Lip Tracking

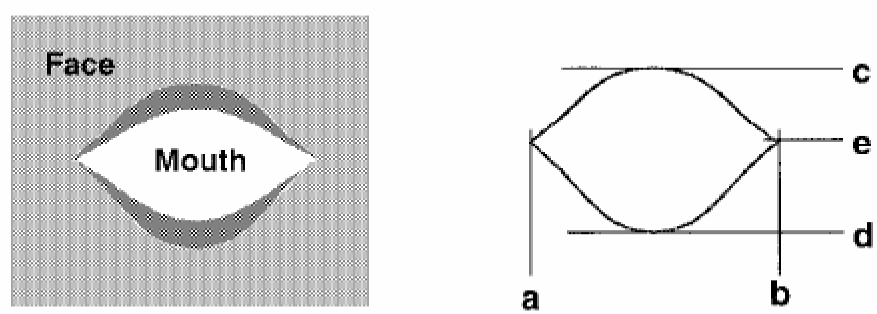
- Visual input is a 3-D video signal with 2 spatial and 1 temporal dimensions
- Visual analysis systems divided into 2 major classes
  - Viseme grouping (VQ & neural networks)
  - Parameter measurement from input image
- We can measure the height between lips and width between corners of the mouth for parameter measurement
- Based on deformable models

Deformable Models

- Deformable templates and snakes
- Basic idea- energy function that relates a parameterized model to an image is formed
- Energy function is minimized and parameter set is obtained
- Snakes model
  - Energy functions in snakes keeps contour smooth and find key features such as edges
  - Can constrain position of snakes to a smaller subspace by Eigen decomposition

Deformable Models contd..

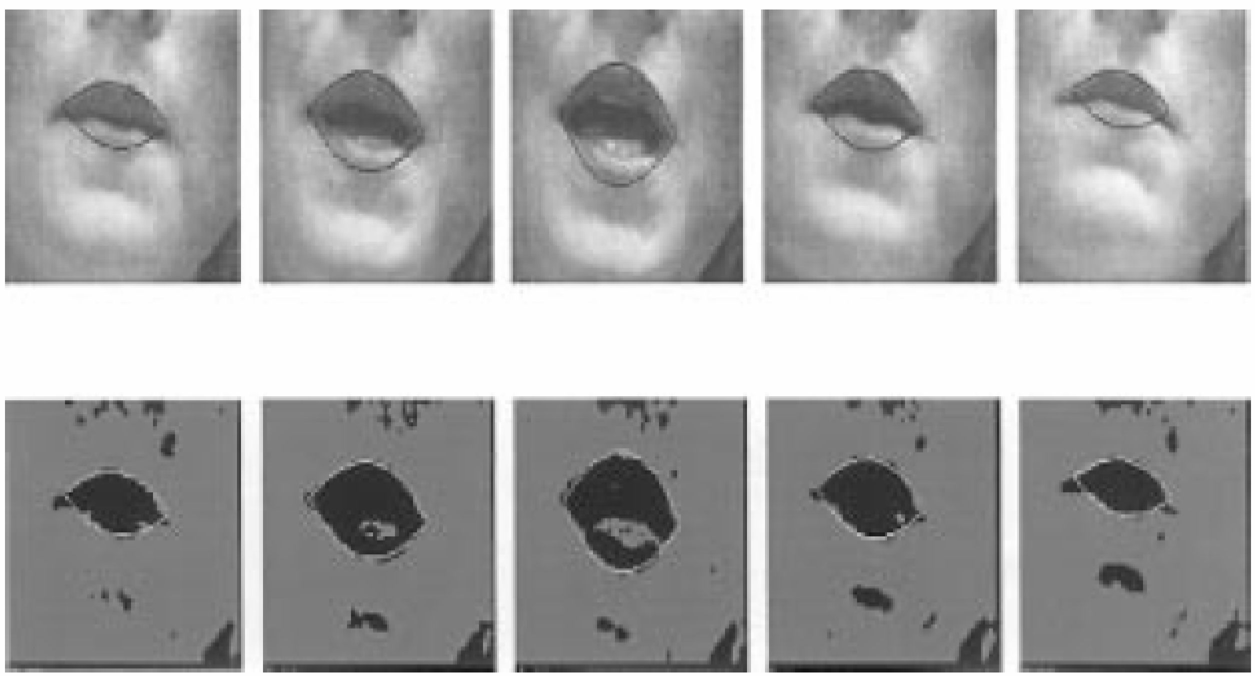
- Deformable Templates
  - Provides both a parameterized model and an energy function
  - More the complexity of model, more the number of parameters
  - Energy function associated with template relates the template to the image
- State-embedded Deformable Templates



(a) (b)

State-embedded deformable templates

- Tracking algorithm results

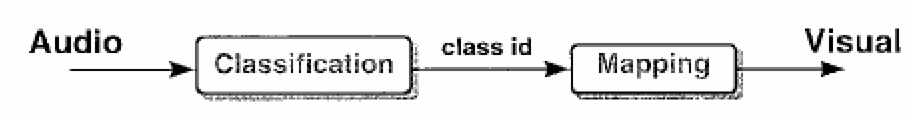


7. Audio-to-Visual (A-V) Mapping

- Acoustic speech to mouth shape parameters
- Can be done from two perspectives
  - Speech as linguistic entity
    - Complete speech recognizer followed by a lookup table
    - computationally intensive
  - Speech as physical phenomenon
    - Functional relationship may exist between speech parameters and visual parameter set
    - Many approaches to perform this task

Different Approaches to A-V mapping

- Classification-based conversion
  - VQ to classify the acoustics
  - Mapping each acoustic class to corresponding visual codewords and averaging them to get visual centroid
  - Averaging results in errors



- Neural networks
  - I/p and o/p patterns presented to the network and Back propagation to train the network weights

Approaches contd..

- Direct Estimation
  - Best estimate of visual parameters derived directly from joint statistics of audio and visual parameters
  - Consider case of 1-D visual parameter and if we can model the joint pdf as Gaussian mixture with k gaussian functions

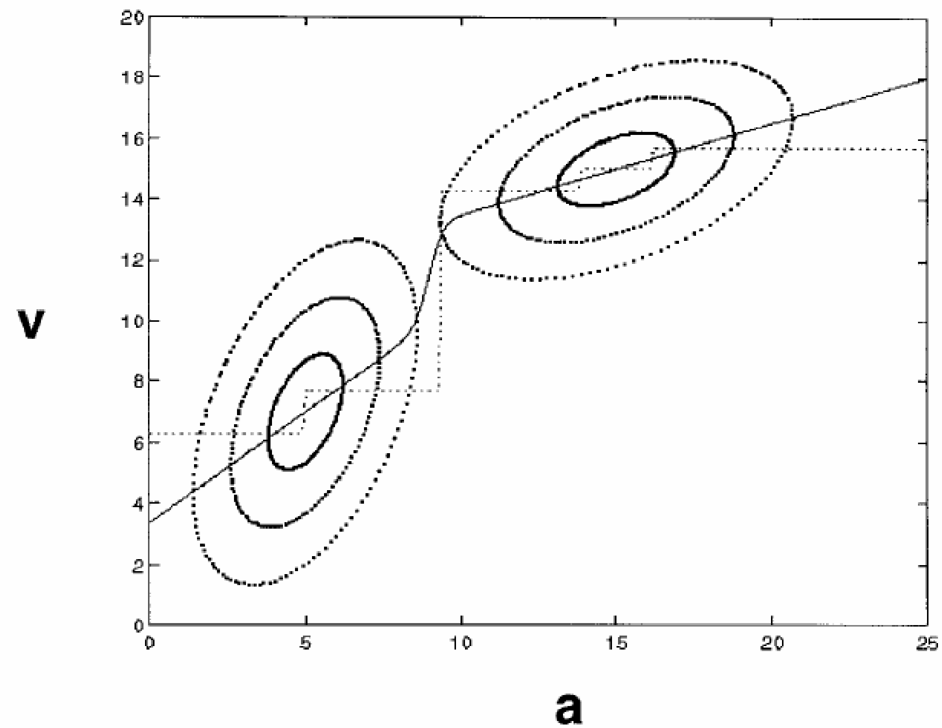
$$f_{av}(\mathbf{a}, v) = \sum_{i=1}^{K} c_i \aleph(\mu_i, \mathrm{R}_i) \qquad \mu = \begin{bmatrix} \mu_{\mathbf{a}} \\ \mu_v \end{bmatrix}, \mathrm{R} = \begin{bmatrix} \mathrm{R}_{\mathbf{a}} & \mathrm{R}_{\mathbf{a}v} \\ \mathrm{R}_{\mathbf{a}v}^T & \sigma_v^2 \end{bmatrix}$$

$$\hat{v} = E\langle v|\mathbf{a}\rangle = \int v \frac{f_{\mathbf{a}v}(\mathbf{a}\, v)}{f_{\mathbf{a}}(\mathbf{a})} dv. \qquad \hat{v} = \sum_{i=1}^{K} \frac{c_i \aleph(\mu_{\mathbf{i},j}, \mathrm{R}_{\mathbf{a},i})|_{\mathbf{a}}}{f_{\mathbf{a}}(\mathbf{a})} \mathbf{b}_i^T \begin{bmatrix} 1 \\ \mathbf{a} \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 1 & \mu_{\mathbf{a}}^T \\ \mu_{\mathbf{a}} & \mathrm{R}_{\mathbf{a}} \end{bmatrix}^{-1} \begin{bmatrix} \mu_v \\ \mathrm{R}_{\mathbf{a}v} \end{bmatrix}.$$

  - Hence, gaussian mixture component yields an optimal linear estimate for v given a. The estimates are non-linearly weighted by $c_i \aleph(\mu_{\mathbf{a},i}, \mathrm{R}_{\mathbf{a},i})|_{\mathbf{a}} / f_{\mathbf{a}}(\mathbf{a})$ to produce final estimate

Direct estimation contd..

- Direct estimation better than classification based method



HMM approach

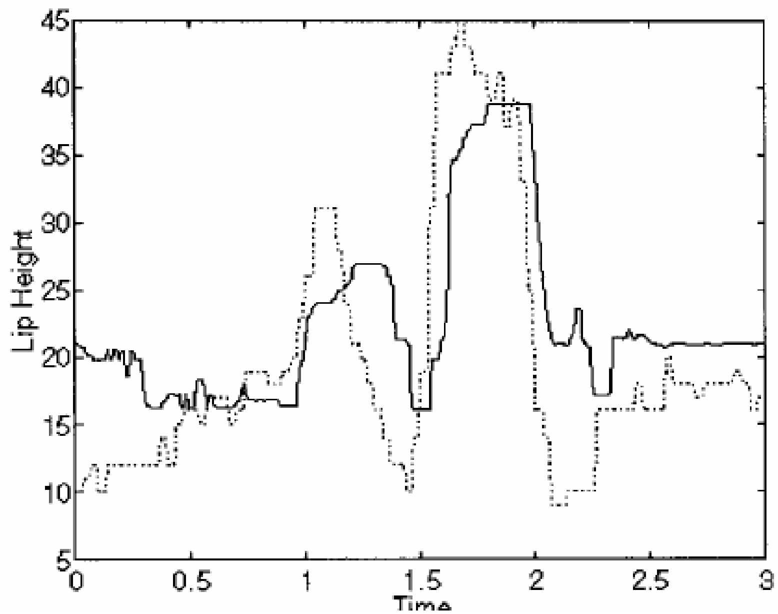
- Let $\mathbf{o} = [\mathbf{a}^T, v]^T$ denote the joint audio-visual parameter
- Process
  - Training
    - N-state left-right HMM on sequence of observations O for each word in vocabulary. This gives A, B and π
    - Extract an acoustic HMM by integrating over the visual parameter

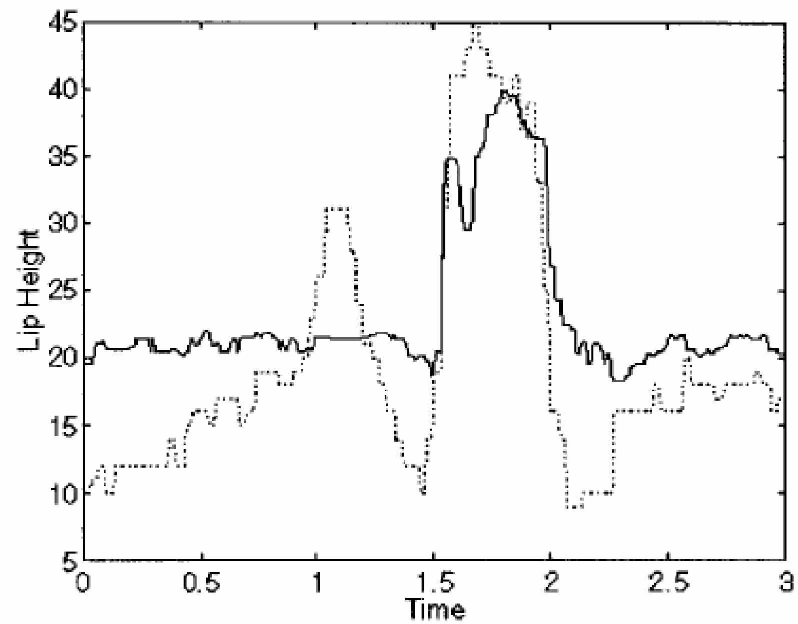
$$b_{\mathbf{a}j}(\mathbf{a}) = \int b_j(\mathbf{o}) dv.$$

    - For each state , one can derive the optimal estimate for the visual parameter given the acoustics $E_j\langle v|\mathbf{a}\rangle$.
  - Conversion
    - Optimal state sequence from acoustic parameters using Viterbi
    - Optimal estimate for visual vector by estimation function $E_j\langle v|\mathbf{a}\rangle$

Comparison of HMM & Neural Network

- HMM found better than neural network in simulations
- HMM Neural network



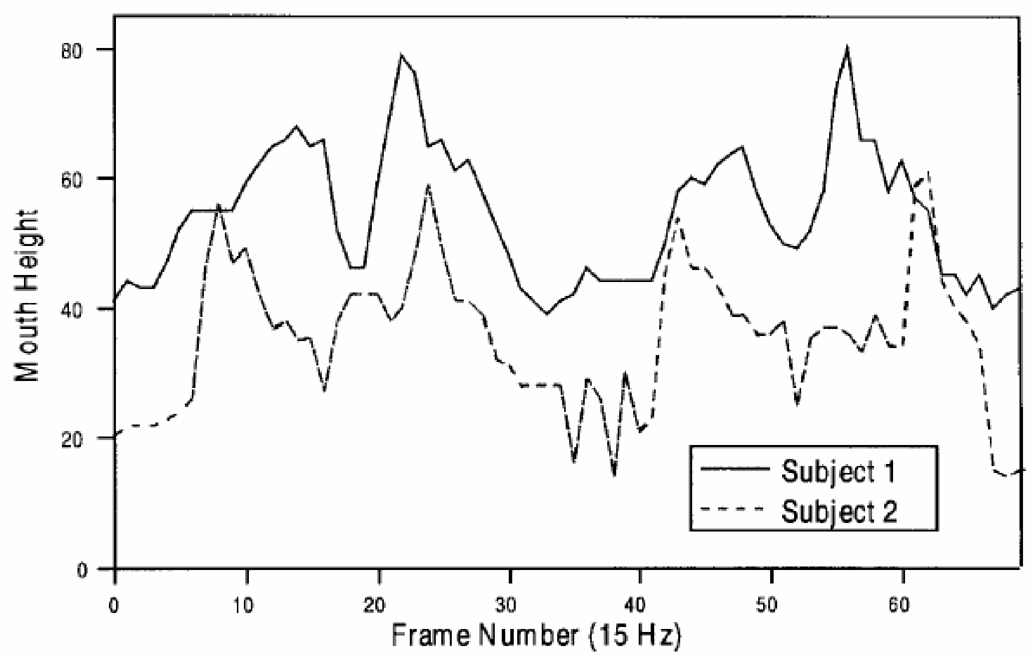


8. Bimodal Person Verification

- Single modality has limitations in both security and robustness
- Combination of voice and visual modalities can be more secure and robust
- Mason et.al., Chen et.al. used an early integration approach and the results were better for bimodal than single modal
- Wassner et.al. used late integration approach

Bimodal Person Verification

- Lip movement can have information about a person's identity
- Time variation of mouth height for two persons



Conclusions

- Joint processing of audio and video provides additional capabilities
- bit-rate allocation between audio and video remains an open issue in audio-visual communication