Northwestern University
Department of Electrical
and Computer Engineering

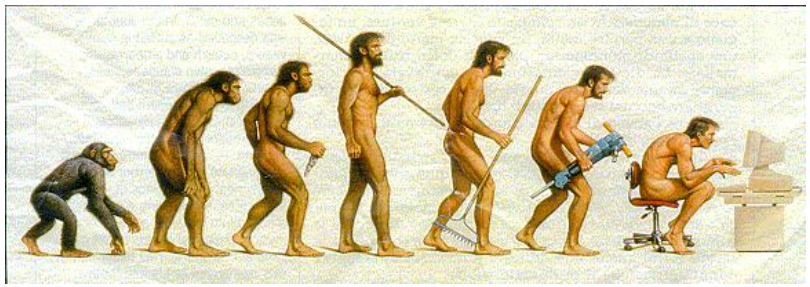
Audio-Visual Processing

Aggelos K. Katsaggelos

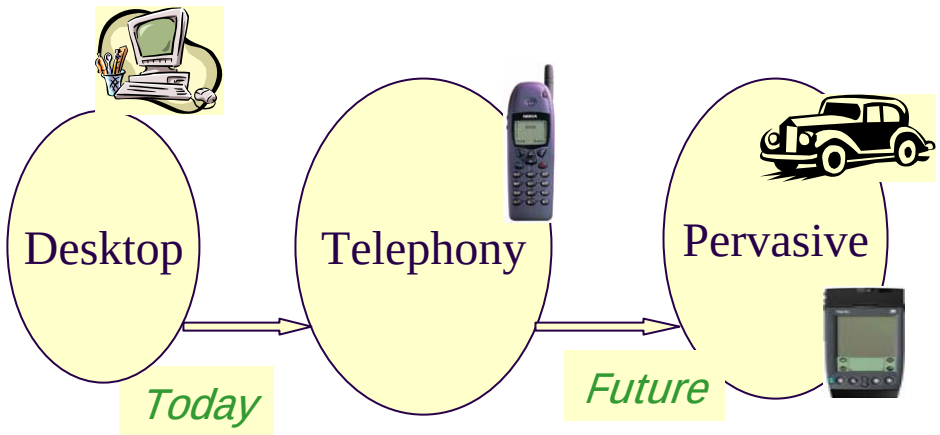
Northwestern University
Department of EECS
aggk@ece.northwestern.edu

Introduction and Motivation

- **Human-computer interaction (HCI):**
  - **Today:** Part of everyday life, but far from natural!



  - **Future:** Pervasive and ubiquitous computing.



Introduction and Motivation – Cont.

- Next generation of HCI will require **perceptual intelligence**:
  - **What** is the environment?
  - **Who** is in the environment?
  - **Who** is speaking?
  - **What** is being said?
  - What is the **state** of the speaker?
  - How can the computer **speak** back?
  - How can the activity be **summarized**, **indexed**, and **retrieved**?

- Operation on basis of traditional audio-only information:
  - **Lacks robustness** to noise.
  - **Lags human performance** significantly, even in ideal environments.

- **Joint audio + visual processing can help bridge the usability gap**!

Introduction and Motivation – Cont.

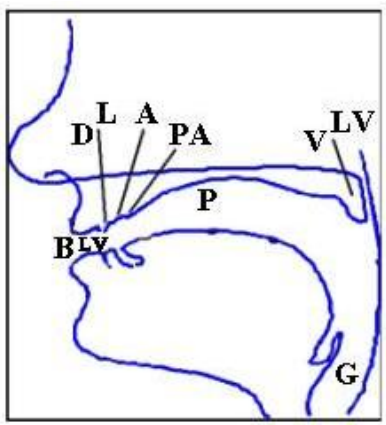
- **Vision of the HCI of the future?**

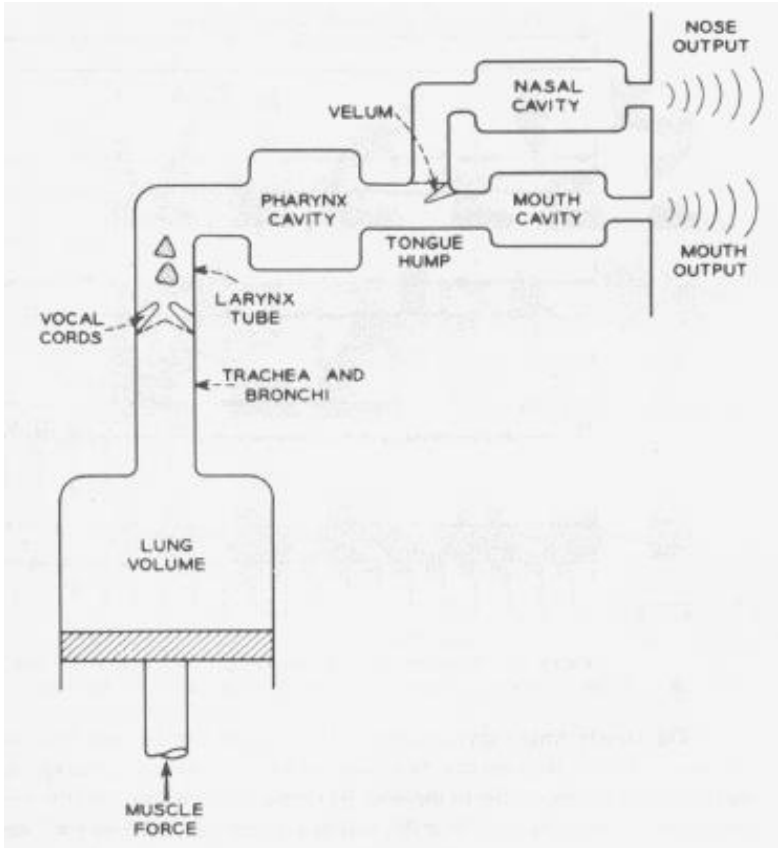
- A famous exchange (HAL's "premature" audio-visual speech processing capability):
  - HAL: I knew that you and David were planning to disconnect me, and I'm afraid that's something I cannot allow to happen.
  - Dave: Where the hell did you get that idea, HAL?
  - HAL: Dave – although you took very thorough precautions in the pod against my hearing you, I could see your lips move.

(From *HAL's Legacy*, David G. Stork, ed., MIT Press: Cambridge, MA, 1997).

Why audio-visual speech?

- **Human speech production** is bimodal:
  - Mouth cavity is part of **vocal tract**.
  - Lips, teeth, tongue, chin, and lower face muscles play part in speech production and are **visible**.
  - Various parts of the vocal tract play different role in the production of the basic speech units. E.g., lips for **bilabial** phone set **B**=/p/,/b/,/m/.

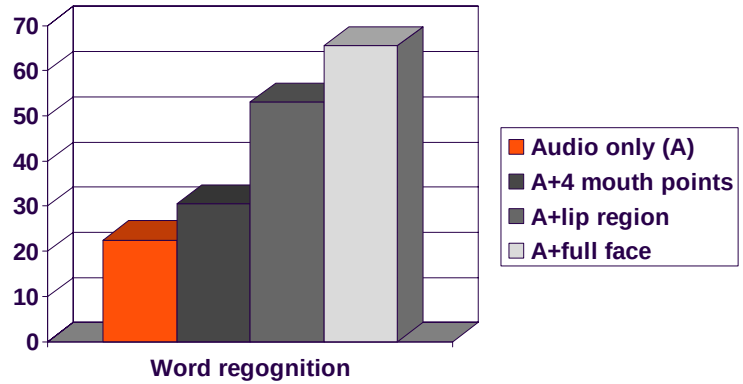




Schematic representation of speech production (J.L. Flanagan, *Speech Analysis, Synthesis, and Perception,* 2nd ed., Springer-Verlag, New York, 1972.)

Why audio-visual speech?

- **Human speech perception** is bimodal:
  - We **lip-read** in noisy environments to improve intelligibility.
    - E.g., human speech perception experiment by Summerfield (1979): Noisy word recognition at low SNR.



  - We integrate audio and visual stimuli, as demonstrated by the **McGurk effect** (McGurk and McDonald, 1976).
    - Audio /ba/ + Visual /ga/ -> AV /da/
    - Visual speech cues can dominate conflicting audio.
      - Audio: My bab pope me pu brive.
      - Visual/AV: My dad taught me to drive.

  - **Hearing impaired** people lip-read.

McGurk Effect

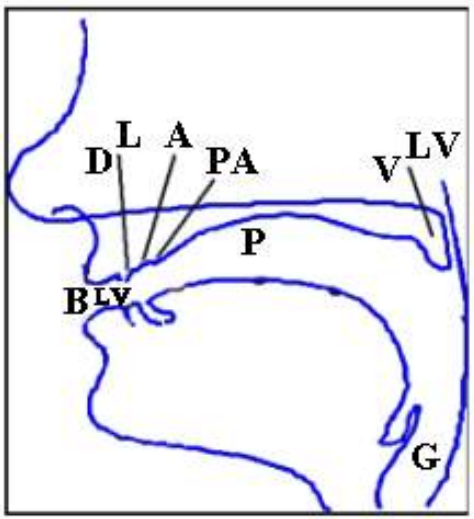
MA

KA

MA (audio) + KA (video)= NA

Why audio-visual speech? – Cont.

- Although the visual speech information content is less than audio …
  - **Phonemes:** Distinct speech units that convey linguistic information; about **47** in English.
  - **Visemes:** Visually distinguishable classes of phonemes: **6-20**.
- … the **visual channel provides important complementary information to audio:**
  - Consonant confusions in audio are due to same **manner** of articulation, in visual due to same **place** of articulation.
  - Thus, e.g., /t/,/p/ confusions drop by 76%, /n/,/m/ by 66%, compared to audio (Potamianos et al., '01).



Place of articulation

G	: Glottal	/ h /
V	: Velar	/ g, k /
P	: Palatal	/ y /
PA	: Palatoalveolar	/ r, dʒ, ʃ, tʃ, ʒ /
A	: Alveolar	/ d, l, n, s, t, z /
D	: Dental	/ θ, ð /
L	: Labiodental	/ f, v /
LV	: Labial-Velar	/ w /
B	: Bilabial	/ b, m, p /

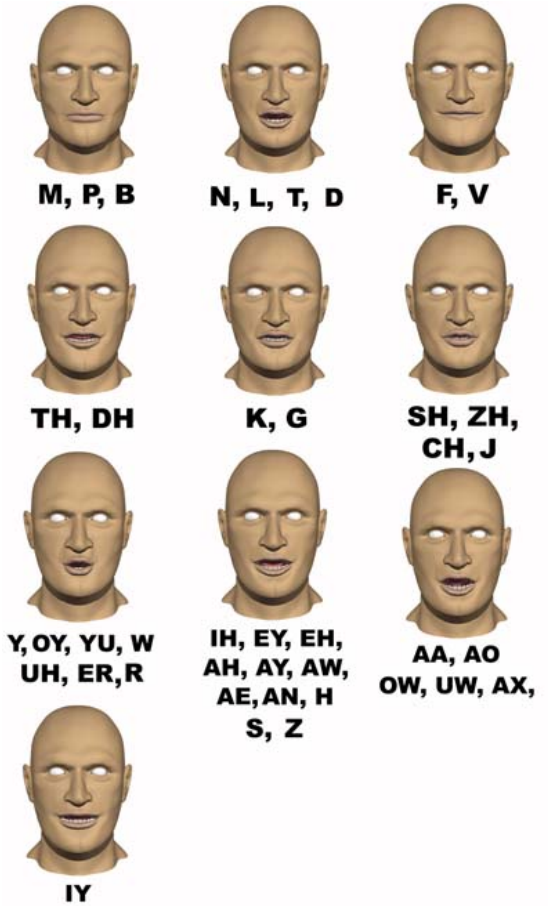
Manner of articulation

AP	: Approximant	/ r, w, y /
LA	: Lateral	/ l /
N	: Nasal	/ m, n /
PL	: Plosive	/ b, d, g, k, p, t /
F	: Fricative	/ f, h, s, v, z, θ, ð, ʃ, ʒ /
AF	: Affricate	/ tʃ, dʒ /

Why audio-visual speech? – Cont.

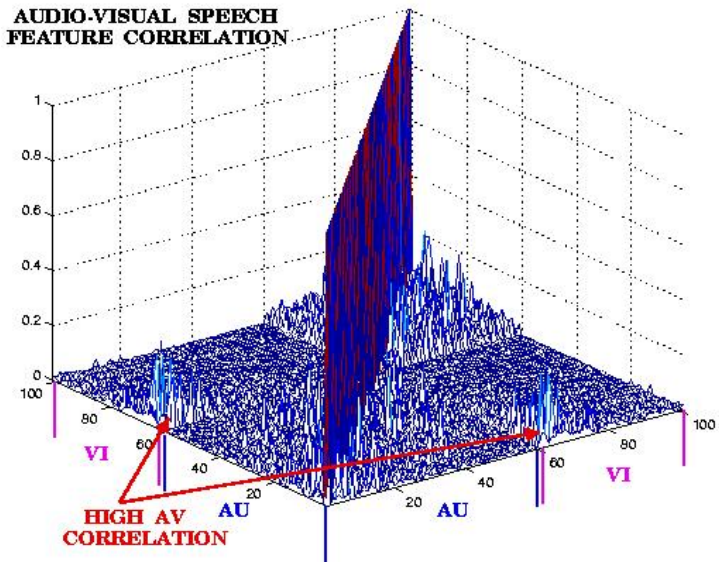
Visemes:

- Commonly agreed viseme categories:
- Confusion sets in the auditory modality are usually distinguishable in the visual modality (i.e., /P/, /t/, and /k/).

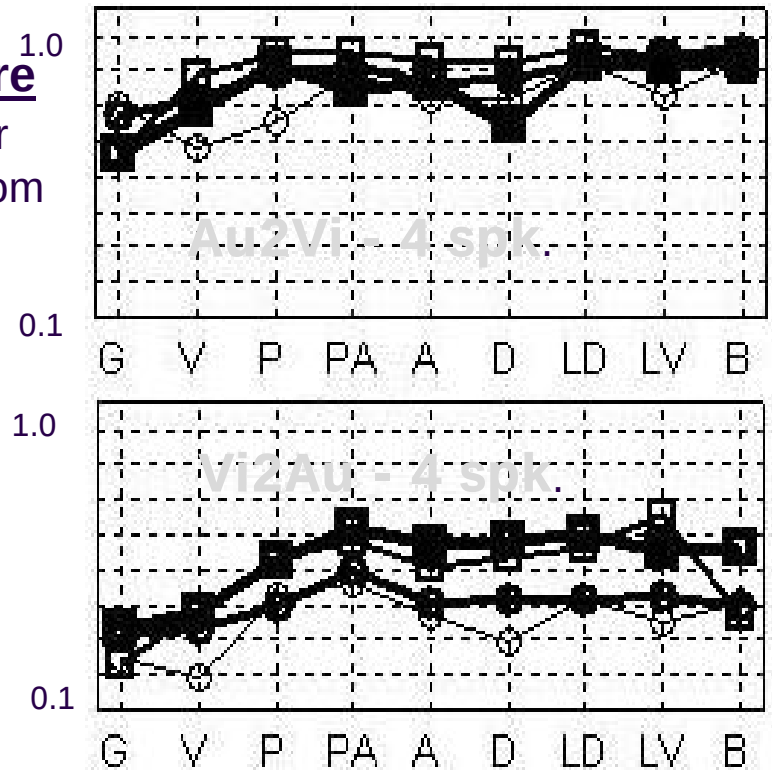


Why audio-visual speech – Cont.

- **Audio and visual speech observations are correlated:** Thus, for example, one can recover part of the one channel from using information from the other.



Correlation between audio and visual features (Goecke et al., 2002).



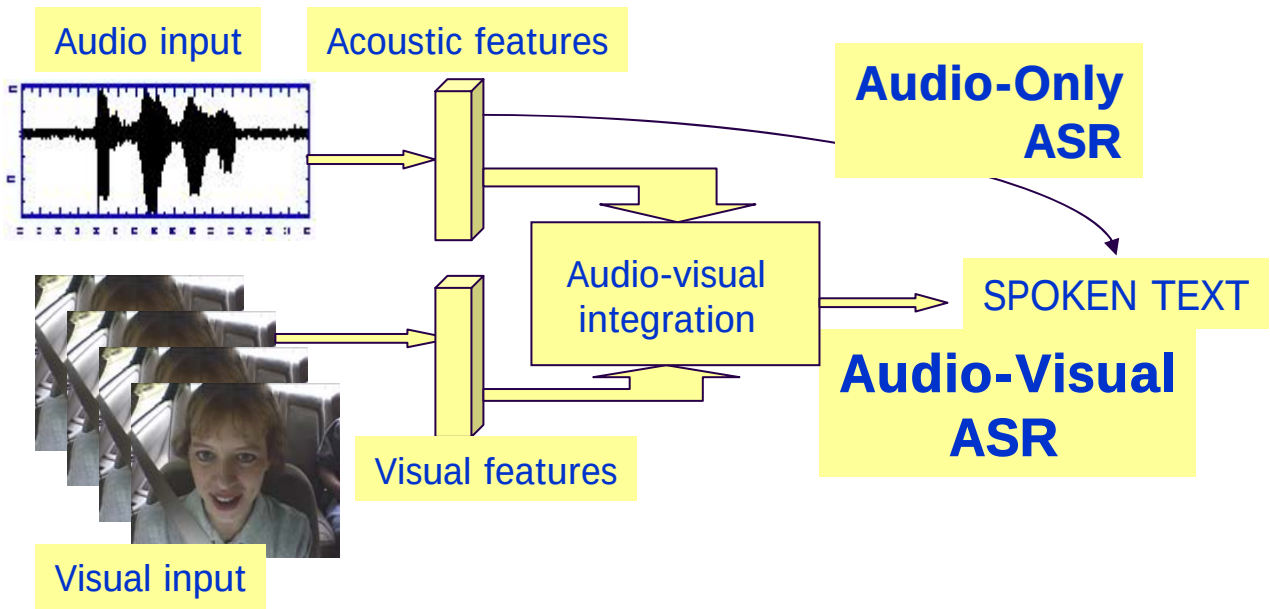
Correlation between original and estimated features; *upper*: visual from audio; *lower*: audio from visual (Jiang et al.,2003).

Why audio-visual speech? – Cont.

- Traditional audio only systems not satisfactory in unconstrained environments:
  - Lack of robustness to noise (far-field microphones, multiple subjects, etc.).
  - Purely acoustic compensation techniques are inadequate.
  - Performance not comparable to human capabilities (recognition, synthesis, etc.).
- Visual capture of information is very feasible and widespread:
  - Cameras are inexpensive, miniature, etc.
  - Cameras in PDAs, cell-phones, toys, etc.
  - Video data storage is becoming cheaper.
  - Large amounts of audio-visual content are available (broadcast video, etc.).
- Increasing computing power allows real-time capture and processing of video.

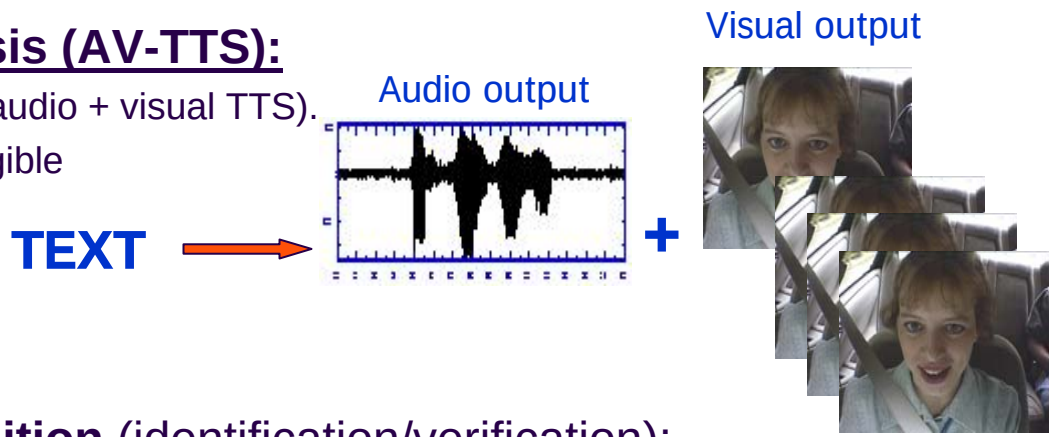
Audio-visual speech used in HCI

- **Audio-visual automatic speech recognition (AV-ASR):**
  - Utilizes both audio and visual signal inputs from the video of a speaker's face to obtain the transcript of the spoken utterance.
  - AV-ASR system performance should be better than traditional audio-only ASR.
  - **Issues:** Audio, visual feature extraction, audio-visual integration.

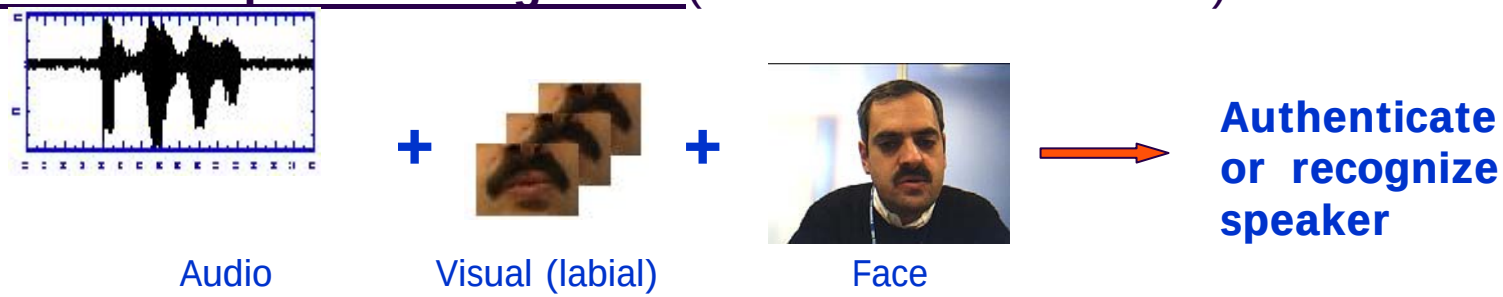


Audio-visual speech used in HCI

- **Audio-visual speech synthesis (AV-TTS):**
  - Given text, create a talking head (audio + visual TTS).
  - Should be more natural and intelligible than audio-only TTS.



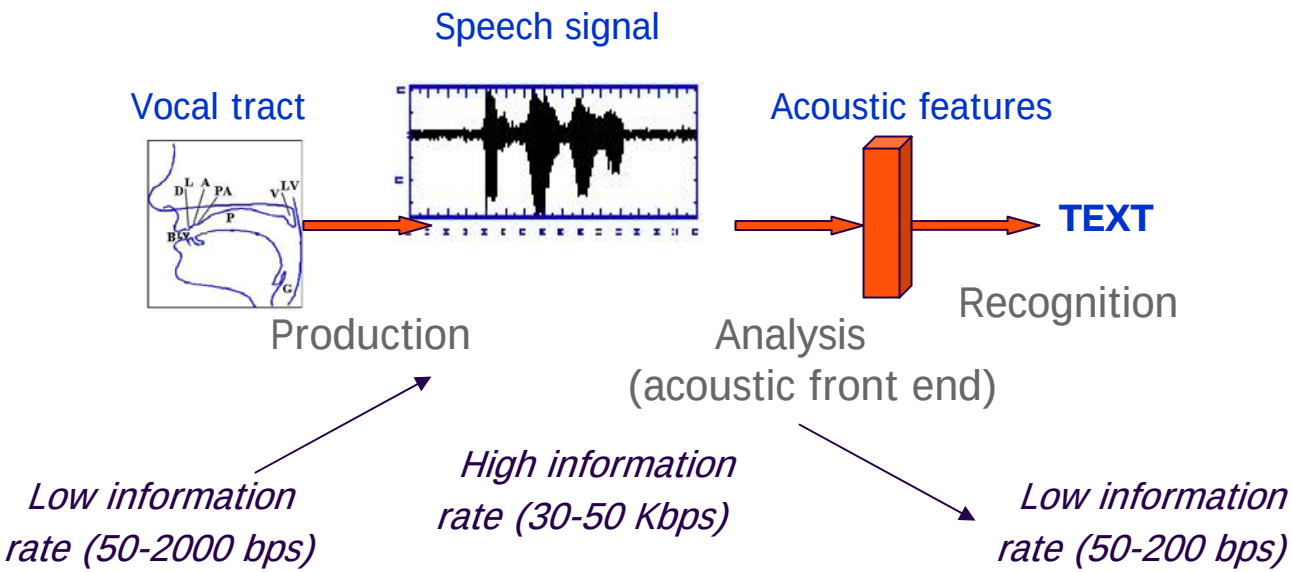
- **Audio-visual speaker recognition** (identification/verification):



- **Audio-visual speaker localization:**
- **Etc…**

Acoustic signal analysis and ASR

- Two components are of interest in automatic speech recognition (ASR) i.e., the speech-to-text process:
  A. Speech signal analysis.
  B. Speech signal statistical modeling and recognition.



Speech signal analysis, feature extraction

- Various approaches exist. Most prevalent ones are low-level, signal based (LPC, MFCC, PLP, etc.). Here, we discuss two popular techniques, based on:
  - The **linear predictive coding** (**LPC**) model of speech.
  - Filter-bank analysis, in particular **mel-frequency cepstral coefficients** (**MFCC**).
- We also discuss:
  - Signal **pre-processing**.
  - Feature **post-processing**.

Signal Pre-processing.

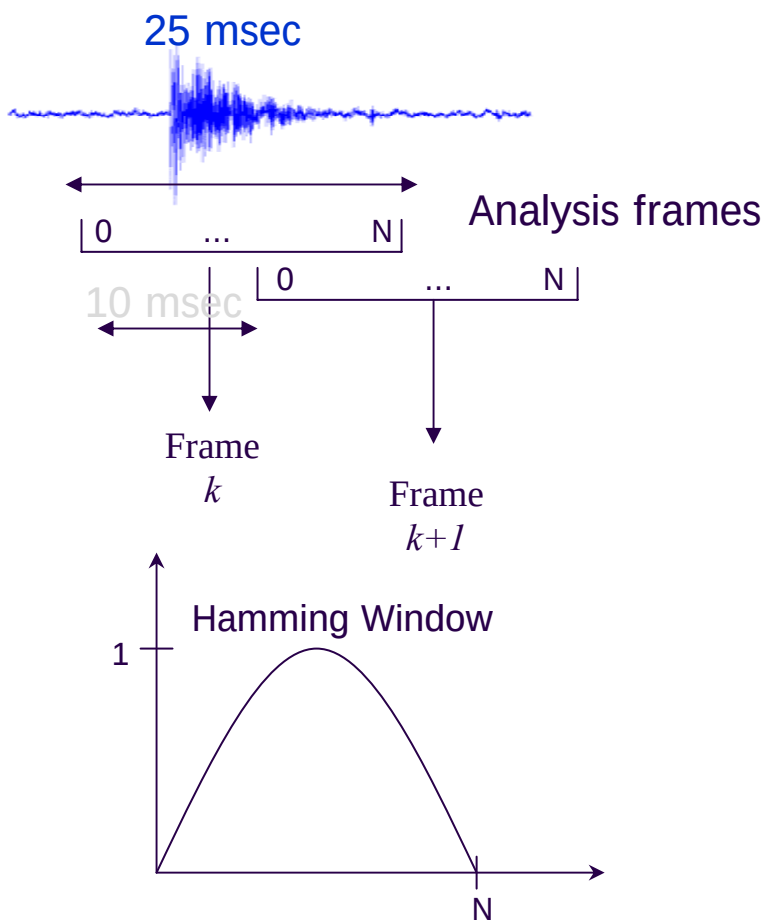
- Processing is applied in **short-duration "frames"**, typically of a 25 msec length, with some overlap (typically 10 ms). Signal in frame is $\{s_n, n=1,\ldots,N\}$.

- The following are applied on frame:
  - **DC signal removal**.
  - **Signal pre-emphasis**:

$$s'_n = s_n - 0.97\, s_{n-1}$$

  - **Hamming windowing**:

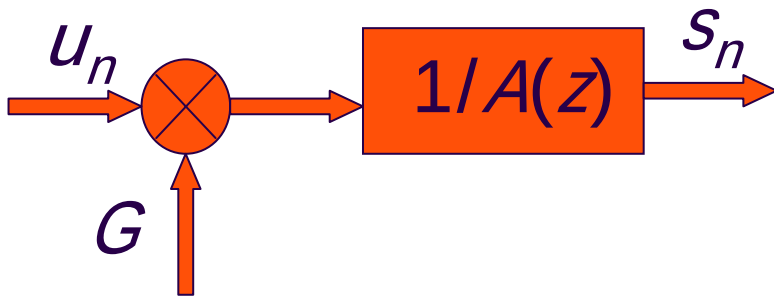
$$s'_n = \left\{0.54 - 0.46\cos\left(\frac{2\pi(n-1)}{N-1}\right)\right\} \times s_n$$



Linear prediction (LP) speech analysis

- Vocal tract is modeled as an all-pole filter, driven by an excitation term:

$$s_n = \sum_{i=1}^{p} a_i s_{n-i} + G u_n$$



- LP analysis aims to minimize the prediction error: and thus is a MSE problem.

$$\mathrm{E}\left[s_n - \sum_{i=1}^{p} a_i s_{n-i} \right]^2$$

- Efficiently solved using **Durbin**'s algorithm for inverting the *pxp* autocorrelation equation system. Results in **LPC** (linear prediction coefficients): $a_1, a_2, ..., a_p$.
- Superior ASR performance is achieved using the **LPCC** (LP cepstral coefficients):

$$c_m = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c_k a_{m-k}; \; m = 1,..., M \le p$$

- Typically, *M* =12, *p* =14.

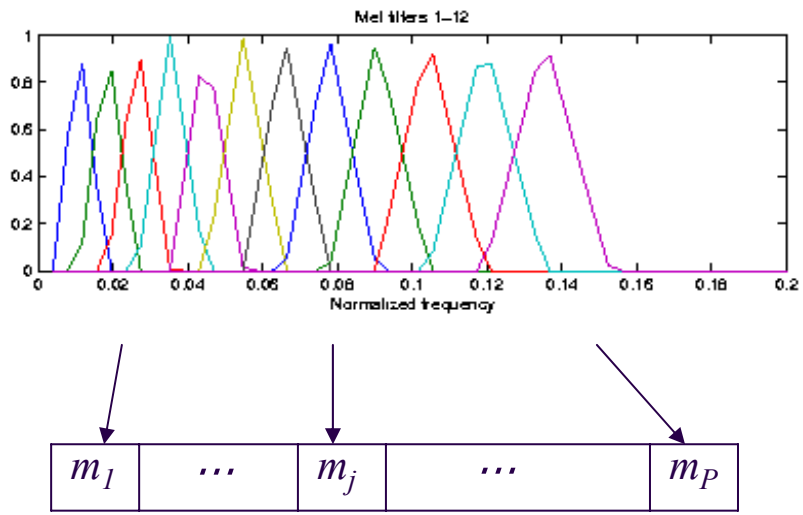
Filter-bank speech analysis

- Computes speech energy in a number of bands, after suitable band-pass filtering.
- Due to human perception, bands are non-uniform. Typically, triangular filters are used, with uniform spacing along the **mel** frequency scale:

$$\mathrm{mel}(f) = 2595 \log_{10}(1 + \frac{f}{700})$$

- **Mel-frequency cepstral coefficients** (**MFCC**) are obtained by a discrete cosine transform of the log filterbank amplitudes m_j .

$$c_i = \sqrt{\frac{2}{p}} \sum_{j=1}^{p} m_j \left(\frac{\pi\, i}{p} (j - 0.5) \right)$$



Feature post-processing

- **Weighting** of the LPC coefficients (also known as “cep-liftering”):

$$c'_n = \left(1 + \frac{L}{2}\sin\frac{\pi n}{L}\right) \times c_n \qquad (\text{e.g., } L = 22).$$

- **Augmentation** of the feature vector (LPCC or MFCC) by log of signal energy:

$$E = \log \sum_{n=1}^{N} s_n^2$$

- **Normalization** by subtracting $E_{\max}$-1 for energy, mean for other features.
- Inclusion of “**dynamic**” information, by augmenting features with first and second derivatives, or “learning” dynamic features as a dimensionality-reduction projection of a **concatenation** of features from consecutive, neighboring frames.
- Feature transformations (rotation) to other spaces for better statistical modeling (**de-correlation**).

Automatic Speech Recognition (ASR)

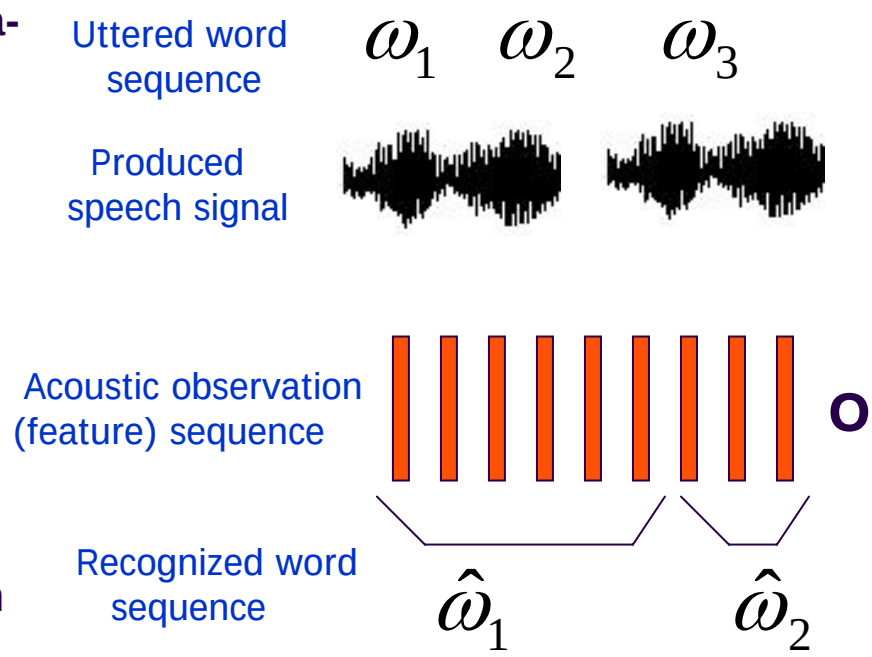
- **Statistical approach** to ASR uses **maximum a-posteriori** (MAP) estimation to obtain optimal word sequence:

$$\hat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} \Pr[\boldsymbol{\omega} \mid \mathbf{O}]$$

  - **"Hidden"** words are **partially observed** through sequence of acoustic features.
- **Two models are needed:**
  - Prior probability of word sequences (**language model**).
  - Generative model of acoustic features from word sequence (**acoustic model**).

$$\Pr[\boldsymbol{\omega} \mid \mathbf{O}] \propto \Pr[\mathbf{O} \mid \boldsymbol{\omega}] \Pr[\boldsymbol{\omega}]$$

AM LM



Hidden Markov models (HMMs)

- HMMs are popular generative models for time series of observations. They are characterized by following:
- States: $C=\{1,2,\ldots,N\}$. Denote q_t state at t.
- Initial state distribution:

$$\boldsymbol{\pi} = \{\pi_i = \Pr[q_1 = i],\ i = 1,\ldots,N\}$$

- State transition probabilities:

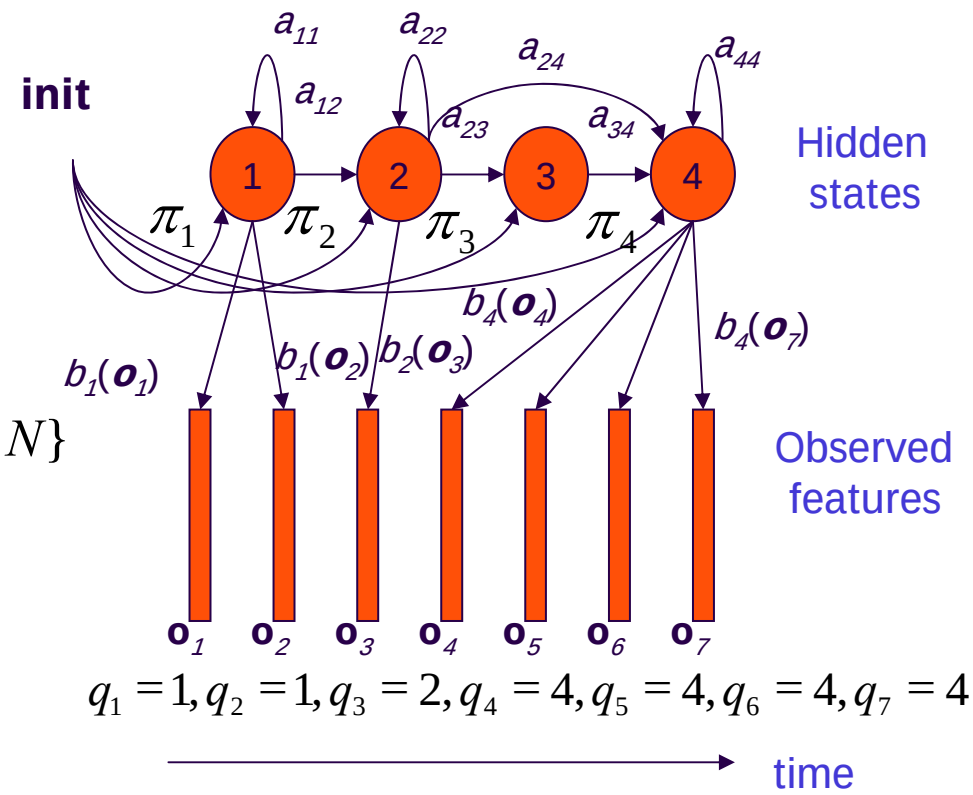
$$\mathbf{a} = \{a_{ij} = \Pr[q_{t+1} = j \mid q_t = i],\ i, j = 1,\ldots,N\}$$

- State conditional observation probability:

$$\mathbf{b} = \text{parametric representation of}$$

$$\{b_j(\mathbf{o}_t) = \Pr[\mathbf{o}_t \mid q_t = j],\ j = 1,\ldots,N\}$$

- Thus, HMM parameters are: $\boldsymbol{\theta} = [\boldsymbol{\pi}, \mathbf{a}, \mathbf{b}]$



HMMs - Cont.

The class-conditional observation probabilities **b** can be:

- **Discrete**, in case that the observation vectors are drawn from a finite set. This can be achieved by vector quantization of the feature space (codebook of size *K*):

$$\mathbf{b} = \{b_j(k) = \Pr[\mathbf{o}_t \approx \mathbf{v}_k \mid q_t = j],\ j = 1,\ldots,N, k = 1,\ldots,K\}$$

- **Continuous**, typically considered as a mixture of multi-dimensional Gaussians:

$$b_j(\mathbf{o}_t) = \sum_{m=1}^{M_j} c_{jm} N(\mathbf{o}_t; \boldsymbol{\mu}_{jm}, \mathbf{U}_{jm}), \quad j = 1,\ldots,N$$

where the d-dimensional Gaussians are

$$N_d(\mathbf{o}; \boldsymbol{\mu}, \mathbf{U}) = \frac{1}{\sqrt{(2\pi)^d |\mathbf{U}|}} \exp\left[-\frac{1}{2}(\mathbf{o}-\boldsymbol{\mu})^{\mathrm{T}} \mathbf{U}^{-1}(\mathbf{o}-\boldsymbol{\mu})\right]$$

and the mixture weights satisfy: $\sum_{m=1}^{M_j} c_{jm} = 1, \quad c_{jm} \geq 0, \quad j = 1,\ldots,N, \quad m = 1,\ldots,M_j$

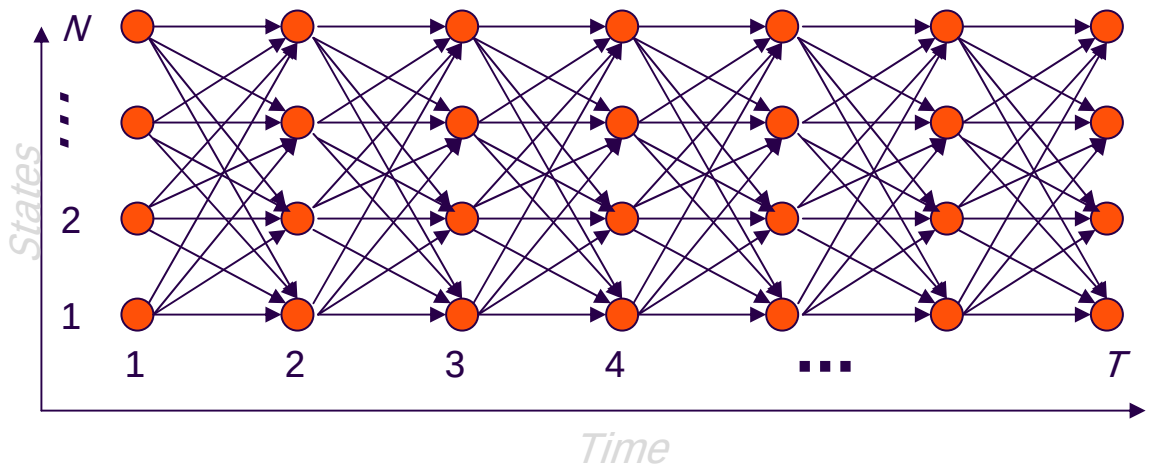
Parameters are then: $\mathbf{b} = \{c_{jm}, \boldsymbol{\mu}_{jm}, \mathbf{U}_{jm},\ j = 1,\ldots,N,\ m = 1,\ldots,M_j\}$

HMMs - Cont.

- **The three basic HMM problems.** Recall:
  - Observation sequence of duration T: $\mathbf{O}=[\mathbf{o}_1,\mathbf{o}_2,\ldots,\mathbf{o}_T]$.
  - State sequence: $\mathbf{q}=[q_1,q_2,\ldots,q_T]$.
  - Model parameters: $\boldsymbol{\theta}=[\boldsymbol{\pi},\mathbf{a},\mathbf{b}]$
- **Problem 1:** Given **O** and model parameters, how do we compute $\Pr(\mathbf{O}\mid\boldsymbol{\theta})$?
  - "Evaluation" of model fit to the data.
  - Solved by the "forward" or "backward" procedure.
- **Problem 2:** Given **O** & model parameters, what is the optimal state seq. **q**?
  - Uncovers the "hidden" states – used in **recognition**!
  - Solved by the **Viterbi** algorithm.
- **Problem 3:** What are the model parameters that optimize $\Pr(\mathbf{O}\mid\boldsymbol{\theta})$?
  - This is the **maximum-likelihood parameter estimation** problem.
  - Solved by the **forward-backward algorithm** (or Baum-Welch), an instance of the expectation-maximization (EM) procedure.

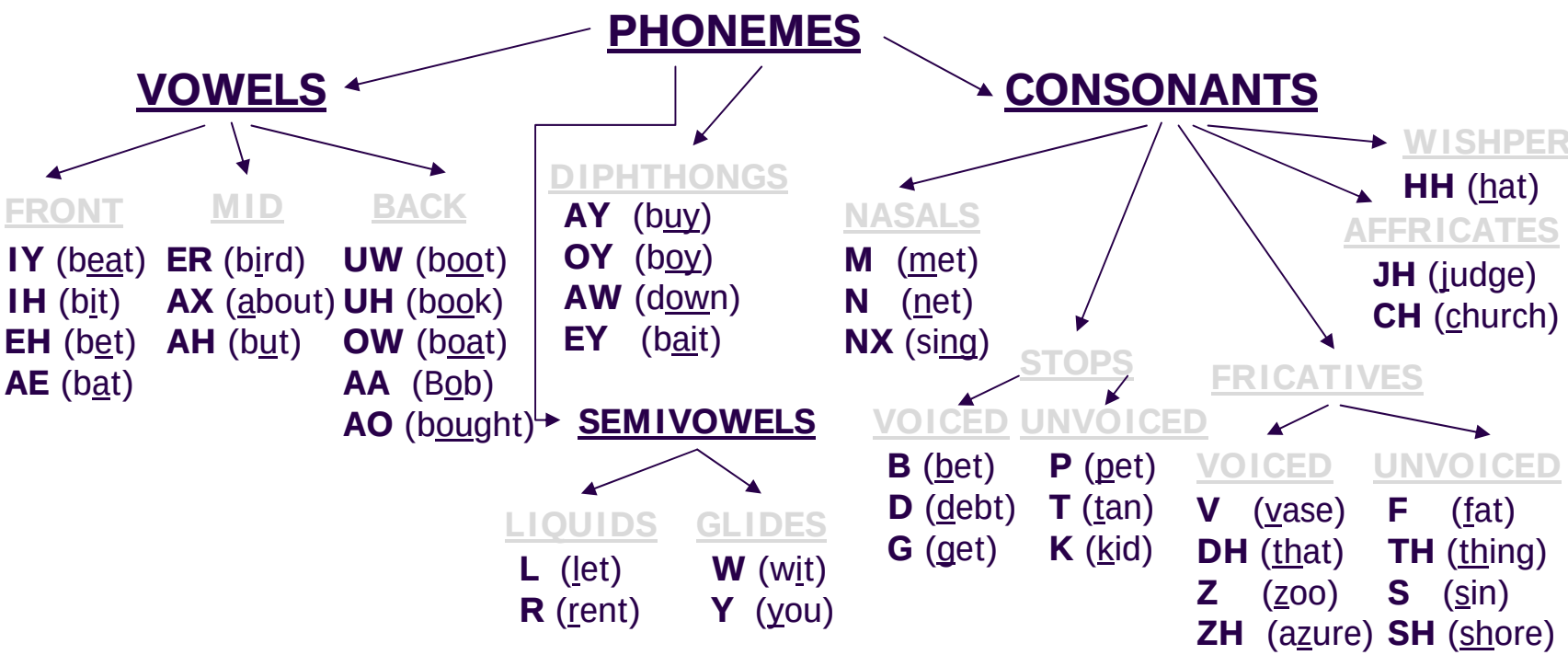
HMMs - Cont.

- Brute force solution to these problems is exponential on T , i.e., $O(TN^T)$.
- Luckily, dynamic programming solutions exist!
- They utilize partial computations on the 2-D **lattice** of TxN states in time.
- Complexity of resulting algorithms is $O(N^2T)$.



Acoustic modeling using HMMs

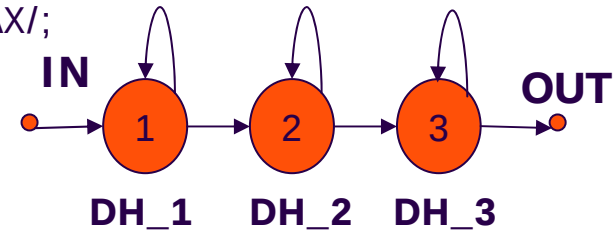
- **Phonemes:** Basic units that describe how speech conveys linguistic information.
- In statistical based ASR (especially large-vocabulary), they constitute the basic **HMM** units.
- **Basic grouping** of the phonemes used in American English (ARPAbet upper case version).



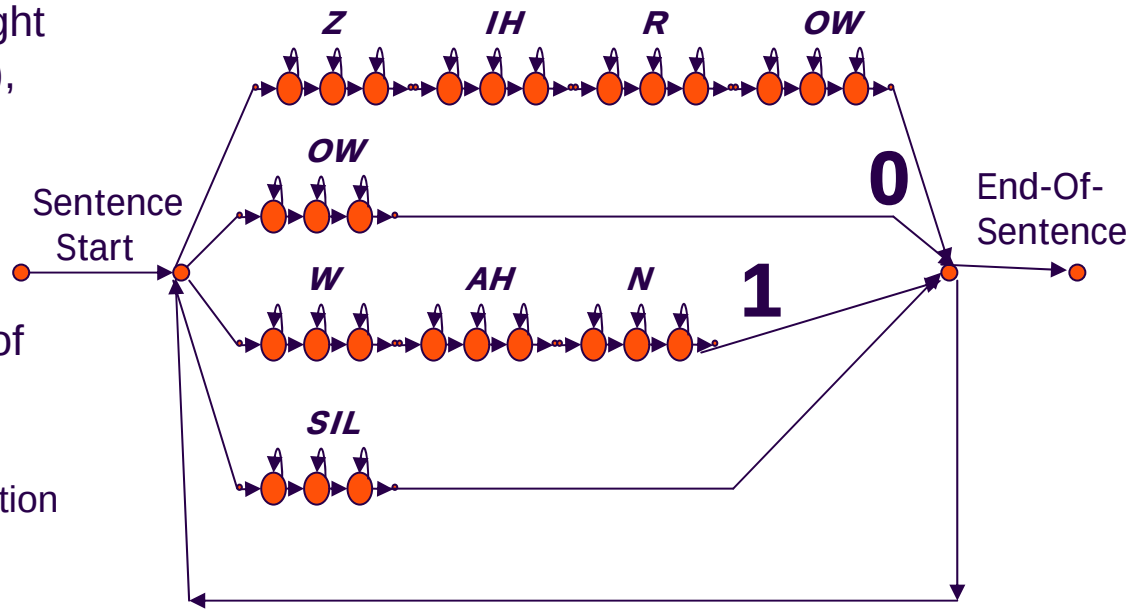
Acoustic modeling using HMMs – Cont.

- Words are modeled as phone sequences (phonetic dictionary).
- Phones are typically modeled as 3-state **left-2-right** HMMs.
- To improve performance, states have **context-dependent** observation pdfs. Contexts are clusters of left and right phonetic sequences (1-5 in length), obtained by a **decision tree**.
- Training and recognition is then performed utilizing the HMM algorithms discussed previously (problems 2 and 3), on a network of HMM states, composed by words, phones, and sub-phonetic units.
  - Example of 0-1 connected recognition using context-independent units.

THE = /DH IY/;/DH AX/;
THEME= /TH IY M/;



HMM for phone **DH**



Language modeling (LM)

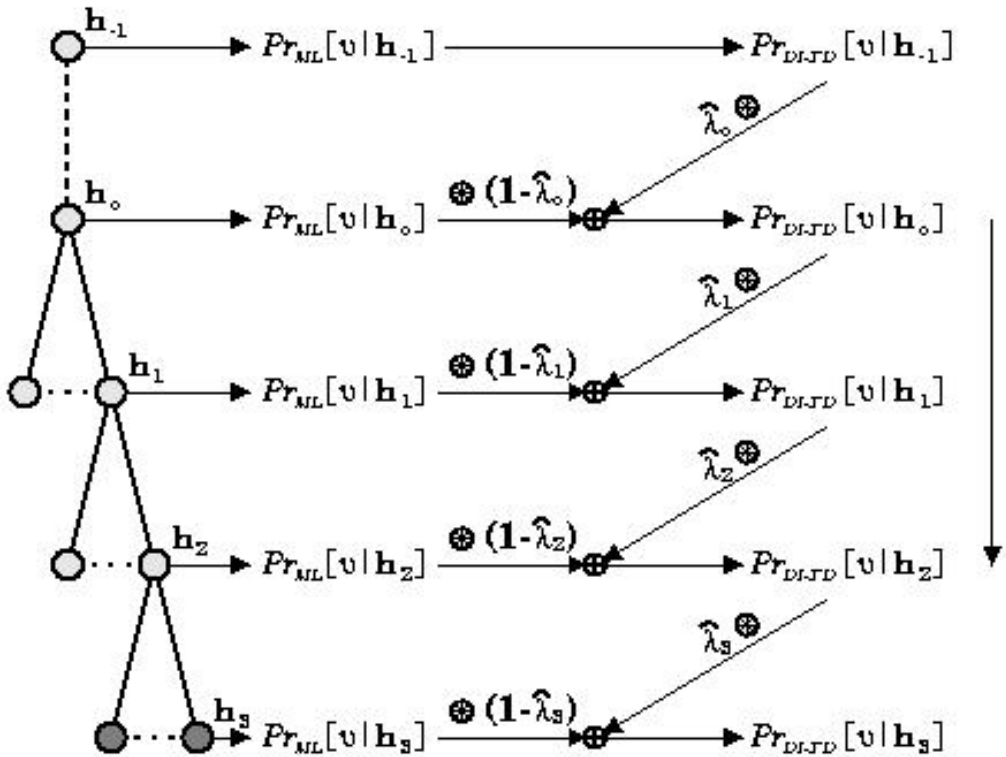
- Aims to provide prior probability for word sequences, thus reducing the "uncertainty" (**perplexity**) in ASR.
- Assumes **causal** model:

$$\Pr[\boldsymbol{\omega}_1^m] = \prod_{i=1}^{m} \overline{\Pr}[\omega_i \mid \boldsymbol{\omega}_1^{i-1}]$$

- Approximation using finite "**history**":

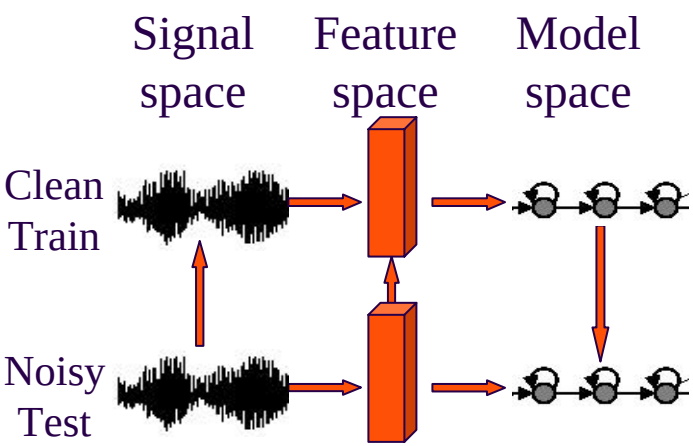
$$\overline{\Pr}[\omega_i \mid \boldsymbol{\omega}_1^{i-1}] \approx \Pr[\omega_o \mid \Phi(\omega_{-1},\ldots,\omega_{-n+1})]$$
$$= \Pr[\upsilon \in \mathrm{Voc} \mid \mathbf{h}_{n-1}].$$

- Two problems:
  - **History classification**: n-grams.
  - **Probability estimation**: ML with parameter "smoothing" on held-out data (deleted interpolation, back-off).



LM probability "smoothing" by top-down deleted interpolation

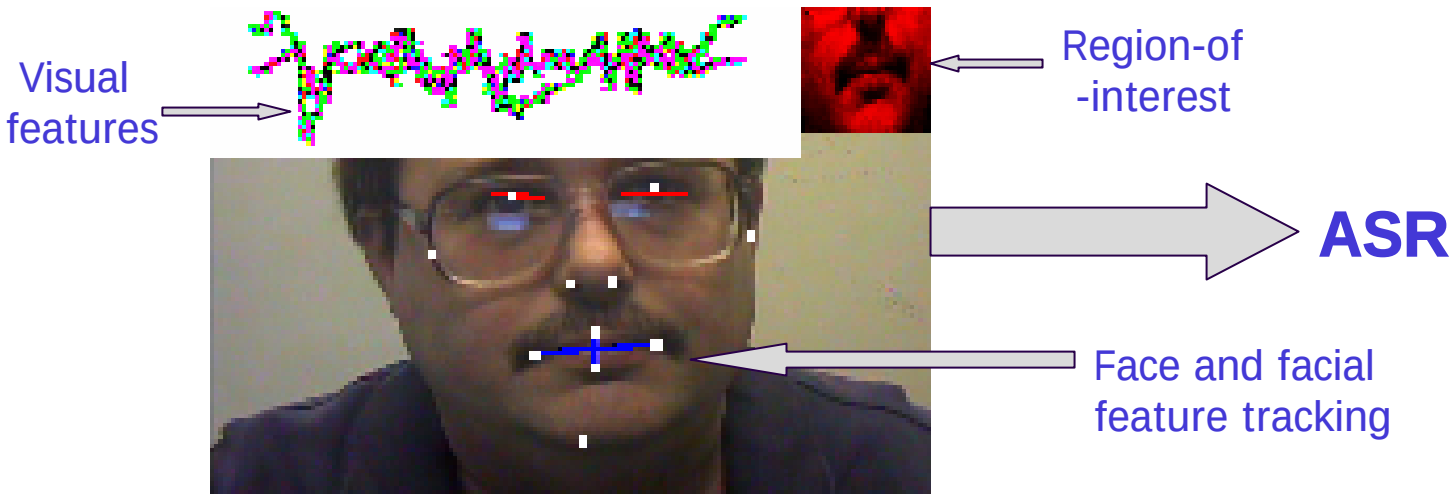
ASR robustness / adaptation



- Typically, ASR performance **degrades** in noisy environments, and **mismatched** conditions and **unseen** speakers in training (lack of **robustness**).
  - Performance can be improved by **noise compensation**, or in case available sample of the new condition / subject, by **adaptation**.
- Three categories of techniques:
  - **Signal** space, **feature** space, **model** based.
  - E.g.: Spectral subtraction, Wiener filtering, vocal tract length normalization (**VTLN**), noise adaptive prototypes, parallel model combination (**PMC**), maximum-a-posteriori adaptation (**MAP**), maximum likelihood linear regression (**MLLR**), speaker-adaptive training (**SAT**), feature-space MLLR (**FMLLR**), etc.
- These techniques are moderately only successful. Lack of robustness remains an issue and motivates the use of the visual modality in ASR!

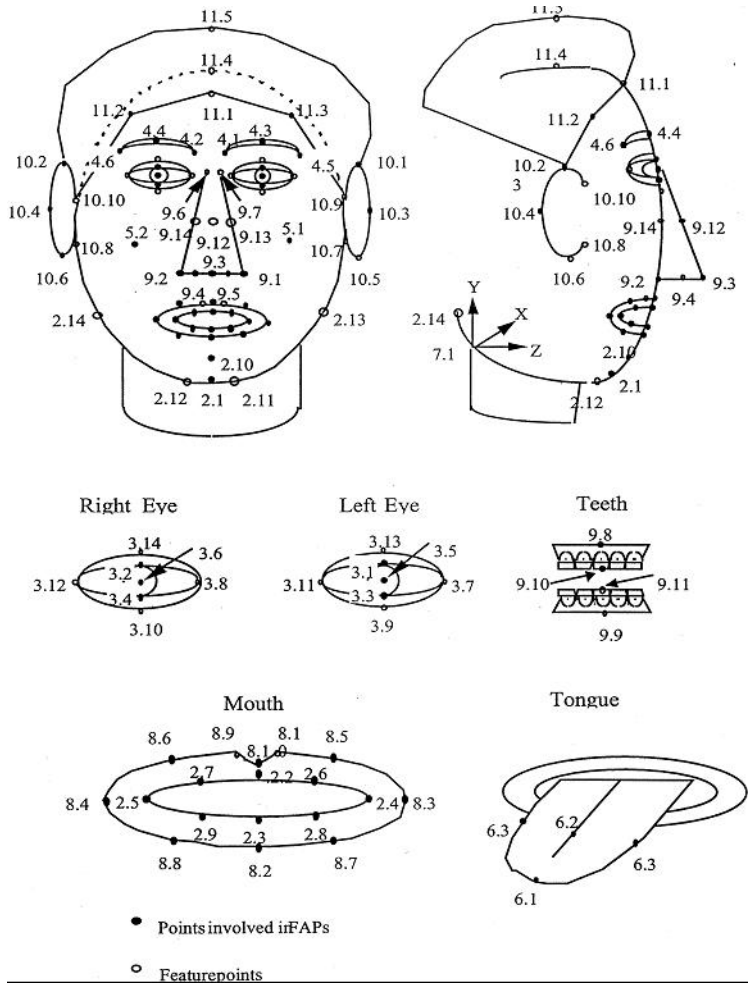
Visual signal analysis of human speech

- **Main questions:**
  A. Where is the talking face in the video?
  B. How to extract the speech informative section of it?
  C. What visual features to extract?
  D. How valuable are they for recognizing human speech?
  E. How do video degradations affect them?



Face and facial feature tracking.

- **Main question:** Is there a ***face*** present in the video, and if so, where? Need:
  - ***Face*** detection.
  - ***Head pose*** estimation.
  - ***Facial feature*** localization (mouth corners). See for example **MPEG-4** facial activity parameters (**FAPs**).
  - Lip/face shape (contour).
- Successful face and facial feature tracking is a prerequisite for incorporating audio-visual speech in HCI.
- In this section, we discuss:
  - **Appearance based** face detection.
  - **Shape** face estimation.



Appearance-based face detection

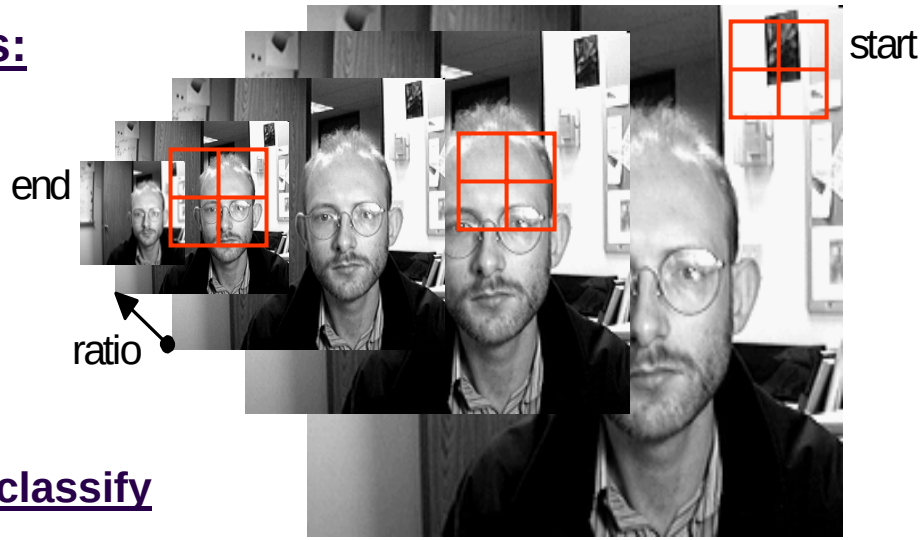
TWO APPROACHES:

- **Non-statistical**:
  - Use **image processing** techniques to detect presence of typical face characteristics (mouth edges, nostrils, eyes, nose), e.g.: Low-pass filtering, edge detection, morphological filtering, etc. Obtain candidate regions of such features.
  - **Score** candidate regions based on their relative location and orientation.
  - Improve robustness by using additional information based on **skin-tone** and **motion** in color videos.

From: Graf, Cosatto, and Potamianos, 1998

Appearance-based face detection – Cont.

- Standard **statistical** approach – **steps:**
  - View face detection as a **2-class** classification problem (into faces/ non-faces).
  - Decide on a "**face template**" (e.g., 11x11 pixel rectangle).
  - Devise a **trainable** scheme to "score"/**classify** candidates into the 2 classes.

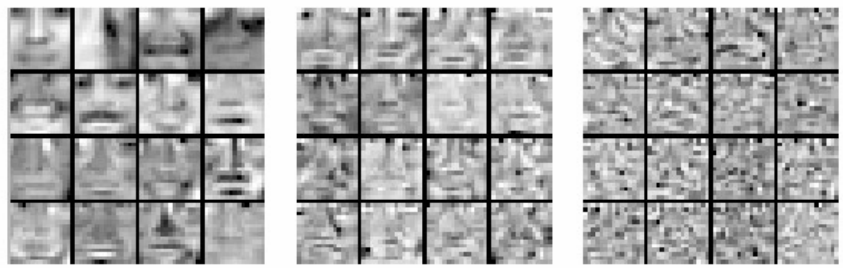


  - **Search** image using a pyramidal scheme (over locations, scales, orientations) to obtain set of face **candidates** and **score** them to detect any faces.
  - Can **speed-up** search by eliminating face candidates in terms of **skin-tone** (based on color information on the *R,G,B* or transformed space), or location/scale (in the case of a video sequence). Use thresholds or statistics.

Appearance-based face detection – Cont.

Statistical face models (for face "vector" $\mathbf{x}$).

- ***Fisher discriminant*** detector (Senior, 1999).
  - Also known as **linear discriminant analysis** – **LDA**
  - One-dimensional projection of 121-dimensional vector $\mathbf{x}$: $y_F = \mathbf{P}_{1 \times 121}\,\mathbf{x}$
  - Achieves best discrimination (separation) between the two classes of interest in the projected space; P is trainable on basis of annotated (face/non-face) data vectors.
- ***Distance from face space*** (DFFS).
  - Obtain a **principal components analysis** (**PCA**) of the training set.
  - Resulting projection matrix $\mathbf{P}_{dx121}$ achieves best information "compression".
  - Projected vectors $\mathbf{y} = \mathbf{P}_{dx121}\,\mathbf{x}$ have a DFFS score: $\mathrm{DFFS} = \left\| \mathbf{x} - \mathbf{y}\,\mathbf{P}^{\mathrm{T}} \right\|$
- Combination of two can score a face candidate vector: $y_F - \mathrm{DFFS} \begin{matrix} > \\ < \end{matrix} \begin{matrix} Face \\ Non-Face \end{matrix}$ th

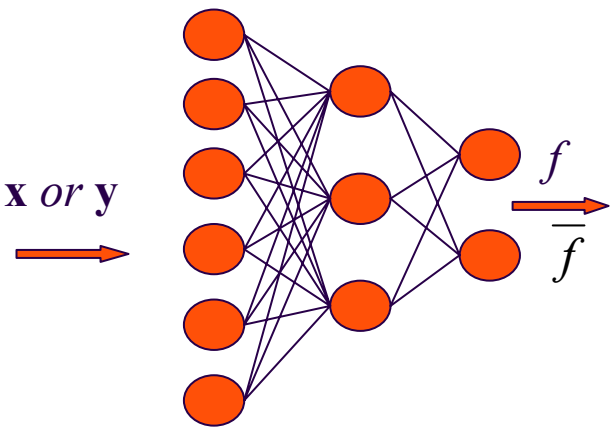


Example PCA eigenvectors

Appearance-based face detection – Cont.

Additional statistical face models:

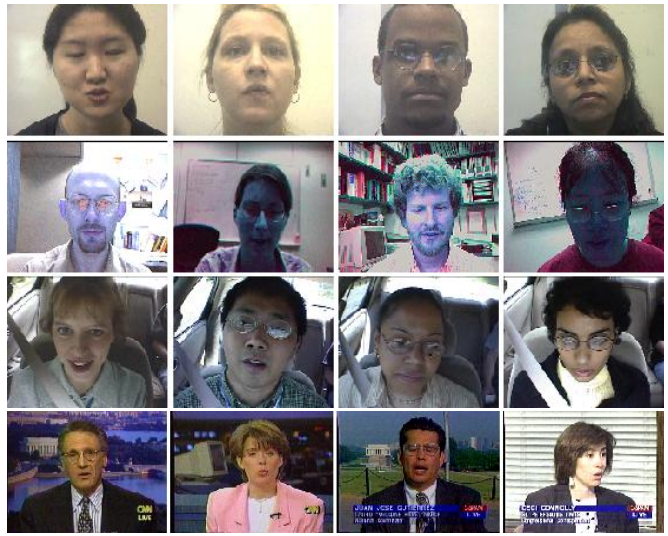
- ***Gaussian mixture*** classifier (**GMM**):
  - Vector $\mathbf{y}$ is obtained by a dimensionality reduction projection of x (PCA, or other image compression transform), $\mathbf{y} = \mathbf{P}\,\mathbf{x}$.
  - Two GMMs are used to model: $\Pr(\mathbf{y} \mid c) = \sum_{k=1}^{K_c} w_{k,c} N(\mathbf{y}, \mathbf{m}_{k,c}, \mathbf{s}_{k,c}), \quad c \in \{f, \overline{f}\}$
  - GMM means/variances/weights are estimated by the EM algorithm.
  - Vector $\mathbf{x}$ is scored by likelihood ratio: $\Pr(\mathbf{y} \mid f) / \Pr(\mathbf{y} \mid \overline{f})$
- ***Artificial neural network*** classifier (**ANN –** Rowley et al., 1998).
- ***Support vector machine*** classifier (**SVM** – Osuna et al., 1997).

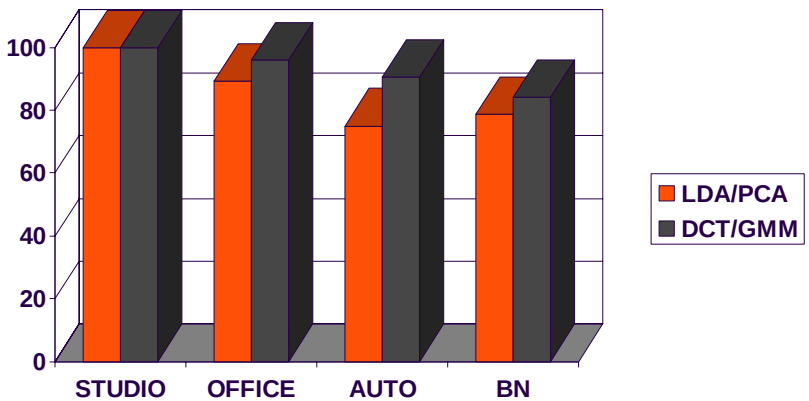


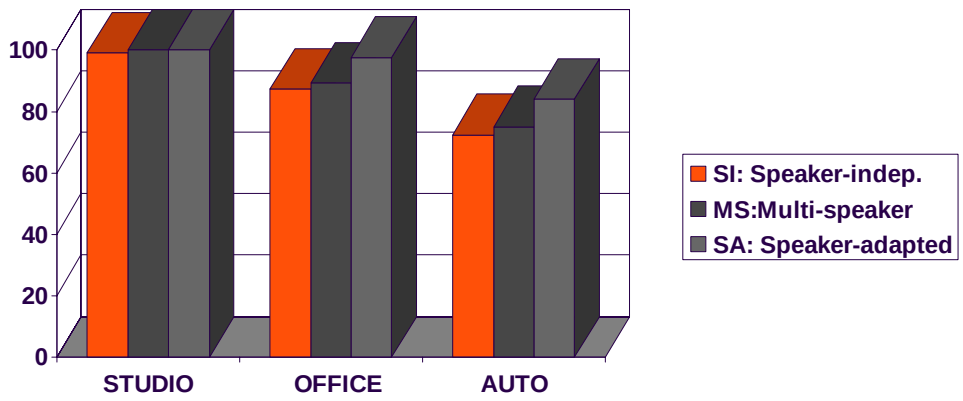
Appearance-based face detection

Face detection experiments:

- Results on 4 in-house **IBM databases**, recorded in:
  - **STUDIO:** Uniform background, lighting, pose.
  - **OFFICE:** Varying background and lighting.
  - **AUTOMOBILES:** Extreme lighting and head pose change.
  - **BROADCAST NEWS:** Digitized broadcast videos, varying head-pose, background, lighting.
- **Face detection accuracy:**



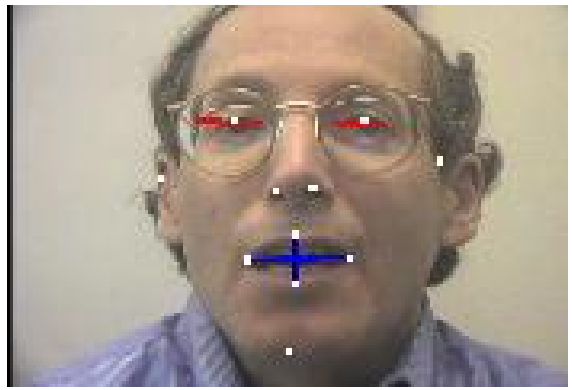




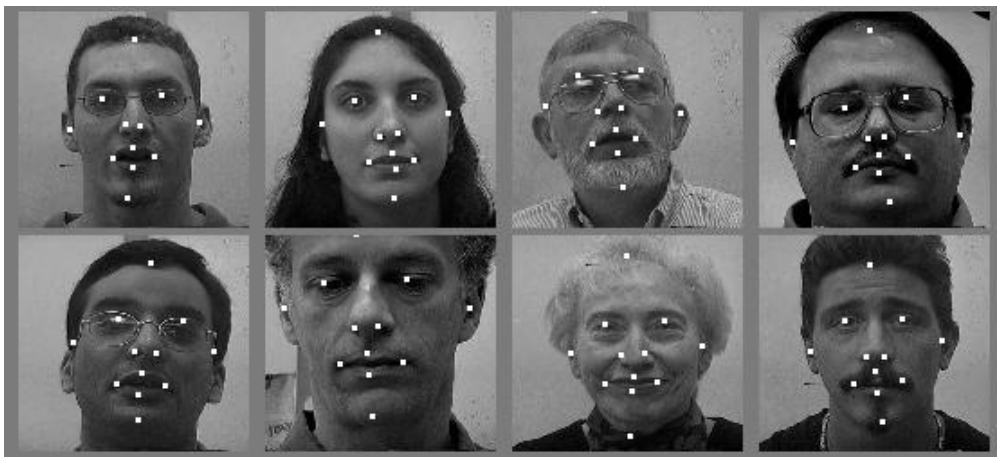
Appearance-based face detection – Cont.

From faces to facial features:

- Facial features are required for visual speech applications!
- Feature detection is similar to face detection:
  - Create *individual* facial feature *templates*. Feature vectors can be scored using trained Fisher, DFFS, GMMs, ANN, etc.
  - *Limited* search, due to *prior* feature location information.
- *Examples of detected facial features:* Remains challenging under varying lighting and head pose variations.



STUDIO



AUTOMOBILE

Face shape & lip contour extraction

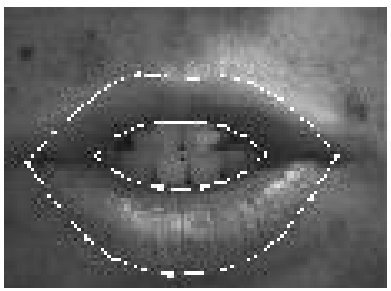
Four popular methods for lip contour extraction:

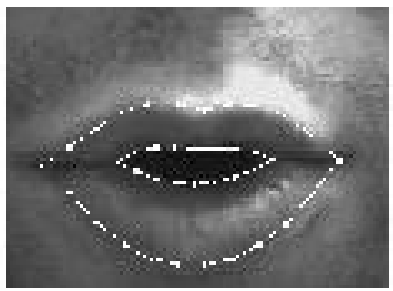
- **Snakes** (Kass, Witkin, Terzopoulos, 1988):
  - A snake is an open or closed ***elastic curve*** defined by ***control points***.
  - An ***energy function*** of the control points and the image / or edge map values is ***iteratively optimized***.
  - Correct snake ***initialization*** is crucial.

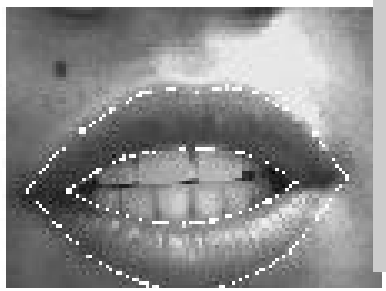
- **Deformable templates** (Yuille, Cohen, Hallinan, 1989):
  - A template is a ***geometric model***, described by ***few parameters***.
  - Minimizing a ***cost function*** (which is the sum of curve and surface integrals) matches the template to the lips.
  - Typically two or more ***parabolas*** are used as the template.

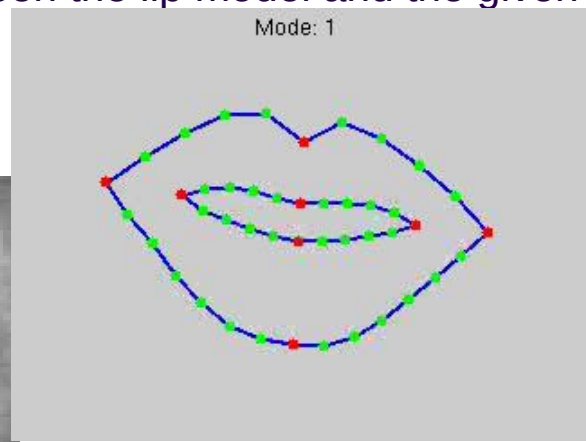
Face shape & lip contour extraction – Cont.

- **Active shape models** (Cootes, Taylor, Cooper, Graham, 1995):
  - A ***point distribution model*** of the lip shape is built.
  - First, a set of images with ***annotated*** (marked) lip contours is given.
  - A ***PCA*** based model of the vector of the lip contour point coordinates is obtained.
  - Lip tracking is based on ***minimizing*** a distance between the lip model and the given image.







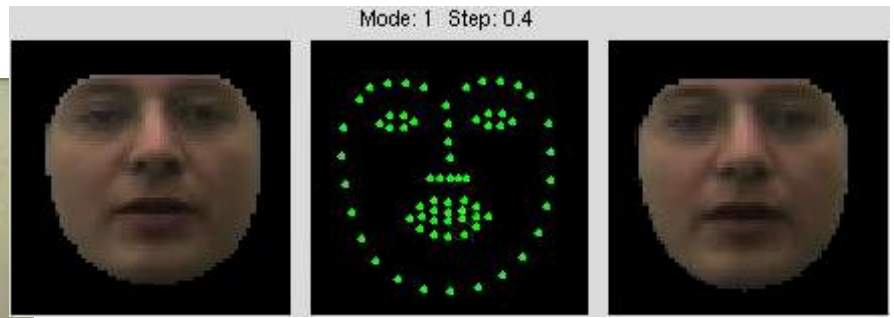


From: Luettin, Thacker, and Beet, 1996.

Face shape & lip contour extraction – Cont.

- **Active appearance models** (**AAM**s- Cootes, Walker, Taylor, 2000):
  - In addition to shape, it also considers a model of face texture (appearance).
  - A ***PCA*** based model of the R,G,B pixel values of normalized face regions is obtained.
  - Thus, a face is ***encoded*** by means of its mean shape, appearance, and the PCA coefficients of both.
  - Facial shape (and face!) detection becomes an ***optimization*** problem where the joint shape/appearance parameters are iteratively obtained, by minimizing a residual error.

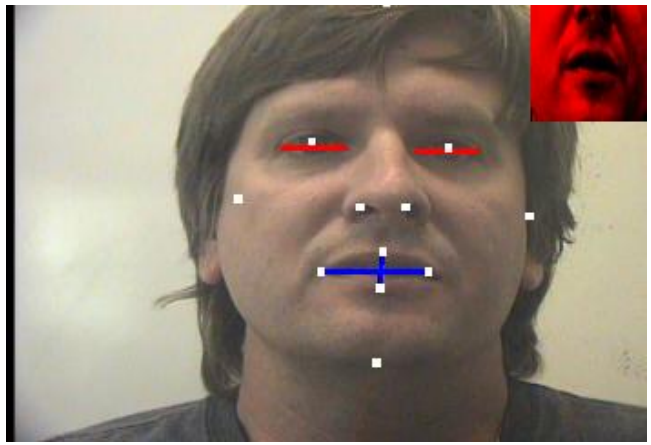
AAM tracking on IBM "studio" data (credit: I. Matthews)

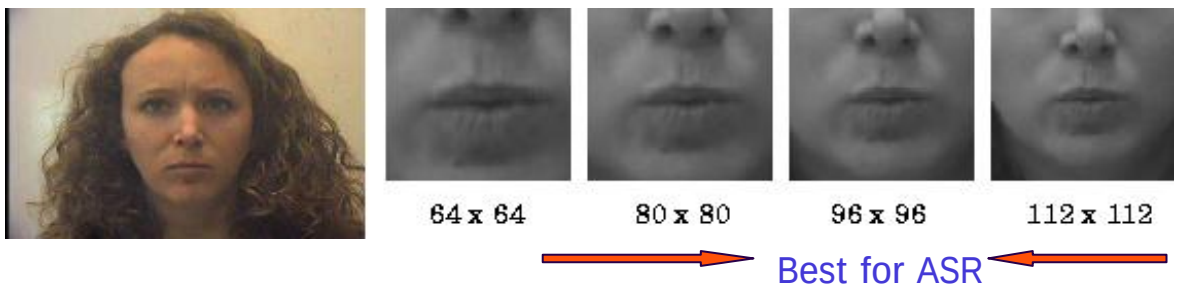


AAM modes trained on IBM data

Region-of-interest for visual speech

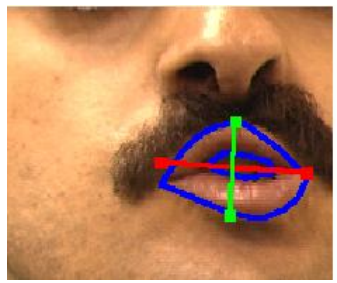
- **Region-of-interest (ROI):**
  - Assumed to contain “all” visual speech information.
  - Key to appearance based visual features, described in III.C.
  - Can be used to limit search of “expensive” shape tracking.
  - Typically is a rectangle containing the mouth, but could be circle, lip profiles, etc.
- **ROI extraction:**
  - Smooth mouth center, size, orientation estimates using median or Kalman filter.
  - Extract size and intensity normalized (e.g., by histogram equalization) mouth ROI.
  - Including parts of “beard region” is beneficial to ASR.
  - ROI “quality” is function of the face tracking accuracy.

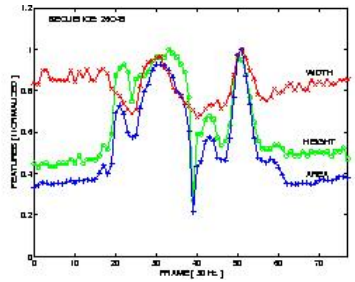




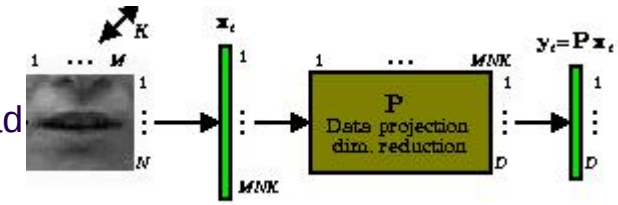
Visual speech features

- What are the right visual features to extract from the ROI?
- Three types of / approaches to feature
- ***Lip- and face-contour (shape) based:***
  - Height, width, area of mouth.
  - Moments, Fourier descriptors.
  - Mouth template parameters.

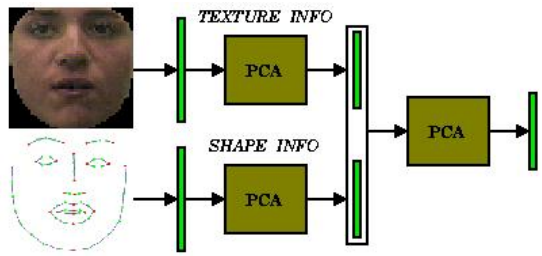




- ***Video pixel (appearance) based features:***
  - Lip contours *do not* capture oral cavity information!
  - Use compressed representation of mouth ROI instead
  - E.g.: DCT, PCA, DWT, whole ROI.



- ***Joint shape and appearance features:***
  - Active appearance models.
  - Active shape models.

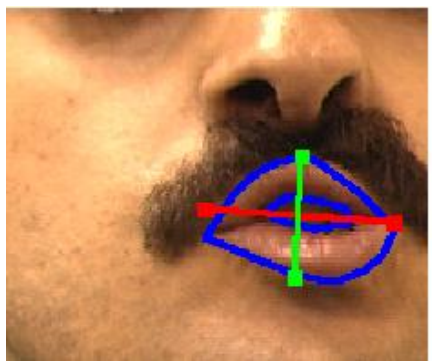


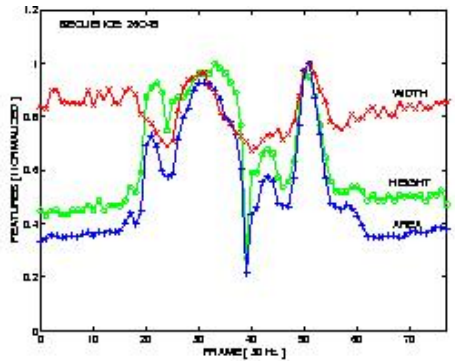
Shape based visual features

- **Geometric lip contour features:** Assume that lip contour (points) are available and are properly normalized using an affine transform (to compensate for head pose and speaker specifics).
- **Feature extraction:**
  - Contour is denoted by $C = \{(x, y)\}$
  - Lip-interior membership function:

$$f(x,y) = \begin{cases} 1, & if\ (x,y) \in C \cup C_{interior} \\ 0, & otherwise \end{cases}$$

  - Some "sensible" lip-features are then:
    - *Height:* $\mathbf{h} = \max_x \sum_y f(x,y)$
    - *Width:* $\mathbf{w} = \max_y \sum_x f(x,y)$
    - *Area:* $\mathbf{a} = \sum_x \sum_y f(x,y)$
    - *Perimeter:* $\mathbf{p} = \sum_i d[C_i, \quad _+]$





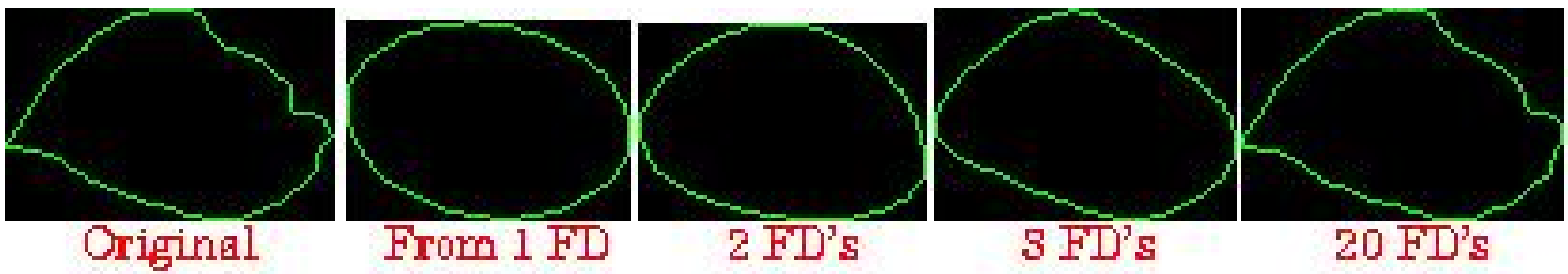
Shape based visual features – Cont.

- **Lip contour Fourier descriptors:** $C = \{(x(t), y(t)) : t \in [0, T]\}$
  - Contour parametrization (encoding):
  - Obtain Fourier series expansion of $\{x(t)\}$ and $\{y(t)\}$:

$$x(t) = A_0 + \sum_{n=1}^{\infty} A_n \cos\frac{2n\pi t}{T} + B_n \sin\frac{2n\pi t}{T}$$

$$y(t) = C_0 + \sum_{n=1}^{\infty} C_n \cos\frac{2n\pi t}{T} + D_n \sin\frac{2n\pi t}{T}$$

  - Use as visual features: $FD_n = \sqrt{A_n^{*2} + B_n^{*2} + C_n^{*2} + D_n^{*2}}$, $n \geq 2$.



Shape based visual features – Cont.

- **Lip image moments:**
  - Create 2D image *f* from contours:

  - Moment functions:

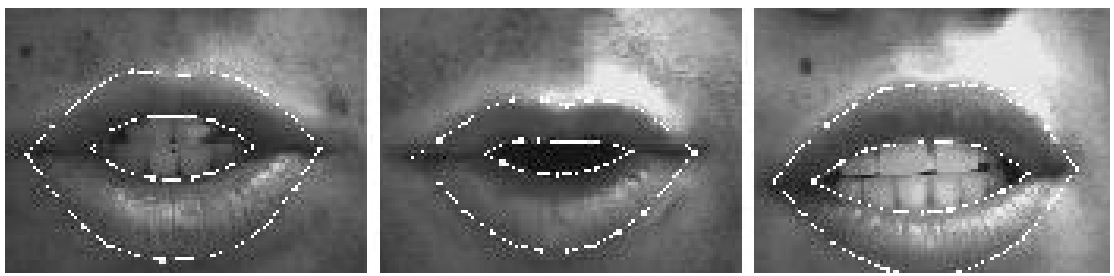
$$m_{pq} = \sum_{x,y} (x - \bar{x})^p (y - \bar{y})^q f(x,y), \quad \text{where}$$

$$\bar{x} = \frac{\mu_{10}}{\mu_{00}}, \qquad y = \frac{\mu_{01}}{\mu_{00}}, \qquad \mu_{pq} = \sum_{x,y} x^p y^q f(x,y).$$

  - Note: Appropriate normalization of moment functions makes them invariant to affine image transforms.

Shape based visual features – Cont.

- **Lip model based features**: Various lip models can be used for lip contour tracking. The resulting lip contour points can be used to derive geometric features, or alternatively, in the case of:
- ***Snakes*** :
  - Use distances or other function of snake *control points* as features.
- ***Deformable templates :***
  - Use the *parabola parameters*.
- ***Active shape models*** :
  - Use the PCA coefficients corresponding to the lip shape as features.



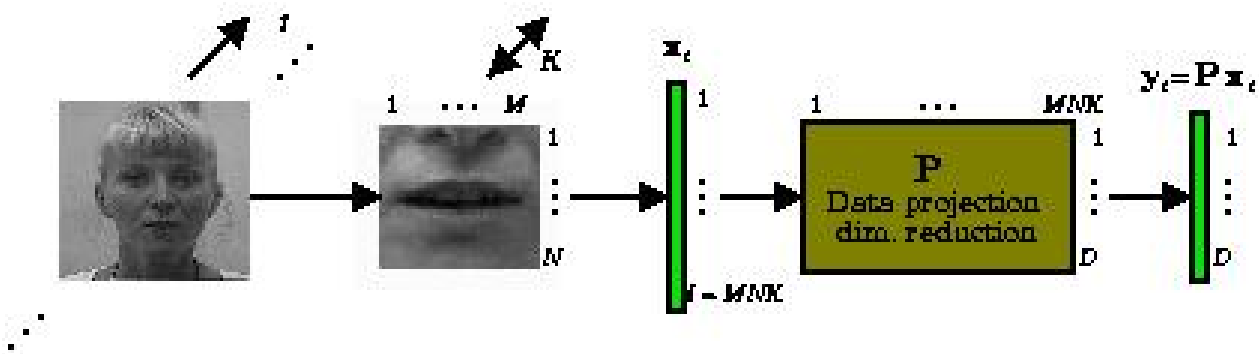
Appearance based visual features

- **Main idea:** Lip contours fail to capture speech information from the oral cavity (tongue, teeth visibility, etc.). Instead, use a compressed representation of the mouth region-of-interest (ROI) as features.
- 2D or 3D **ROI vector** consists of d=*MNK* pixels, lexicographically ordered in:

$$\mathbf{x}_t \leftarrow \{V_t(m,n,k) : m_t - \lfloor M/2 \rfloor \le m < m_t + \lceil M/2 \rceil ,$$
$$n_t - \lfloor N/2 \rfloor \le n < n_t + \lceil N/2 \rceil ,$$
$$k_t - \lceil K/2 \rceil \le k < k_t + \lceil K/2 \rceil \}.$$

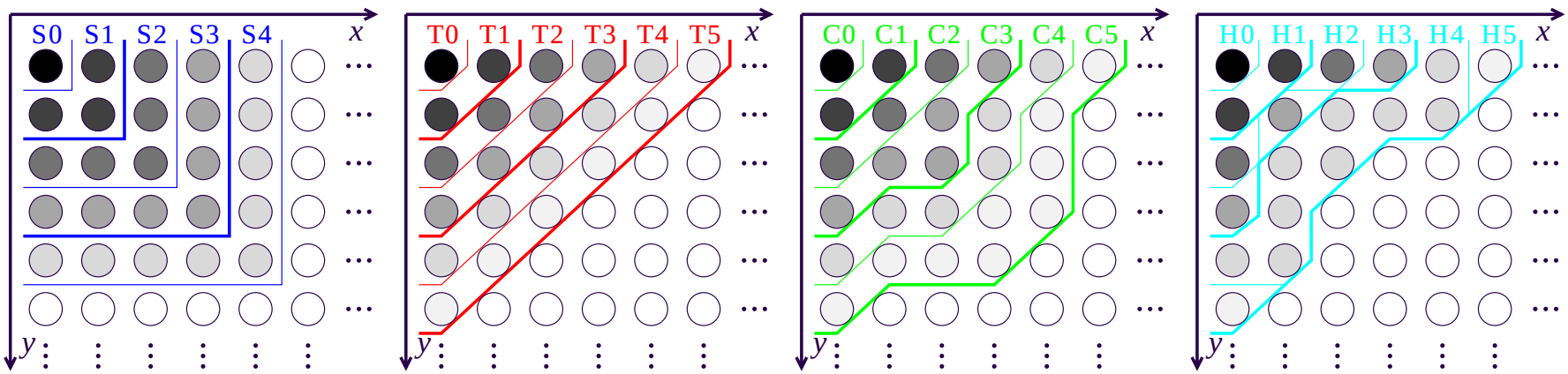
- Seek dimensionality reduction transform:

$$\mathbf{y}_t = \mathbf{P}\ \mathbf{x}_t , \quad \text{with}$$
$$\mathbf{P} \in R^{D \times d}, \ D << d$$



Appearance based visual features – Cont.

- **Image compression transforms** can be used for feature extraction:
  - **Discrete cosine** transform (**DCT**).
  - **Discrete wavelet** transform (**DWT**); e.g., Daubechies wavelet of order 3.
- In both cases, place a small number of transform coefficients into the feature vector $\mathbf{y}_t$ .These can be located on *predefined lattices* (see Fig.), or estimated on basis of largest training data *energy*.
- Both DCT & DWT are *separable* and *fast* when M,N,K are powers of 2!



Appearance based visual features – Cont.

- **Principal component analysis** (**PCA**) is also a good candidate for feature selection. Achieves *optimal compression*, but requires *expensive training*, and *does not* allow fast implementation.

- STEPS:
  - Compute training data *covariance / correlation* matrix **R** (of d x d size).
  - *Diagonalize* **R**: $\mathbf{R} = \mathbf{A}\,\Lambda\,\mathbf{A}^T$
  - Select $D<<d$ *largest eigenvalues*, of Λ located in $j_1,\ldots,j_D$ positions.
  - Then, $\mathbf{P}_{PCA} = [\mathbf{a}_{j1},\ldots,\mathbf{a}_{jD}]$, where $\mathbf{a}_j$ are the eigenvectors of **R**.
  - *Final features* are: $\mathbf{y}_t = \mathbf{P}_{PCA}\,\mathbf{x}_t$

Appearance based visual features – Cont.

- Note: Image transforms provide a simple feature extraction mechanism. However, they aim at *compression*, not *classification* of the resulting vectors among competing (speech) classes!
- **Linear discriminant analysis** (**LDA**) is more appropriate for the latter!
  - Assumes a *set of classes* C chosen a-priori. Training data $\mathbf{x}_l$ are *labeled*: $c(l) \in C$
  - Seeks matrix $\mathbf{P}_{LDA}$ so that projected training data are *well-separated* into C.
  - Formally, it *maximizes*: $\det(\mathbf{P}\mathbf{S}_B\mathbf{P}^{\mathrm{T}}) / \det(\mathbf{P}\mathbf{S}_W\mathbf{P}^{\mathrm{T}})$ wrt $\mathbf{P}$, where the data **within/ between** class scatter is: $\mathbf{S}_W = \sum_{c \in C} \Pr(c)\Sigma^{(c)}, \quad \mathbf{S}_B = \sum_{c \in C} \Pr(c)(m^{(c)} - m)(m^{(c)} - m)^{\mathrm{T}}$
  - Then, it solves the *generalized eigen-value/vector* problem: $\mathbf{S}_B\mathbf{F} = \mathbf{S}_W\mathbf{F}\,\Lambda$
  - *Features* are: $\mathbf{y}_t = \mathbf{P}_{LDA}\,\mathbf{x}_t$ where $\mathbf{P}_{LDA} = [\mathbf{f}_{j1}, \ldots, \mathbf{f}_{jD}]$ contains *D* e-vecs of $\mathbf{F}$.

Appearance based visual features – Cont.

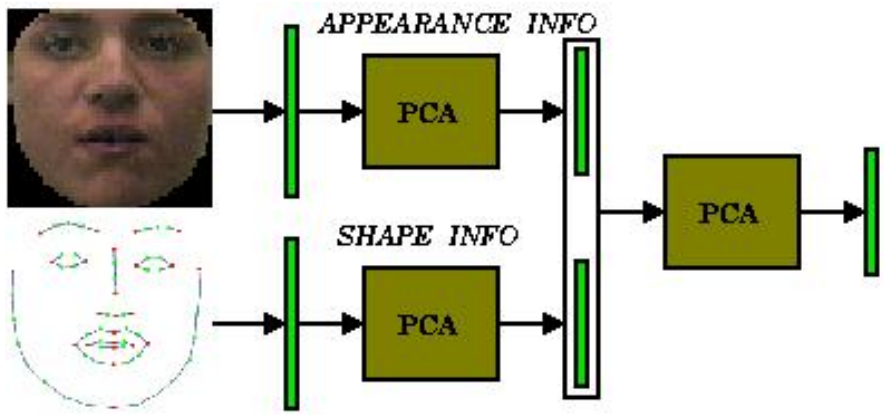
- Note: Typical statistical modeling of speech feature vectors assumes that their elements are uncorrelated (per-class). In practice, this does not hold!
- A data rotation based on the **maximum likelihood linear transform** (**MLLT**) can remedy this (Gopinath, 1998).
  - MLLT maximizes the observation data likelihood in the *original* feature space, under the assumption of diagonal data covariance in the *transformed* space.
  - Desired rotation matrix is obtained by solving:

$$\mathbf{P}_{MLLT} = \arg\max_{\mathbf{P}} \{\det(\mathbf{P})^{L} \prod_{c \in C} (\det(\mathrm{diag}(\mathbf{P}\Sigma^{(c)}\mathbf{P}^{\mathrm{T}})))^{-L_c/2}\}$$

Joint shape and appearance features

- **Main idea:** Combine the "best" of the two types of features.
- **Two approaches** for doing so:
  - *Concatenate* shape + appearance features into new vectors, e.g.:
    - Active shape models + PCA of image intensity values along stripes perpendicular to lip contour (Dupont and Luettin, 2000).
    - Snake parameters combined with PCA of color image ROI (Chiou and Hwang, 1997).
  - Or, build a *joint model* of shape and appearance by PCA on the concatenated vector of shape and appearance features (Matthews, 1998).

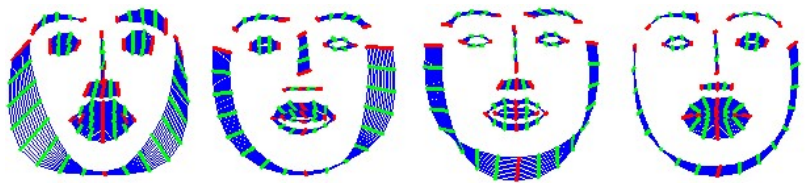
- **Active appearance models** (**AAMs** – Matthews, 1998):
  - Use two stages of PCA.
  - Three steps (next).



Joint shape and appearance features – AAMs - Cont.

- **STEP 1**: Shape modeling.
  - Shape vector of landmark point coordinates.

$$\mathbf{x}^{(S)} = [x_1, y_1, x_2, y_2, ..., x_K, y_K]^{\mathrm{T}}$$



  - Shape PCA (68 to 11 dims): $\mathbf{x}^{(S)} = \bar{\mathbf{x}}^{(S)} + \mathbf{P}^{(S)}\mathbf{y}^{(S)}$

- **STEP 2:** Appearance modeling.
  - Normalized color appearance vector.

$$\mathbf{x}^{(A)} = [r_1, g_1, b_1, ..., r_{MN}, g_{MN}, b_{MN}]^{\mathrm{T}}$$

  - Appearance PCA (6k to 186): $\mathbf{x}^{(A)} = \bar{\mathbf{x}}^{(A)} + \mathbf{P}^{(A)}\mathbf{y}^{(A)}$

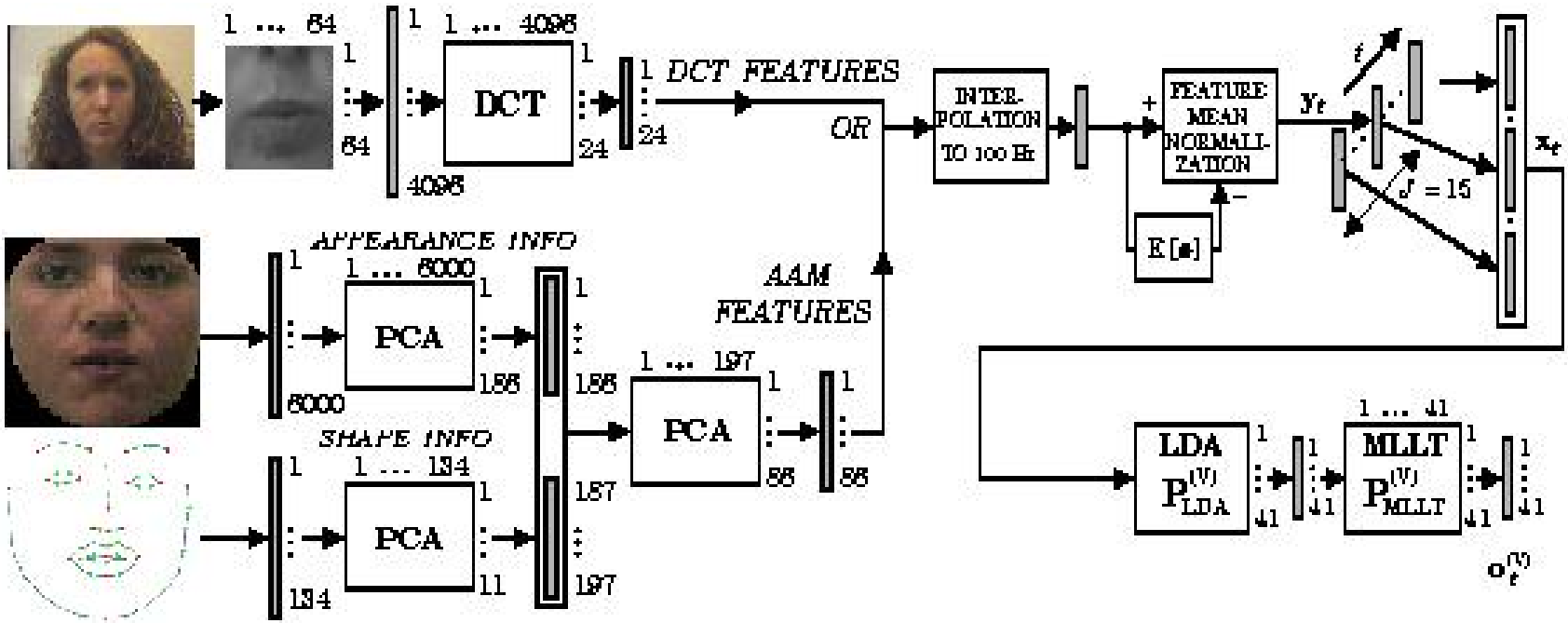
- **STEP 3:** Joint modeling.
  - Concatenated features. $\mathbf{x}^{(A,S)} = [\mathbf{y}^{(A)\mathrm{T}}\mathbf{W}, \mathbf{y}^{(S)\mathrm{T}}]^{\mathrm{T}}$
  - PCA on joint vector (197 to 86 dims):

$$\mathbf{x}^{(A,S)} = \bar{\mathbf{x}}^{(A,S)} + \mathbf{P}^{(A,S)}\mathbf{y}^{(A,S)}$$

- Feature extraction: AAM tracking + 3 PCAs give $\mathbf{y}^{(A,S)}$.

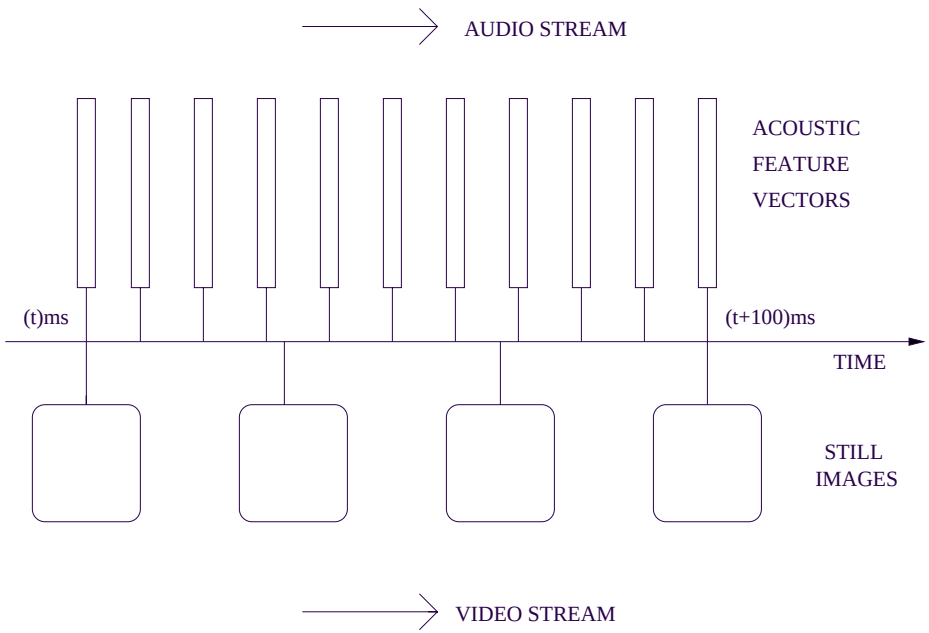
Visual feature post-processing

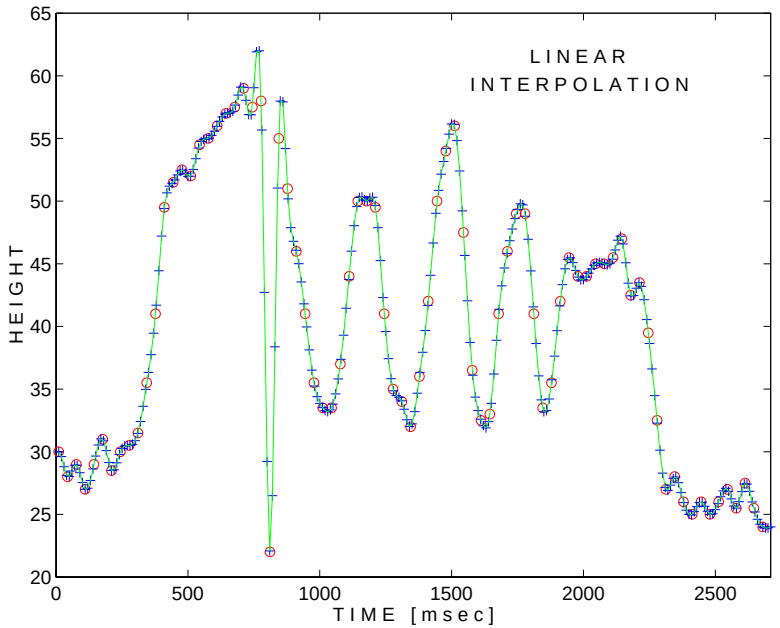
- Visual feature post-processing is desirable before presenting vectors for ASR:
  - ***Normalization*** (e.g., *cepstral mean subtraction*-**CMS**): Reduces variability due to illumination.
  - Incorporation of ***dynamic*** information (e.g., LDA on concatenation of *neighboring* features).
  - ***Up-sampling*** to the audio stream feature rate (30 or 60 to 100 Hz).



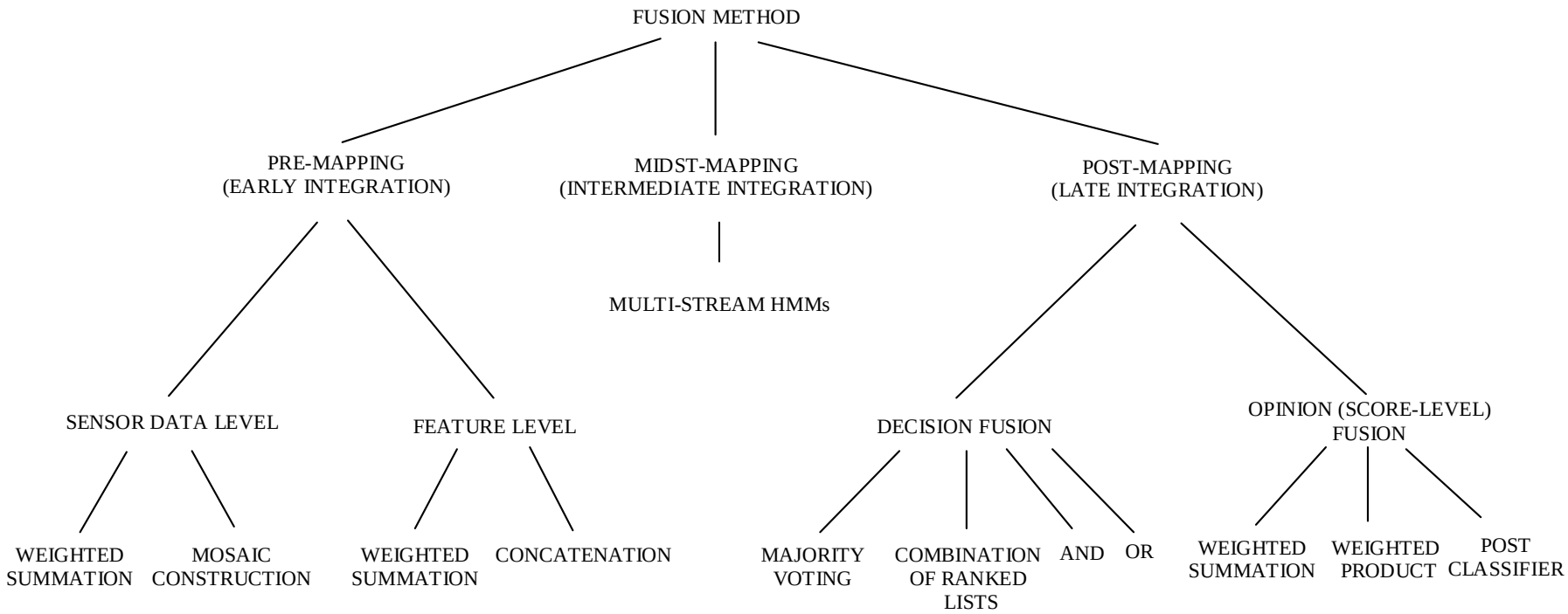
Visual feature post-processing - Cont.

- Static to dynamic features:
  - LDA/MLLT on concatenation of neighboring features.
  - Augmenting of visual features by their ***first*** and ***second*** time-derivatives.
- Visual feature up-sampling to 100 Hz by **linear interpolation.** Simplifies visual only model training for ASR and audio-visual fusion.



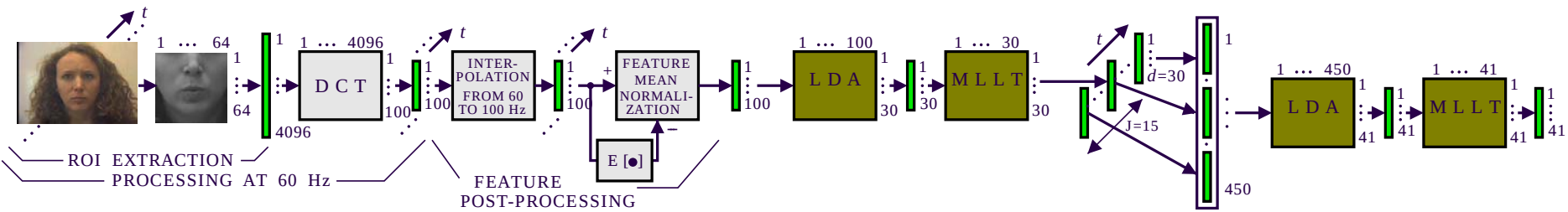


Audio-Visual Fusion



The IBM system visual front end

- **Face tracking:** 2-level statistical detection of faces and facial features; mouth location estimates are **smoothed** over time.
- **ROI extraction:** Enlarged ROI contains “beard region”; normalized for head pose and illumination variations. ROI size is **64 x 64** pixels.
- **Static features:** **100**-dimensional compressed representation of ROI using **DCT**.
- **Post-processing:** Intra-frame + inter frame **LDA/MLLT** for better within and across frame discrimination and statistical modeling; **CMS** and **up-sampling.**
- **Final features:** **41**-dimensional at 100 Hz.



Visual feature comparisons

- Let's now address these **issues:**
  - How much (if any?) visual speech information is ***captured*** by the above features?
  - How do these features ***compare*** to each other?
- **Visual-only ASR performance** provides answers to these questions.
  - ***Single-subject***, ***connected-digit*** ASR experiments.
  - Modeling: Whole-word HMMs, unknown string length.
- **Feature comparisons** (Potamianos et al., 1998):

Outer lip features	%, Word accuracy
h , w	**55.8**
+ a	**61.9**
+ p	**64.7**
+ $FD_{2\text{-}5}$	**73.4**

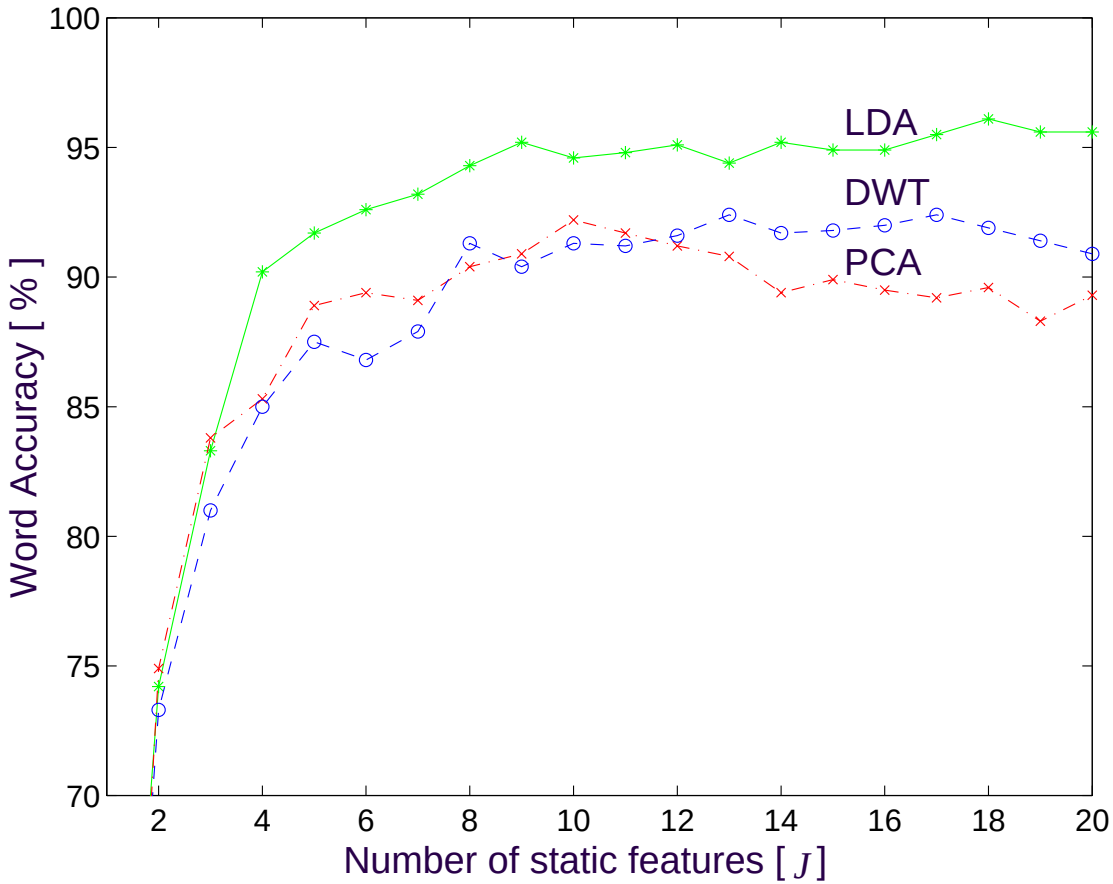
Lip contour features	%, Word accuracy
Outer-only	**73.4**
Inner-only	**64.0**
2 contours	**83.9**

Feature type	%, Word accuracy
Lip-contour based	**83.9**
Appearance (LDA)	**97.0**

- Thus, appearance based modeling is preferable!

Visual feature comparisons – Cont.

- Performance of various **appearance** based features (LDA, DWT, PCA) vs. static feature size (Potamianos et al, 1998).



Visual feature comparisons – Cont.

- Appearance (**DCT**) vs. joint (**AAM**) features:
  - Speaker-independent, LVCSR corpus.
  - Word error rate, after "rescoring" of lattices, that have been generated based on noisy audio features (Neti et al., 2000).

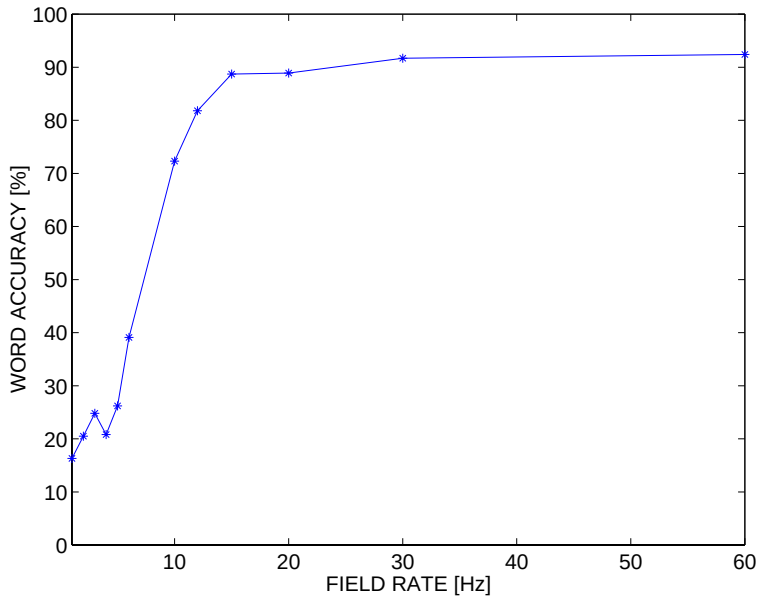
VI-feats	+ Derivs	+ LDA/MLLT
DCT	61.80	**58.14**
DWT	n/a	58.79
PCA	n/a	58.86
AAM	65.90	**64.00**

- Thus, using the particular implementation, *DCT features were the best.*

Video degradation effects

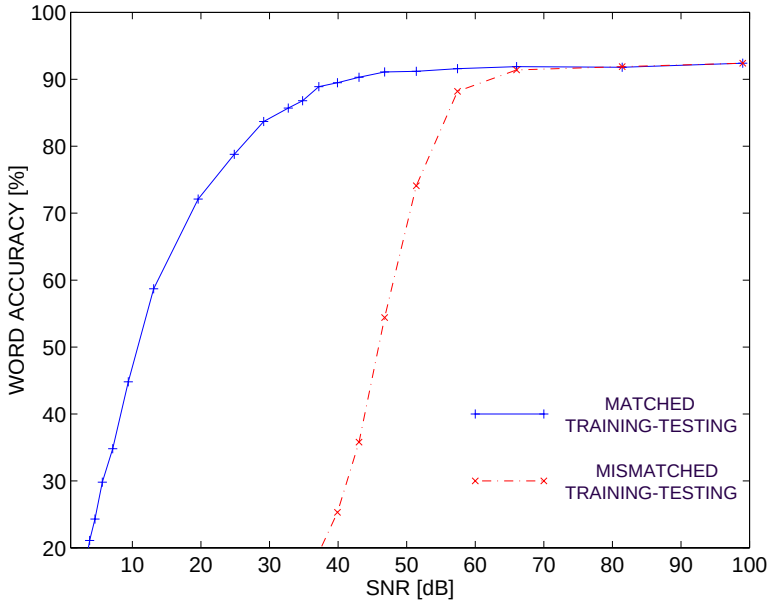
- Frame rate decimation:

  Limit of acceptable video rate for automatic speechreading is 15 Hz.



- Video noise:

  Robustness to noise only in a matched training/testing scenario.

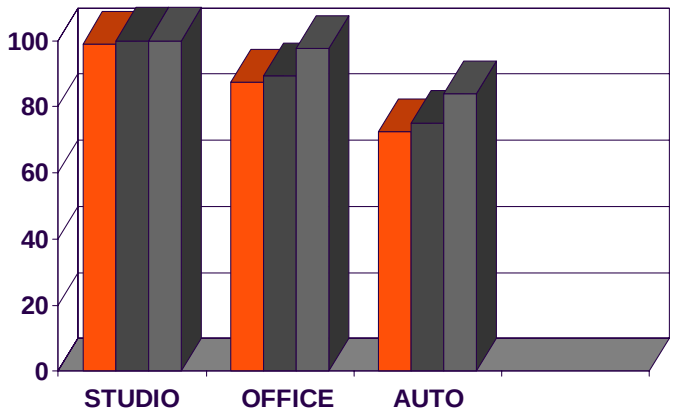


Both cases: *DWT visual features – connected digits recognition* (Potamianos et al., 1998).

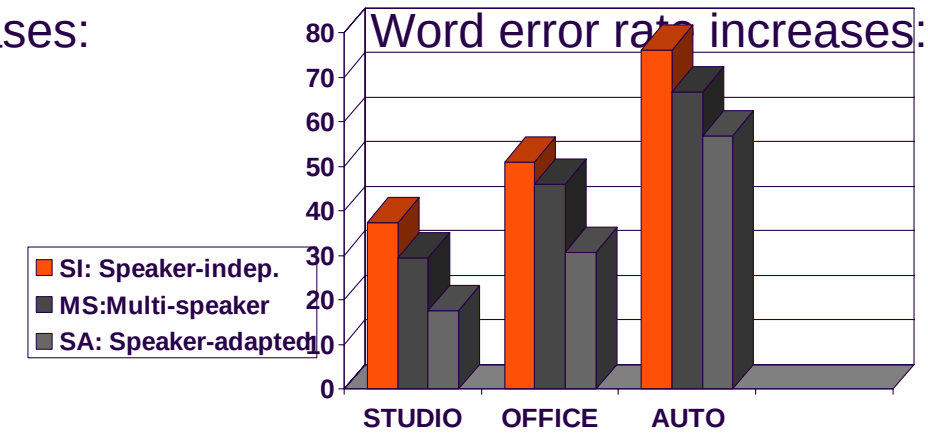
Video degradation effects – Cont.

- **Unconstrained visual environments** remain challenging, as they pose difficulties to robust visual feature extraction.
- **EXAMPLE:** Recall our three "increasingly-difficult" domains: ***Studio***, ***office***, and ***automobile*** environments (multiple-speakers, connected digits).

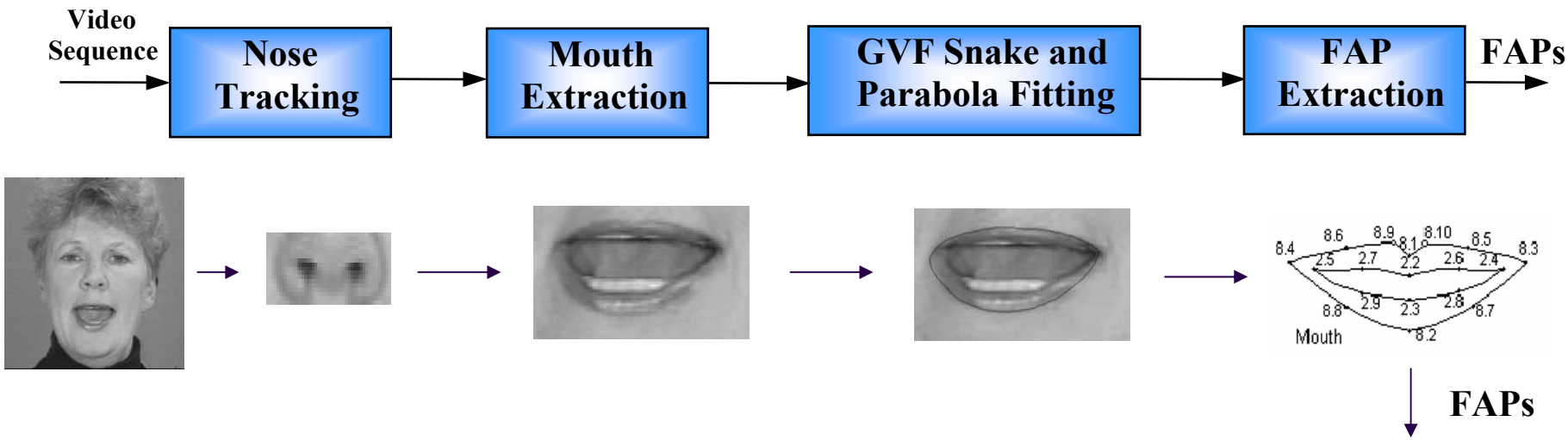
Face detection accuracy decreases:



Word error rate increases:



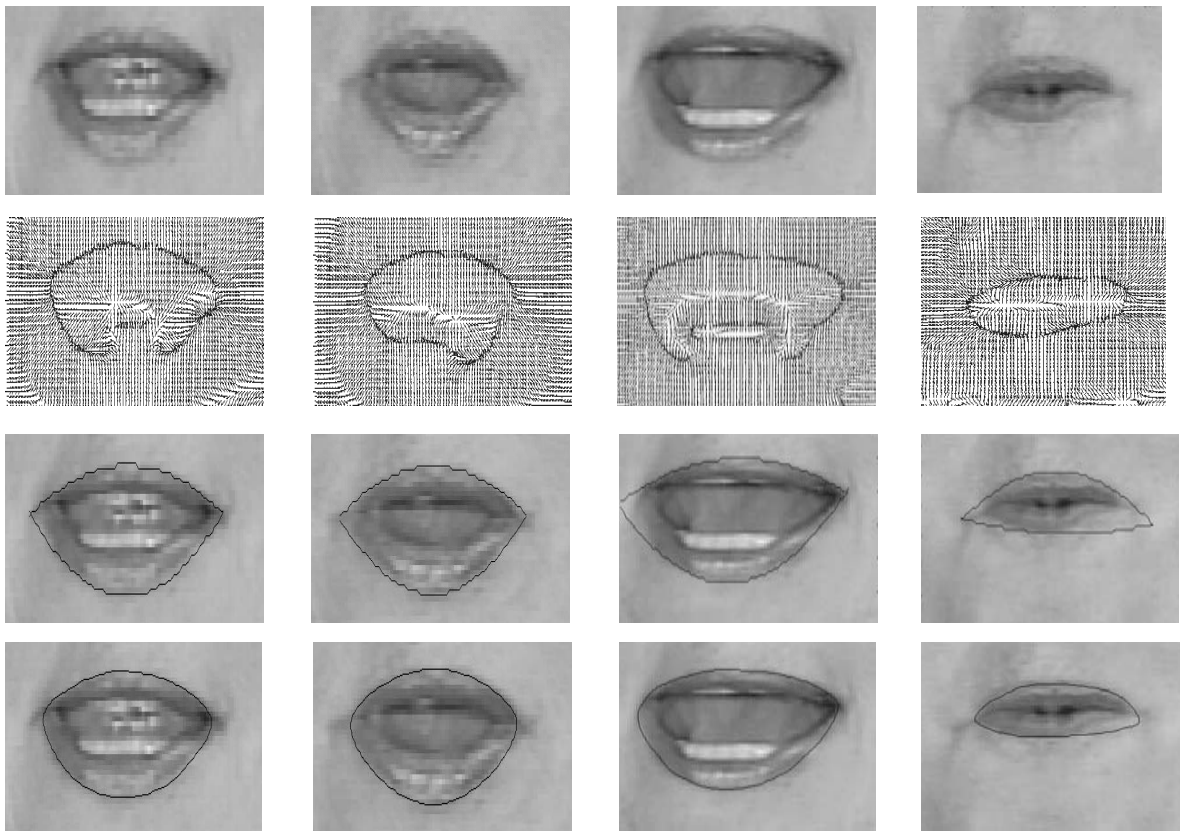
A Visual Front End Example



- Bernstein audio-visual database
  - 474 sentences (average length ~ 4 seconds)
  - vocabulary size around 1000 words
  - video (320x240 - frame size) time-synchronized with speech

MPEG-4 decoder

Outer Lip Modeling



Original mouth images

GVFs

Fitted parabolas

Final results

Outer Lip FAP Extraction

Original image frames, and MPEG-4 facial animations driven by the extracted FAPs

PCA FAP Analysis

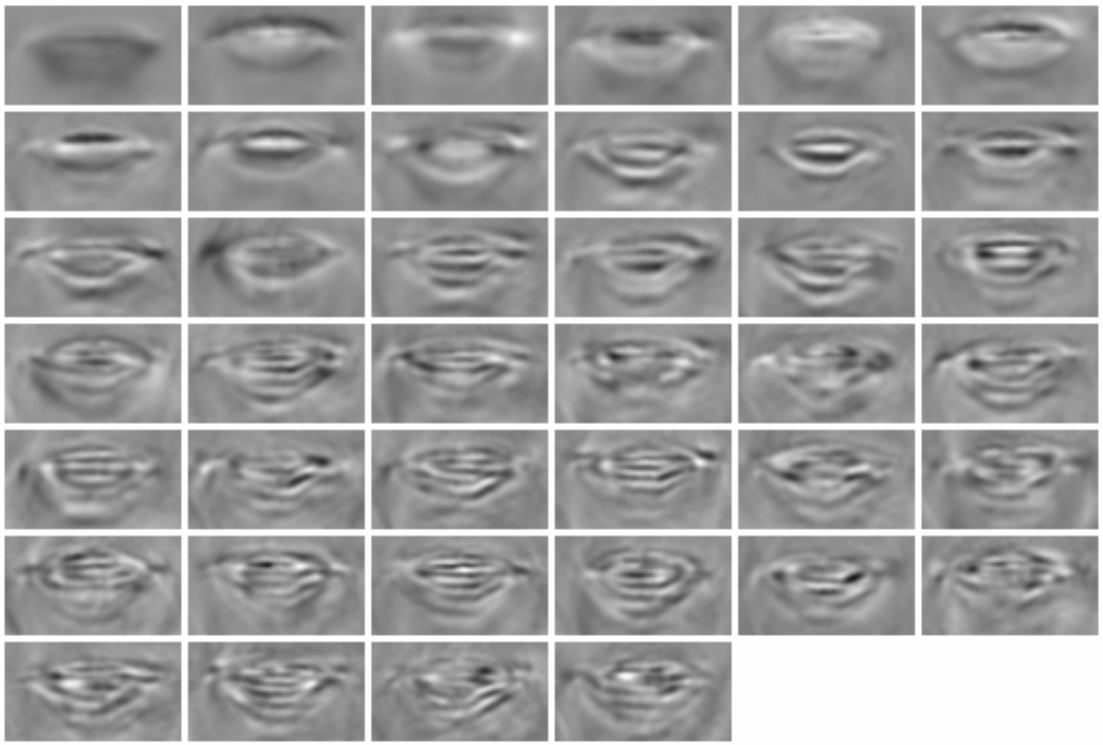
Mean lip shape: the middle images; lip shapes obtained by varying the projection weights corresponding to the first eigenvector (upper) and second eigenvector (bottom) by +2 standard deviations (left) and –2 standard deviations (right)

# of principal components	Percentage of variance
6	99.7%
2	93%
1	81%

Distribution of the variance among eigenvectors

Eigen-Lips

- Ensemble contained 1460 images of dim 80x45 (3600 pixels)
  - 20 Eigen-lips - 90.92% Statistical variance
  - 40 Eigen-lips - 95.52% Statistical variance



Compression Examples

- What do all these marks mean? (f0046)

Compressed 20:

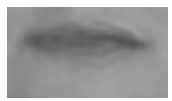
Compressed 40:

Typical Spatial Quality
35.08dB using 20 Coef.
37.56dB using 40 Coef.

- Do you like the length of this skirt? (f0338)

Compressed 20:

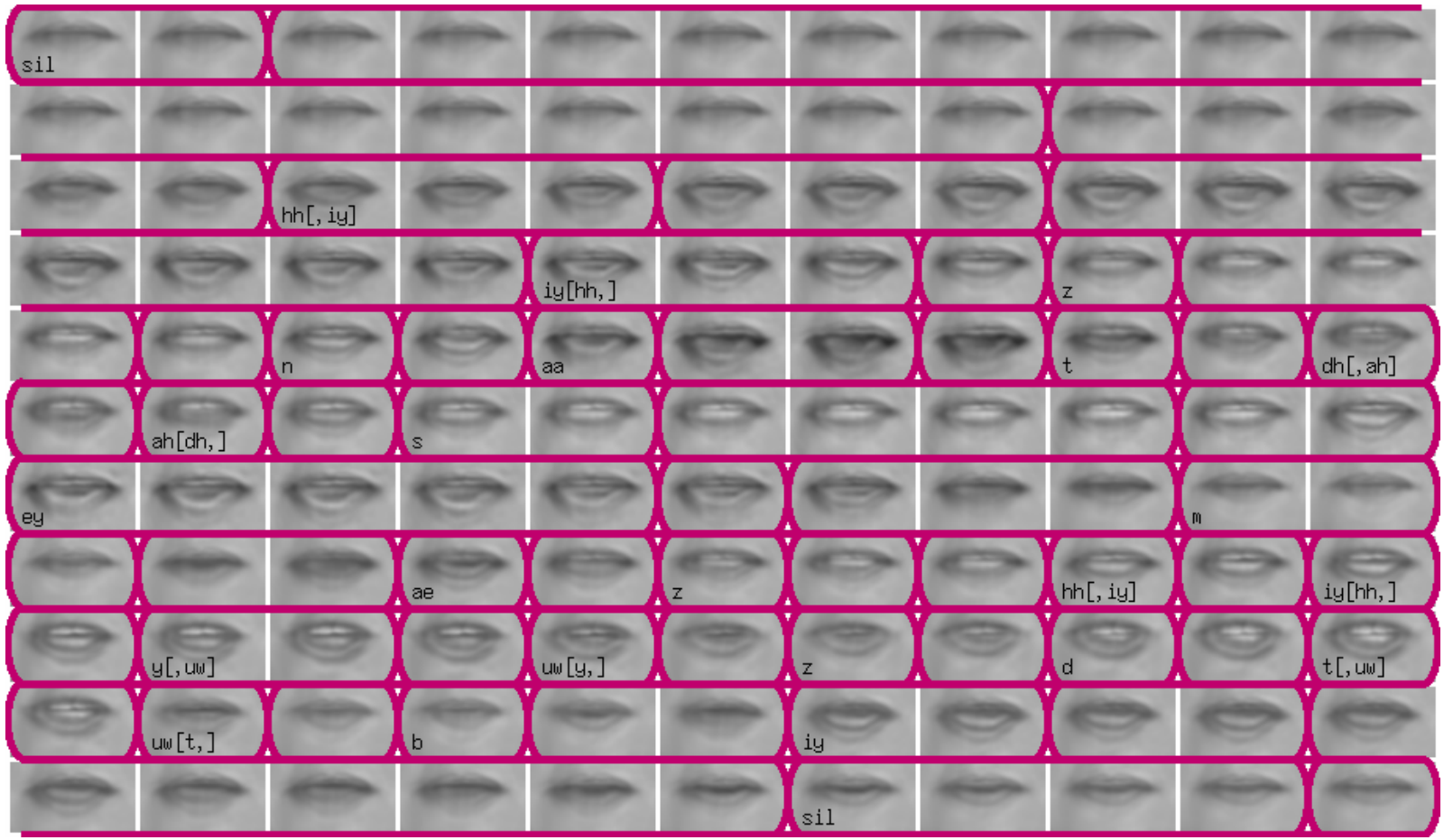
Compressed 40:



Worse Spatial Quality
28.42dB using 20 Coef.
30.69dB using 40 Coef.

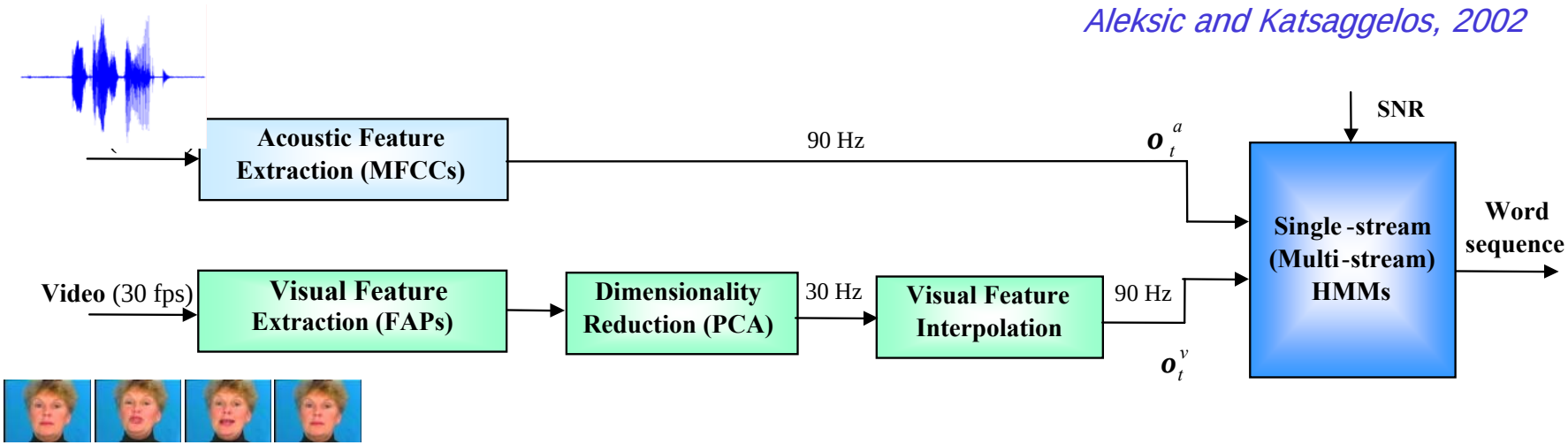
Segmentation Example

- **He’s not the same as he used to be (f0386)**



Summary of AV-ASR experiments

Aleksic and Katsaggelos, 2002



o_t^a - acoustic observation vector at time t

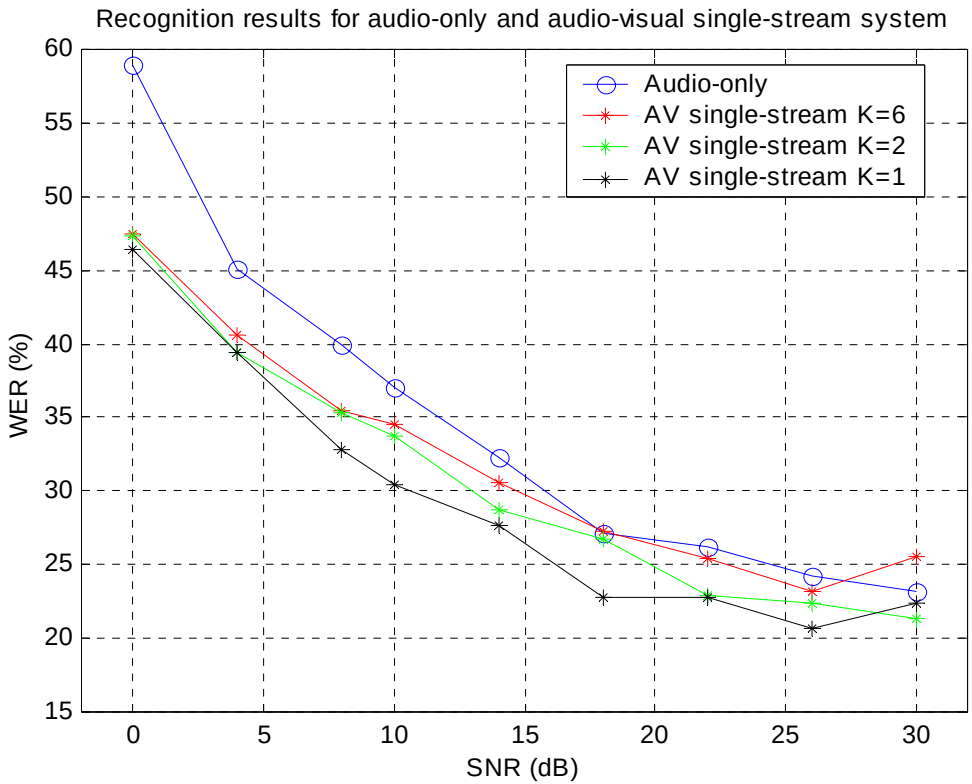
o_t^v - visual observation vector at time t

PCA - Principal Component Analysis
MFCC - Mel-frequency Cepstral Coefficients
HMM - Hidden Markov Models

- Bernstein audio-visual database
  - 474 sentences (average length ~ 4 seconds)
  - vocabulary size around 1000 words
  - video (320x240 - frame size) time-synchronized with speech

Summary of AV-ASR experiments - Cont.

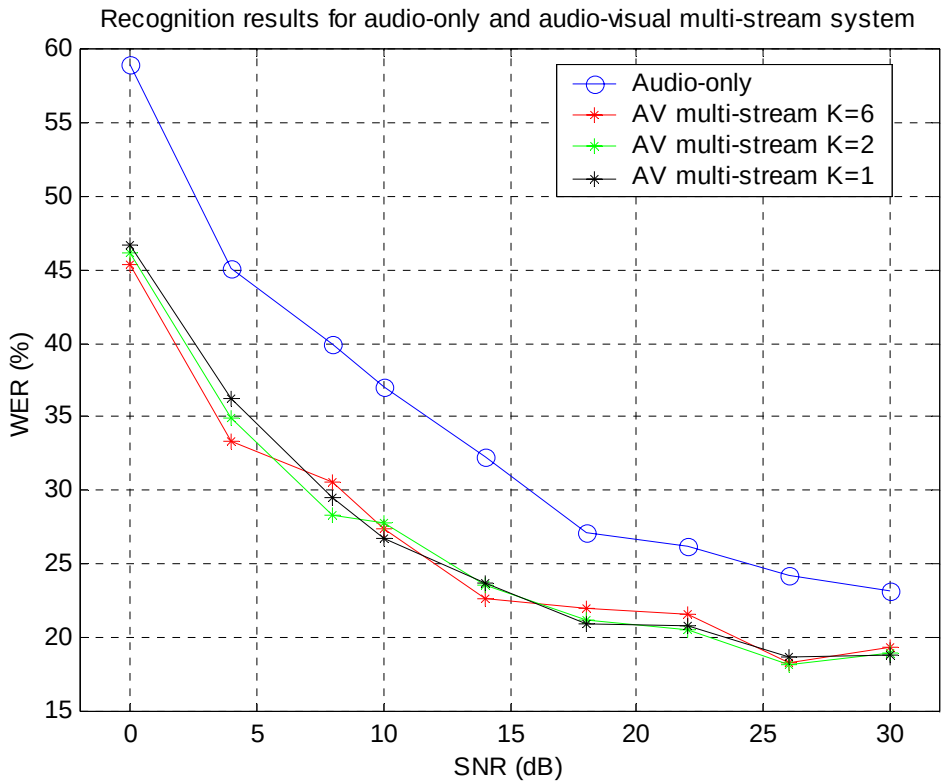
Single-stream HMMs



Aleksic and Katsaggelos, 2002

Summary of AV-ASR experiments - Cont.

Multi-stream HMMs



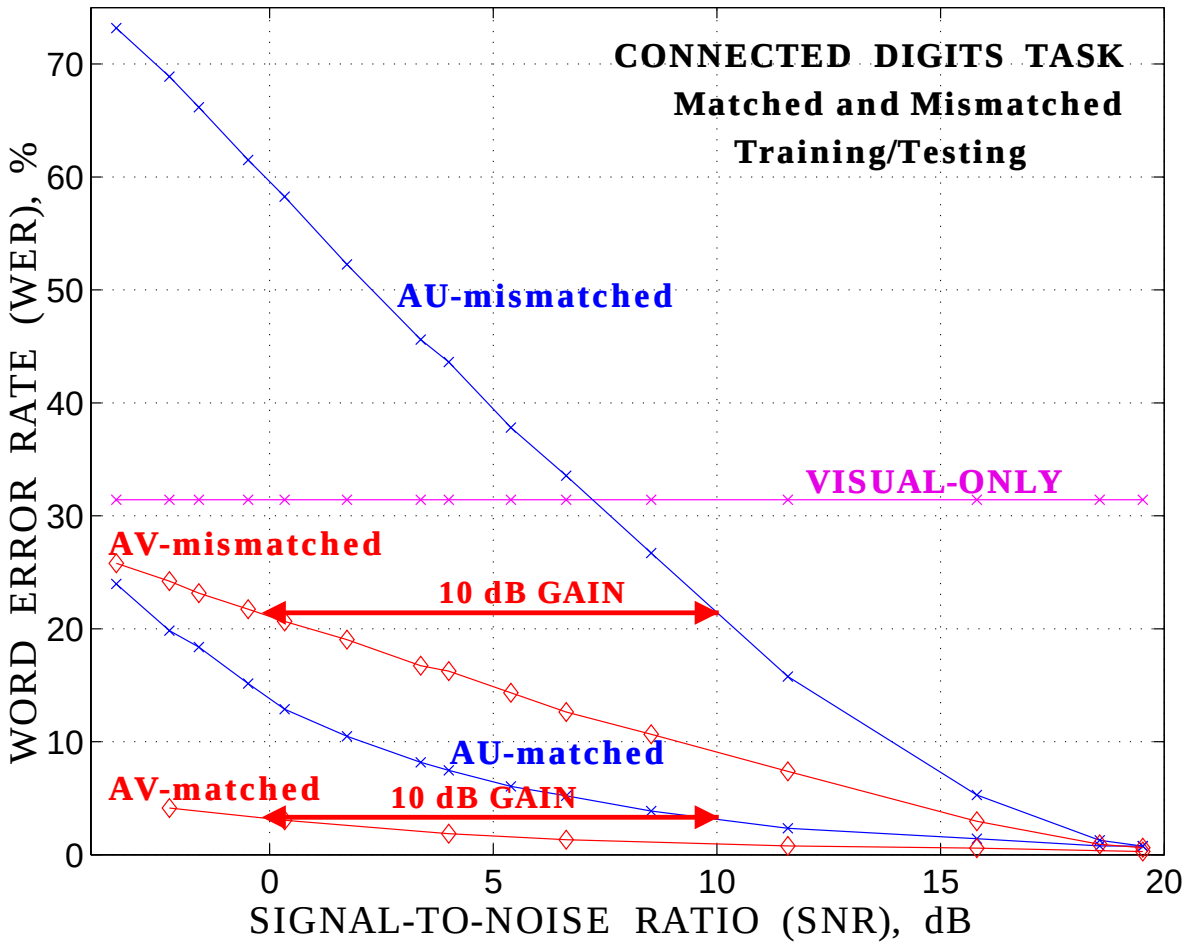
Summary of AV-ASR experiments - Cont.

Audio-only and audio-visual systems performance in clean audio conditions		WER [%]	Audio-stream Weight
Audio-only system		22.19	
Audio-visual system (Single-stream HMM method)	K=1	21.48	
	K=2	21.34	
	K=6	24.47	
Audio-visual system (Multi-stream HMM method)	K=1	18.21	0.75
	K=2	18.07	0.7
	K=6	18.16	0.85

Summary of AV-ASR experiments (IBM)

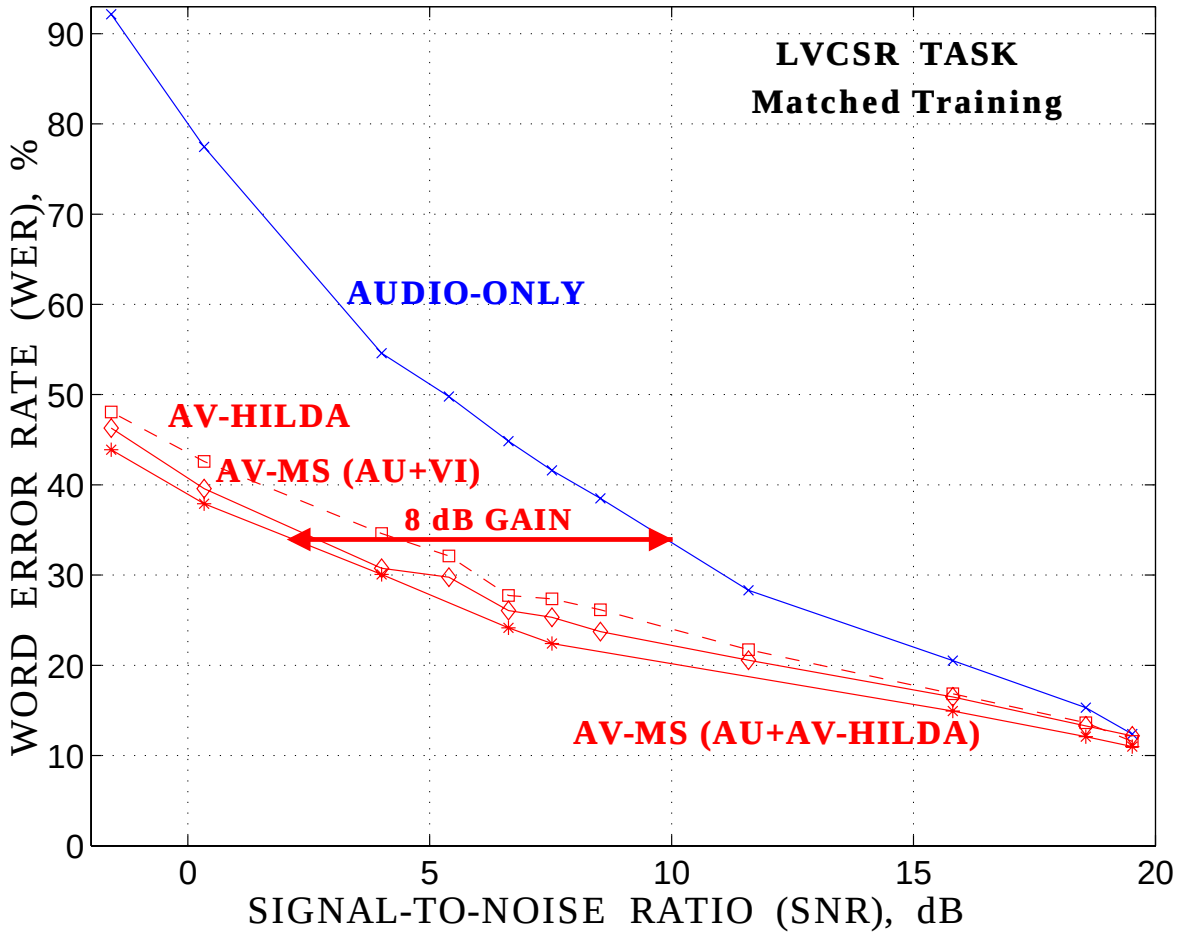
- **Summary** of AV-ASR results for **connected-digit** recog.

- Multi-speaker training/testing.
- 50 subjects, 10 hrs of data.
- Additive noise at various SNRs.
- Two training/testing scenarios:
  - **Matched** (same noise in training and testing).
  - **Mismatched** (trained in clean, tested in noisy).
- **10 dB** effective SNR **gain** for both, using **product HMM**.



Summary of AV-ASR experiments - Cont.

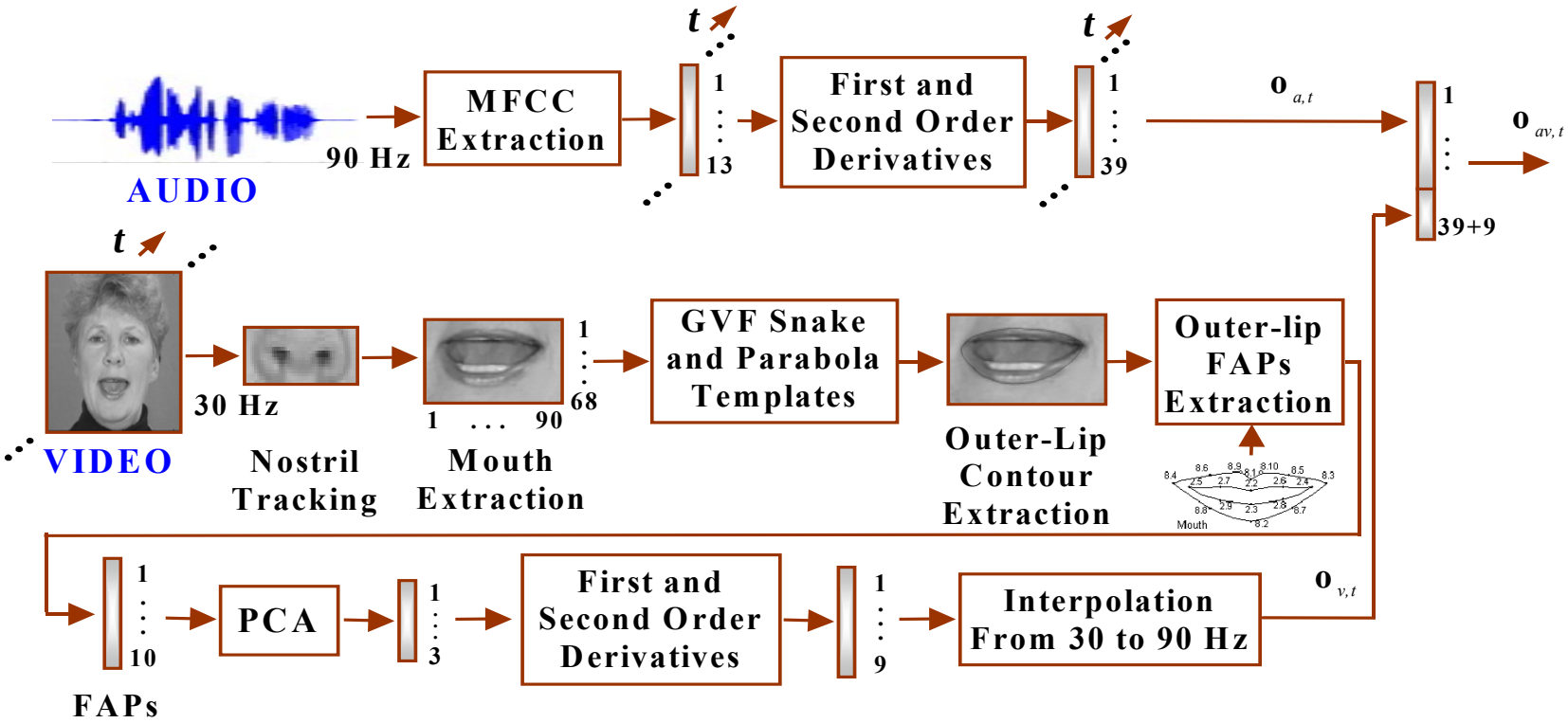
- **Summary** of AV-ASR results for large-vocabulary continuous speech(**LVCSR**).
- Speaker-independent training (**239** subj.) testing (**25** subj.).
- **40** hrs of data.
- **10,400**-word vocabulary.
- 3-gram LM.
- Additive noise at various SNRs.
- Matched training/testing.
- **8 dB** effective SNR **gain** using hybrid fusion.
- Product HMM did not help.



Audio-Visual Person Verification and Identification

- Biometrics-Automated methods for recognizing a person based on a physiological or behavioral characteristic
- Measured Features
  - Face, fingerprints, handwriting, hand geometry
  - Iris, retina, voice
  - Ear, vein (unique vein patterns on the back of one's hand)
- High security person recognition technologies for many applications
  - Network security, secure electronic banking, financial transactions
  - Local, state, federal government, law enforcement
- Considerably more accurate than current methods like passwords, PINs (can be used by somebody else)
- Biometrics approaches link to a particular individual, nothing to carry or remember
- It is becoming inexpensive and socially acceptable

Audio-Visual Speaker Recognition Using FAPs as Visual Features



Speaker Recognition

- Two important problems are **speaker verification (authentication) and identification**
- **Speaker verification:**
  - Verify claimed identity based on observations **O**
  - A **two-class** problem; True claimant vs. impostor (general population).
  - **Based on:**

$$\frac{\Pr(c_{claim} \mid \mathbf{O})}{\Pr(c_{all} \mid \mathbf{O})} \begin{matrix} > \\ < \end{matrix} \begin{matrix} Accept \\ Reject \end{matrix} thresh$$

Threshold-a priori determined
$\mathbf{C}_{\mathbf{all}}$ – world model
$\mathbf{C}_{\mathbf{claim}}$ – claimed client model

  - More important problem in practice, most applications require identity claim verification
- **Speaker identification:**
  - Obtain speaker identity $\hat{c}$ within a closed set of known subjects ***C*** based on observations **O**

$$\hat{c} = \arg\max_{c \in C} \Pr(c \mid \mathbf{O})$$

$$\Pr(c \mid \mathbf{O}), \quad c \in C.$$

- **Text-dependent (TD) vs. text-independent (TI) approaches**
  - **TD –** speech used for testing and training is constrained to be the same
  - **TI –** speech used for testing is unconstrained

Verification Performance Measures

- Two error measures are False Acceptance (**FA**) and False Rejection (**FR**)
  - **FA** – an impostor, claiming the identity of a client, is accepted
  - **FR** – a client, claiming his true identity, is rejected

$$FA = \frac{I_A}{I} \times 100\ \% \qquad FR = \frac{C_R}{C} \times 100\ \%$$

I - impostor claims
I_A - impostors accepted
C - client claims
C_R - clients rejected

  - Trade-off between the two errors is controlled by the **threshold**
    - Threshold chosen a-priori on an evaluation set to meet certain FA and FR requirements
    - Minimum FA, minimum FR, or FA=FR
- Equal Error Rate (**EER**)
  - Obtained after a full authentication experiment by choosing the threshold for which FA=FR
  - Is an unrealistic measure
- Receiver Operator Curve (**ROC**)
  - Plots either FA or FR against the other or against the verification threshold

Sample Audio-Visual Person Recognition Systems

System	Features		Database	Non-ideal Conditions	Expert	AV Fusion Method	Recognition Mode*
	Acoustic	Visual					
Luettin *et al.* [135]	none	shape- and appearance-based, and joint (concatenation)	Tulips1	none	HMMs GMMs	none	TD+TI/ID
Chibelushi *et al.* [5]	MFCCs	shape-based (PCA,LDA, concatenation)	10 speakers [5]	white noise at different SNRs	ANNs	opinion fusion (weighted summation)	TD/ID
Brunelli and Falavigna [6]	MFCCs+Δ**+ΔΔ	appearance-based	89 speakers 3 sessions	none	VQ	opinion fusion (weighted product)	TI/ID
Ben-Yacoub *et al.* [7]	LPCs	appearance-based	XM2VTS	none	HMMs, sphericity measure [7]	post classifier using binary classifiers (SVM, Bayesian classifier, FLD, decision tree and MLP)	TD+TI/VER
Sanderson and Paliwal [8]	MFCCs+Δ	appearance-based (PCA)	VidTIMIT	white and operations-room noise at different SNRs	GMMs	weighted summation, concatenation, adaptive weighted summation, SVM, Bayesian classifier	TI/VER
Hazen *et al.* [9]	MFCCs	appearance-based	35 speakers [9]	data recorded on a handheld device	SVMs	opinion fusion (weighted summation)	TD/ID
Jourlin *et al.* [10]	LPCs+Δ+ΔΔ	appearance- and shape-based features	M2VTS	none	HMMs	opinion fusion (weighted summation)	TD/VER
Wark *et al.* [11-13]	MFCCs	shape-based (PCA and LDA)	M2VTS	white noise at different SNRs	GMMs	opinion fusion (weighted summation)	TI/ID+VER

* TD: text-dependent; TI: text-independent
VER: verification: ID: identification

** Δ – first derivative
ΔΔ – second derivative

Sample Audio-Visual Person Recognition Systems

System	Features		Database	Non-ideal Conditions	Expert	AV Fusion Method	Recognition Mode*
	Acoustic	Visual					
Aleksic and Katsaggelos [14]	MFCCs+Δ+ΔΔ	shape-based (PCA applied on lip-contours)	AMP/CMU	white noise at different SNRs	HMMs	feature-level concatenation	TD/ID+VER
Chaudhari *et al.* [15]	MFCCs	appearance-based (DCT applied on ROI)	IBM	none	GMMs	feature-level concatenation, opinion fusion	TI/ID+VER
Bengio *et al.* [32, 33]	MFCCs+Δ	shape-based and appearance-based	M2VTS	white noise at different SNRs	asynchronous HMMs	midst-mapping fusion	TD /VER
Fox *et al.* [34,35]	MFCCs+ Δ	appearance-based (DCT)	XM2VTS	white noise at different SNRs	HMMs	feature-level concatenation, opinion fusion (weighted summation)	TD/ID
Nefian *et al.* [30]	MFCCs+Δ+ΔΔ	appearance-based (PCA+LDA)	XM2VTS	white noise at different SNRs	Coupled HMMs embedded HMMs	midst-mapping fusion, opinion fusion (weighted summation)	TD/ID
Kanak *et al.* [38]	MFCCs+Δ+ΔΔ	appearance-based (PCA)	38 speakers [38]	white noise at different SNRs	HMMs	concatenation, opinion fusion (Bayesian fusion)	TD/ID

* TD: text-dependent; TI: text-independent
VER: verification: ID: identification

** Δ – first derivative
ΔΔ – second derivative

Audio-Visual Database

- Carnegie Melon Audio-Visual Database
  - **10 speakers** (7 male, 3 female) uttering the digit sequence **“234567”** ten times
- In AV experiments audio (Mel Frequency Cepstral Coefficients-MFCC) and video (FAPs) features were appended to form joint AV feature vectors
- The AV feature vectors were used in all AV experiments

Audio-Visual Speaker Recognition Using FAPs as Visual Features

Person Identification Error [%]		
SNR [dB]	Audio only	Audio-visual
clean	5.13	5.13
20	19.51	7.69
10	38.03	10.26
0	53.10	12.82

SNR [dB]	Audio only [%]			Audio-visual [%]		
	FA	FR	EER	FA	FR	EER
clean	2.85	25.64	2.56	0	12.82	1.71
20	2.85	41.03	3.99	2.85	20.51	2.28
10	0	53.85	4.99	0	23.08	2.71
0	5.7	61.54	8.26	2.85	28.21	3.13

Person Verification Results

- Tests performed under different acoustic noise conditions (SNRs 0-20 dB), and for clean speech

Automatic Facial Expression Recognition

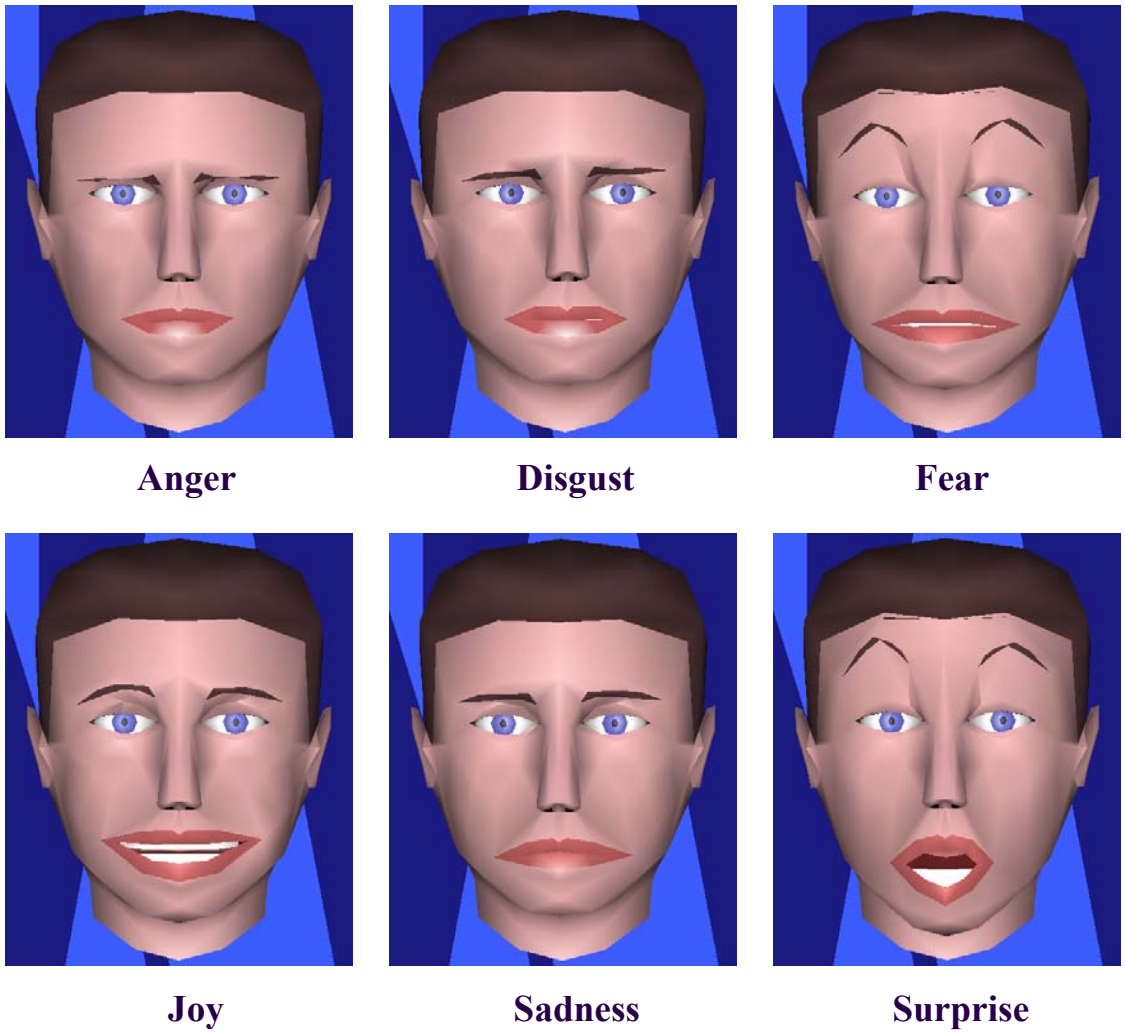
- Track visual features
- Extract outer lip and eyebrow FAPs
- Train HMMs for each of six basic expressions
  - Spatio-temporal approach
  - Takes temporal evolution of facial expressions into account
  - Provides improved recognition results
- Recognize expressions using HMMs
  - Use extracted FAPs (outer lip, eyebrow) for HMM training
  - Use multi-stream HMMs with stream weights for outer lip and eyebrow FAPs to improve recognition performance

Audio-visual emotion estimation

- Automatic detection of human emotional state is important in HCI.
- Emotional state reflects in **both facial expression and voice**.
- Facial expressions are universal and represent **happiness**, **anger**, **sadness**, **fear**, **surprise**, and **disgust/dislike**.
- Facial action coding system (**FACS**) codes facial expressions as sets of action units (**AU**).
- Various features can be used for **facial** emotion recognition, such as optical flow, ASMs.
- **Audio** features can be pitch contour statistics, energy, etc.
- **Example accuracies** (Chen et al., 1998):
  - Audio-only: 77.8%
  - Visual-only: 69.4%
  - Audio-visual: **97.2**%

Two examples of facial expressions (left-2-right): Anger, dislike, fear (from Cohen et al., 2003).

Six Basic Facial Expressions



Expression Recognition Results

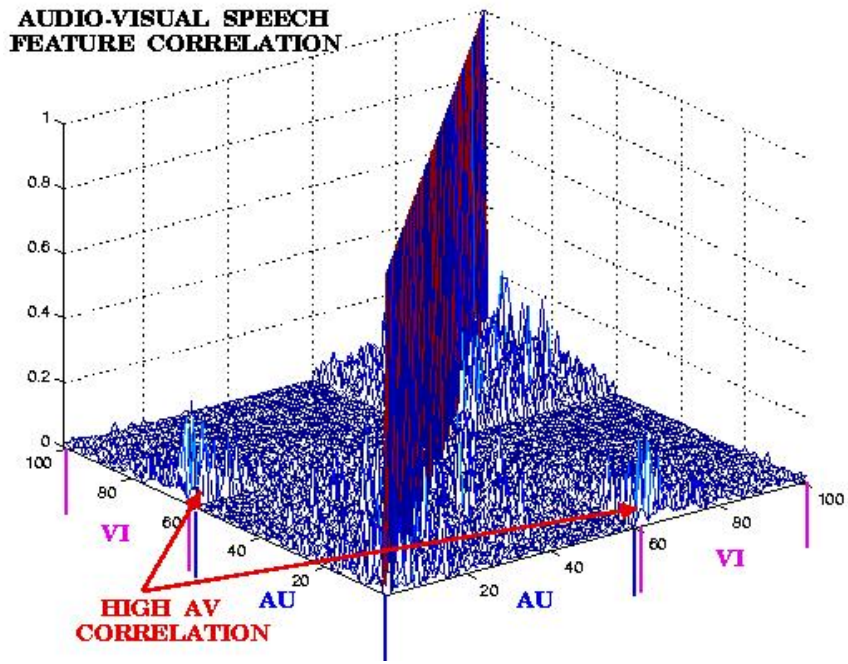
FAPS / EXP	Eyebrow (E) [%]	Outer Lip (OL) [%]	E and OL [%]	E and OL [%] (Multi-stream)	OL stream weight
Anger	52.9	64.7	64.7	70.6	0.7
Disgust	75.7	91.9	97.3	97.3	0.7
Fear	5.9	82.4	76.5	88.2	0.8
Joy	84.1	95.2	98.4	98.4	0.6
Sadness	17.0	81.1	81.1	96.2	0.7
Surprise	90.6	96.9	100	100	0.6
Total	62.68	87.32	88.73	**93.66**	

Confusion matrix
for Multi-stream system

	Anger	Disgust	Fear	Joy	Sadness	Surprise	Corr [%]
Anger	**24**	**4**	0	0	**6**	0	70.6
Disgust	0	**36**	**1**	0	0	0	97.3
Fear	0	0	**30**	**2**	**1**	**1**	88.2
Joy	0	0	0	**61**	0	**1**	98.4
Sadness	**2**	0	0	0	**51**	0	96.2
Surprise	0	0	0	0	0	**64**	100

Bimodal enhancement of audio

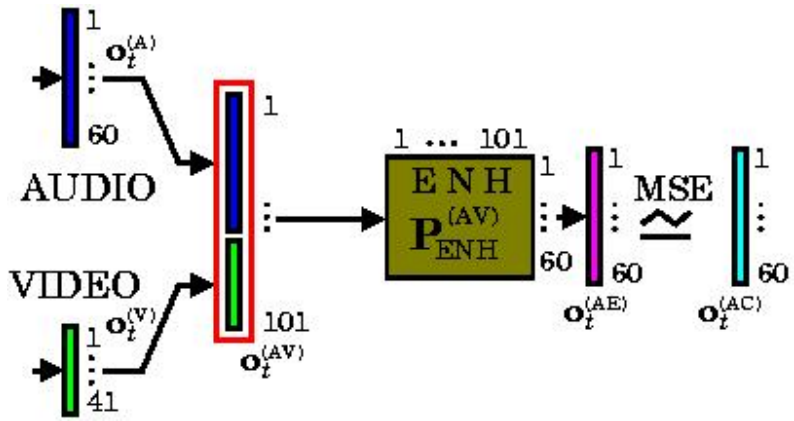
- **Main idea:**
  - Recall that the audio and visual features are **correlated**. E.g., for 60-dim audio features ($\mathbf{o}_{At}$) and 41-dim visual ($\mathbf{o}_{Vt}$):
  - Thus, one can hope to exploit visual input to **restore** acoustic information from the video and the corrupted audio signal.
- **Enhancement** can occur in the:
  - **Signal** space (based on **LPC** audio feats.).
  - Audio **feature** space (discussed here)
- **Main techniques:**
  - **Linear** (min. mean square error est.).
  - **Non-linear** (neural nets., CDCN).
- **Result:** Better than audio-only methods.



Linear bimodal audio enhancement.

- **Paradigm:**
- Training on noisy AV features

$$\mathbf{o}_{AV,t} = [\mathbf{o}_{A,t}, \mathbf{o}_{V,t}], \text{ and clean AU } \mathbf{o}_{A,t}^{(C)}, \ t \in T.$$



- Seek linear transform **P**, s.t:

$$\mathbf{o}_{A,t}^{(E)} = \mathbf{P}\,\mathbf{o}_{AV,t} \approx \mathbf{o}_{A,t}^{(C)}, \ t \in T.$$

- Can **estimate** P by minimizing the **mean square error** (**MSE**) between $\mathbf{o}_{A,t}^{(E)}, \mathbf{o}_{A,t}^{(C)}$.
- Problem separates per audio feature dimension (i=1,…,d_A):

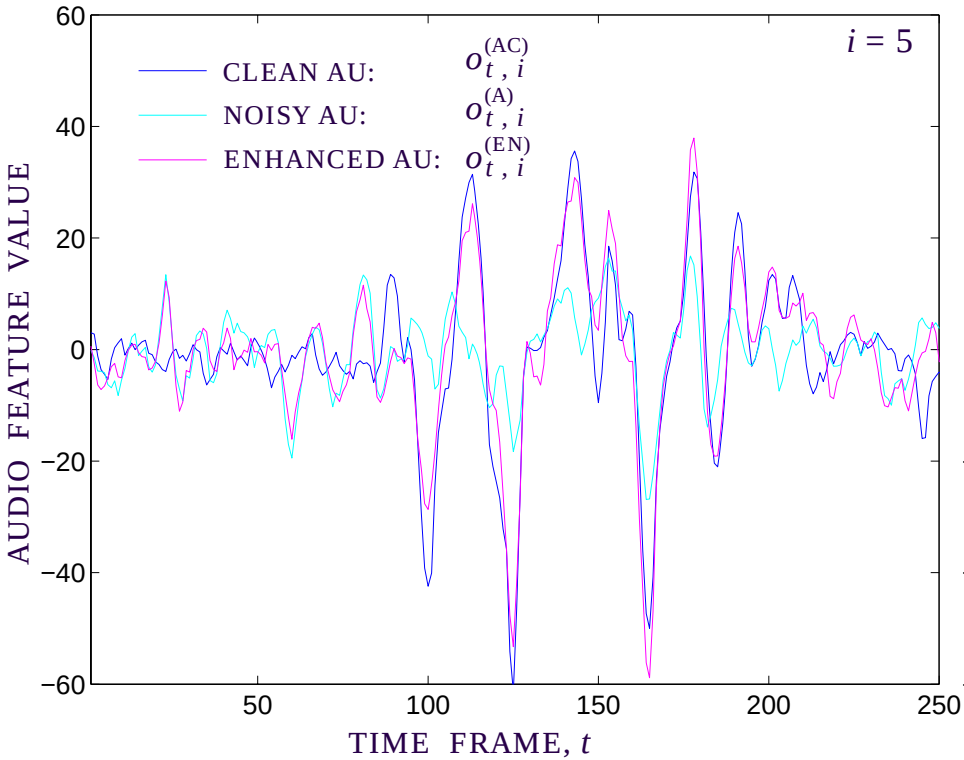
$$\mathbf{p}_i = \arg\max_{\mathbf{p}} \sum_{t \in T} [\, o_{A,t,i}^{(C)} - < \mathbf{p}, \mathbf{o}_{AV,t} >]^2, \quad i = 1,..., d_A$$

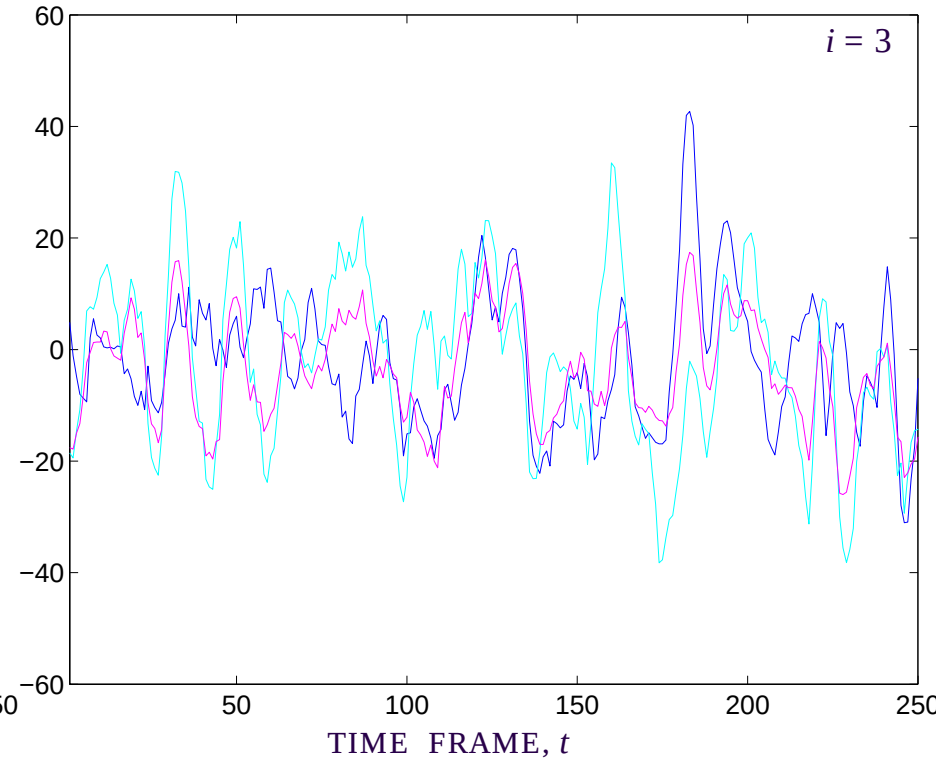
- Solved by d_A systems of Yule-Walker equatiions:

$$\sum_{j=1}^{d} [\sum_{t \in T} o_{AV,t,i} o_{AV,t,k}]\, p_{i,j} = \sum_{t \in T} o_{A,t,i}^{(C)} o_{AV,t,k}, \quad k = 1,..., d$$

Linear bimodal audio enhancement – Cont.

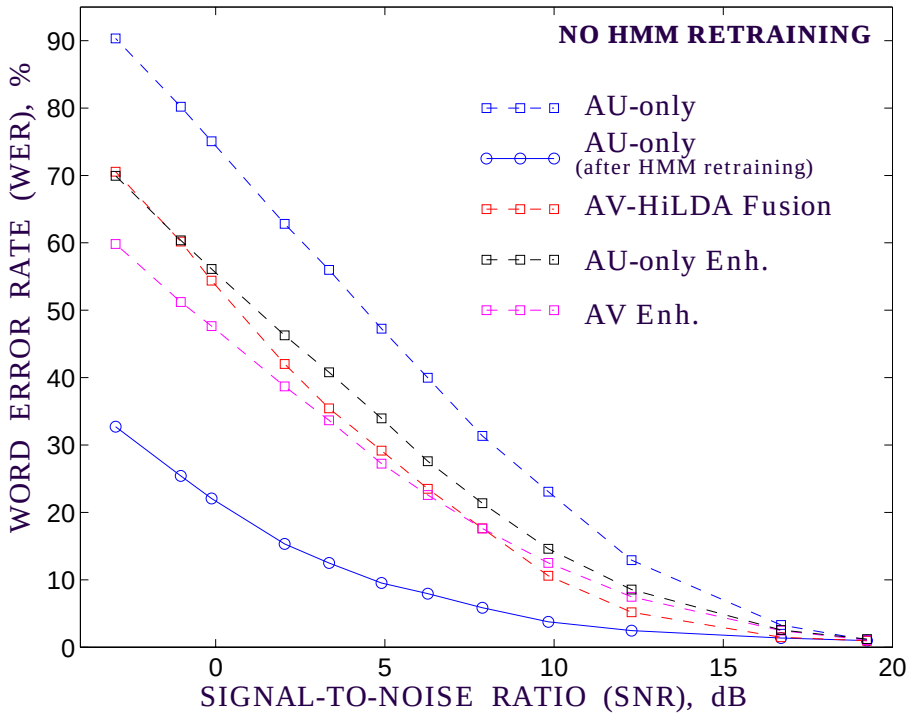
- Examples of **audio feature estimation** using bimodal enhancement (additive speech babble noise at **4 dB SNR**): Not perfect, but better than noisy features, and helps ASR!

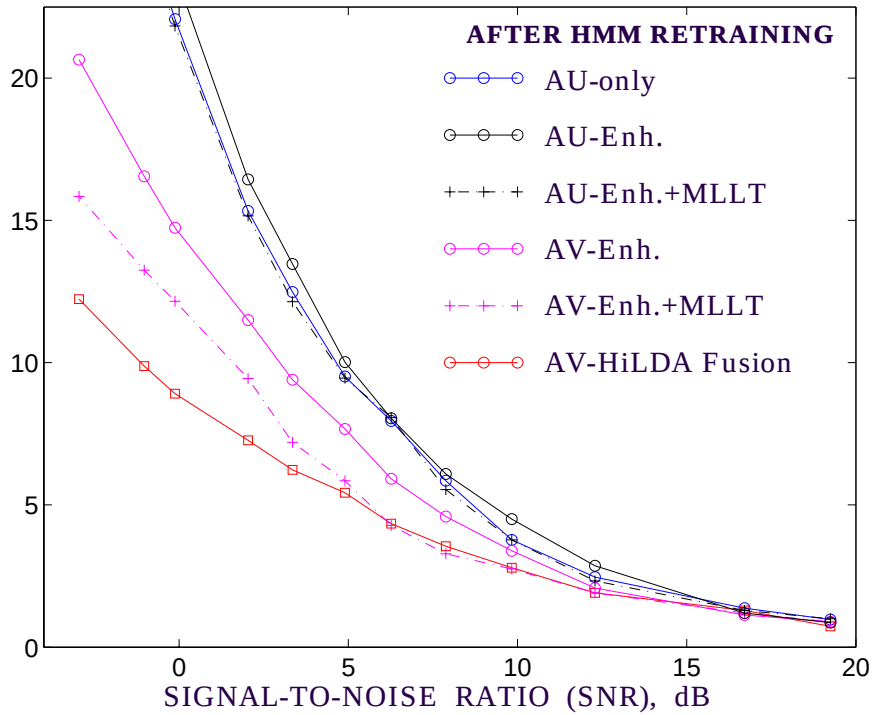




Linear bimodal audio enhancement – Cont..

- **Linear enhancement and ASR** (digits task – automobile noise):
  - Audio-based enhancement is inferior to bimodal one.
  - For mismatched HMMs at low SNR, AV-enhanced features outperform AV-HiLDA feature fusion.
  - After HMM retraining, HiLDA becomes superior.
  - Linear enhancement creates within-class feature correlation - MLLT can help.





Audio-visual speaker detection

Applications/problems:

- **Audio-visual speaker tracking** in 3D-space (e.g., meeting rooms). Signals are available from microphone arrays and video cameras. Three approaches:
  - Audio-guided active camera (Wang and Brandstein, 1999).
  - Vision-guided microphone arrays (Bub, Hunke, and Waibel, 1995).
  - Joint audio-visual tracking (Zotkin, Duraiswami, and Davis, 2002).
- **Audio-visual synchrony** in video: Which (if any) face in the video corresponds to the audio track? Useful in broadcast video.
  - Joint audio-visual speech activity can be quantified by **mutual information** of the audio and visual observations (Nock, Iyengar, and Neti, 2000):

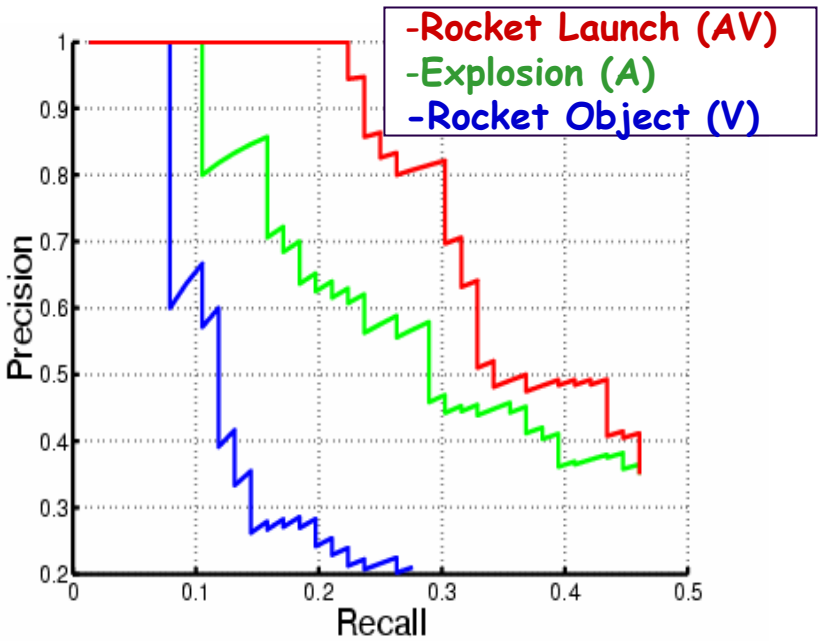
$$I(A;V) = \sum_{\mathbf{a}\in A;\mathbf{v}\in V} P(\mathbf{a},\mathbf{v}) \log \frac{P(\mathbf{a},\mathbf{v})}{P(\mathbf{a})P(\mathbf{v})} = \frac{1}{2}\log \frac{|s_{\mathbf{a}}||s_{\mathbf{v}}|}{|s_{\mathbf{a},\mathbf{v}}|}$$

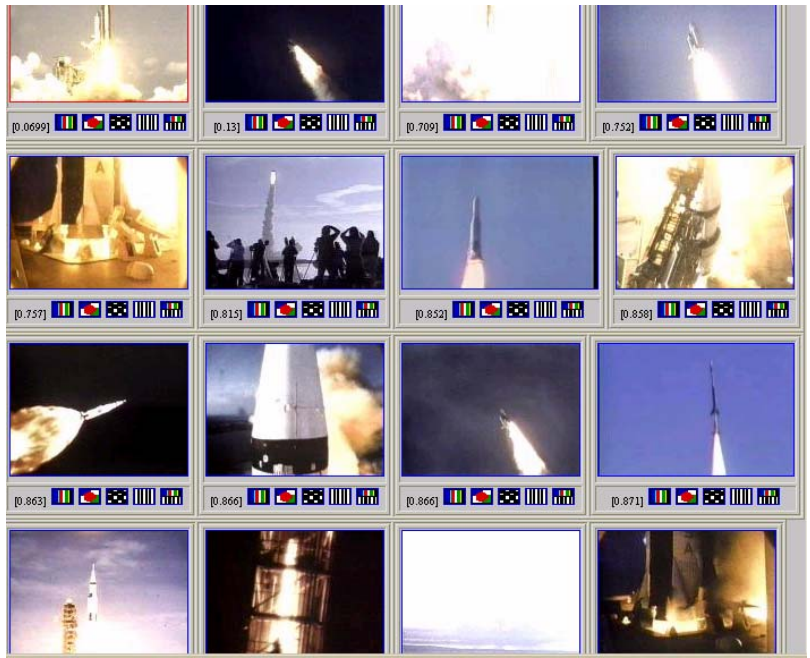
- **Speech intent detection:** User pose, proximity, and visual speech activity indicate speaker intent for HCI. Visual channel improves robustness compared to audio-only system (De Cuetos and Neti, 2000).

Audio-visual synchrony and tracking (Nock, Iyengar, and Neti, 2000).

Audio-visual mining of multimedia data.

- **Main idea:** Utilize both audio and visual channels to represent, search, and retrieve content from video corpora (news, etc.).
- **Example** from Adams et al., 2003: Retrieve videos with rocket launch content using audio, visual, or bimodal cues.



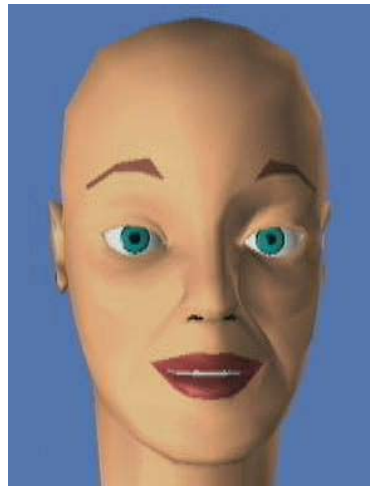


Audio-Visual Speech Synthesis

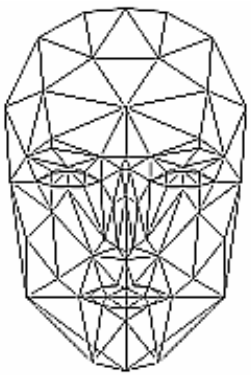
- Deals with
  - the automatic generation of voice and facial animation from arbitrary text, or
  - the automatic generation of facial animation from arbitrary speech
- Applications include human communication and perception, tools for the hearing impaired, spoken and multimodal agent-based user interfaces (newscasters, helpers on desktops, messenger with emails, personal friends), aid in education, and synthetic actors in entertainment
- A view of the face can improve intelligibility of both natural and synthetic speech significantly, especially under degraded acoustic conditions
- Moreover, facial expressions can signal emotion, add emphasis to the speech and support the interaction in a dialogue situation

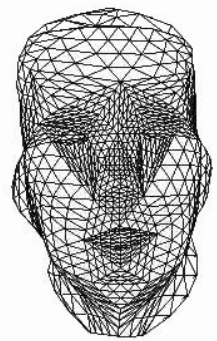
Direct Parameterization

- **Parke's Descendants**
  - *Baldi* (Cohen and Massaro, 1993, UCSC)
  - Finnish talking head (Olives et al 1999)
  - *Kattis, Holger, August* (Beskow, KTH 1995)
  - Eisert (2000)



Candide

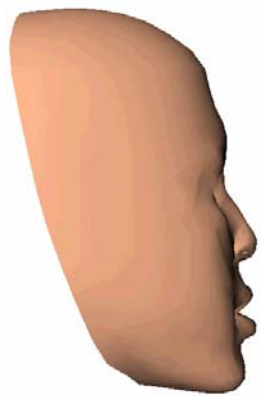


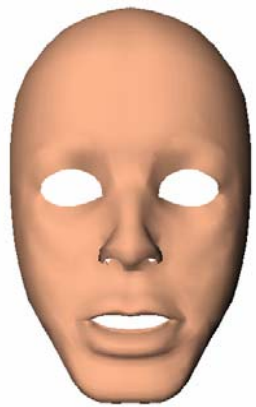


Muscular and Physiological Models

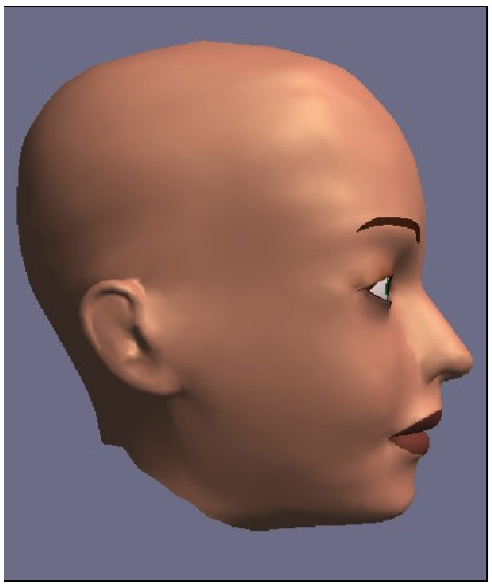
Water's descendants

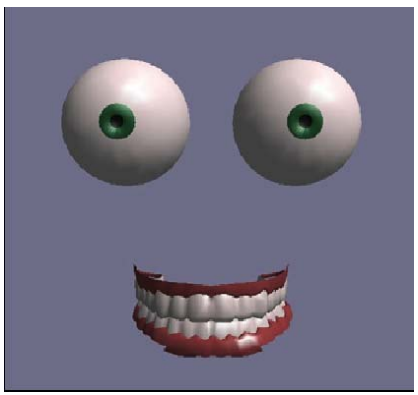
- Pseudo-muscular models: direct parameterization models that use the human facial muscle structure for modeling deformations
- The study of the natural anatomical limitations of the human face reduces the space of allowable configurations
- Muscles are modeled (as, e.g., linear contractors) with one end affixed to the bone structure of the skull and the other end attached to the skin
- More detailed physiological models have also been developed by modeling the skin with three spring-mass layers





GRETA (S. Pasquariello, C. Pelachaud, 2001)

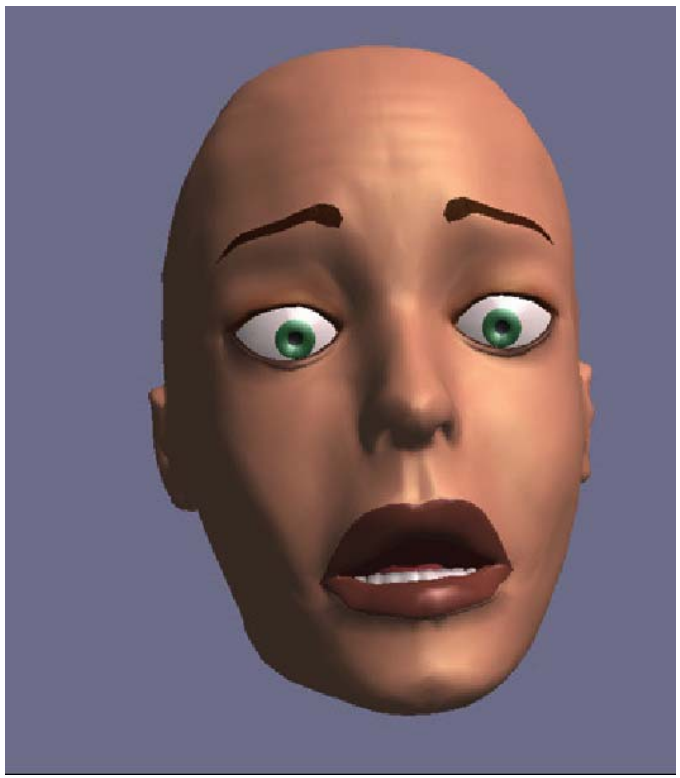




- Realization of a Simple Facial Animation Engine
- core of an MPEG-4 decoder and compliant with the “Simple Facial Animation Object Profile” of the standard
- A 3D facial model consisting of 15,000 polygons
- Able to generate the structure of a 3D model, animate it, and render it in real time
- Uses a pseudo-muscular approach to emulate the behavior of the face
- Includes particular features, such as wrinkles and furrow, to enhance realism

Internal anatomic components

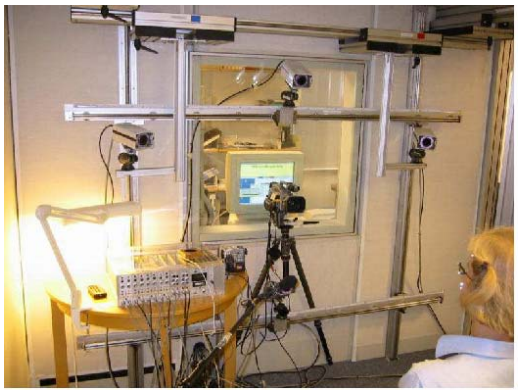
GRETA: Able to express emotions

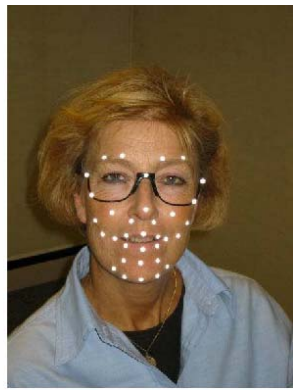


Data sources for talking heads

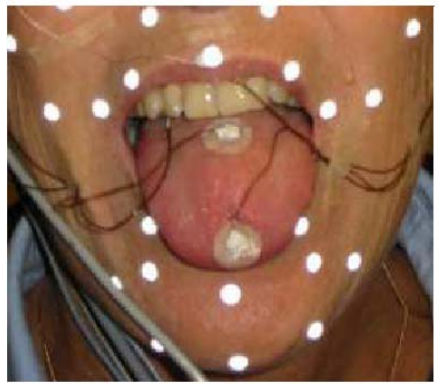
- How is the initial 3D mesh obtained
- Static methods
  - 3D photogrammetry
  - Laser-based scanning
  - Internal static methods
- Dynamic methods
  - Video-based methods
  - Systems for optical tracking
  - Non-optical internal dynamic methods (ultrasound, EPG, x-ray micro-beam, MRI, cineradiography)

Examples





Four camera measurement setup and subject with reflexive markers (Beskow 2003)



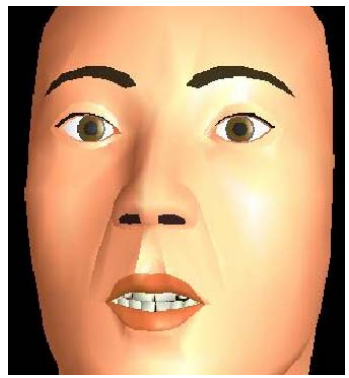
EMA coils glued to the tongue of the Subject (Beskow 2003)

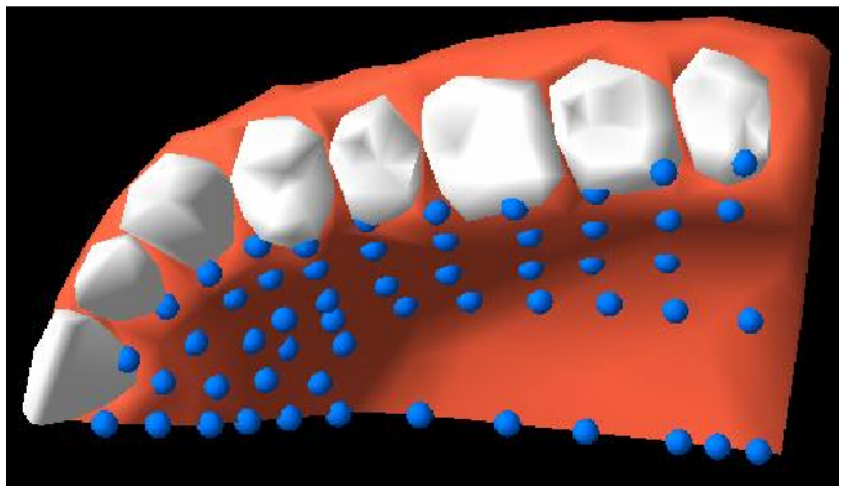
Gathering fleshpoint positions Using a photogrammetric method (Bailly 2001)

Modeling of Internal Articulators

- ***visual speech synthesis should be driven by detailed studies of how humans produce speech***
- ***highly realistic palate, teeth, and tongue models using 3D ultrasound data and electropalatography (EPG)***
- ***describe correct articulation for pedagogical purposes (e.g., provide visible speech targets for the hearing impaired)***
- ***should speech production be multimodal, as is speech perception?***
- ***use of MRI and uptrasound data for modeling the tongue***

Transparent talking head



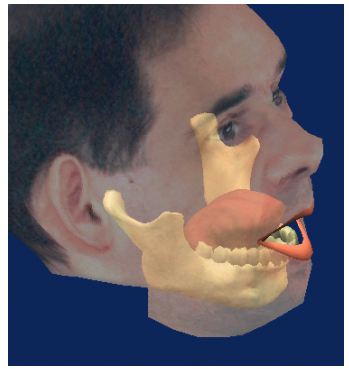


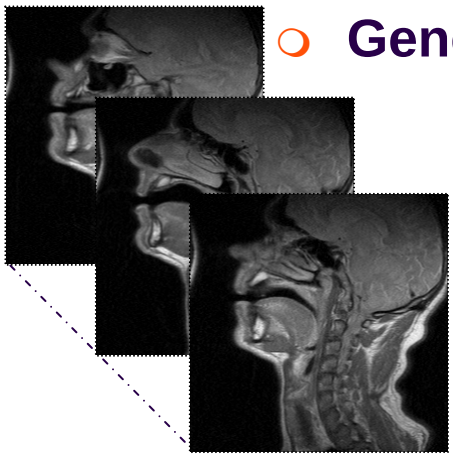
100 electrodes detect contact between Tongue and palate at 100 times per sec

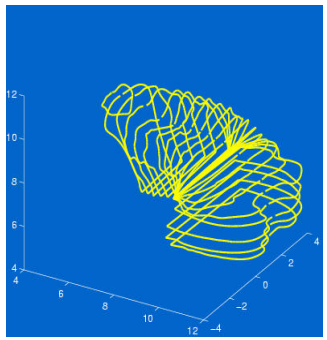
M. Cohen, J Beskow, D. Massaro, "Recent Developments in Facial animation: An Inside View", *Proc. Int. Conf. Auditory-Visual Speech Proc., AVSP'98*, pp. 201-206, Terrigal, Australia, 1998.

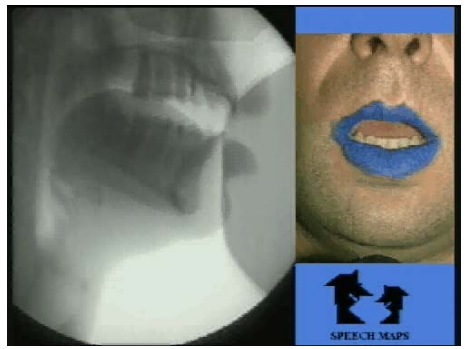
Building internal shape models

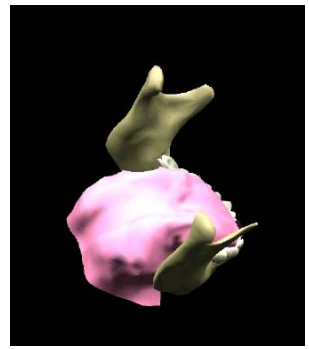
- **Speaker-specific tongue model**
  - Constrained by a generic model (adaptive grid)
- **Generic model of the jaw and teeth…**

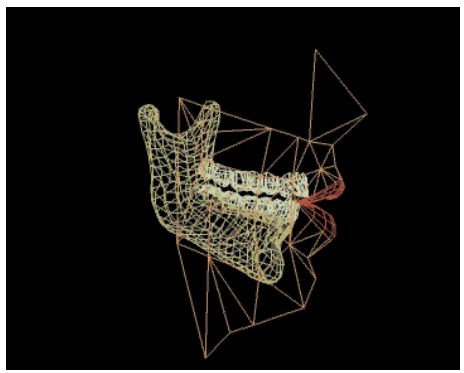


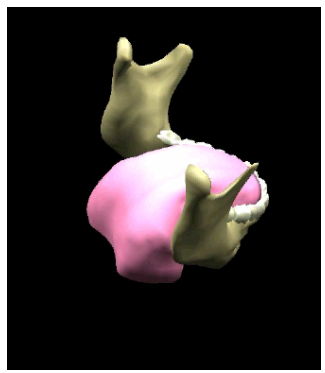


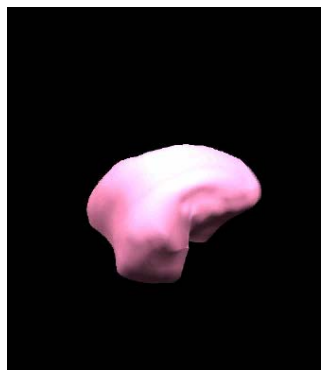


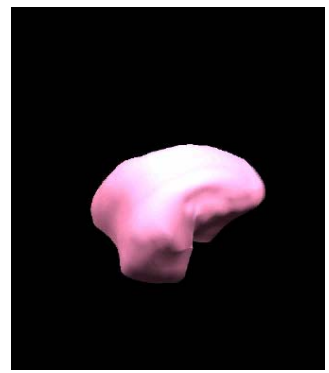












G. Bailly