Lipreading by Humans and Machines

6.870 Intelligent Multimodal Interfaces, Spring 2008

Guest Lecture by Kate Saenko

Outline

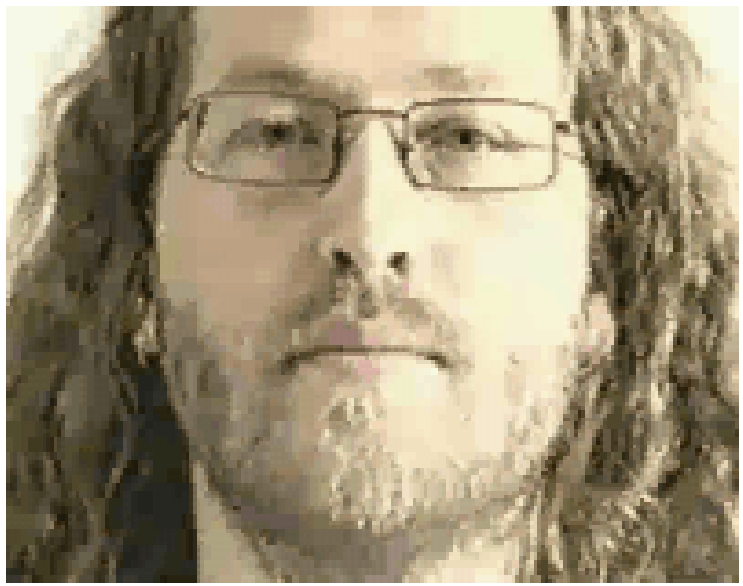
- **Lipreading by humans**
- Forensic lipreading
- Speech recognition basics
- Visual feature extraction
- Comparison of two feature types

Lipreading by Humans

- Who can read lips?

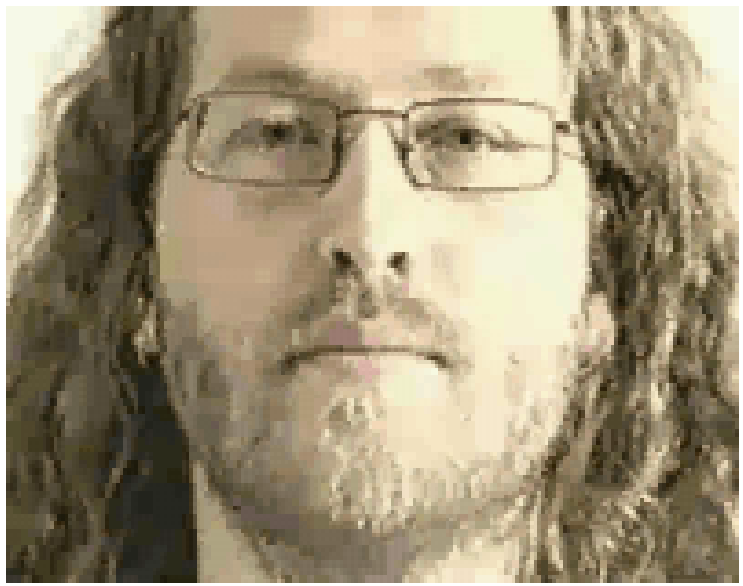
What is he saying?

- You will see and hear 6 syllables
- Listen and watch the face closely



Now, close your eyes and listen…

audio /ba/ + video /ga/ = /da/



McGurk effect

- First published by McGurk & McDonald, 1976
- Shows we cannot help but integrate visual and acoustic speech

Generality of McGurk effect

- Works on perceivers with **all language backgrounds** (e.g., Massaro, Cohen, Gesi, Heredia, & Tsuzaki, 1993; Sekiyama. & Tokhura, 1993)
- Works on **young infants** (Rosenblum, Schmuckler, & Johnson, 1997).
- Works when the visual and auditory components are from speakers of **different genders** (Green, Kuhl, Meltzoff, & Stevens, 1991).
- Works with highly **reduced face images** (Rosenblum & Saldaña, 1996).
- Works when observers **touch**—rather than look—at the face (Fowler & Dekle, 1991).
- Works **less well with vowels** than consonants (Summerfield & McGrath, 1984).
- Works **less well with nonspeech** pluck & bow stimuli (Saldaña & Rosenblum, 1994).

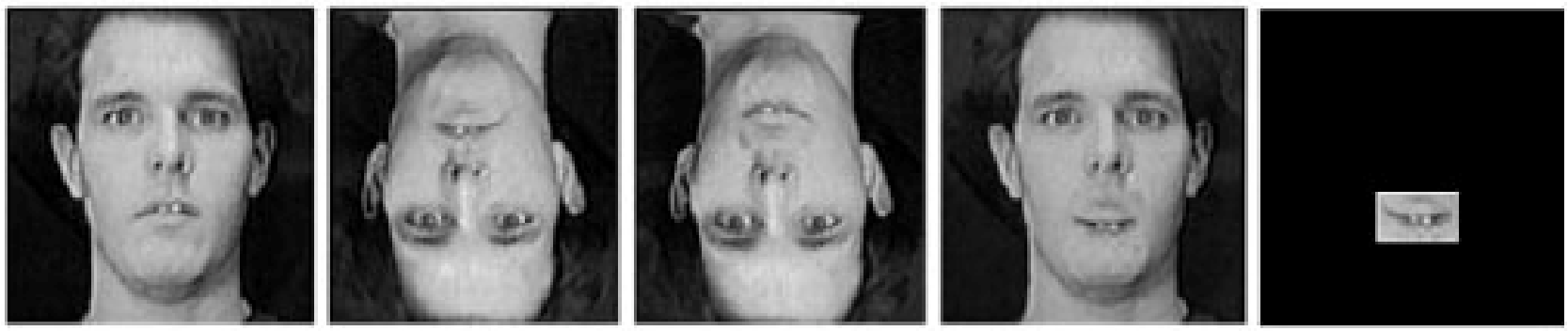
From http://www.faculty.ucr.edu/~rosenblu/lab-index.html

Recall: Margaret Thatcher Effect

- Lips and mouth are inverted in the right photo
- Effect shows that it is hard to recognize upside-down faces

“Margaret Thatcher” + “McGurk”

Do upside-down lips and/or face interfere visual speech perception?



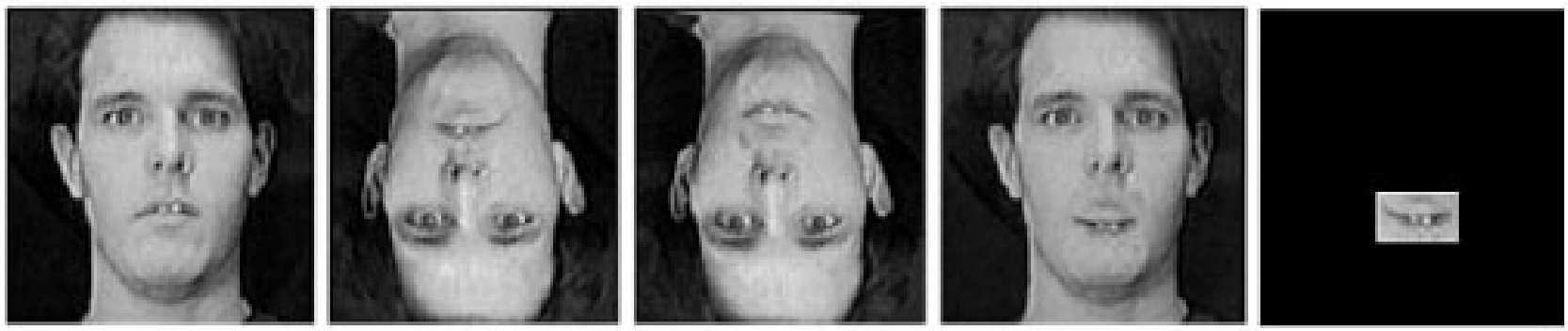
From http://www.faculty.ucr.edu/~rosenblu/lab-index.html

“McThatcher” effect

“McThatcher” effect

Most people perceive:

/va/ **/va/** **/va/** **/ba/** **/va/**



Definitions from Psychology

Lipreading

the person relies ONLY on the visual signal provided by the talker's face for recognizing speech

Speechreading

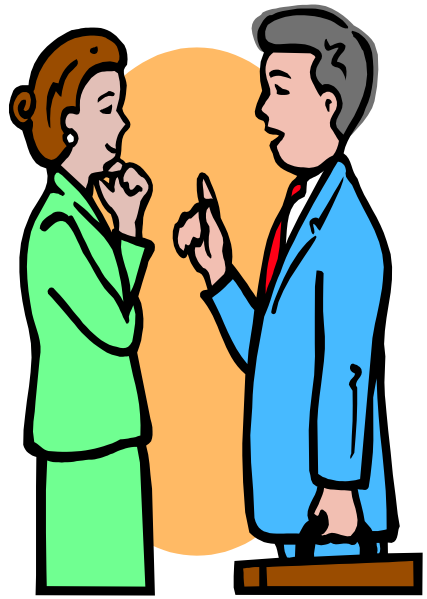
the person attends to both the talker's facial expressions AND gestures, and any other available cues

Note: terms used interchangeably in CS literature

slide credit: Perry C. Hanavan

Speechreading

- Lip cues
- Facial expression cues
- Gesture cues
- Body language cues
- Linguistic
- Situational cues
- Auditory cues



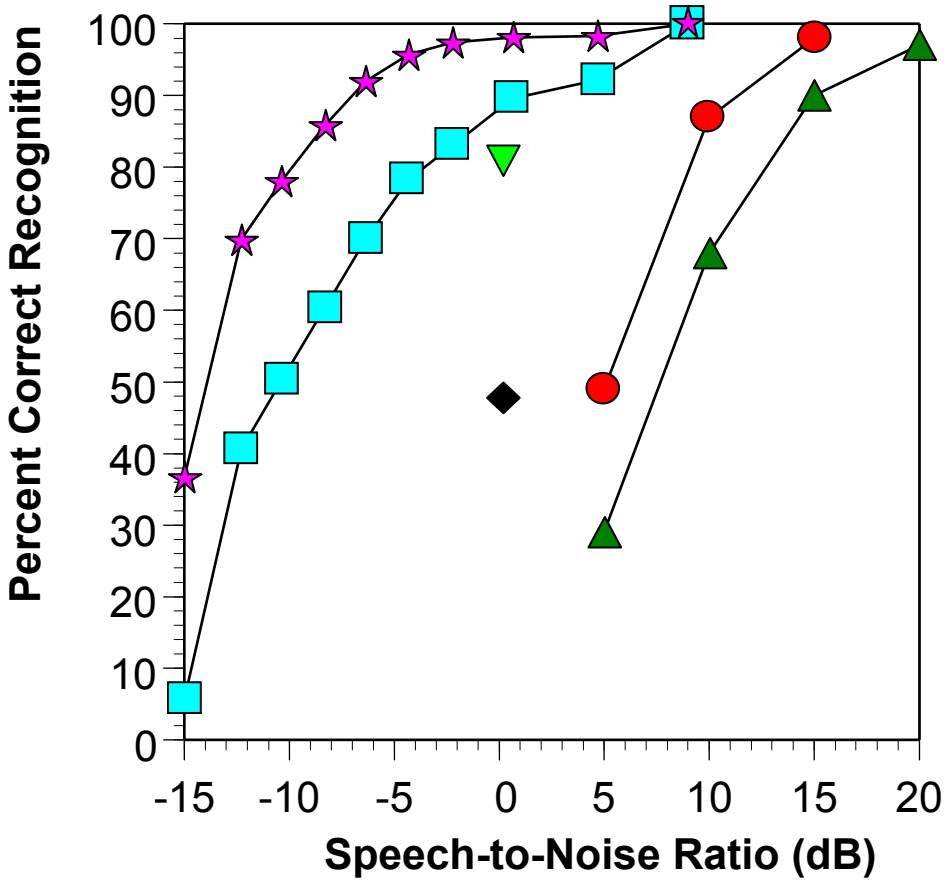
slide credit: Perry C. Hanavan

Speechreading for Communication

- Normal hearing adults
- Infants

slide credit: Perry C. Hanavan

Human Audio-Visual Recognition



- NH - Audiovisual Consonants
- NH - Auditory Consonants
- HI - Audiovisual Consonants
- HI Auditory Consonants
- HI-Auditory Sentences
- ASR Sentences

NH = Normal Hearing
HI = Hearing Impaired

slide credit: S. Greenberg and K. Grant

Difficulty Lipreading

- One third speech sounds visible
- mid and back consonants invisible
- Vowels not highly visible
- Rapidity of speech – 150 to 250 word/min
- Coarticulation
- Stress can change appearance of word
- Talker variability

slide credit: Perry C. Hanavan

Difficulty Lipreading: Talkers

- **Familiarity** – easier to lipread someone familiar
  - Family members, teachers, etc.
- **Gender** – Females are easier to lipread than males
  - However, auditory plus vision may be more difficult as females are less audible to person with hearing loss

slide credit: Perry C. Hanavan

Difficulty Lipreading: Message

- **Structure** – complexity of message, frequency of use, linguistic context
- **Frequency of usage** – how often a word occurs in everyday conversations
- **Neighborhoods** – fewer lexical neighbors can be beneficial
- **Context** – words specified by context are easier

slide credit: Perry C. Hanavan

Difficulty Lipreading: Environment

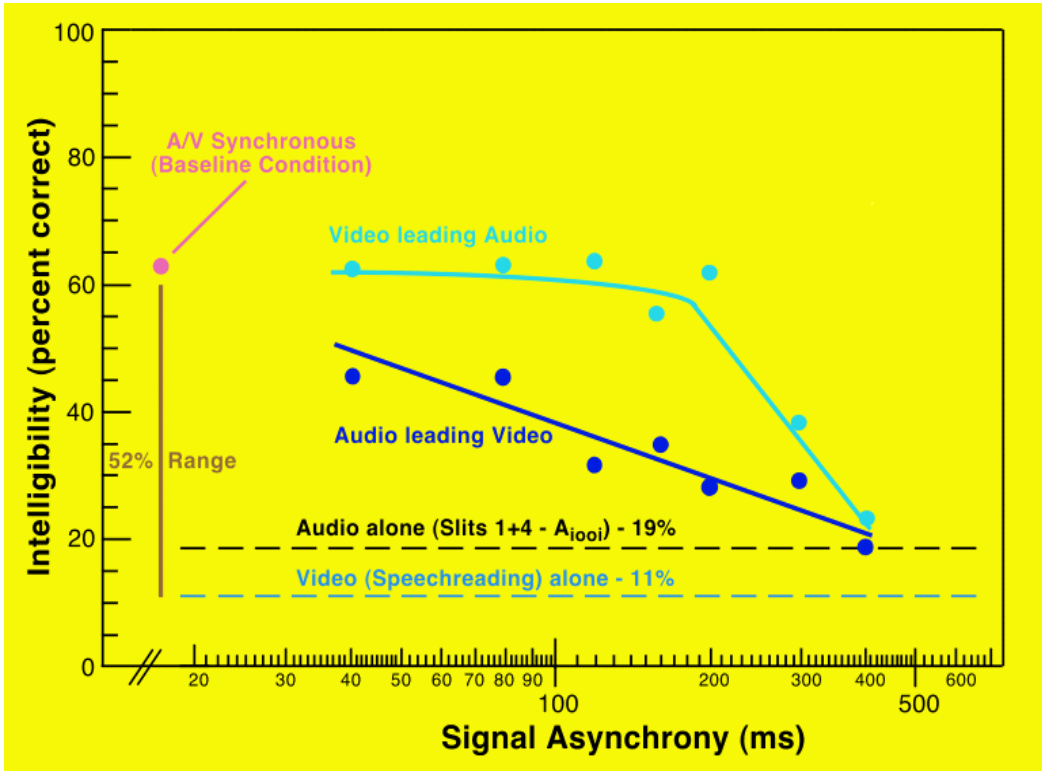
- **Viewing angle** – face to face
- **Distance** – favorable seating
- **Room conditions** – lighting, lighting angle, shining light, interfering objects, room noise

slide credit: Perry C. Hanavan

Humans Tolerate A-V Asynchrony

When the VIDEO signal LEADS the AUDIO, intelligibility is preserved for asynchrony intervals as large as 200 ms

When AUDIO leads the VIDEO, intelligibility declines at a constant rate



slide credit: Perry C. Hanavan

Outline

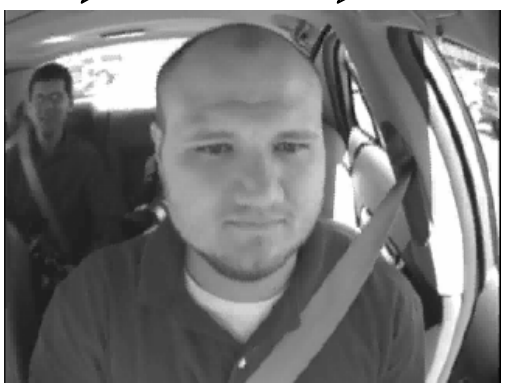
- Lipreading by humans
- **Forensic lipreading**
- Speech recognition basics
- Visual feature extraction
- Comparison of two feature types

Forensic Lipreading

- watch video:

  http://video.google.com/videoplay?docid=1896087054259916 17&hl=en

Visual Cues for Conversational Interfaces

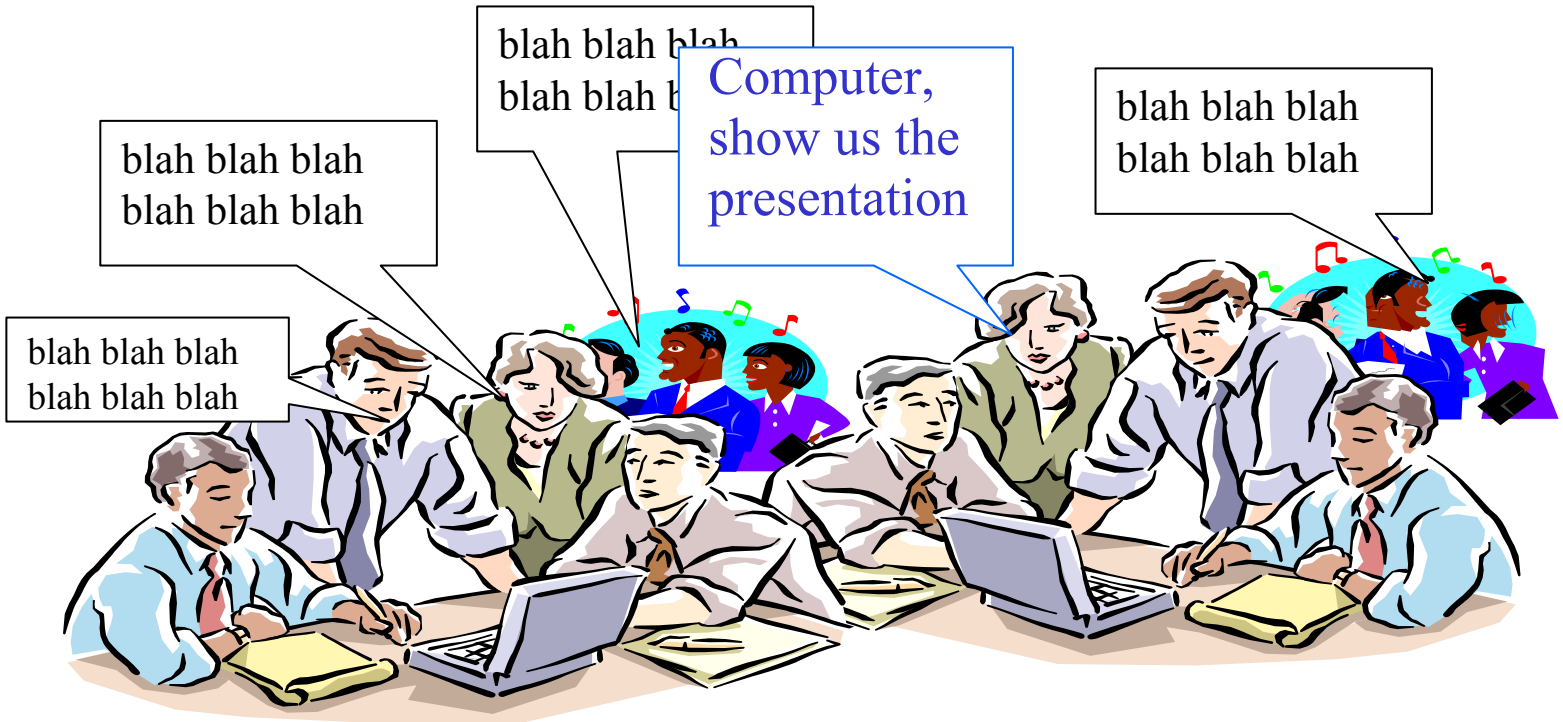


Who's speaking?

What are they saying?

What is their emotional state?

Challenge for ASR*: Noise



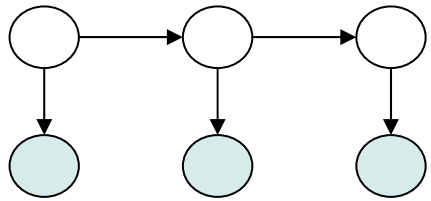
*ASR = Automatic Speech Recognition

Outline

- Lipreading by humans
- Forensic lipreading
- **Speech recognition basics**
- Visual feature extraction
- Comparison of two feature types

ASR in 1 slide

- Extract speech information from waveform
  - Mel-frequency cepstral coefficients (MFCC) are popular features based on how humans hear sound
- Recognize phonemes or subword units
  - Phonemes are smallest units of speech
  - Popular subword units are tri-phones
  - Hidden Markov Models used to model units
- Identify the most likely words based on hypothesized sequences of subword units
  - Grammar is used to constrain search



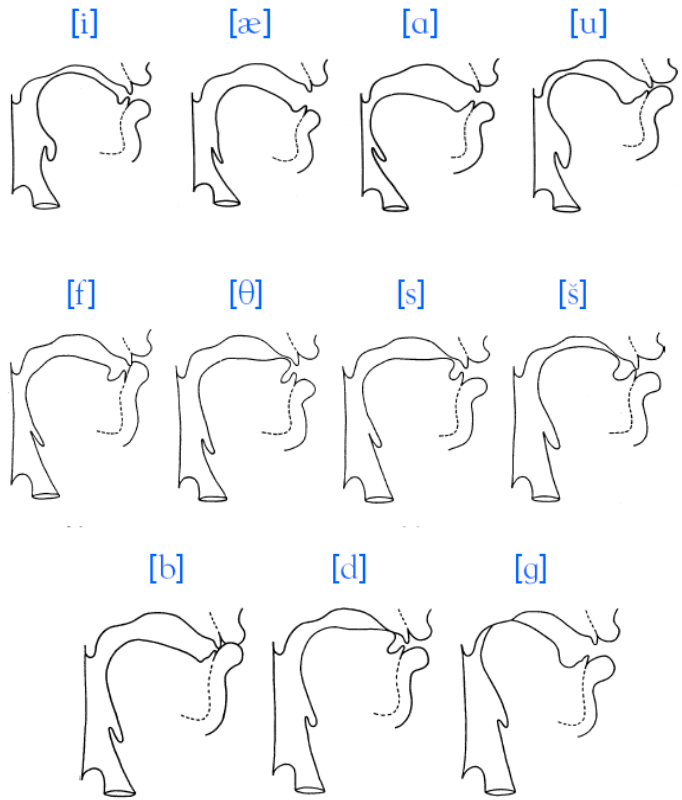
HMM

Phonemes and Homophones

- **Phonemes** are the smallest units of speech that distinguish one word from another:
  - **seat, meat, beat, feet** only differ in 1st phoneme
  - about 40-50 phonemes in English
- **Homophones** are words that have the same pronunciation
  - **one** and **won**
  - **their** and **they're**

Production of Phonemes

- **Manner** – how the lips, tongue and other speech organs make contact, e.g.
  - vowel
  - fricative
  - stop
- **Place** – where the constriction occurs in the vocal tract, e.g.
  - bilabial for b-p-m
  - labiodental for f-v
- **Voicing**
  - voiced
  - unvoiced

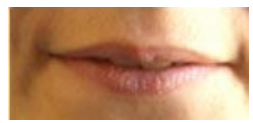


figures from 6.345 Lecture Notes

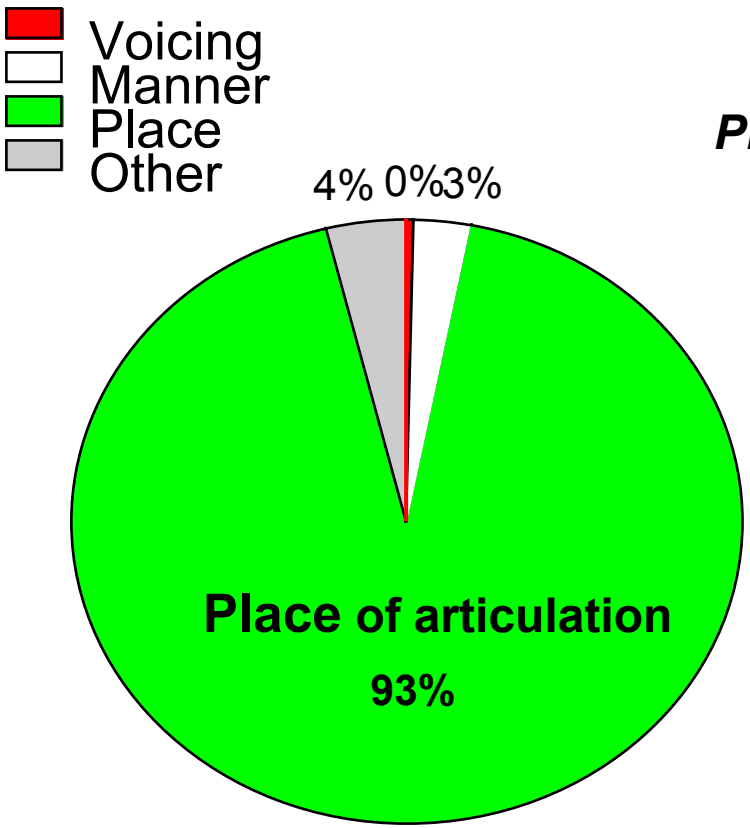
Visemes and Homophenes

- **Visemes** are groups of phonemes that appear identical on the lips
  - identified through human studies or statistical clustering of phonemes
  - 10-20 visemes in English
- **Homophenes** are words that look identical on the mouth
  - **pan, ban, man**
  - **tug, tongue, tuck**
  - 40-60% of words

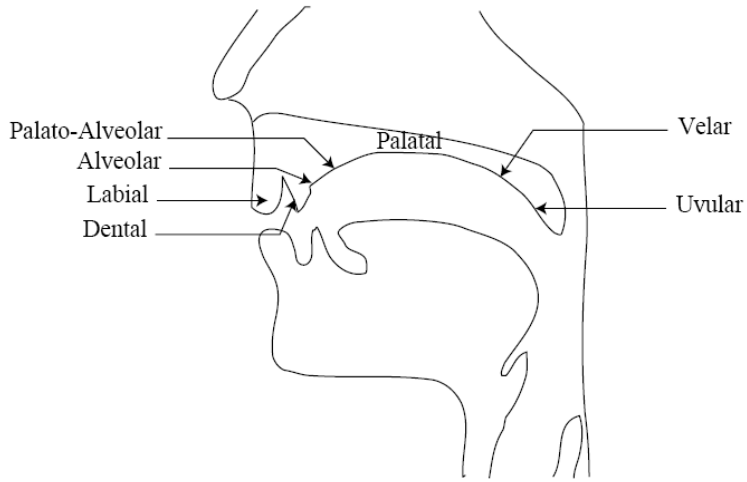
Examples of English Visemes

aa ah ao aw er oy hh 	ch ih sh zh 
uh uw ow w 	b p m 
ae eh ey ay 	f v 
ih iy ax axr 	g k ng 

What is Conveyed by Visual Cues?



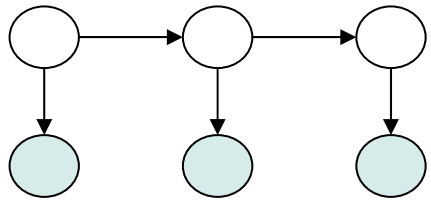
Place of Articulation Most Important



slide credit: Perry C. Hanavan

VSR in 1 slide

?



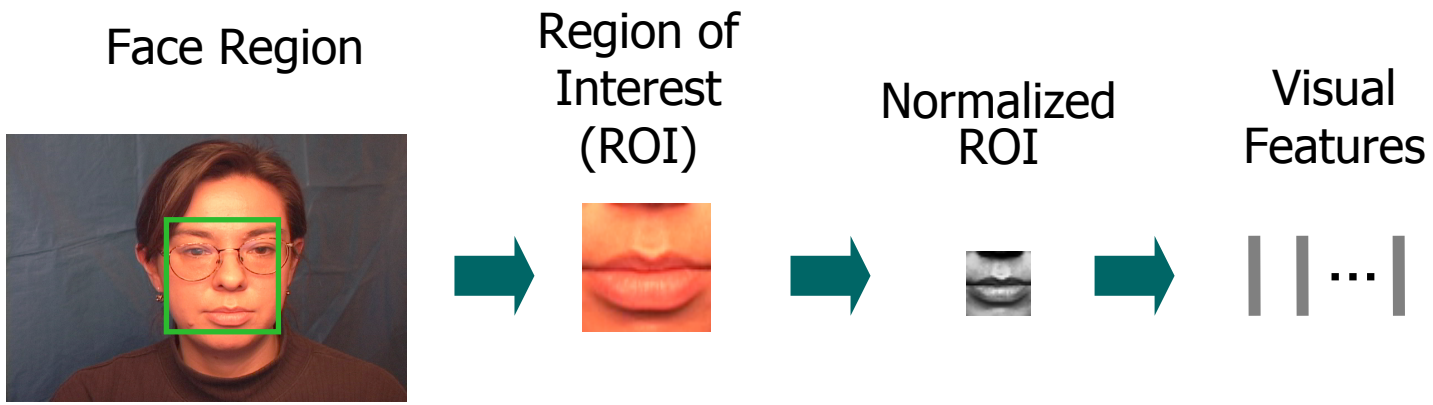
HMM

- Recognize phonemes or subword units
  - Phonemes are smallest units of speech
  - Popular subword units are tri-phones
  - Hidden Markov Models used to model units
- Identify the most likely words based on hypothesized sequences of subword units
  - Grammar is used to constrain search

Outline

- Lipreading by humans
- Forensic lipreading
- Speech recognition basics
- **Visual feature extraction**
- Comparison of two feature types

Video Processing



- System detects and tracks facial landmarks
- Only relevant speech information is retained
- Typically, face detection is the easy part

ROI Tracking

- Important to have good tracking algorithm
- Tracking noise can affect recognition

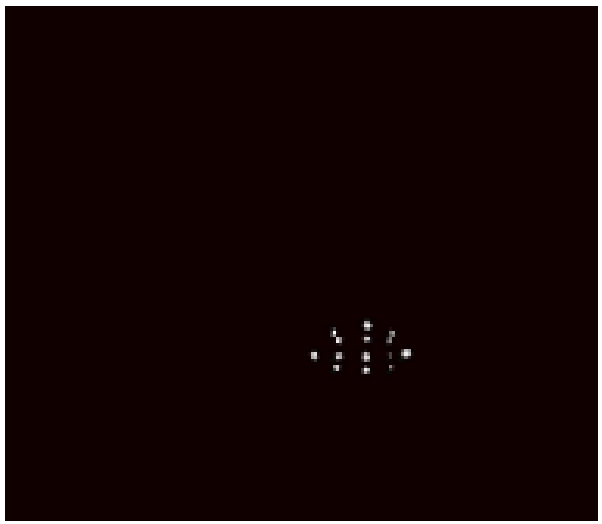
Unstable ROI

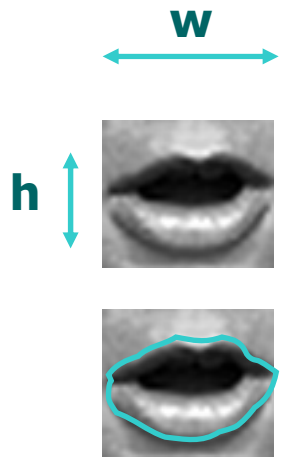
Stable ROI

- Can sometimes completely miss the mouth!
- In many AVSR experiments, tracking involves manual steps

Types of Visual Features

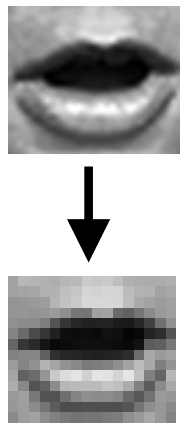
- **Lip shape features**
  - Width, height, area (Benoit)
  - Lip contour, moments, Fourier descriptors (Potamianos)
  - Template, B-spline parameters (Silsbee, Blake)





Types of Visual Features

- **Lip shape features**
  - Width, height, area (Benoit)
  - Lip contour, moments, Fourier descriptors (Potamianos)
  - Template, B-spline parameters (Silsbee, Blake)
- **Appearance features**
  - all pixels in the image are deemed important
  - PCA (Bregler); whole ROI (Waibel); DCT (Duchnowski); DWT (Potamianos); LDA (Duchnowski)



Types of Visual Features

- **Lip shape features**
  - Width, height, area (Benoit)
  - Lip contour, moments, Fourier descriptors (Potamianos)
  - Template, B-spline parameters (Silsbee, Blake)
- **Appearance features**
  - all pixels in the image are deemed important
  - PCA (Bregler); whole ROI (Waibel); DCT (Duchnowski); DWT (Potamianos); LDA (Duchnowski)
- **Lip shape-driven appearance features**
  - Active appearance models (AAMs) (Matthews)
  - Active shape models (Luettin)

Outline

- Lipreading by humans
- Forensic lipreading
- Speech recognition basics
- Visual feature extraction
- **Comparison of two feature types**

Case Study: AVSR Workshop 2000

- Brought many researchers in the field together at Johns Hopkins University in the summer of 2000
- Among other things, compared two types of visual features
  - Active Appearance Model (AAM)
  - Discreet Cosine Transform (DCT)

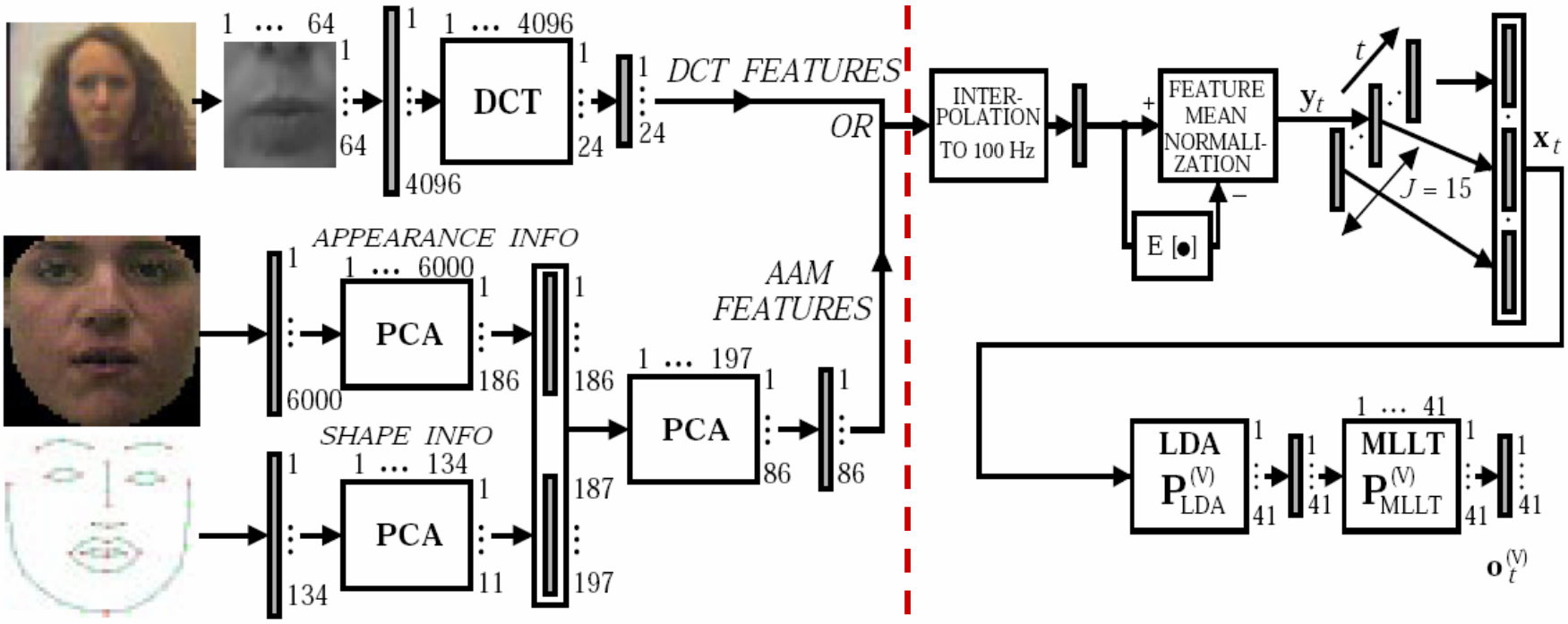
THE AUDIO-VISUAL DATABASE

- The IBM ViaVoice audio-visual database:
  - 290 subjects.
  - 50 hours, large vocabulary (10.5 K words), continuous speeech.
  - Frontal face color video, 704 × 480 pixels, 30 Hz, MPEG2.
  - 16 KHz audio (clean).

- Experimental setting:
  - *Speaker independent* (SI) data partitioning: 35 hrs *training*, 5 hrs *held-out*, 2 hrs *test* (in addition: 1 hr SI *adaptation* set, and *multi-speaker held-out* and *test* sets).
  - Audio: (a) *clean*; (b) *matched noisy* ("babble", 10 db SNR).
  - Transcriptions, dictionary, language model scores are provided.
  - Clean and matched noisy audio lattices are provided.
  - Lattice audio-only WER: 14.44 % clean, 48.10 % noisy.

slide credit: AVSR Workshop Final Report '00

Feature Extraction

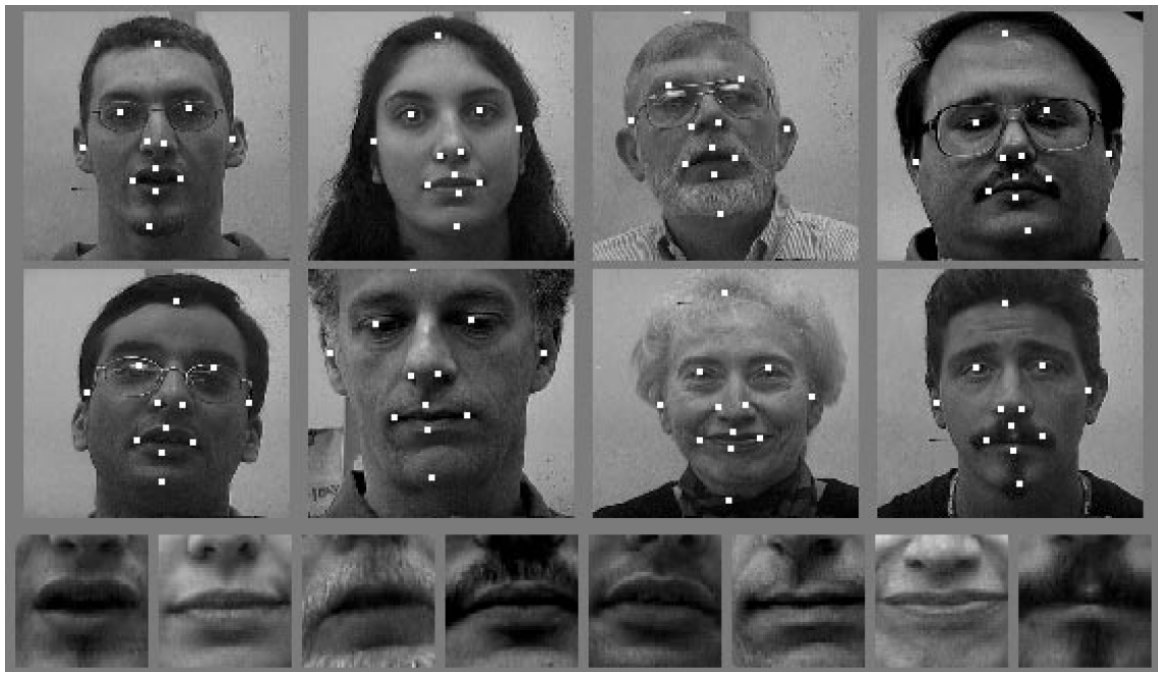


Extract either DCT or AAM features

Further processing done on extracted features

Figure from Neti et. al., 2000

DCT features: ROI extraction

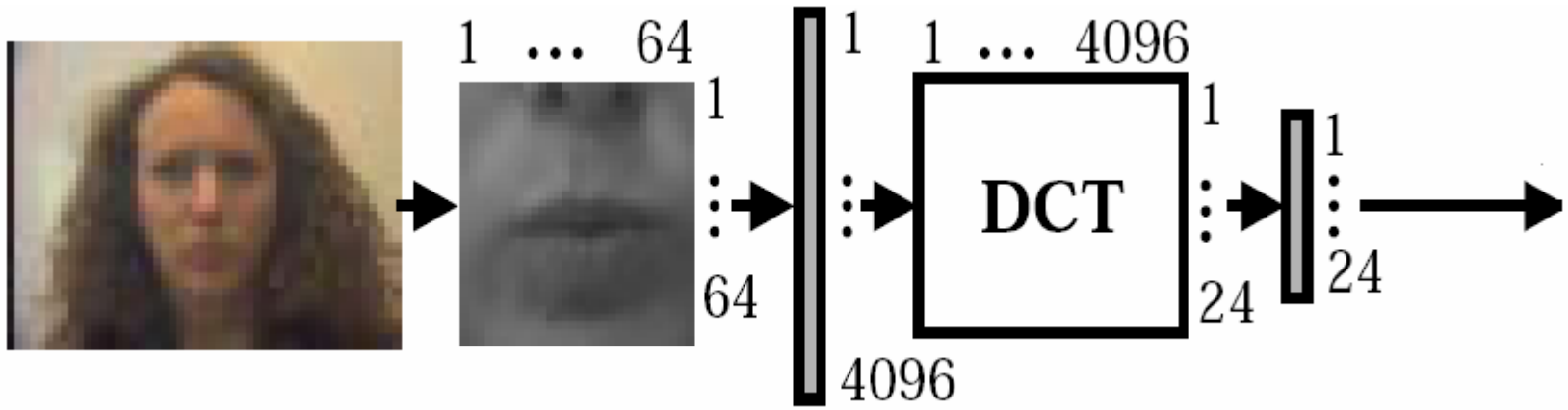


- Detect face and 26 facial features
- Obtain mouth center and size
- Smooth coordinates, normalize size
- Extract a 64 x 64 pixel ROI

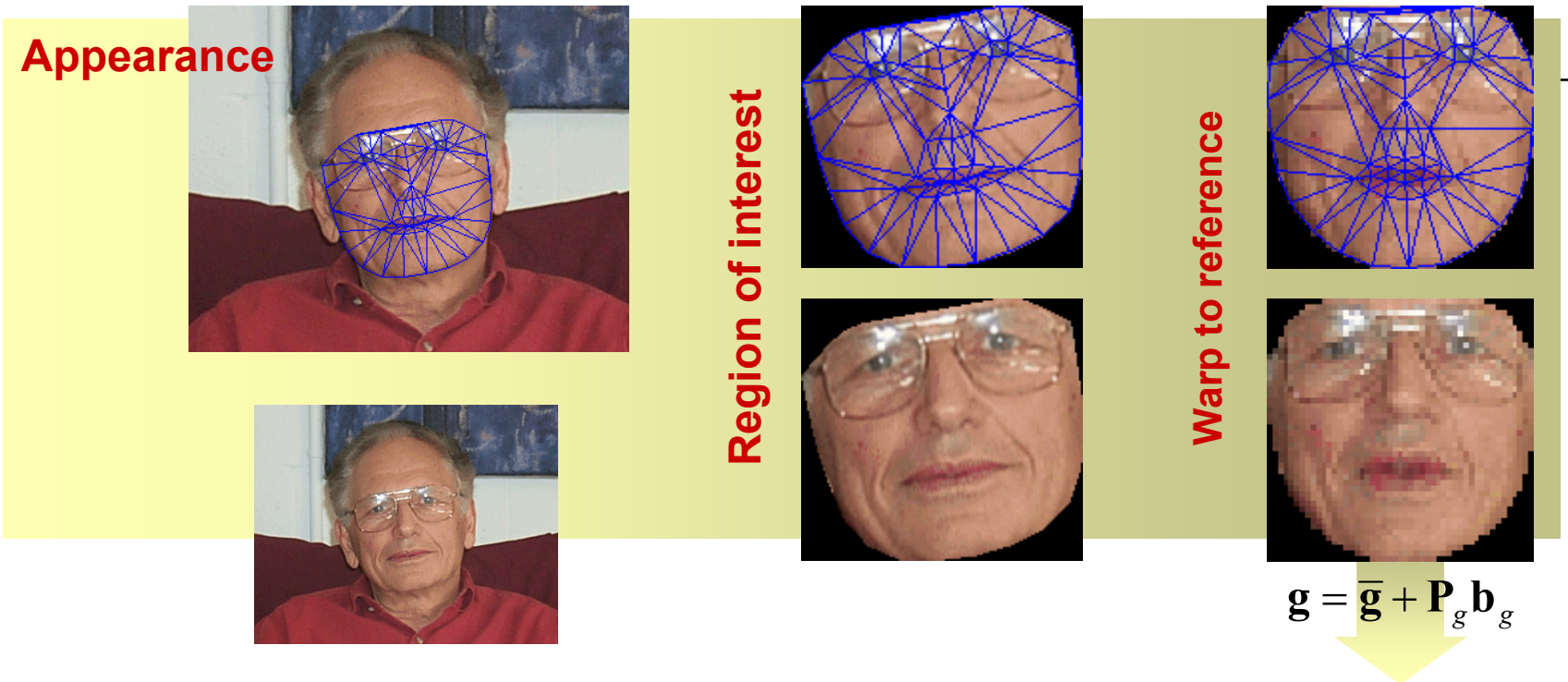
Figure from Neti et. al., 2000

Discreet Cosine Transform

- Achieves information compression
- Computationally efficient
- Retain transform coefficients at high energy lattice locations



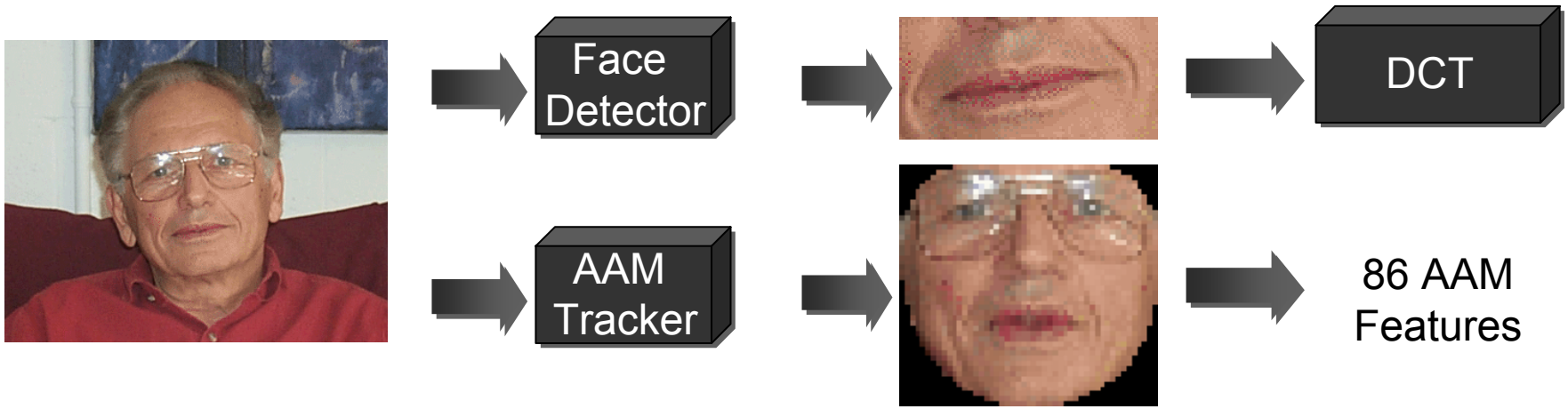
Active Appearance Model (AAM)



slide credit: Iain Matthews

DCT vs. AAM Features

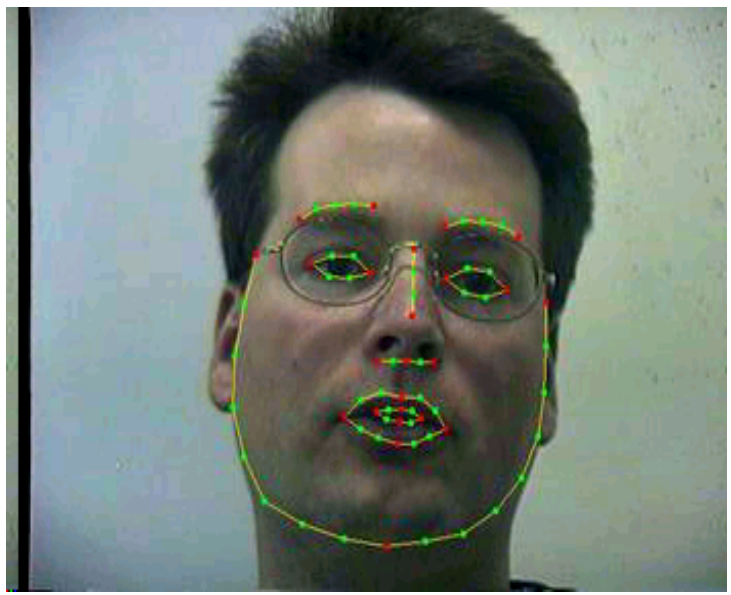
- External feature detector vs. model-based learned tracking
- ROI 'box' vs. explicit shape + appearance modeling



slide credit: Iain Matthews

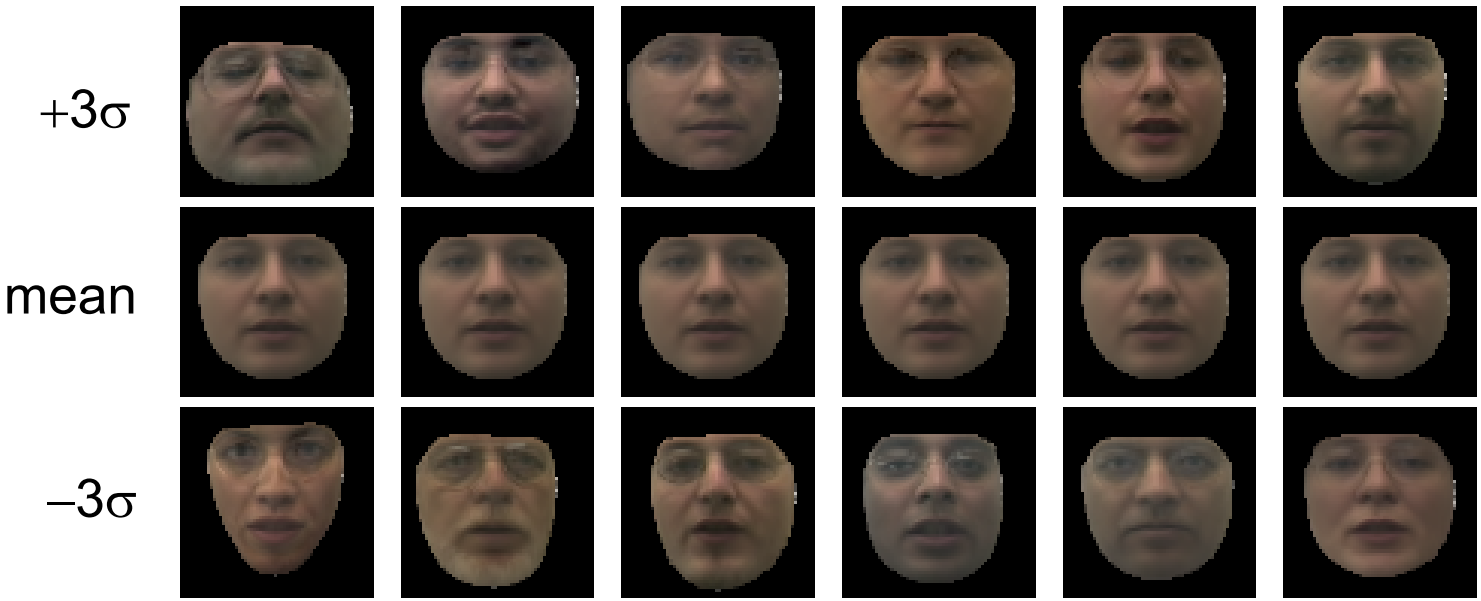
AAM Training Data

- 4072 hand labeled images = 2m 13s (out of 50h)
- Question: which feature extraction technique (DCT or AMM) requires more manual labeling?



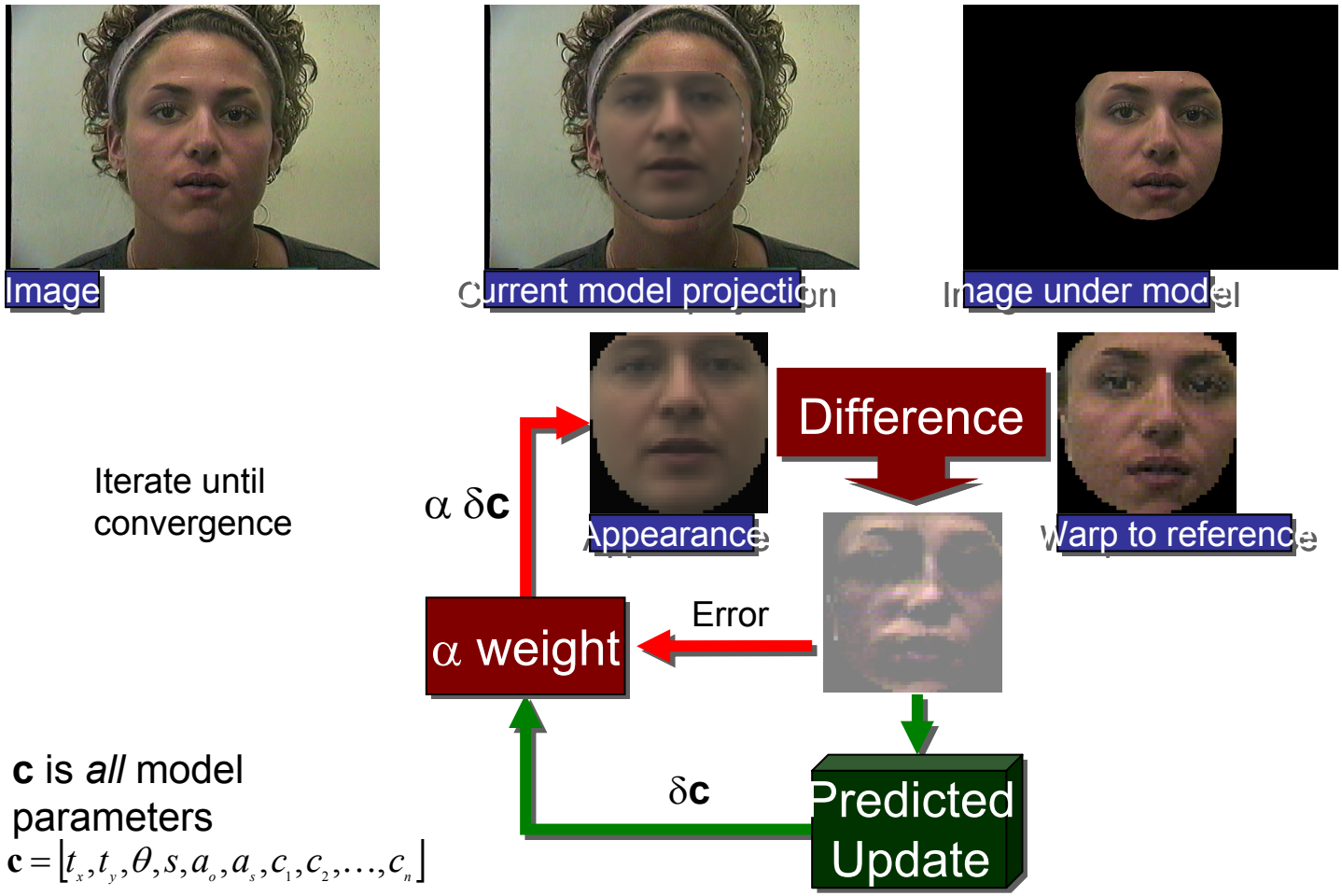
slide credit: Iain Matthews

AAM Final Model



slide credit: Iain Matthews

AAM Fitting Algorithm



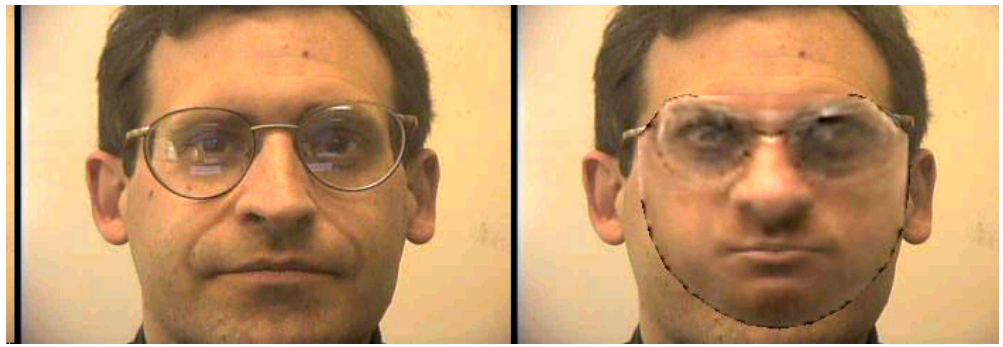
c is *all* model parameters

$$\mathbf{c} = \left[t_x, t_y, \theta, s, a_o, a_s, c_1, c_2, \ldots, c_n\right]$$

slide credit: Iain Matthews

AAM Tracking Results

- Worst sequence - mean, mean square error = 548.87



- Best sequence - mean, mean square error = 89.11

- Mean, mean MSE per sentence = 254.21

slide credit: Iain Matthews

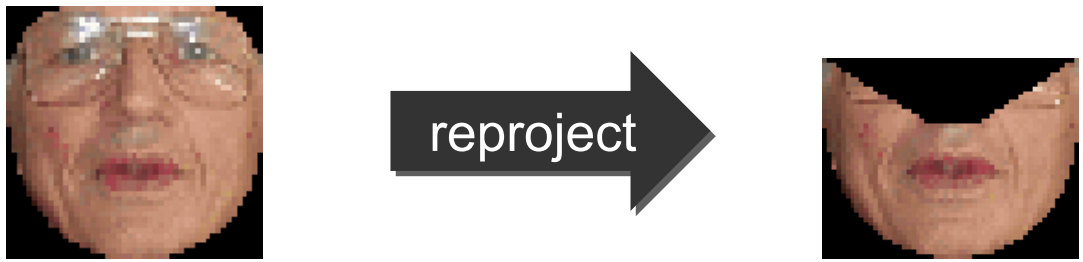
Comparison of AAM and DCT

Modality	Remarks	WER	Modality	Remarks	WER
Visual	DCT	58.1	Acoustic	MFCC (noisy)	55.0
	DWT	58.8	None	Oracle	31.2
	PCA	59.4		Anti-oracle	102.6
	AAM	64.0		LM best path	62.0

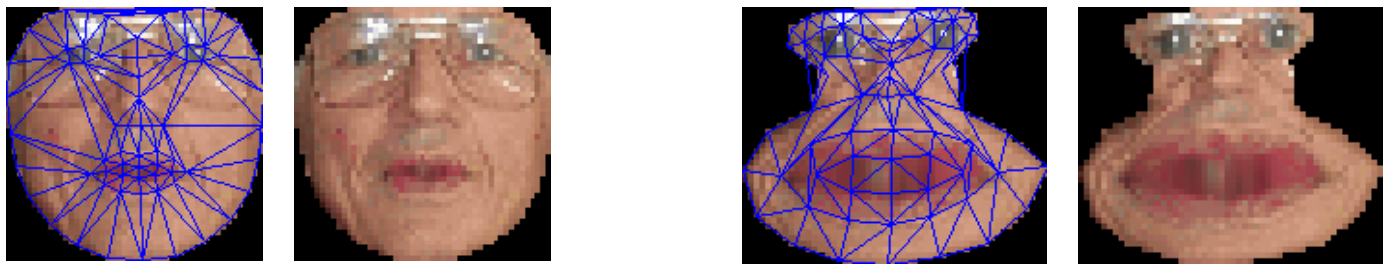
Table: Comparisons of various visual features (three appearance based features, and one joint shape and appearance feature representation) for speaker-independent LVCSR (Neti et al., 2000; Matthews et al., 2001). Word error rate (WER),%, is depicted on a subset of the IBMViaVoice database. **Visual performance** is obtained after rescoring of lattices, that have been previously generated based on noisy (at 8.5 dB SNR) audio-only MFCC features. For comparison, characteristic lattice WERs are also depicted (oracle, anti-oracle, and best path based on language model scores alone). Among the visual speech representations considered, the DCT based features are superior and contain significant speech information.

AAM analysis

- Models are under trained
  - Little more than face detection on 2m of training
- Project face through a more compact model
  - Retain only useful articulation information?

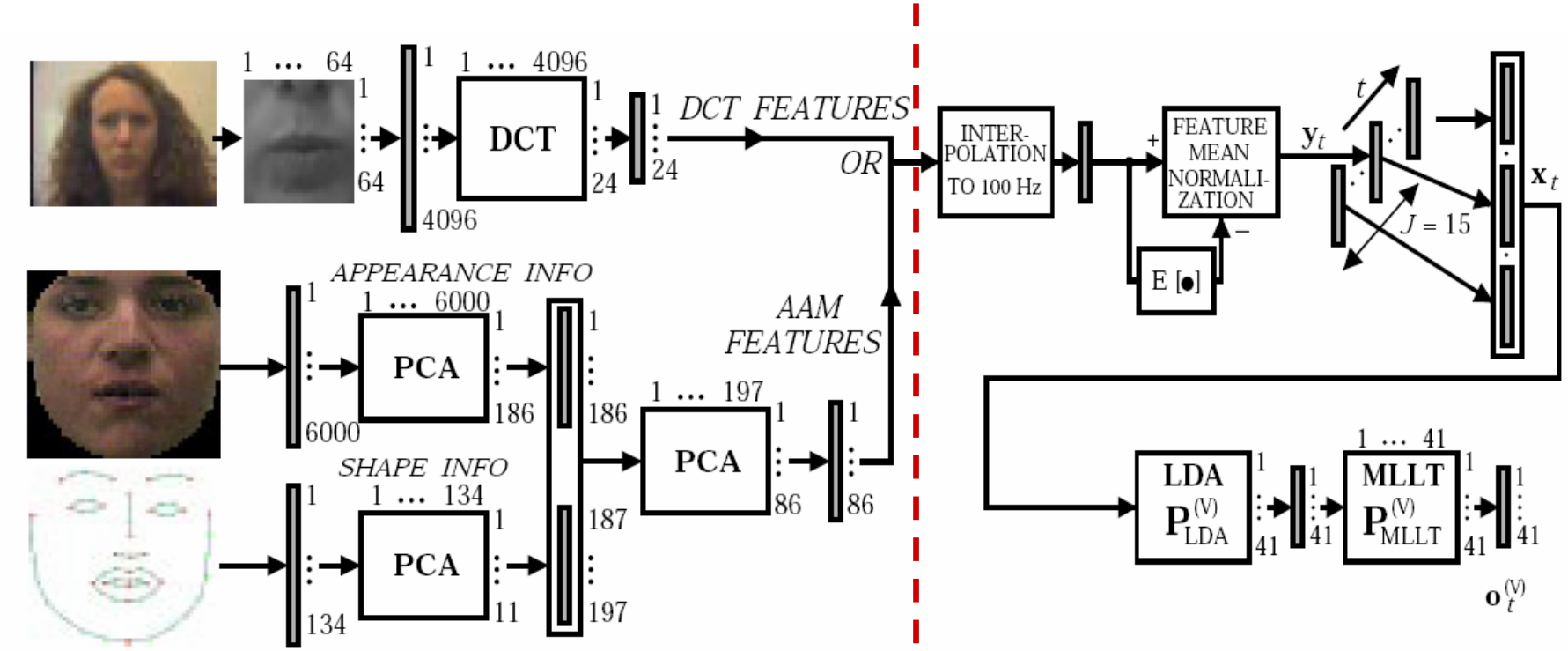


- Improve the reference shape
  - Minimal information loss through the warping?



slide credit: Iain Matthews

Recall: Feature Extraction

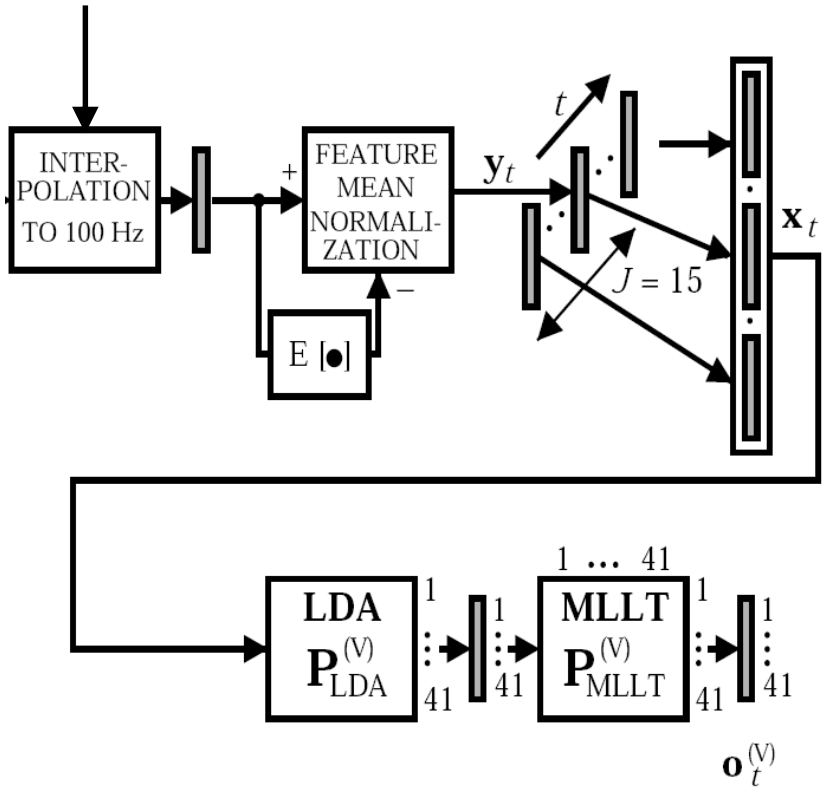


Extract either DCT or AAM features

Further processing done on extracted features

Figure from Neti et. al., 2000

Feature Post-processing



- Linear Discriminant analysis (LDA)
  - Assumes a known set of classes C = {1, 2, …, |C|}
  - Training data feature vectors must be labeled with class labels
  - Achieves optimal separation of projected sample into the set of classes C
- Maximum Likelihood Linear Transform (MLLT)
  - HMM models assume that observations have diagonal covariance, which is not true in general
  - MMLT maximizes data likelihood such that transformed class covariances are diagonal

Example of LDA Dimensions

- The original image vectors (raw pixels) were stacked and mean-normalized
- These four "movies" were reconstructed from the first four columns of the LDA projection matrix
- Each "movie" corresponds to several consecutive images

Summary

- Lipreading by humans
- Forensic lipreading
- Speech recognition basics
- Visual feature extraction
- Comparison of two feature types