Audio-Visual Automatic Speech Recognition: Theory, Applications, and Challenges

Gerasimos Potamianos

IBM T. J. Watson Research Center
Yorktown Heights, NY 10598
USA

http://www.research.ibm.com/AVSTG

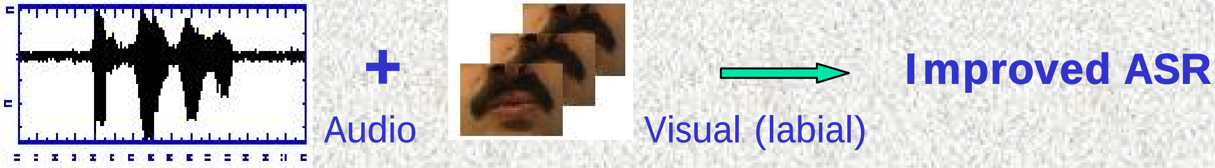
12.01.05

I. Introduction and motivation

- Next generation of Human-Computer Interaction 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; e.g:



Introduction and motivation – Cont.

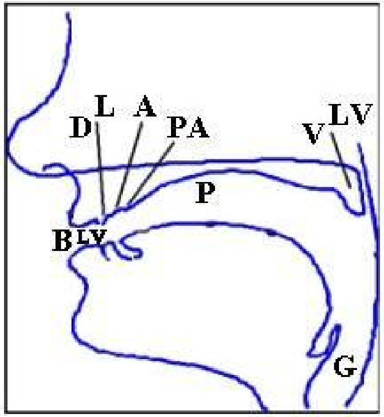
- **Vision of the HCI of the future?**

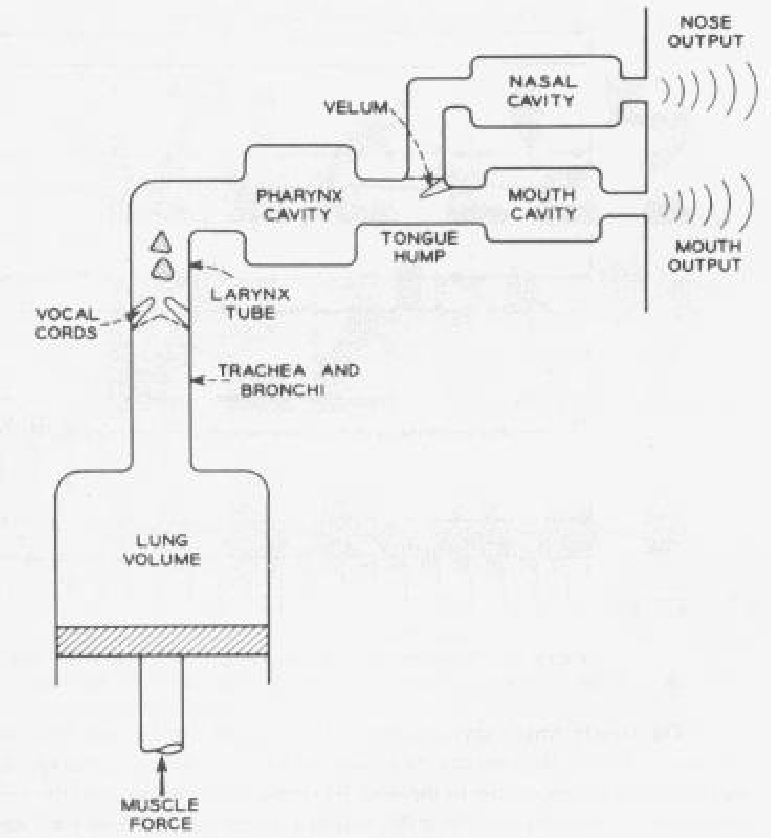
- A famous exchange (HAL’s “premature” audio-visual speech processing capability):
  - HAL: I knew that you and David were planning to disconnect me, and I’m afraid that’s something I cannot allow to happen.
  - Dave: Where the hell did you get that idea, HAL?
  - HAL: Dave – although you took very thorough 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).

I.A. Why audio-visual speech?

- **<u>Human speech production</u>** <u>is bimodal:</u>
  - 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/.

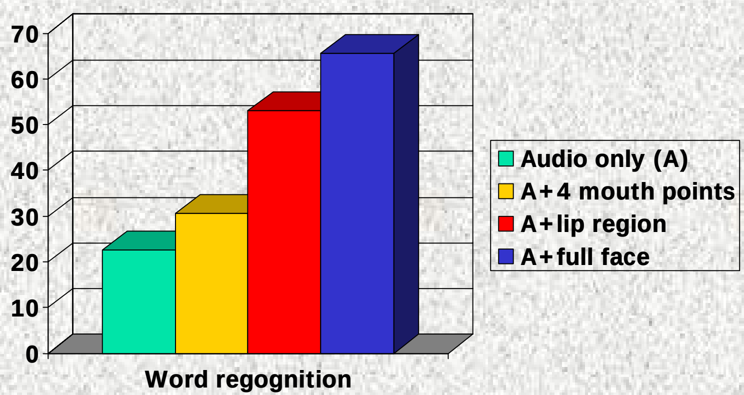




Schematic representation of speech production (J.L. Flanagan, *Speech Analysis, Synthesis, and Perception*, 2nd ed., Springer-Verlag, New York, 1972.)

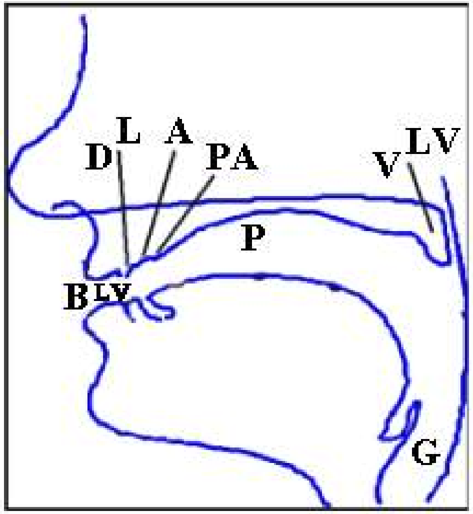
Why audio-visual speech – Cont.

- **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.



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

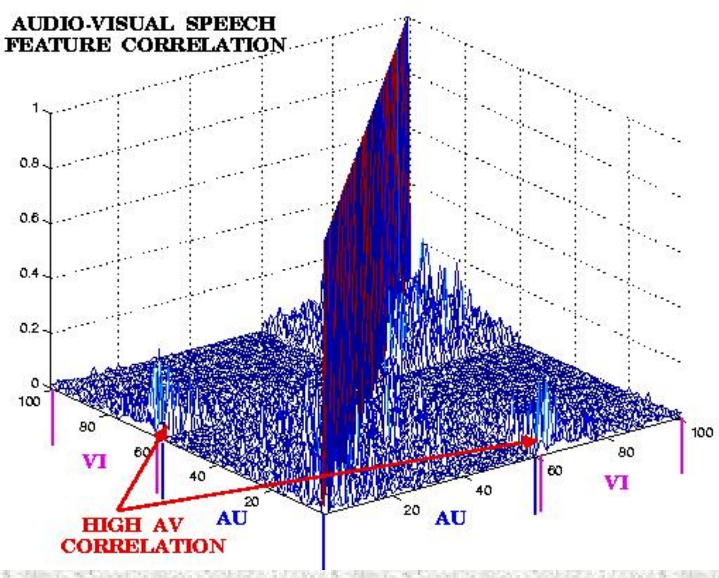
G	: Glottal	/ h /
V	: Velar	/ g, k /
P	: Palatal	/ y /
PA	: Palatoalveolar	/ r, dʒ, ʃ, tʃ, ʒ /
A	: Alveolar	/ d, l, n, s, t, z /
D	: Dental	/ θ, ð /
L	: Labiodental	/ f, v /
LV	: Labial-Velar	/ w /
B	: Bilabial	/ b, m, p /

Manner of articulation

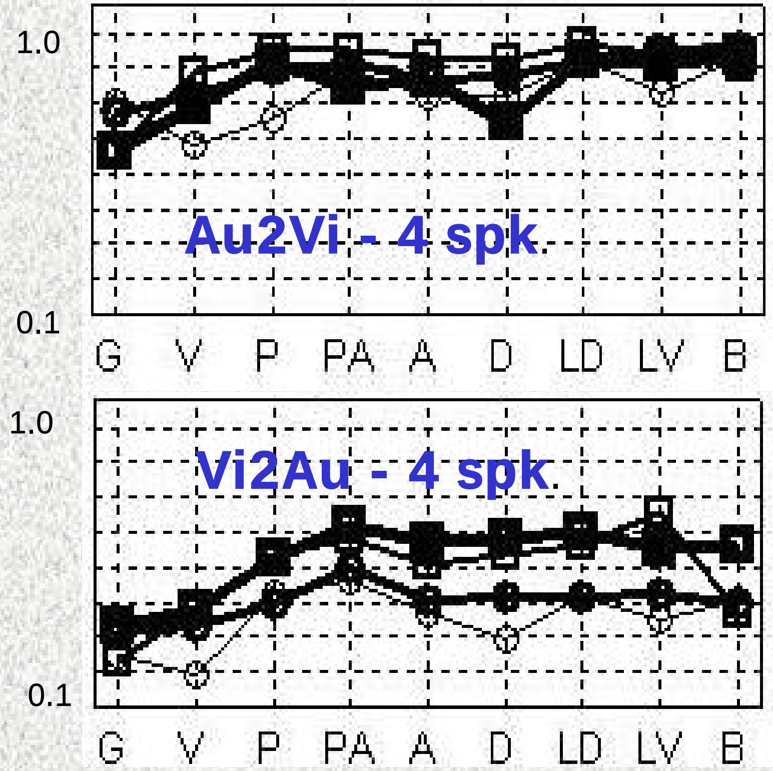
AP	: Approximant	/ r, w, y /
LA	: Lateral	/ l /
N	: Nasal	/ m, n /
PL	: Plosive	/ b, d, g, k, p, t /
F	: Fricative	/ f, h, s, v, z, θ, ð, ʃ, ʒ /
AF	: Affricate	/ tʃ, dʒ /

Why audio-visual speech – Cont.

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



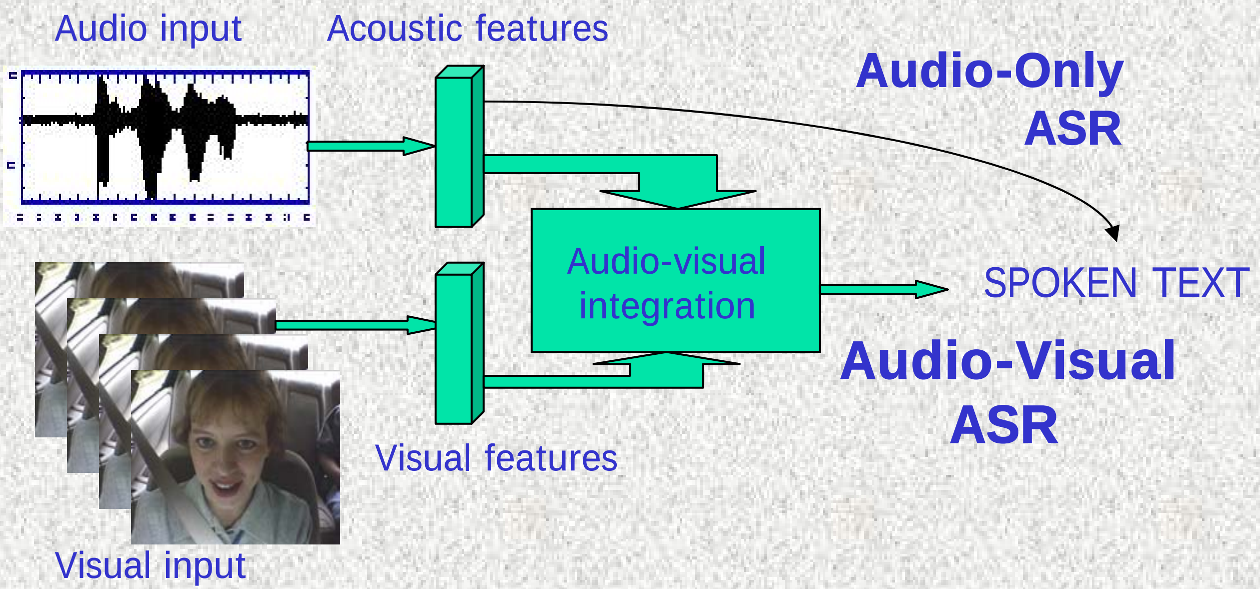
Correlation between audio and visual features (Goecke et al., 2002).



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

I.B. 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.

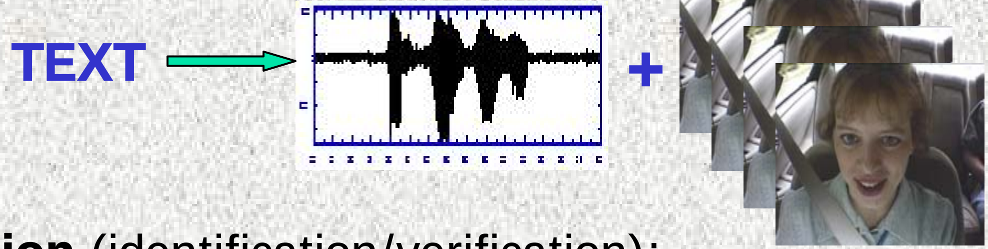


Audio-visual speech used in HCI

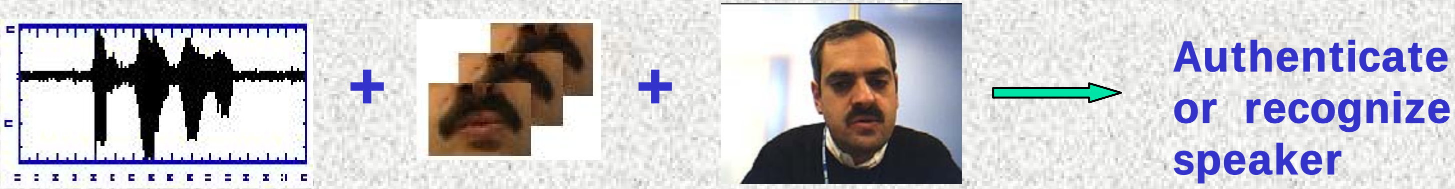
- **Audio-visual speech synthesis (AV-TTS):**
  - Given text, create a talking head (audio + visual TTS).
  - Should be more natural and intelligible than audio-only TTS.

Visual output

Audio output



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



Authenticate or recognize speaker

Audio

Visual (labial)

Face

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

Who is talking?

I.C. Outline

I. Introduction / motivation for AV speech.

II. Visual feature extraction for AV speech applications.

III. Audio-visual combination (fusion) for AV-ASR.

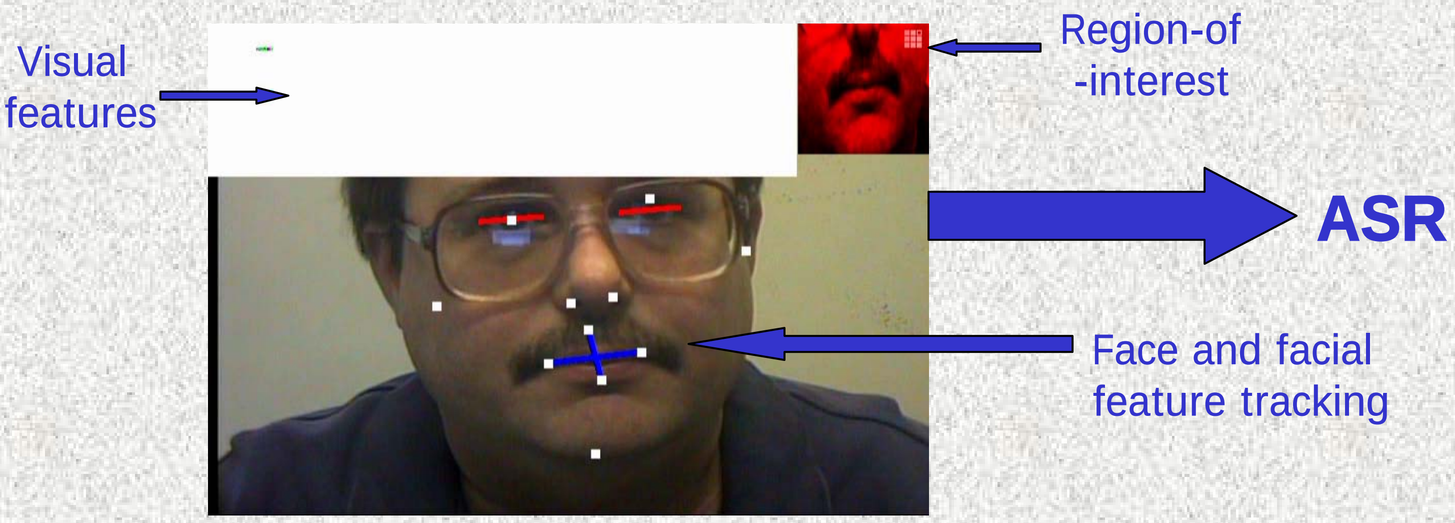
IV. Other AV speech applications.

V. Summary.

Experiments will be presented along the way.

II. Visual speech feature extraction.

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?



Visual features

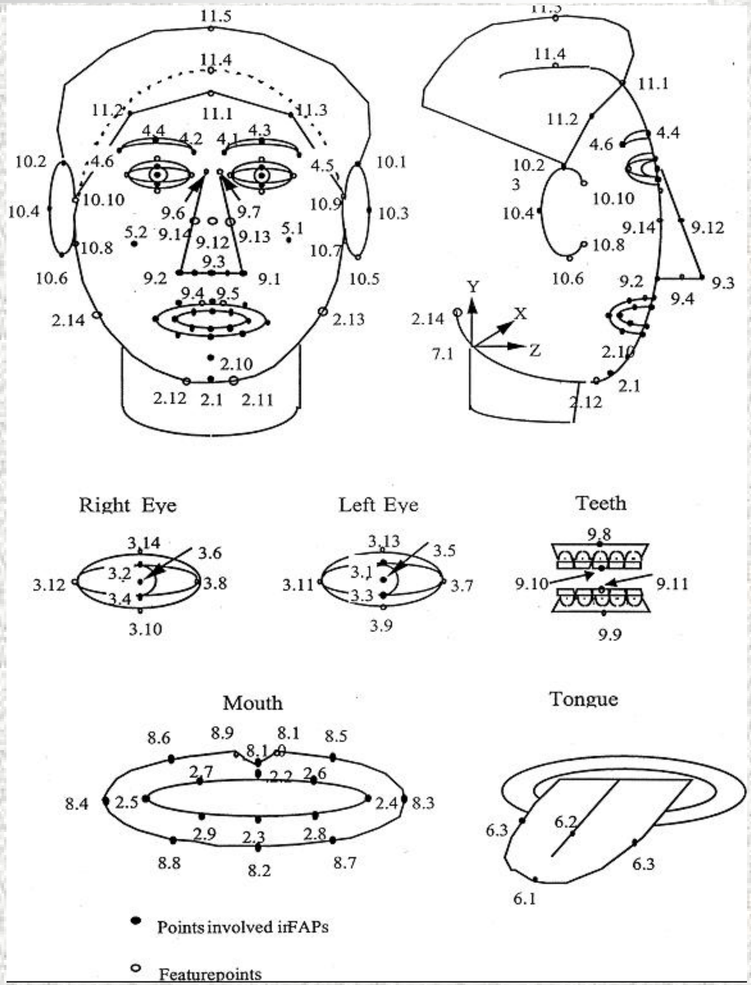
Region-of -interest

ASR

Face and facial feature tracking

II.A. Face and facial feature tracking.

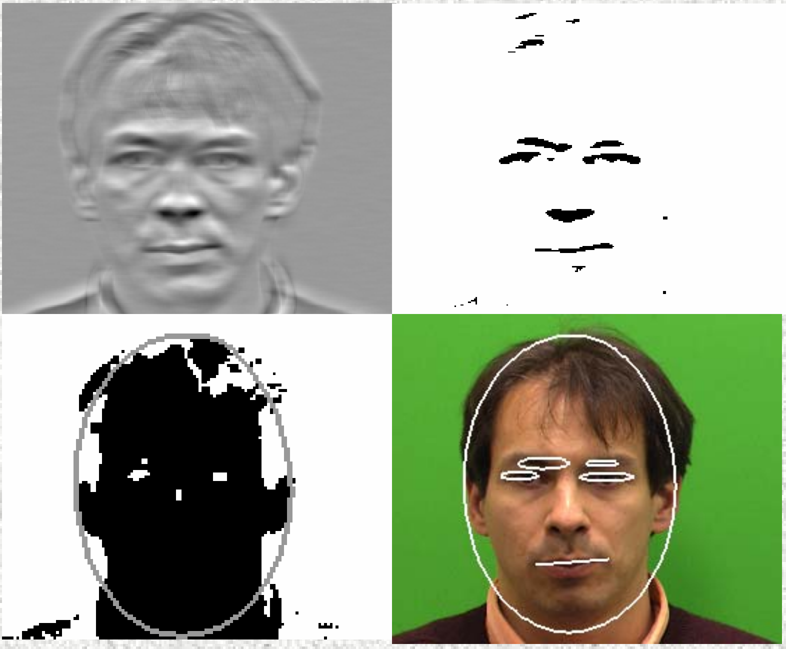
- **Main question:** Is there a ***face*** present in the video, and if so, where? Need:
  - ***Face*** detection.
  - ***Head pose*** estimation.
  - ***Facial feature*** localization (mouth corners). See for example **MPEG-4** facial activity parameters (**FAPs**).
  - Lip/face shape (contour).
- Successful 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.



II.A.1 Appearance-based face detection.

TWO APPROACHES:

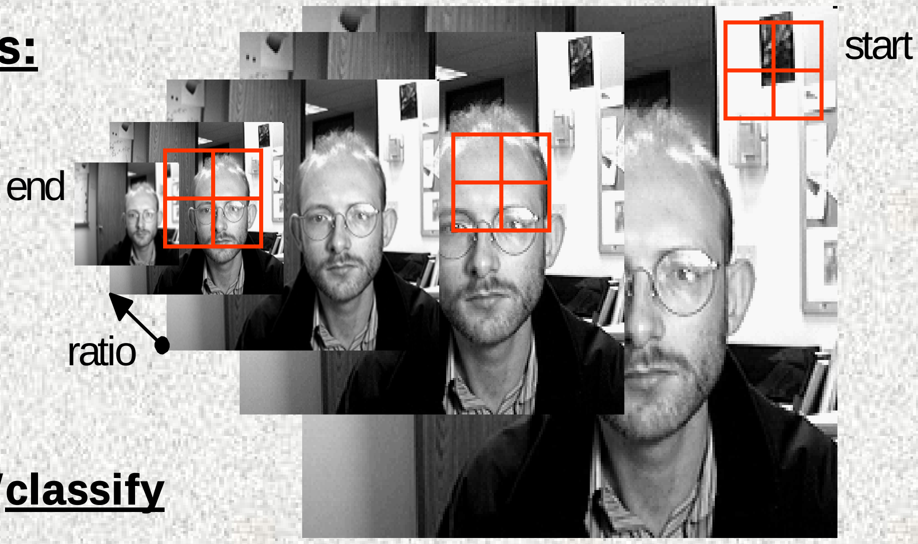
- **Non-statistical** (not discussed further):
  - Use **image processing** techniques to detect presence of typical face characteristics (mouth edges, nostrils, eyes, nose), e.g.: Low-pass filtering, edge detection, morphological filtering, etc. Obtain candidate regions of such features.
  - **Score** 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.

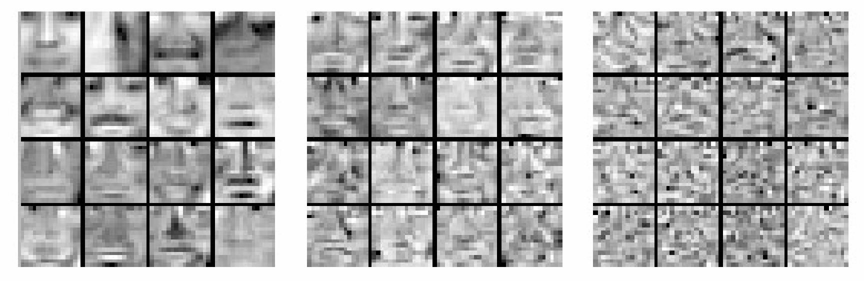
- Standard **statistical** approach – **steps:**
  - View face detection as a **2-class** classification problem (into faces/ non-faces).
  - Decide on a “**face template**” (e.g., 11x11 pixel rectangle).
  - Devise a **trainable** scheme to “score”/**classify** candidates into the 2 classes.
  - **Search** 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** (discussed in Section III.C).
  - 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 (Section III.C).
  - Resulting projection matrix $\mathbf{P}_{dx121}$ achieves best information "compression".
  - Projected vectors $\mathbf{y} = \mathbf{P}_{dx121}\, \mathbf{x}$ have a DFFS score: $\mathrm{DFFS} = \| \mathbf{x} - \mathbf{y}\, \mathbf{P}^{\mathrm{T}} \|$
- Combination of two can score a face candidate vector: $y_F - \mathrm{DFFS} \underset{Non-Face}{\overset{Face}{\gtrless}} th$

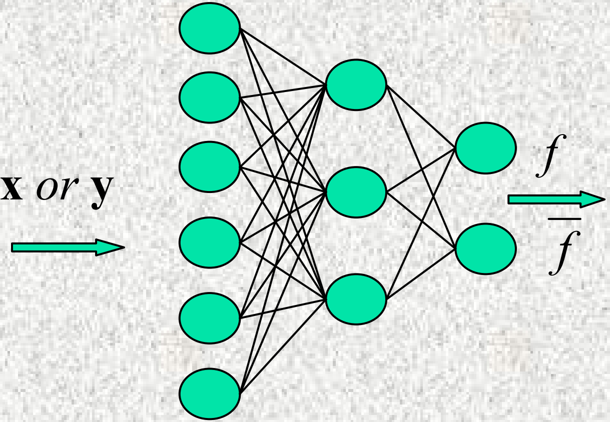


Example PCA eigenvectors

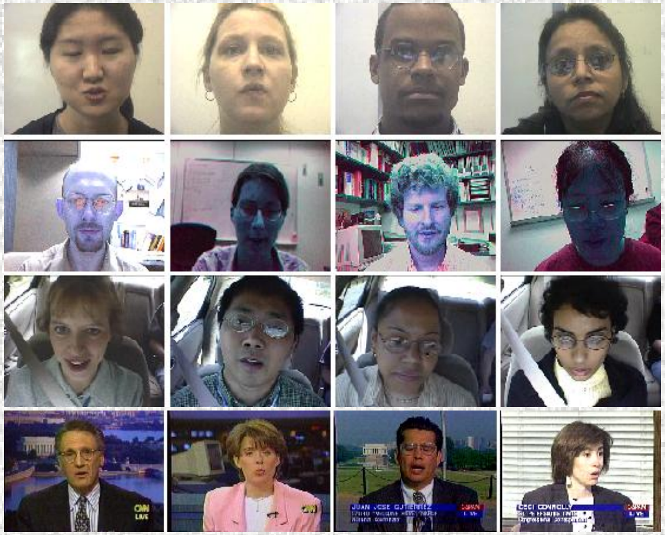
Appearance-based face detection – Cont.

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).

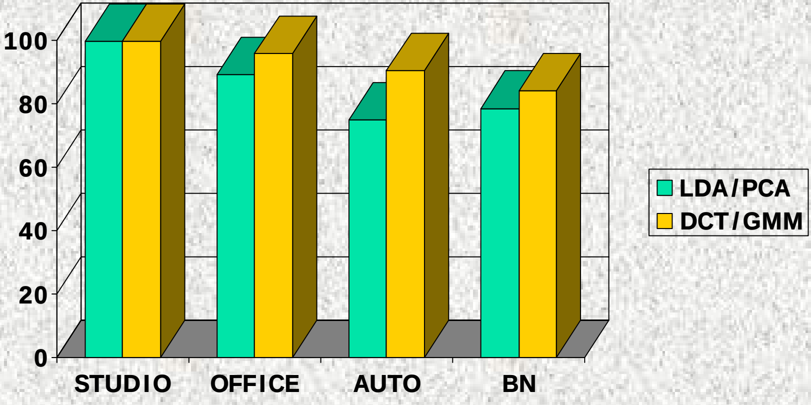


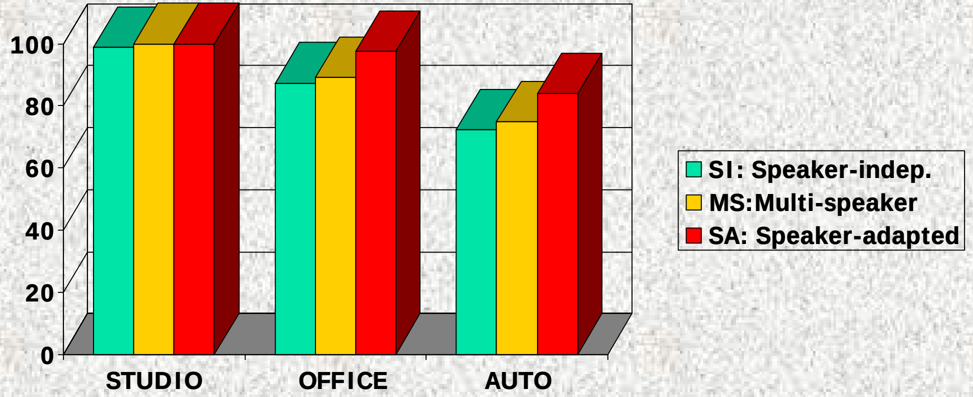
Appearance-based face detection – Cont.



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:**

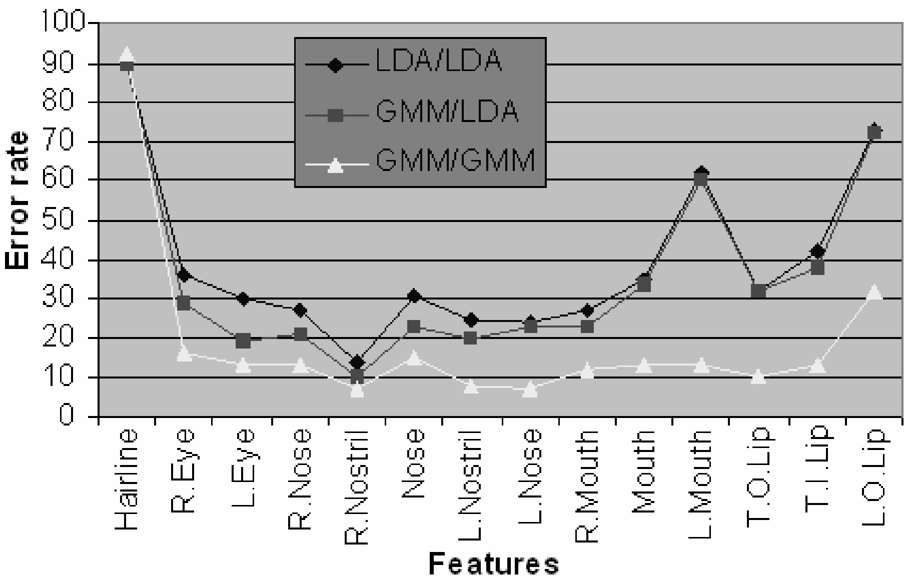


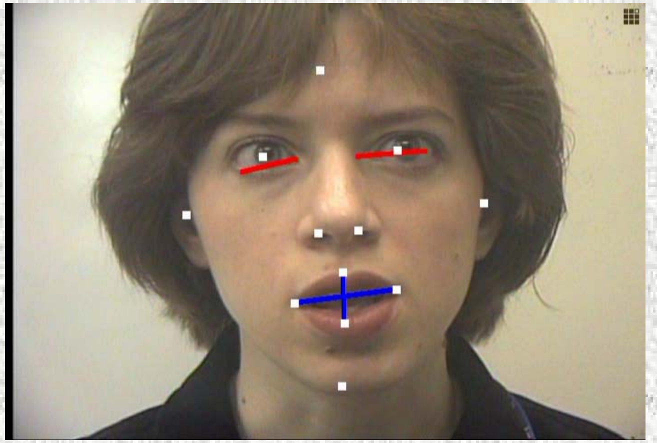


Appearance-based face detection – Cont.

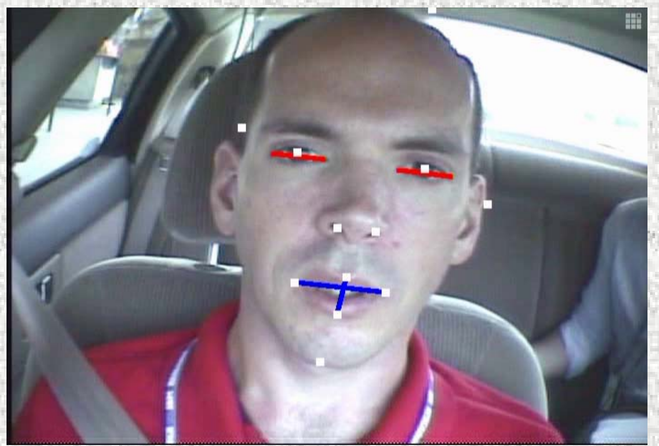
From faces to facial features:

- 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

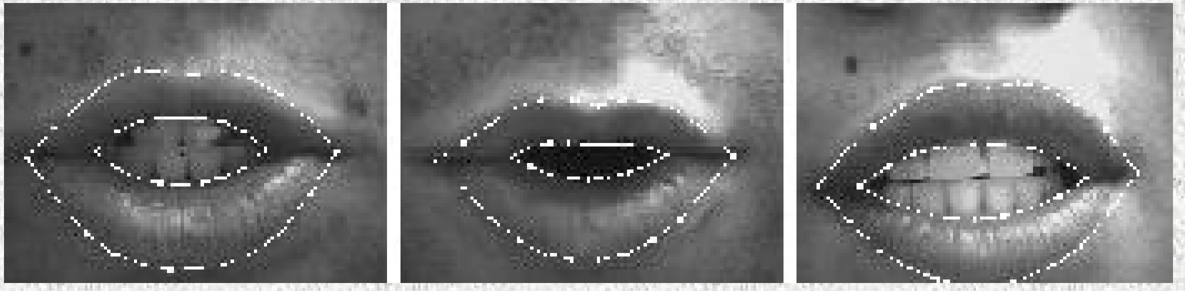
II.A.2. Face shape & lip contour extraction

Four popular methods for lip contour extraction:

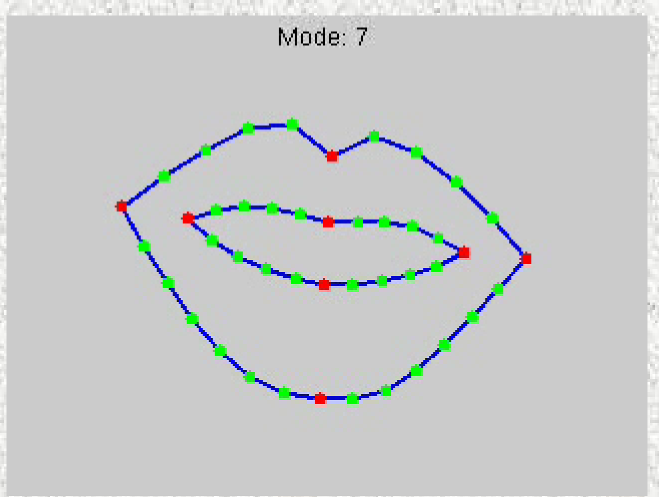
- **Snakes** (Kass, Witkin, Terzopoulos, 1988):
  - A snake is an open or closed ***elastic curve*** defined by ***control points***.
  - An ***energy function*** of the control points and the image / or edge map values is ***iteratively optimized***.
  - Correct snake ***initialization*** is crucial.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

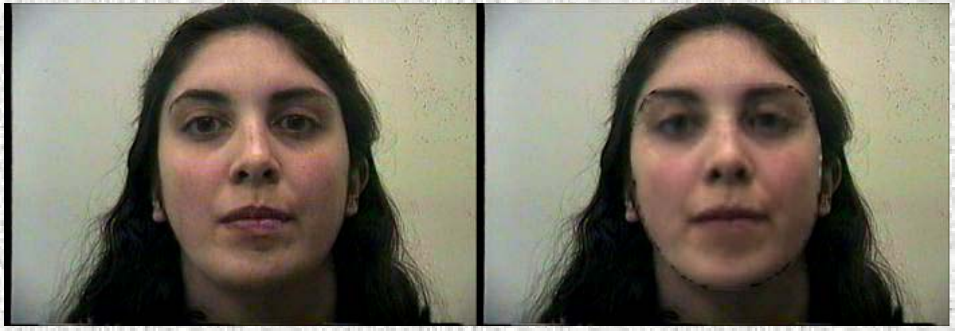


From: Luettin, Thacker, and Beet, 1996.

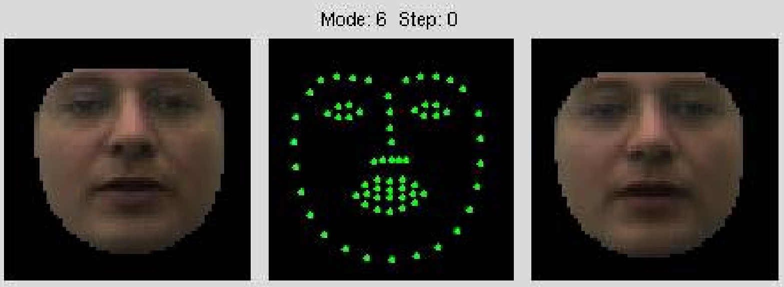


Face shape & lip contour extraction – Cont.

- **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.
  - We will re-visit AAMs in the next section.



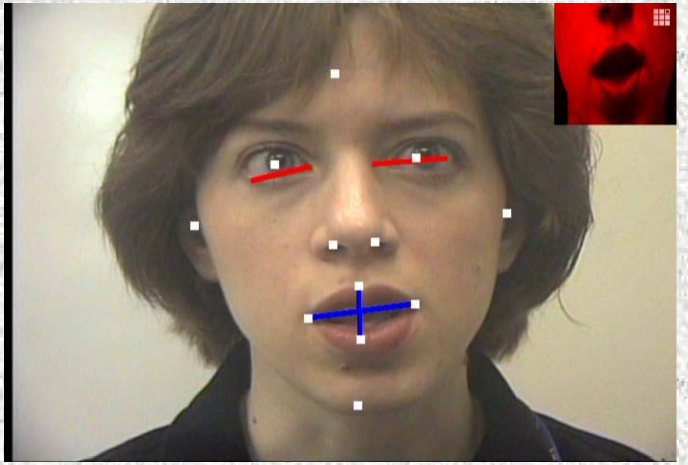
AAM tracking on IBM "studio" data (credit: I. Matthews)

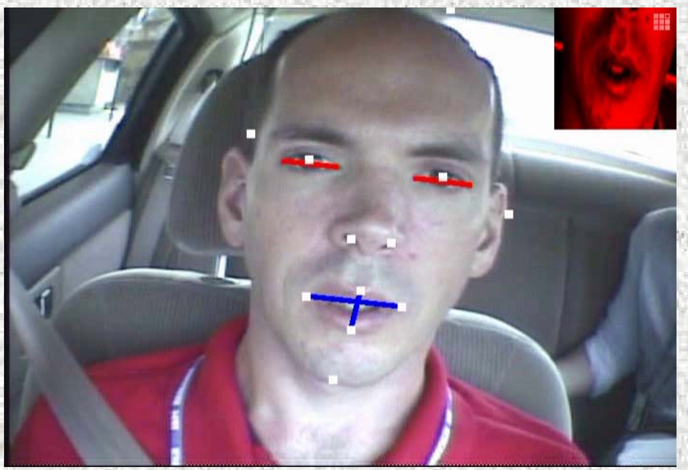


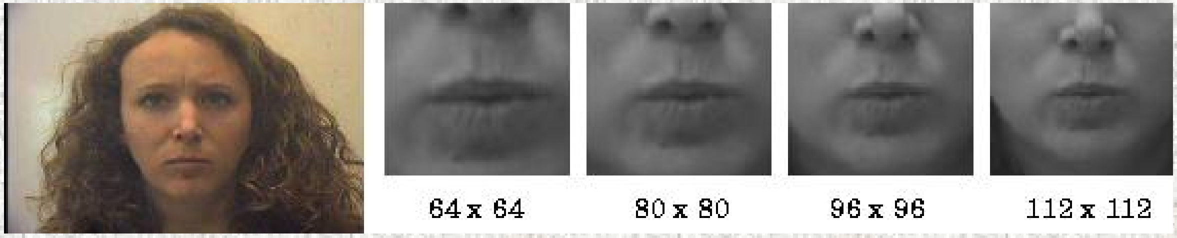
AAM modes trained on IBM data

II.B. 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 II.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.





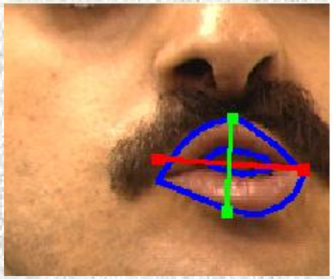


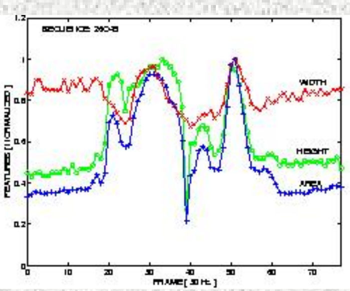
Best for ASR

II.C. Visual speech features.

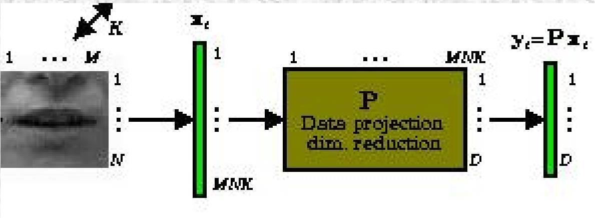
- What are the right visual features to extract from the ROI?
- Three types of / approaches to feature extraction:

- ***Lip- and face-contour (shape) based:***
  - Height, width, area of mouth.
  - Moments, Fourier descriptors.
  - Mouth template parameters.

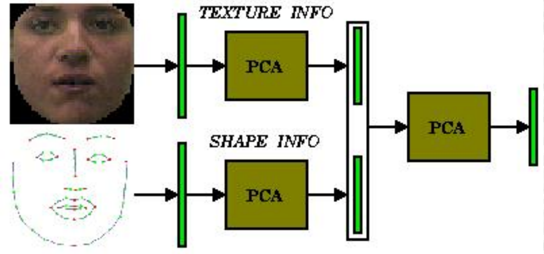




- ***Video pixel (appearance) based features:***
  - Lip contours *do not* capture oral cavity information!
  - Use compressed representation of mouth ROI instead.
  - E.g.: DCT, PCA, DWT, whole ROI.



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



II.C.1. Shape based visual features

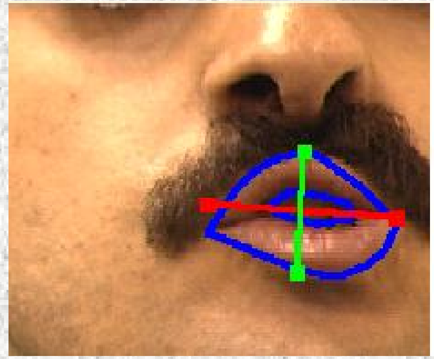
- **Geometric lip contour features:** Assume that lip contour (points) are available (extracted as discussed in III.A), and are properly normalized using an affine transform (to compensate for head pose and speaker specifics).
- **Feature extraction:**

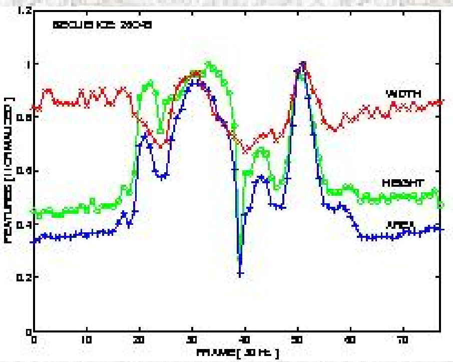
❑ Contour is denoted by $C = \{(x, y)\}$

❑ Lip-interior membership function: $f(x, y) = \begin{cases} 1, & \mathit{if}\ (x,y) \in C \cup C_{interior} \\ 0, & \mathit{otherwise} \end{cases}$

❑ Some “sensible” lip-features are then:

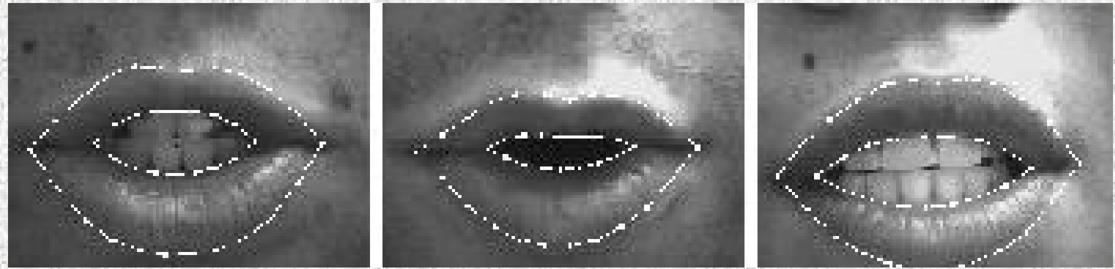
- *Height:* $\mathbf{h} = \max_x \sum_y f(x, y)$
- *Width:* $\mathbf{w} = \max_y \sum_x f(x, y)$
- *Area:* $\mathbf{a} = \sum_x \sum_y f(x, y)$
- *Perimeter:* $\mathbf{p} = \sum_i d[C_i, C_{i+1}]$
- *Lip-contour Fourier descriptors.*





Shape based visual features – Cont.

- **Lip model based features**: Various lip models can be used for lip contour tracking (as discussed in III.A). 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.



II.C.2. 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:

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

- Seek dimensionality reduction transform:

$$\mathbf{y}_t = \mathbf{P}\, \mathrm{x}_t \,, \quad \text{with}$$

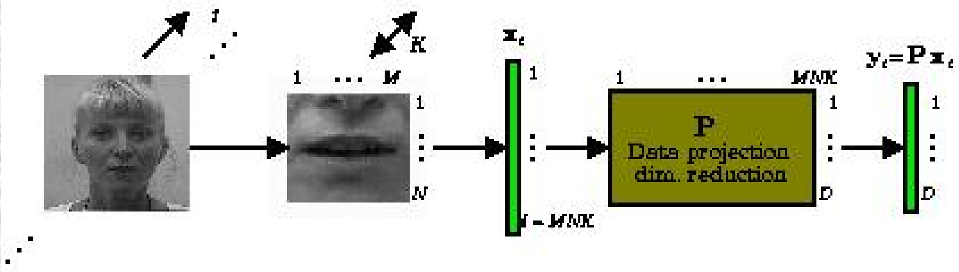
$$\mathbf{P} \in R^{D \times d}, \; D << d$$

E.g.: DCT: Discrete cosine transform.

DWT: Discrete wavelet transform.

PCA: Principal components analysis.

LDA: Linear discriminant analysis.



II.D. Visual feature comparisons.

- **Geometric** (shape) vs. appearance features (Potamianos et al., 1998).
- Comparisons are based on ***single-subject***, ***connected-digit*** ASR experiments.

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

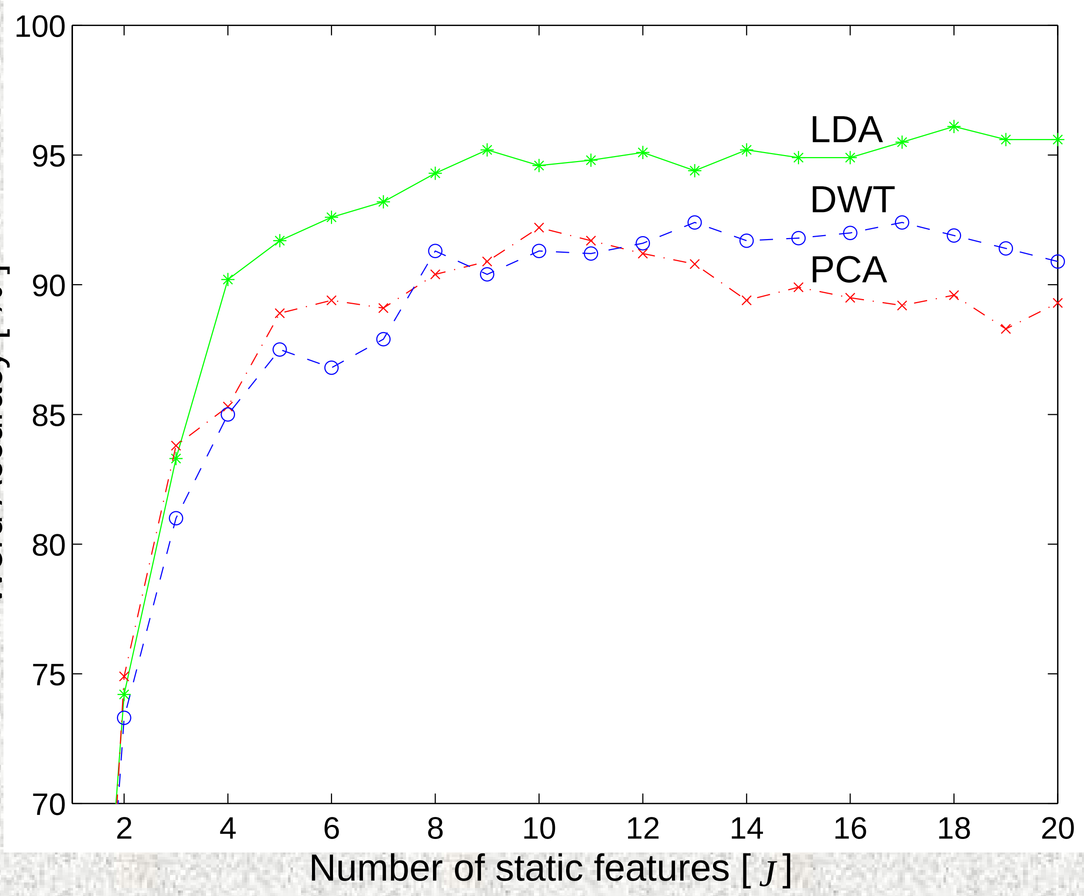
Lip contour features	%, Word accuracy
Outer-only	**73.4**
Inner-only	**64.0**
2 contours	**83.9**

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

- *Thus, appearance based modeling is preferable!*

Visual feature comparisons – Cont.

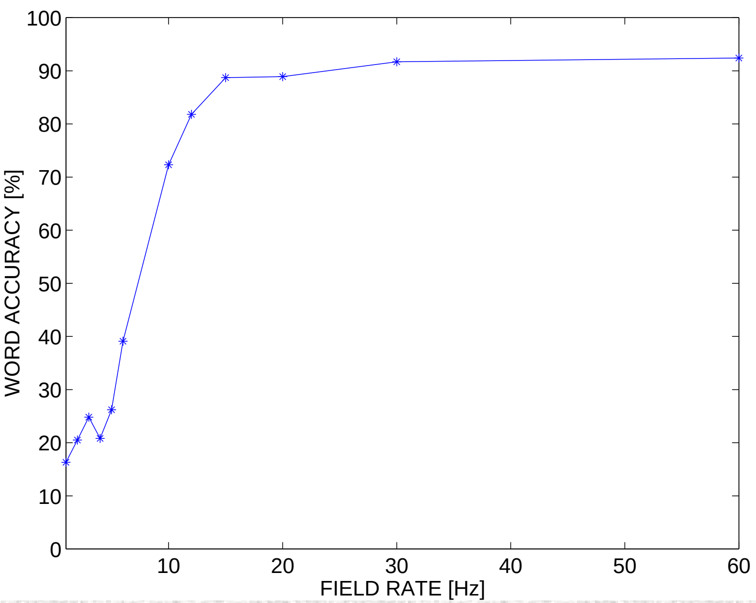
- Performance of various **appearance** based features (LDA, DWT, PCA) vs. static feature size (Potamianos et al, 1998).



II.E. Video degradation effects.

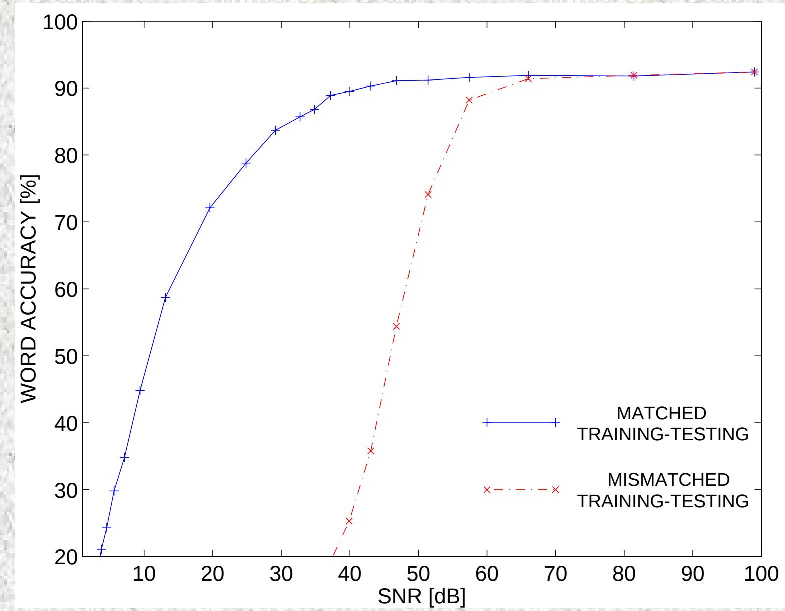
- Frame rate decimation:

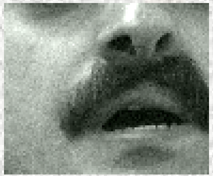
  Limit of acceptable video rate for automatic speechreading is 15 Hz.



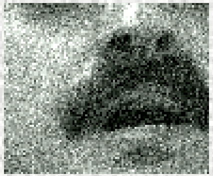
- Video noise:

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

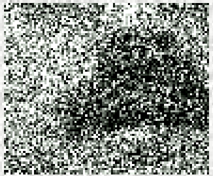




SNR = 60 dB



SNR = 30 dB



SNR = 10 dB

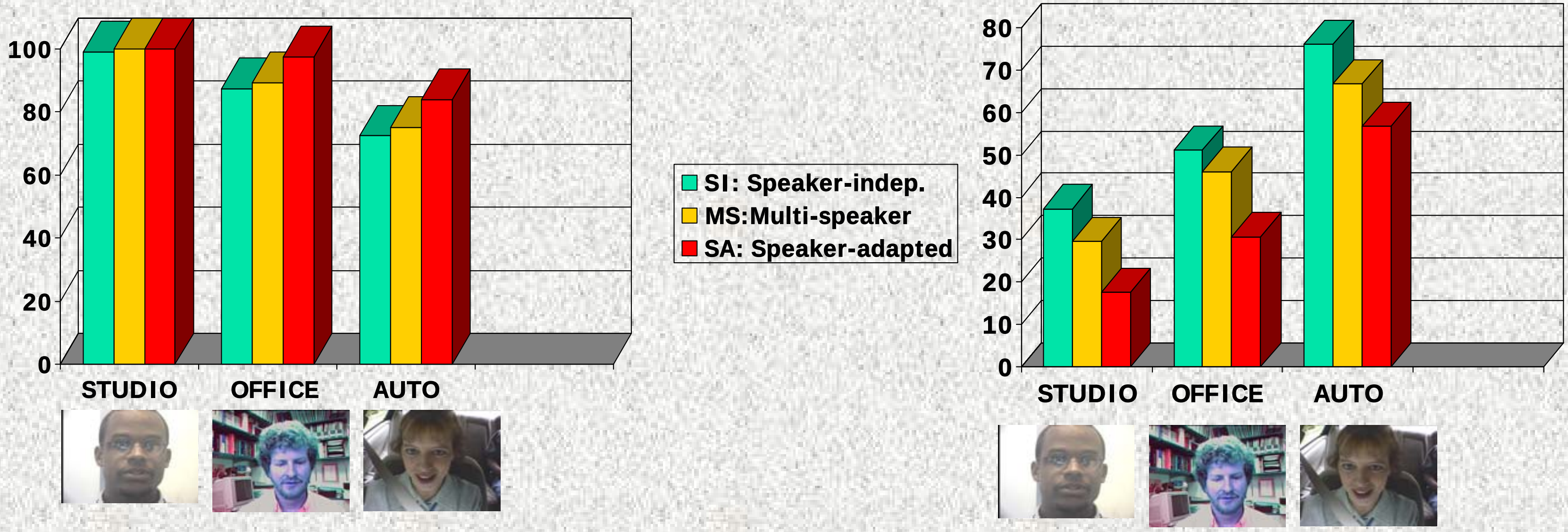
<u>Both cases:</u> *DWT visual features – connected digits recognition* (Potamianos et al., 1998).

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 – Potamianos et al., 2003).

Face detection accuracy decreases:

Word error rate increases:

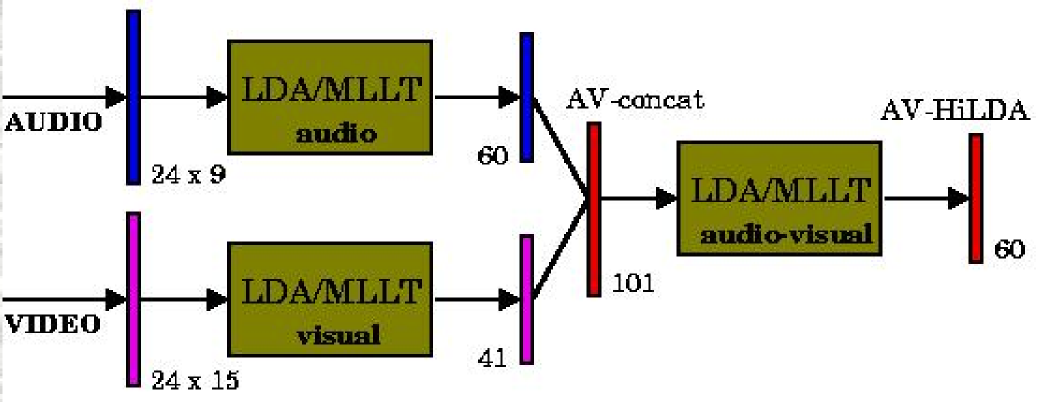


III. Audio-visual fusion for ASR.

- Audio-visual ASR:
  - ***Two*** observation streams. Audio, $\mathbf{O}_A = [\mathbf{o}_{t,A} \in R^{d_A},\ t \in T]$ Visual: $\mathbf{O}_V = [\mathbf{o}_{t,V} \in R^{d_V},\ t \in T]$
  - Streams assumed to be at ***same rate*** – e.g., 100 Hz. In our system, $d_A = 60$, $d_V = 41$.
  - We aim at ***non-catastrophic*** fusion: $\mathrm{WER}(\mathbf{O}_A, \mathbf{O}_V) \le \min[\mathrm{WER}(\mathbf{O}_A), \mathrm{WER}(\mathbf{O}_V)]$
- Main points in audio-visual fusion for ASR:
  - **Type** of fusion:
    - Combine audio and visual info at the feature level (**feature fusion**).
    - Combine audio and visual classifier scores (**decision fusion**).
    - Could envision a combination of both approaches (**hybrid fusion**).
  - Decision **level** combination:
    - **Early** (frame, HMM state level).
    - **Intermediate** integration (phone level – coupled, product HMMs).
    - **Late** integration (sentence level – discriminative model combination).
  - **Confidence** estimation in decision fusion:
    - **Fixed** (global).
    - **Adaptive** (local).
  - Fusion algorithmic performance / **experimental results**.

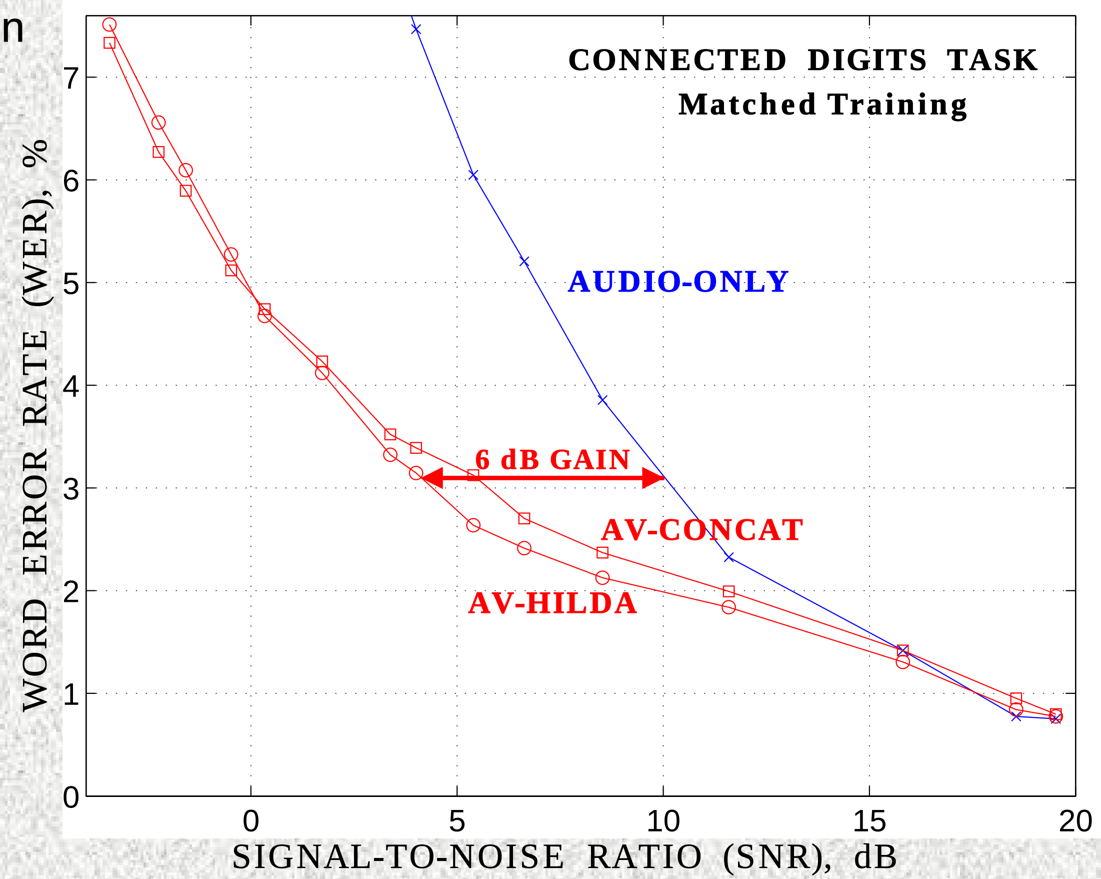
III.A. Feature fusion in AV-ASR.

- **Feature fusion:** Uses a *single classifier* (i.e.. of the same type as the audio-only and visual-only classifiers – e.g., *single-stream HMM*) to model the *concatenated* audio-visual features, or any *transformation* of them.
- **Examples:**
  - Feature **concatenation** (also known as **direct identification**).
  - Hierarchical discriminant features: LDA/MLLT on concatenated features (**HiLDA**).
  - **Dominant** and **motor recording** (transformation of one or both feature streams).
  - Bimodal **enhancement** of audio features (discussed in Section V).
- **HiLDA fusion advantages:**
  - Second LDA learns audio-visual **correlation.**
  - Achieves discriminant **dimensionality reduction.**



Feature fusion in AV-ASR – Cont.

- **AV-ASR results:** Multiple subjects (50), **connected-digits** (Potamianos et al., 2003).
  Discriminant feature fusion is superior – results in an **effective SNR gain** of **6 dB** SNR.
- Additive babble noise is considered at various SNRs.

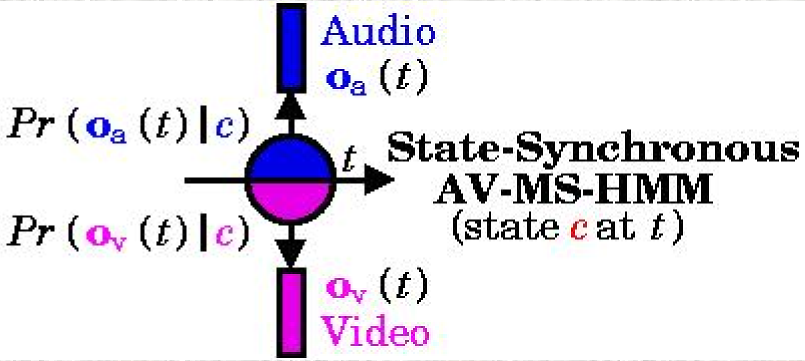


III.B. Decision fusion in AV-ASR.

- **Decision fusion:** Combines two ***separate*** classifiers (audio-, visual-only) to provide a ***joint*** audio-visual score. Typical example is the ***multi-stream HMM***.
- The **multi-stream HMM** (**MS-HMM**):
  - Combination at the frame (HMM state) level.
  - Class-conditional ($c \in C$) observation **score**:

$$\text{Score}(\mathbf{o}_{AV,t} \mid c) = \Pr(\mathbf{o}_{A,t} \mid c)^{\lambda_{A,t,c}} \Pr(\mathbf{o}_{V,t} \mid c)^{\lambda_{V,t,c}}$$

$$= \prod_{s \in \{A,V\}} \left[\sum_{k=1}^{K_{s,c}} w_{s,c,k} N_{d_s}(\mathbf{o}_{s,t}; \mathbf{m}_{s,c,k}, \mathbf{s}_{s,c,k}) \right]^{\lambda_{s,t,c}}$$



  - Equivalent to log-likelihood linear combination (**product rule** in classifier fusion).
  - Exponents (weights) capture stream reliability: $0 \leq \lambda_{s,c,t} \leq 1; \ \sum_{s \in \{A,V\}} \lambda_{s,c,t} = 1$
  - MSHMM parameters: $\boldsymbol{\theta} = [\boldsymbol{\theta}_A, \boldsymbol{\theta}_V, \boldsymbol{\lambda}], \text{ where:}$

$$\boldsymbol{\theta}_s = [(w_{s,c,k}, \mathbf{m}_{s,c,k}, \mathbf{s}_{s,c,k}), \ c \in C, \ k = 1, \ldots, K_{s,c}]$$

$$\boldsymbol{\lambda} = [\lambda_{A,c,t}, c \in C, t \in T]$$

Decision fusion in AV-ASR - Cont.

Multi-stream HMM parameter estimation:

- Parameters $[\boldsymbol{\theta}_A, \boldsymbol{\theta}_V]$ can be obtained by **ML** estimation using the **EM** algorithm.
  - **Separate estimation** (separate E,M steps at each modality):

$$\boldsymbol{\theta}_s^{(k+1)} = \arg\max_{\theta_s} Q(\boldsymbol{\theta}_s^{(k)}, \boldsymbol{\theta}_s \,|\, \mathbf{O}_s)\,, \quad \text{for } \; s \in \{A, V\}$$

  - **Joint estimation** (joint E step, M steps factor per modality):

$$\boldsymbol{\theta}_s^{(k+1)} = \arg\max_{\theta_s} Q(\boldsymbol{\theta}_s^{(k)}, \boldsymbol{\theta} \,|\, \mathbf{O})\,, \quad \text{for } \; s \in \{A, V\}$$

- Parameters λ can be obtained discriminatively – as discussed in Section IV.D.
- MS-HMM transition probabilities:
  - Scores are dominated by observation likelihoods.
  - One can set:

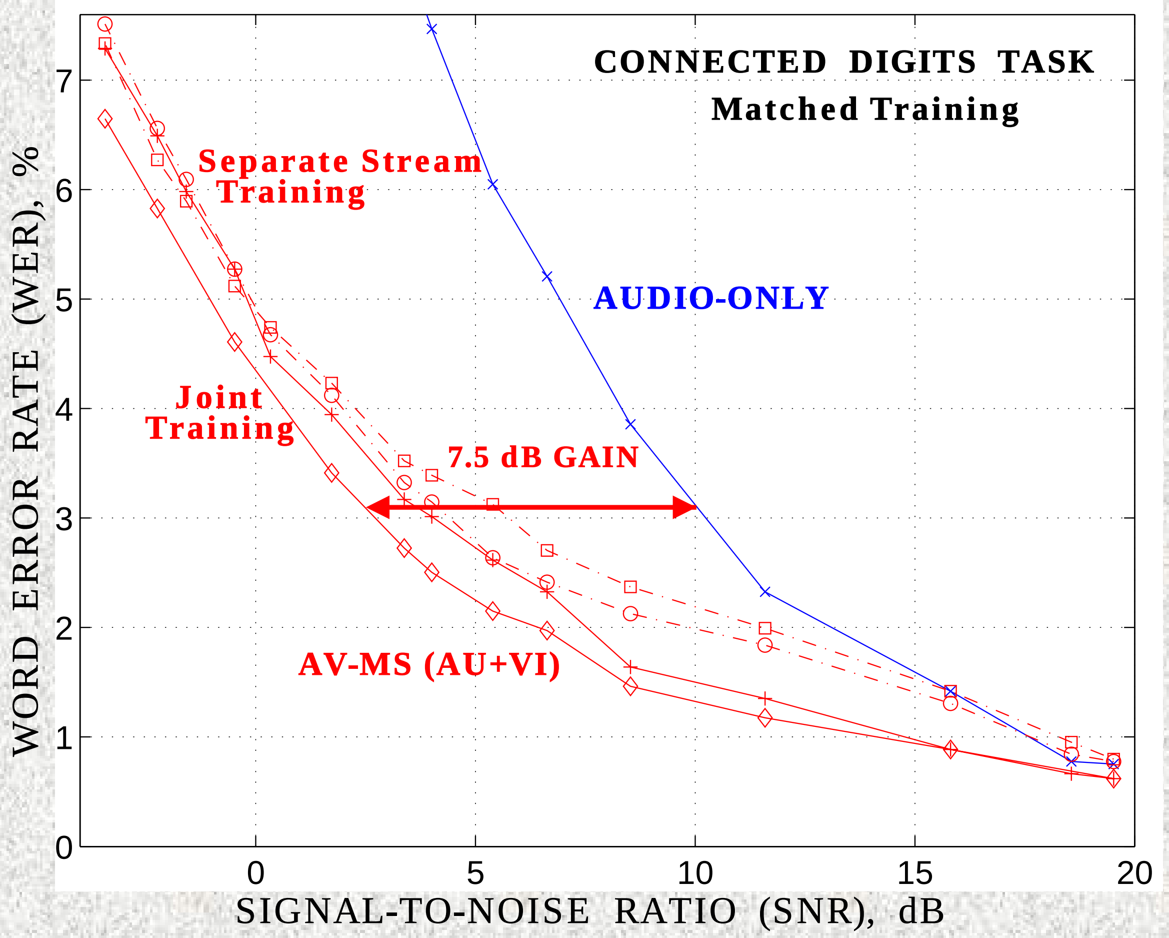
$$\mathbf{a}_{AV} = \mathbf{a}_A\,, \;\text{ or }\; \mathbf{a}_{AV} = diag(\mathbf{a}_A^{\mathrm{T}} \mathbf{a}_V)\,,$$

$$\text{where} \quad \mathrm{a}_s = [\,\Pr_s(c \,|\, c')\,, \; c, c' \in C]$$

Decision fusion in AV-ASR - Cont.

AV-ASR results:

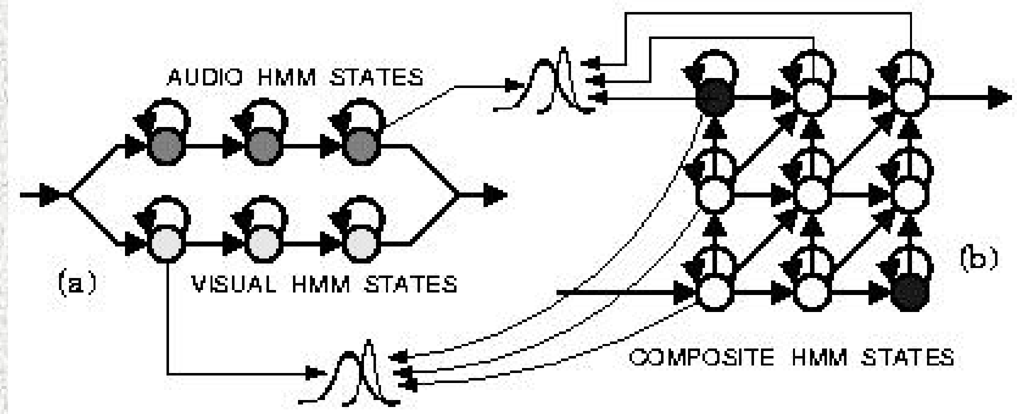
- Recall the **connected-digit** ASR paradigm.
- MSHMM-based **decision fusion** is superior to feature fusion.
- **Joint** model training is superior to **separate** stream training.
- Effective SNR gain: **7.5 dB** SNR.



III.C. Asynchronous integration

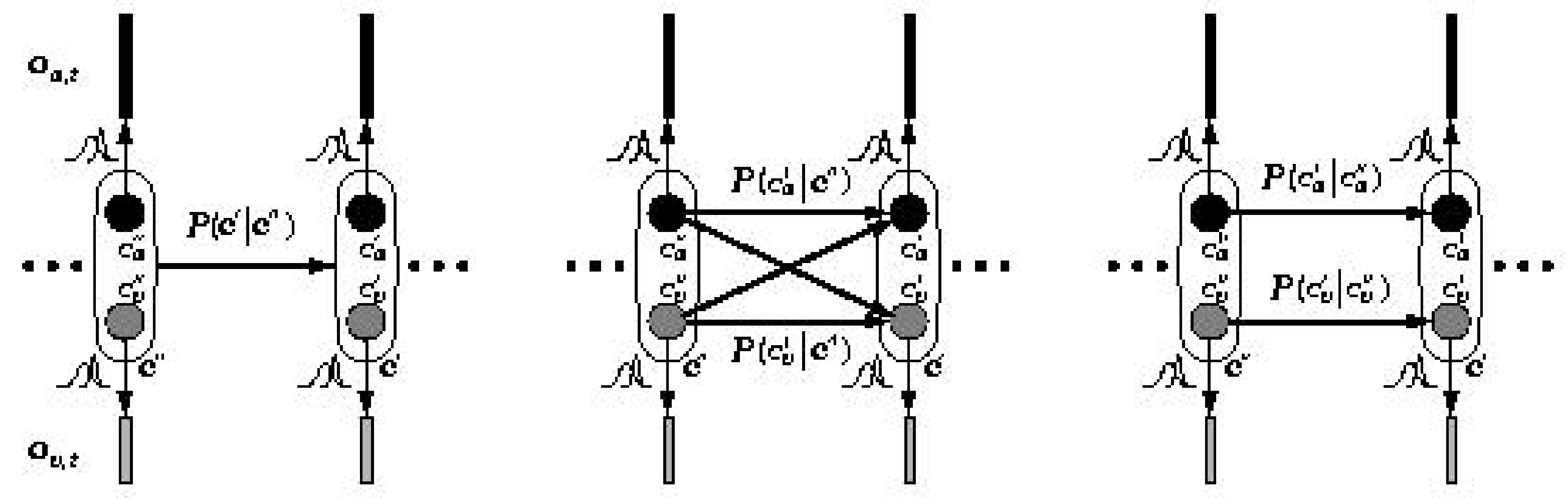
- **Intermediate integration** combines stream scores at a **coarser** unit level than HMM states, such as **phones**. This allows state-asynchrony between the two streams, within each phone.
- Integration model is <u>equivalent to the **product HMM**</u> (Varga and Moore, 1990).
  - Product HMM has "<u>**composite**</u>" (audio-visual) states: $\mathbf{c} = \{ c_s \ , \ s \in S \} \, ,$ i.e., $\mathbf{c} \in C^{|S|}$
  - Thus, state space becomes larger, e.g., $|C| \text{x} |C|$ for a 2-stream model.
  - Class-conditional observation probalities can follow the MS-HMM paradigm, i.e.:

$$\text{Score}\,(\mathbf{o}_{AV,t} \,|\, \mathbf{c}) \;= \prod\nolimits_{s \in S} \Pr(\mathbf{o}_{s,t} \,|\, c_s)^{\lambda_{s,t,c}} \ .$$



Intermediate integration - Cont.

- **Product HMM – Cont.:**
    - If properly tied, the observation probabilities have **same number** of parameters as state-synchronous MS-HMM.
    - Transition probabilities may be more. Three possible models. The miiddle is known as the **coupled HMM**.

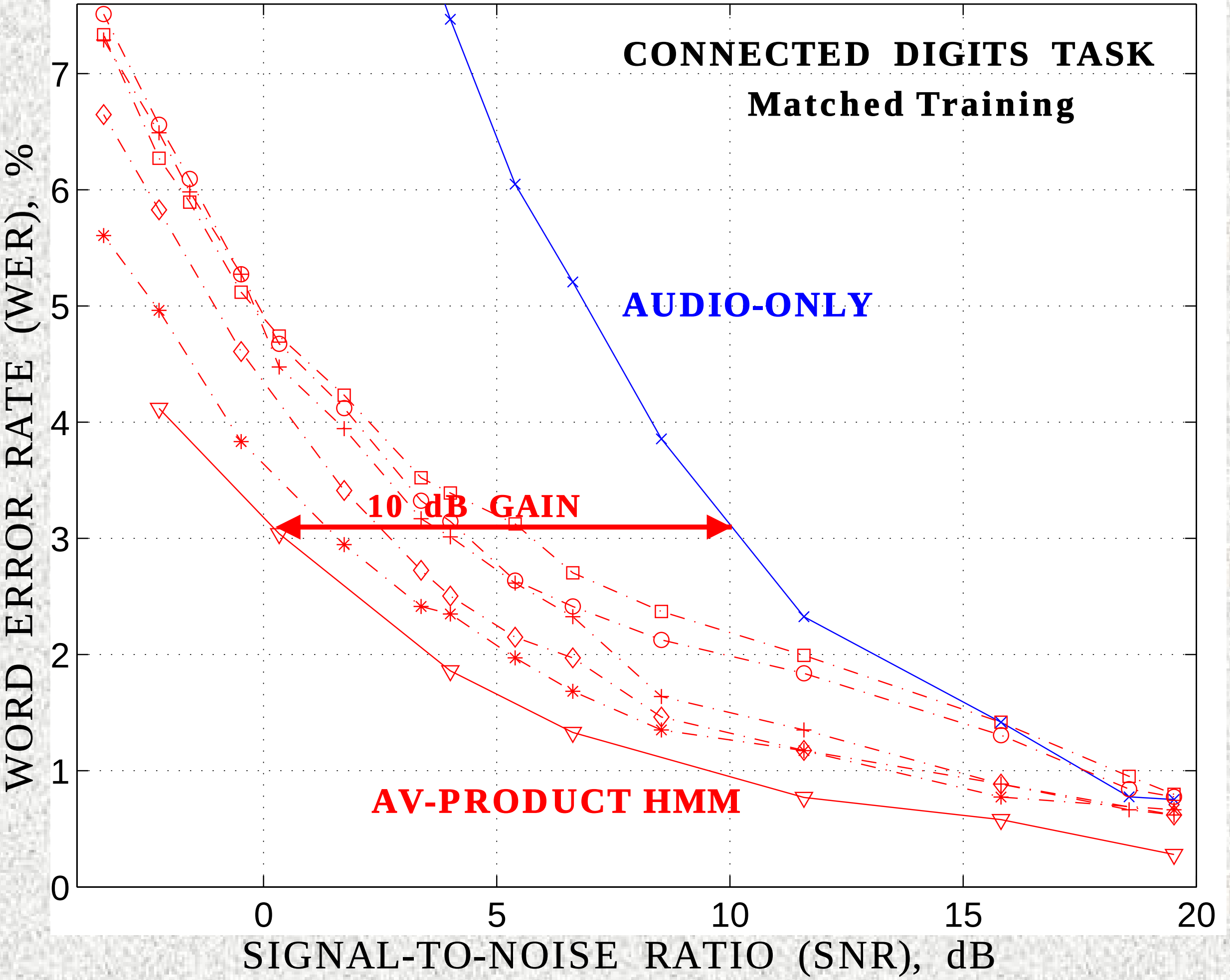


Asynchrony; Intermediate integration - Cont.

AV-ASR results:

- Recall the **connected-digit** ASR paradigm.
- **Product HMM fusion** is superior to state-synchronous fusion.
- Effective SNR gain:

  10 dB SNR.



III.D. Stream reliability modeling

- We revisit the MS-HMM framework, to discuss weight (exponent) estimation.
- Recall the MS-HMM observation score (assume 2 streams):

$$\text{Score}(\mathbf{o}_{AV,t}\,|\,c) = \Pr(\mathbf{o}_{A,t}\,|\,c)^{\lambda_{A,t,c}}\ \Pr(\mathbf{o}_{V,t}\,|\,c)^{\lambda_{V,t,c}}$$

- Stream exponents model reliability (information content) of each stream.
- We can consider:
  - **Global weights**: Assumes that audio and visual conditions do not change, thus global stream weights properly model the reliability of each stream for all available data. Allows for state-dependent weights. $\lambda_{s,c,t} \longrightarrow \lambda_{s,c}$
  - **Adaptive weights** at a **local** level (**utterance** or **frame**): Assumes that the environment varies locally (more practical). Requires stream reliability estimation at a local level, and mapping of such reliabilities to exponents.

$$\lambda_{s,c,t} \longrightarrow \lambda_{s,t} = f(\mathbf{o}_{s,t'},\ s \in \{A,V\},\ t' \in [t - t_{\text{win}}, t + t_{\text{win}}]).$$

III.D.1. Global stream weighting.

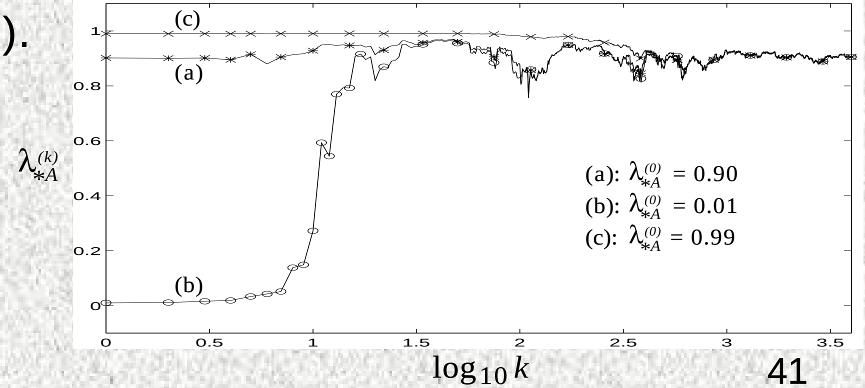
- Stream weights **cannot** be obtained by **maximum-likelihood** estimation, as:

$$\lambda_{s,c} = \begin{cases} 1, & \text{if} \quad s = \arg\max_{s \in \{A,V\}} \mathbf{L}_{s,c,F} \\ 0, & \quad \text{otherwise} \end{cases}$$

  where $\mathbf{L}_{s,c,F}$ denotes the training set log-likelihood contribution due to the s-modality, c-state (obtained by forced-alignment F).

- Instead, one needs to **discriminatively** estimate the exponents:
  - Directly minimize **WER** on a held-out set – using brute force grid search.
  - Minimize a function of the misrecognition error by utilizing the **generalized probabilistic descent** algorithm (**GPD**).
  - Example of exponent convergence →

    (GPD based estimation)



III.D.2. Adaptive stream weighting.

- In practice, stream reliability varies **locally**, due to audio and visual input degradations (e.g., acoustic noise bursts, face tracking failures, etc.).
- **Adaptive weighting** can capture such variations, by using:
  - **Estimate** of the environment and/or input stream **reliabilities.**
  - **Mapping** such estimates to stream exponents.
- Stream reliability indicators:
  - **Acoustic** signal based: SNR, voicing index.
  - **Visual** processing: Face tracking confidence.
  - **Classifier** based reliability indicators (either stream):
    - Consider N-best most likely classes for observing $\mathbf{o}_{s,t}\, ,\; c_{s,t,n} \in C, \quad n = 1,2,\ldots,N.$
    - N-best log-likelihood **difference**: $L_{s,t} = \dfrac{1}{N-1} \sum_{n=2}^{N} \log \dfrac{\Pr\,(\mathbf{o}_{s,t} \,|\, c_{s,t,1})}{\Pr\,(\mathbf{o}_{s,t} \,|\, c_{s,t,n})}$
    - N-best log-likelih. **dispersion**: $D_{s,t} = \dfrac{2}{N(N-1)} \sum_{n=2}^{N} \sum_{n'=n+1}^{N} \log \dfrac{\Pr\,(\mathbf{o}_{s,t} \,|\, c_{s,t,n})}{\Pr\,(\mathbf{o}_{s,t} \,|\, c_{s,t,n'})}$

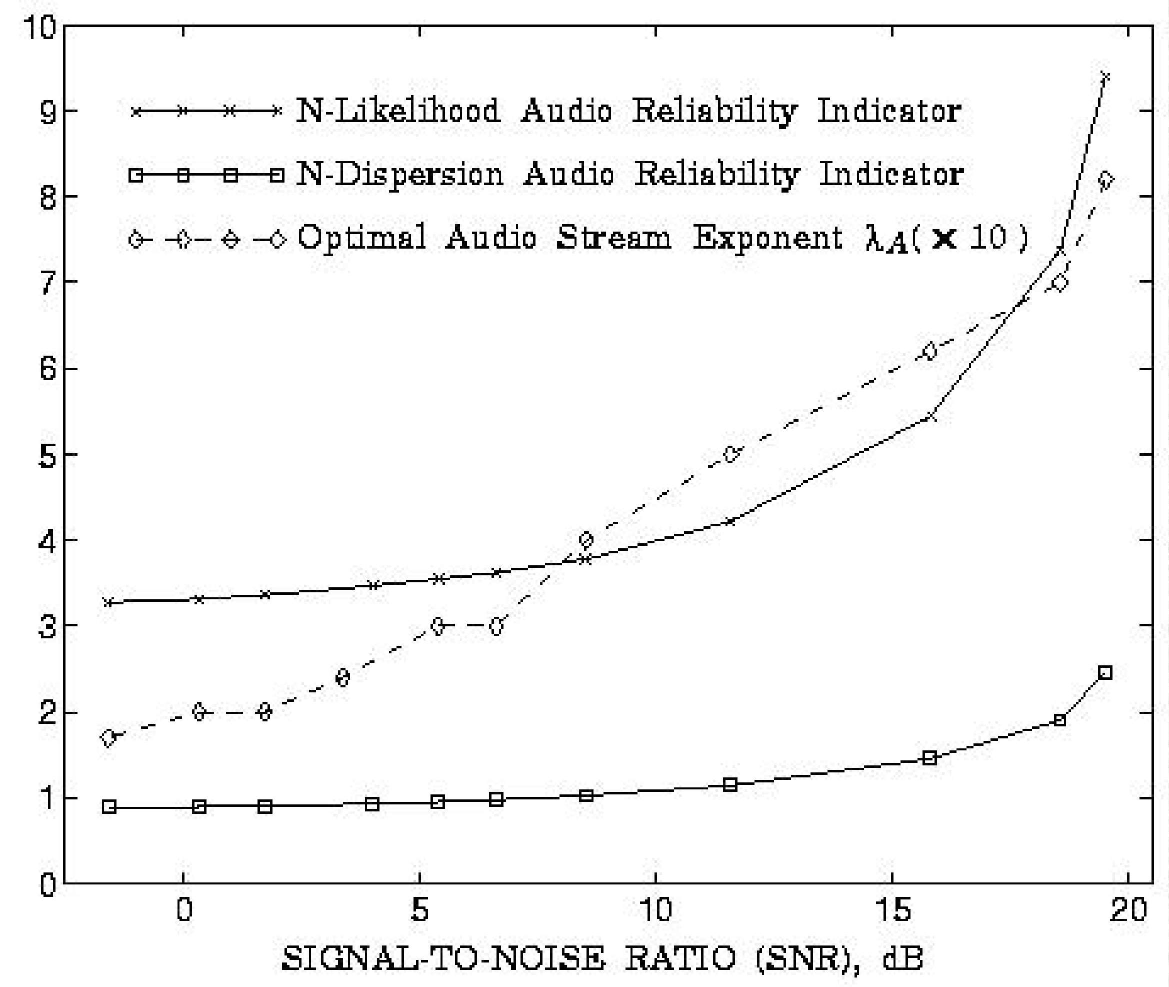
Adaptive stream weighting – Cont.

- Stream reliability indicators and exponents vs. SNR →

- Then estimate exponents as:

$$\lambda_{A,t} = [\,1 + \exp\,(-\textstyle\sum_{i=1}^{4} w_i \, d_i)\,]^{-1}$$

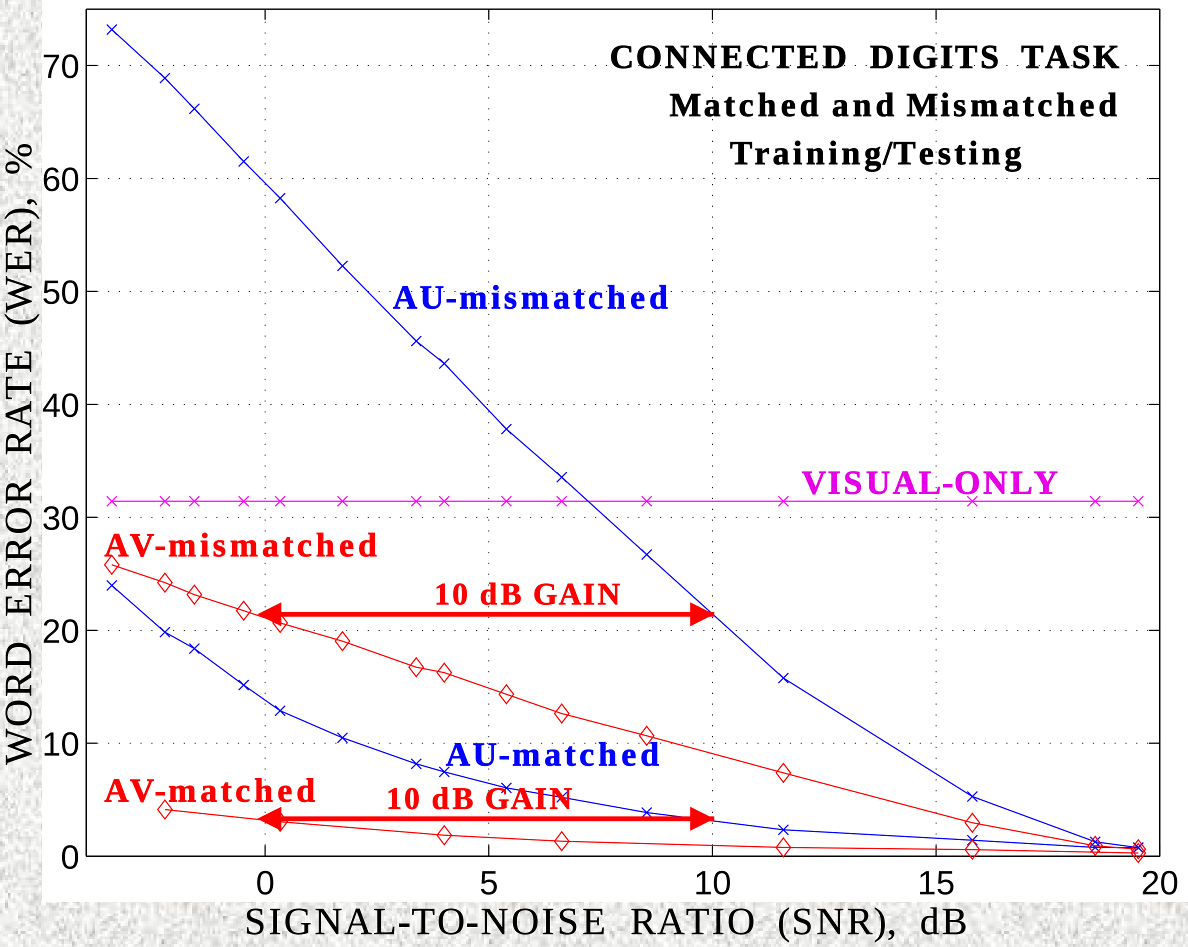
- Weights w_i are estimated using MCL or MCE on basis of frame error (Garg et al., 2003).



III.E. Summary of AV-ASR experiments.

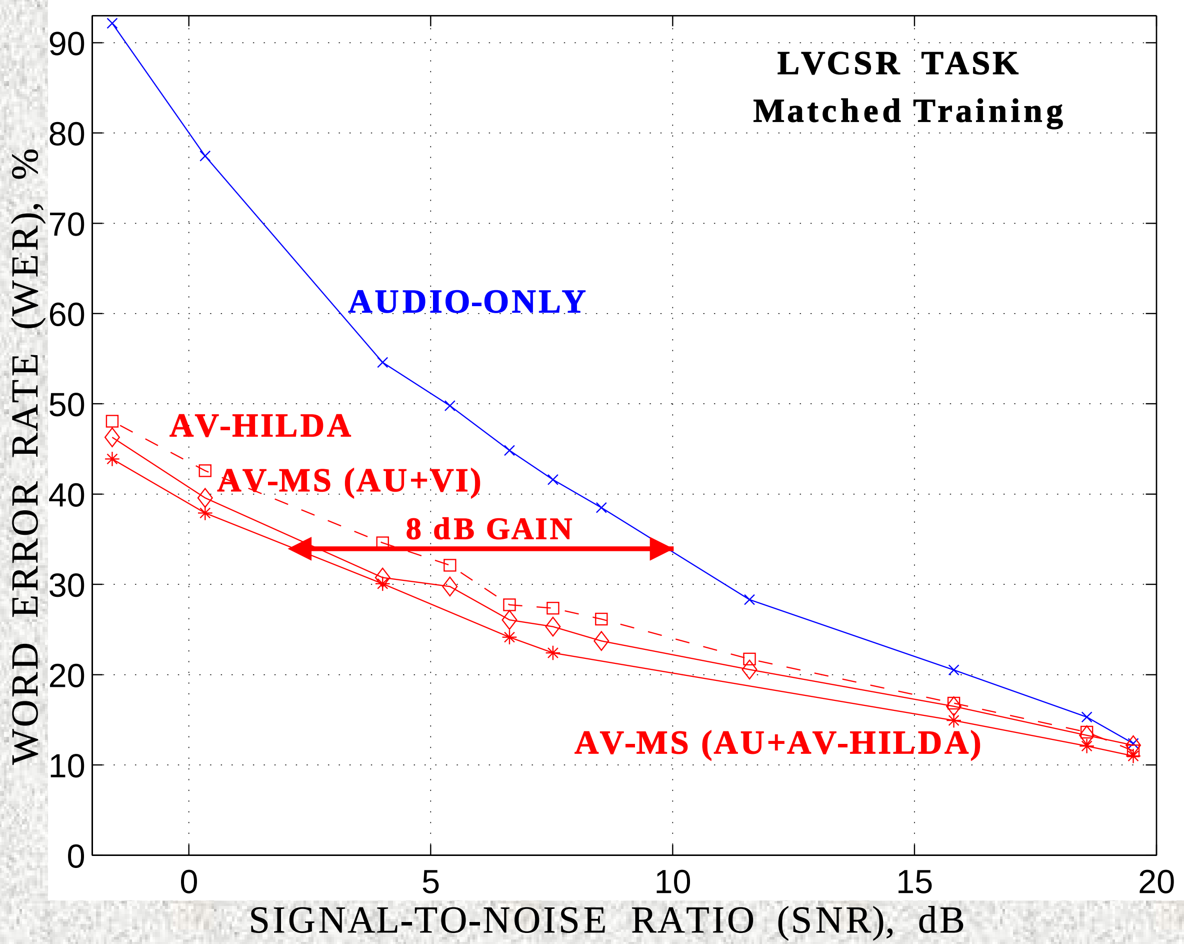
- **Summary** of AV-ASR results for **connected-digit** recog.

- 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**.



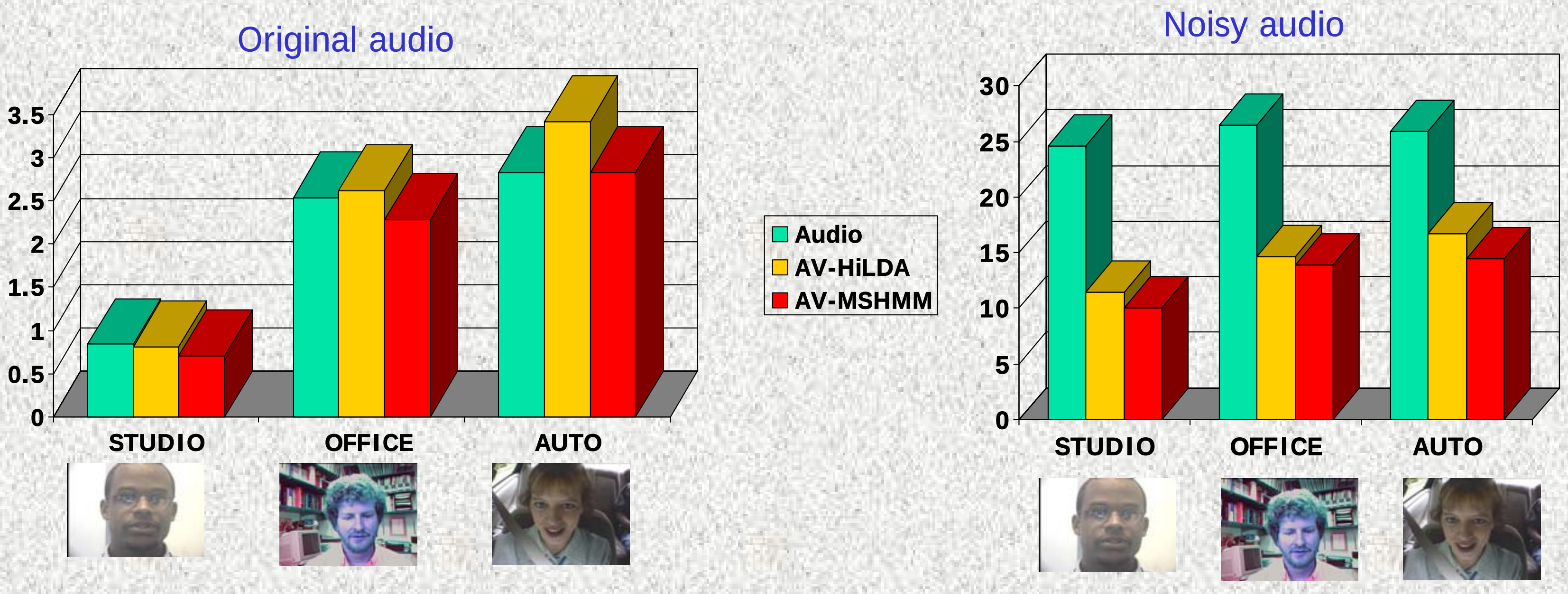
Summary of AV-ASR experiments - Cont.

- **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.



Summary of AV-ASR experiments - Cont.

- **<u>AV-ASR in challenging domains:</u>**
  - Office and automobile environments (challenging) vs. studio data (ideal).
  - Feature fusion hurts in challenging domains (clean audio).
  - Relative improvements due to visual information diminish in challenging domains.
  - Results reported in WER, %.



IV. Other audio-visual speech applications.

- Next generation of speech-based human-computer-interfaces require natural interaction & perceptual intelligence, i.e.:

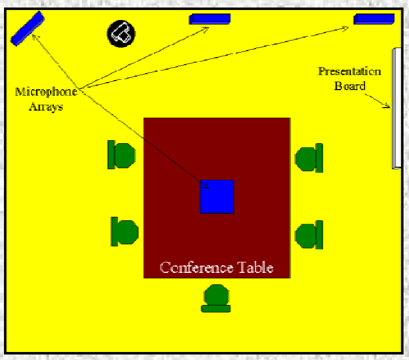
A. Speech **synthesis** (AV Text-To-Speech).

B. Detection of who is speaking (**speaker recognition**).

C. What is being spoken (ASR/**enhancement**).

D. Where is the **active speaker** (speech event detection).

E. How can the audio-visual interaction be segmented, labeled, and retrieved? (**mining**).

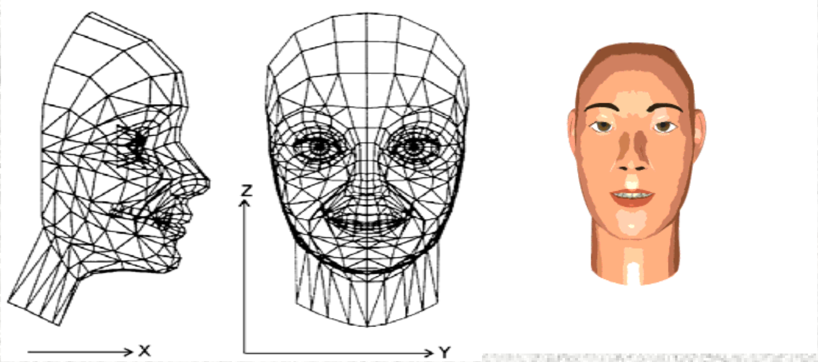


IV.A. Audio-Visual Speech Synthesis

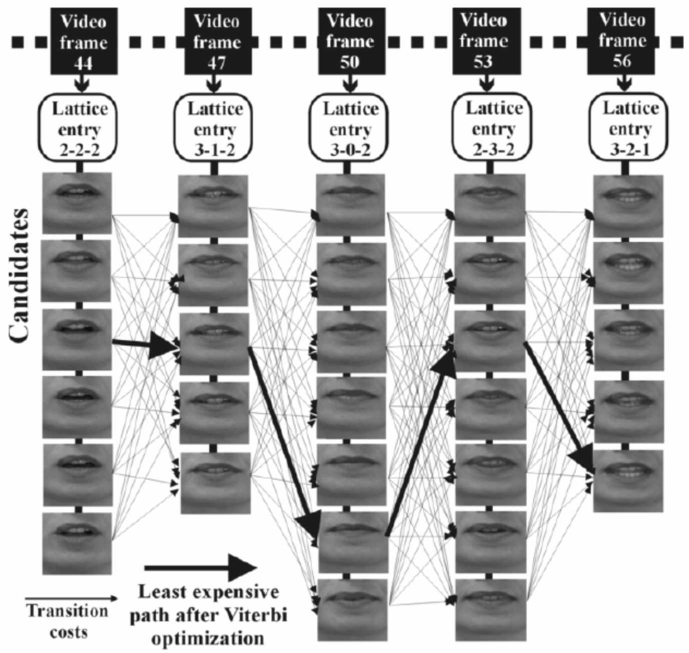
- **What is it:**
  - Automatic generation of voice and facial animation from arbitrary text (**AV-TTS**).
  - Automatic generation of facial animation from arbitrary speech.
- **Applications:**
  - Tools for the hearing impaired.
  - Spoken and multimodal agent-based user interfaces.
  - Educational aids.
  - Entertainment.
  - Video conferencing.
- **Benefits:**
  - Improved speech intelligibility.
  - Improved naturalness of HCI.
  - Less bandwidth.

AV-TTS – Two approaches.

- **Model-based**:
  - Face is modeled as a 3D object.
  - Control parameters deform it using
    - Geometric;
    - Articulatory;
    - Muscular models.
- **Sample based** (Photo-realistic).
  - Video segments of a speaker are:
    - Acquired
    - Processed
    - Concatenated



Viterbi search for best mouth sequence (Cosatto and Graf, 2000).



IV.B. Audio-visual speaker recognition

Two important problems are **speaker verification (authentication) and identification**

- **Speaker verification:**
  - Verify claimed identity based on audio-visual observations **O**
  - A **two-class** problem; True claimant vs. impostor (general population).
  - **Based on:**

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

- **Speaker identification:**
  - Obtain speaker identity $\hat{c}$ within a closed set of known subjects ***C*** based on observations **O** :

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

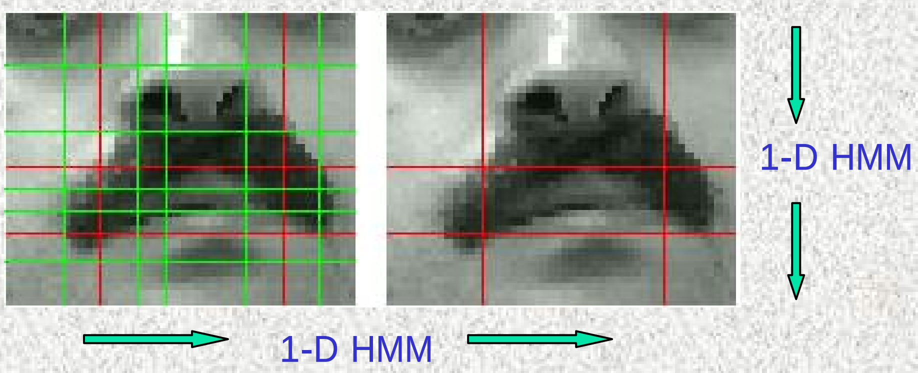
- Multi-modal systems better than single-modality!

IV.B.1. Single-modality speaker recognition

- **Audio-only:** Traditional acoustic features are used, such as LPC, MFCCs (Section II).
- **Visual-labial:** Mouth region visual features can be used, such as lip contour geometric and shape features, or appearance based features.
  - **visual-labial** features: shape (S), intensity (I), shape and intensity (SI) (Luettin et al., 1996):

    ID-error: TD: S: 27.1, I: 10.4, SI: 8.3 %

    TI: S: 16.7, I: 4.2, SI: 2.1 %
- **Visual-face (face recognition):** Features can be characterized as:
- **Shape vs. appearance based:**
  - **Shape** based: Active shape models, vector of facial feature geometry, profile histograms, dynamic link architecture, elastic graphs, Gabor filter jets.
  - **Appearance** based: LDA (“Fisher-faces”), PCA (“eigen-faces”), other image projections.
- **Global vs. local/hierarchical:**
  - **Global:** Single feature vector is classified (e.g., single PCA representation of entire face)
  - **Local/hierarchical:** Multiple feature vectors are classified (each representing local information, possibly organized in a hierarchy) and classification results are cumulated (e.g., **embedded HMMs**).



IV.B.2. Multi-modal speaker recognition

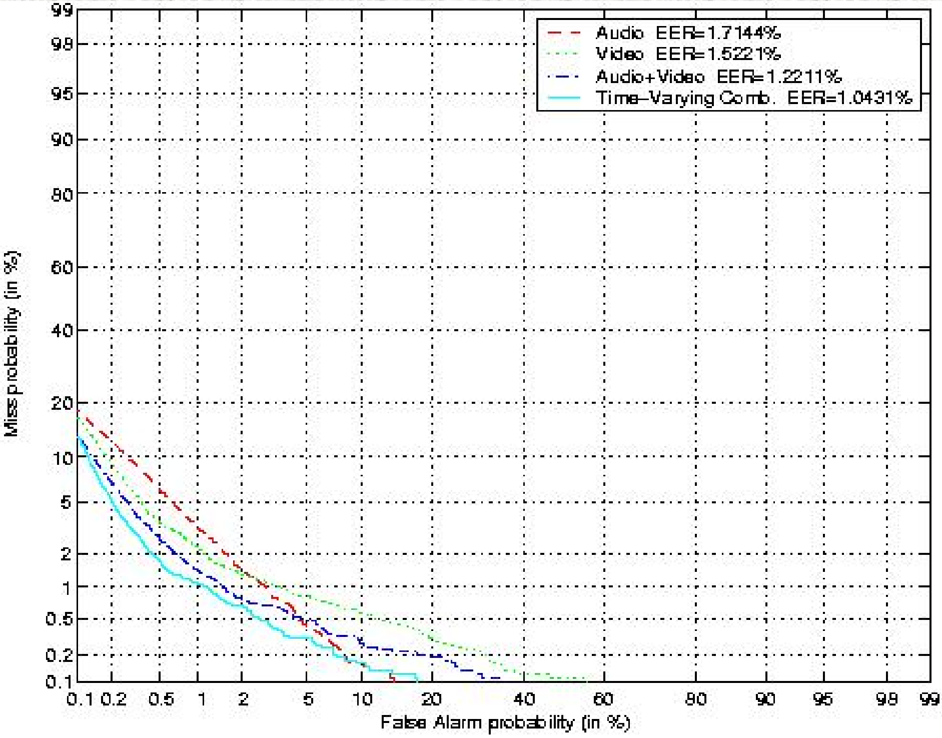
- Fusion of two or three single-modality speaker-recognition systems

- **Examples:**

- **Audio + visual-labial** (Chaudhari et al., 2003 ->)
  ID-error: A: 2.01, V: 10.95, AV: **0.40** %
  VER-EER: A:1.71, V: 1.52, AV: **1.04** %

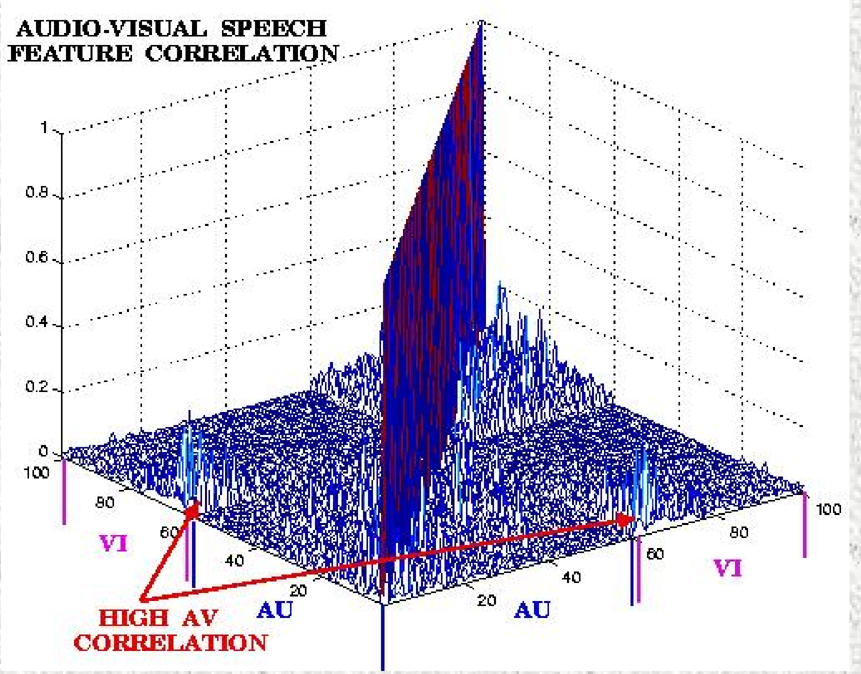
- **Audio + face** (Chu et al., 2003):
  ID-error: A: 28.4, F: 28.8, AF: **9.12** %

- **Audio + visual + face** (Dieckmann et al., 1997):
  ID-error: A: 10.4, V: 11.0, F: 18.7, AVF: **7.0** %

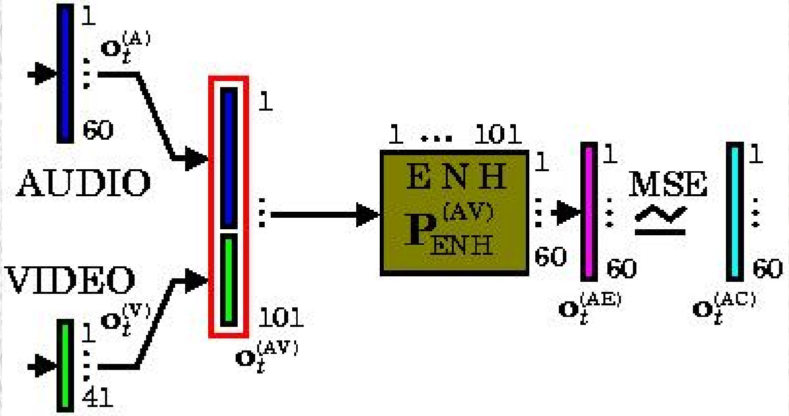


IV.C. Bimodal enhancement of audio

- **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.



IV.C.1. Linear bimodal audio enhancement.



- **Paradigm:**
  - Training on noisy AV features $\mathbf{o}_{AV,t} = [\mathbf{o}_{A,t}, \mathbf{o}_{V,t}]$, and clean AU $\mathbf{o}_{A,t}^{(C)}$, $t \in T$.
  - Seek linear transform $\mathbf{P}$, 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):

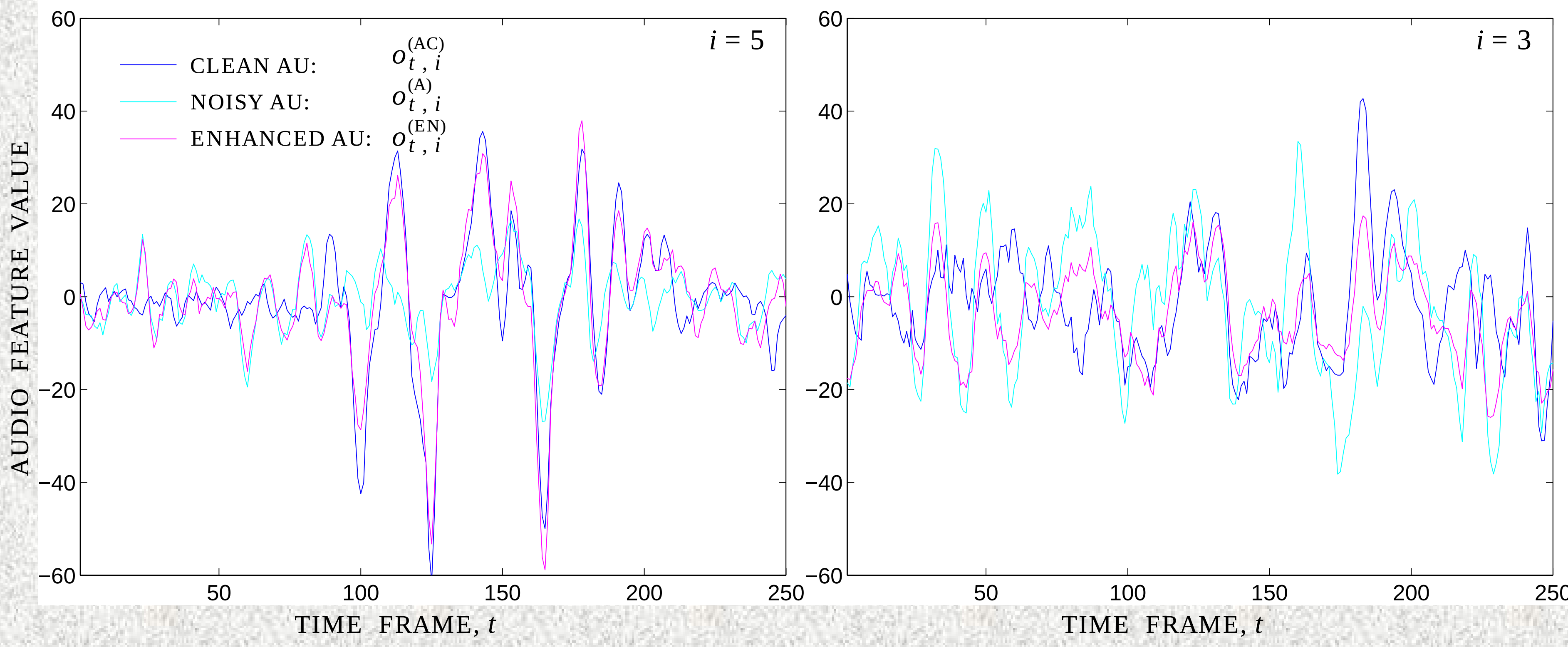
$$\mathbf{p}_i = \arg\max_{\mathbf{p}} \sum_{t \in T} [\, o_{A,t,i}^{(C)} - < \mathbf{p}, \mathbf{o}_{AV,t} >]^2, \quad i = 1,\ldots, d_A$$

  - Solved by d_A systems of Yule-Walker equatiions:

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

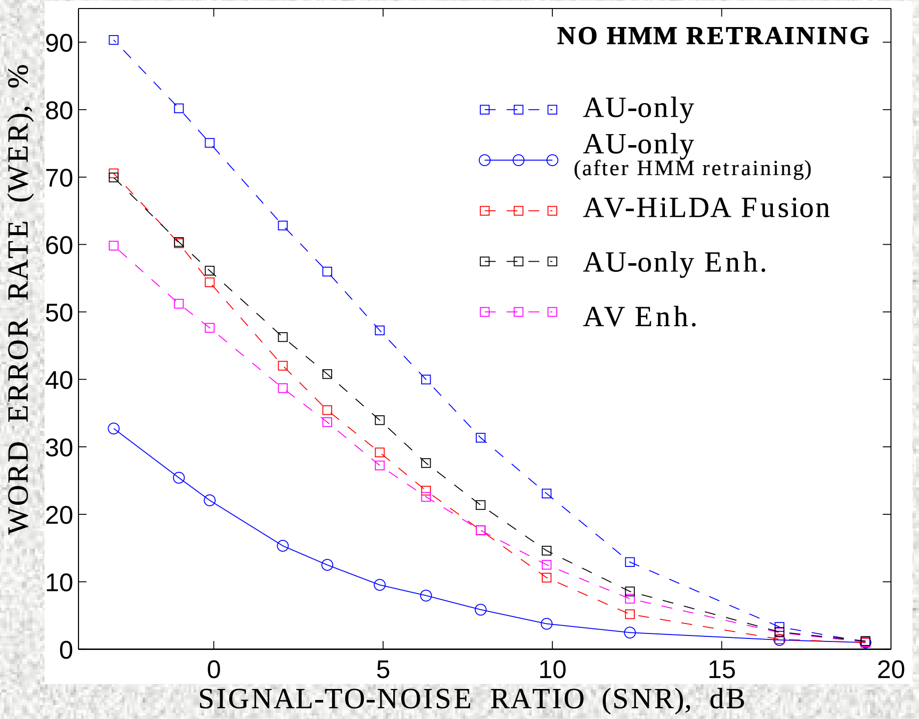
Linear bimodal audio enhancement – Cont.

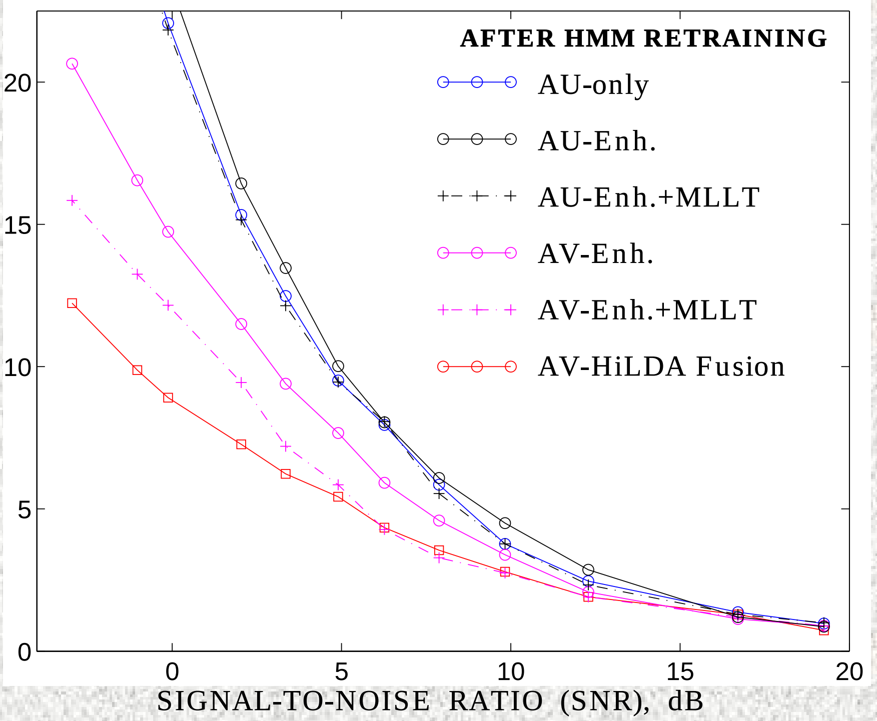
- Examples of **audio feature estimation** using bimodal enhancement (additive speech babble noise at **4 dB SNR**): Not perfect, but better than noisy features, and helps ASR!



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.





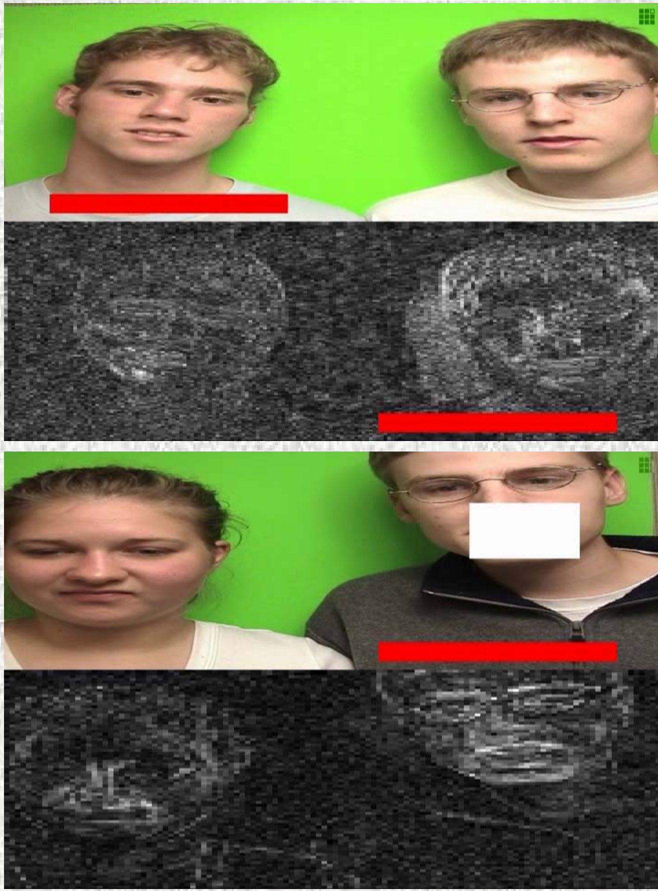
IV.D. Audio-visual speaker detection

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).

V. Summary / Discussion

- **We discussed and presented:**
  - The **need** to augment the acoustic speech with the visual modality in HCI.
  - How to **extract** and represent visual speech information.
  - How to **combine** the two modalities within the HMM based statistical **ASR** framework.
  - Additional examples of **how to utilize** the visual modality in HCI; for example, speech synthesis, speaker authentication, identification, and localization, speech enhancement.
  - **Experimental results** demonstrating its significant benefit to many of these areas.

Summary / Discussion – Cont.

- Much progress has been accomplished in including visual speech in HCI. Still however, visual speech is not in wide-spread use in main-stream HCI, due to:
  - Visual signal processing lack-of-robustness to typical, challenging HCI environments.
  - Cost for high-quality video capture, storage, and processing.

- However, with the explosion of camera miniaturization and hardware speed, as well as the associated drastic cost reduction, we believe that audio-visual speech is becoming ready for targeted applications !

- The field is clearly multi-disciplinary, presenting many research and development opportunities and challenges.

- ***THANK YOU FOR YOUR ATTENTION !***

References

- M.E. Hennecke, D.G. Stork, and K.V. Prasad, “Visionary speech: Looking ahead to practical speechreading systems,” in *Speechreading by Humans and Machines*, D.G. Stork and M.E. Hennecke eds., Springer, Berlin, pp. 331-349, 1996.
- T. Chen, “Audiovisual speech processing. Lip reading and lip synchronization,” *IEEE Signal Process. Mag.*, 18(1): 9-21, 2001.
- G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, “Recent advances in the automatic recognition of audio-visual speech,” *Proc. IEEE*, 91(9): 1306-1326, 2003.
- S. Dupont and J. Luettin, “Audio-visual speech modeling for continuous speech recognition,” *IEEE Trans. Multimedia,* 2(3): 141-151, 2000.