Audio-Visual Speech Processing: Progress & Challenges

Gerasimos Potamianos

Nov. 3, 2006

www.research.ibm.com/AVSTG/makis.html

Outline

Overview / Introduction:

- Why audio-visual speech in human-computer interaction.
- Audio-visual speech technologies.
- Potential applications.

Audio-visual speech components with emphasis on ASR:

- Visual feature representation for speech applications.
- Audio-visual combination (fusion).

Other audio-visual speech technologies:

- Speech enhancement.
- Speaker recognition.
- Speech detection.
- Speech synthesis.

Summary & Conclusions.

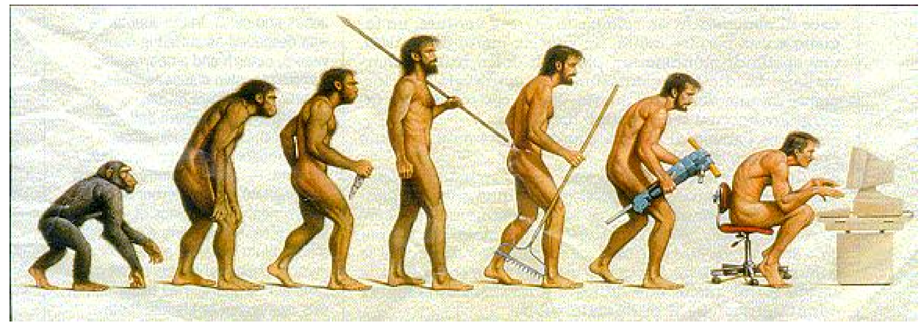
Audio-visual HCI – Motivation

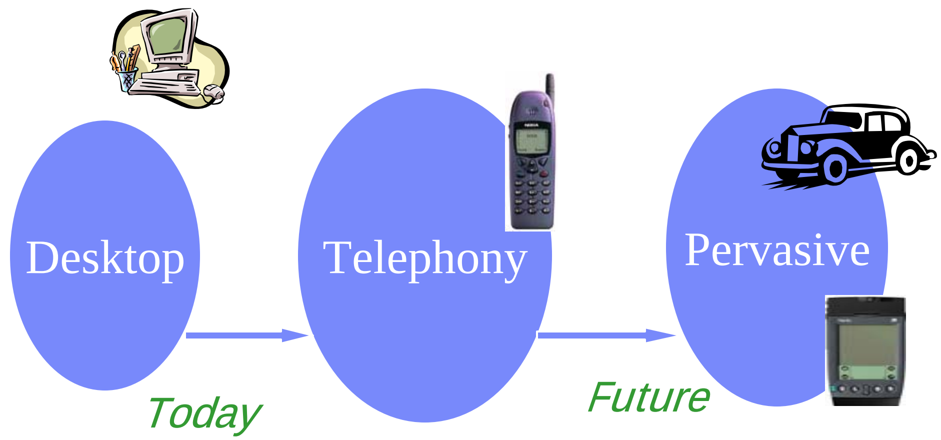
- **Human-computer interaction (HCI):**
  - **Today:** Part of everyday life, but far from natural!
  - **Future:** Pervasive and ubiquitous computing.

- Next generation HCI → **<u>perceptual intelligence</u>**:
  - **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.

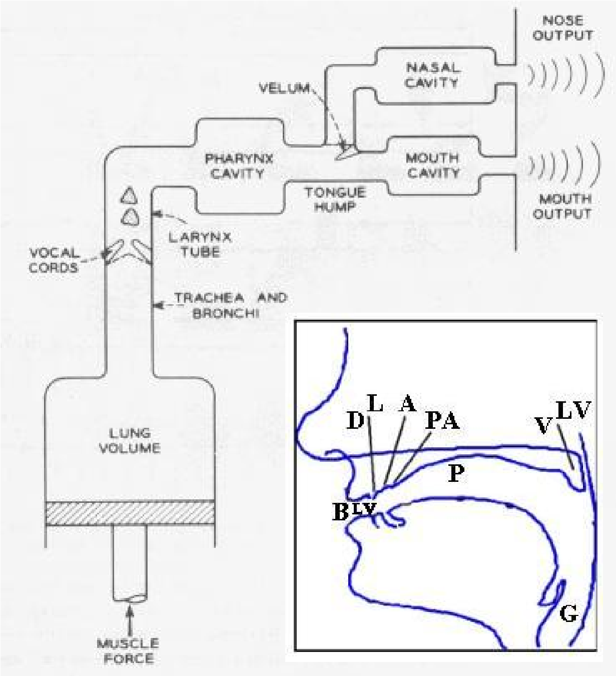
- **Joint audio + visual processing can help bridge the usability gap**!





Audio-Visual Speech – Motivation

- **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/.

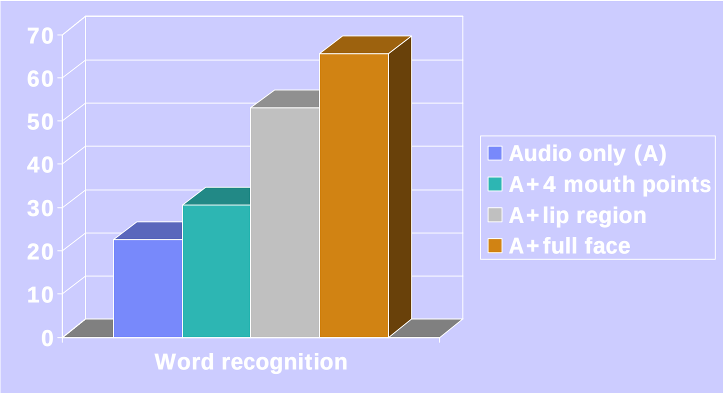


Schematic representation of speech production (J.L. Flanagan, *Speech Analysis, Synthesis, and Perception,* 2nd ed., Springer-Verlag, New York, 1972.)

- **Human speech perception** is bimodal:
  - We **lip-read** in noisy environments to improve intelligibility.

    E.g., human speech perception experiment by Summerfield (1979): Noisy 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/
  - **Hearing impaired** people lip-read.

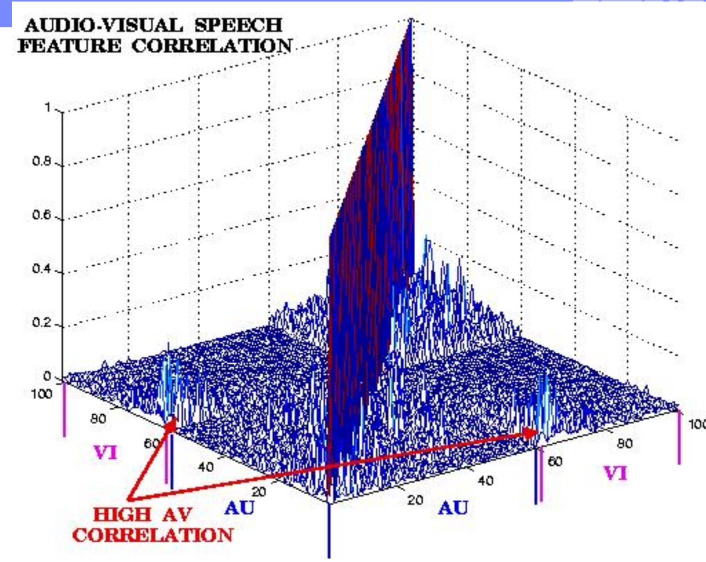


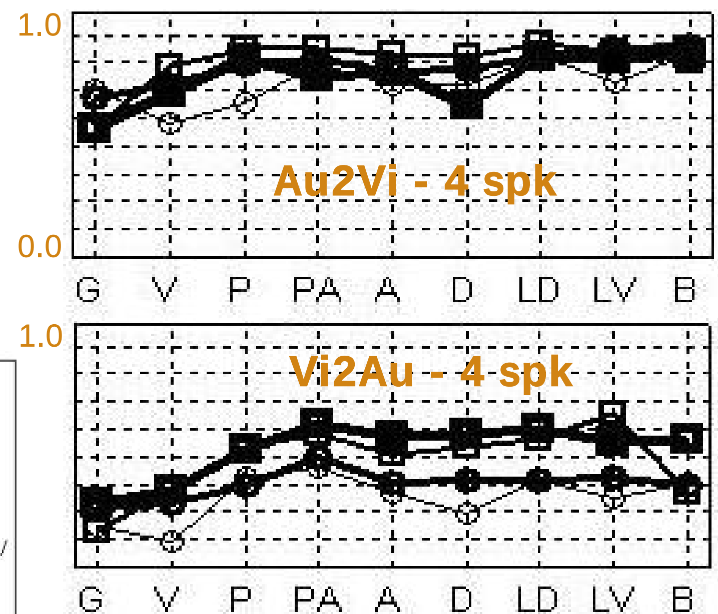
Audio-Visual Speech Motivation – Cont.

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

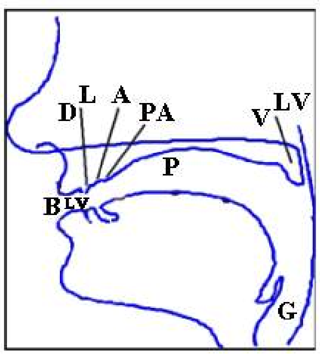
- Although the visual speech information is less than audio …
  - **Phonemes:** Distinct speech units that convey linguistic information; about **47** in English.
  - **Visemes:** Visually distinguishable classes of phonemes: **6-20**.

- … the **<u>visual channel provides important complementary information to audio:</u>**
  - 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).





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



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/

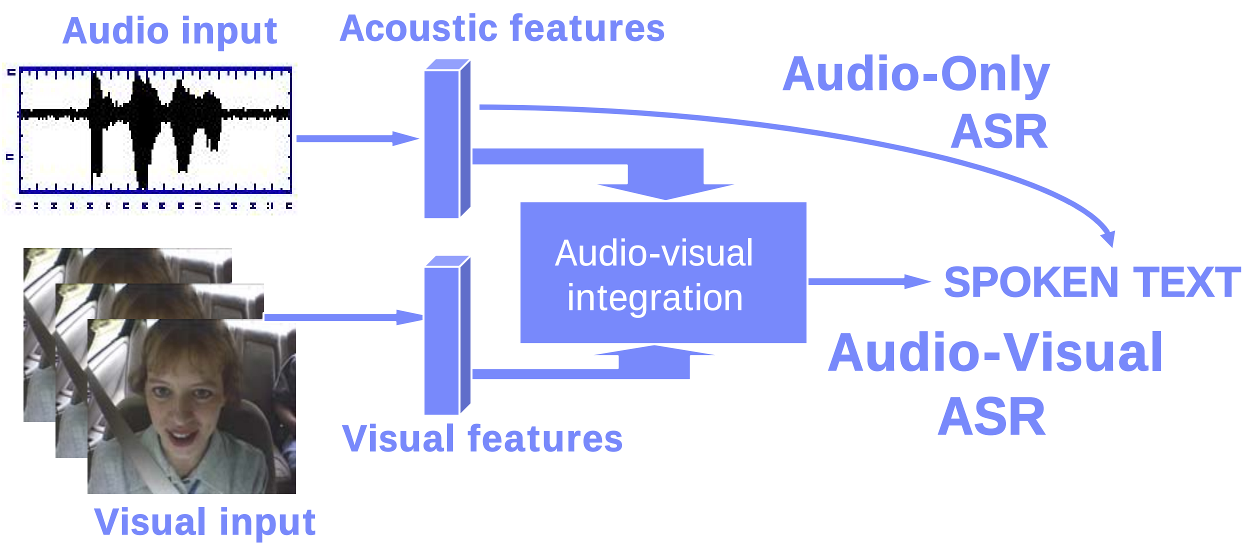
Manner of articulation

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/

Audio-Visual Speech Technologies

All major speech technologies can benefit from the visual modality:

- **Automatic speech recognition (ASR).**



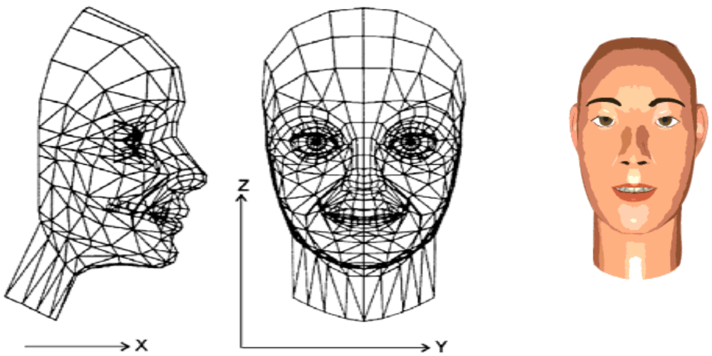
- **Automatic speaker identification / verification.**

Audio-Visual Speech Technologies – Cont.

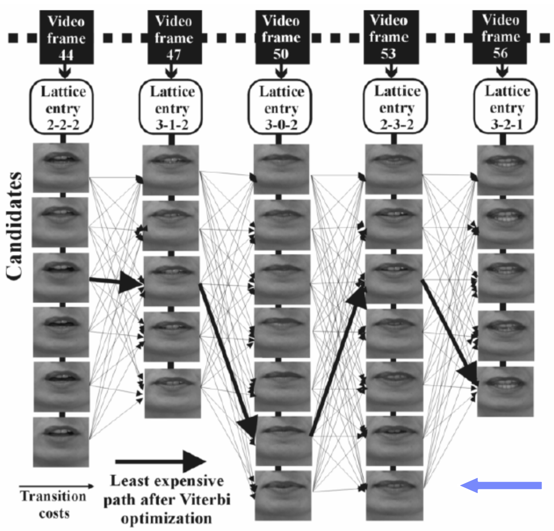
- **Speaker localization / speech activity detection / speech separation.**

- **Speech synthesis:**

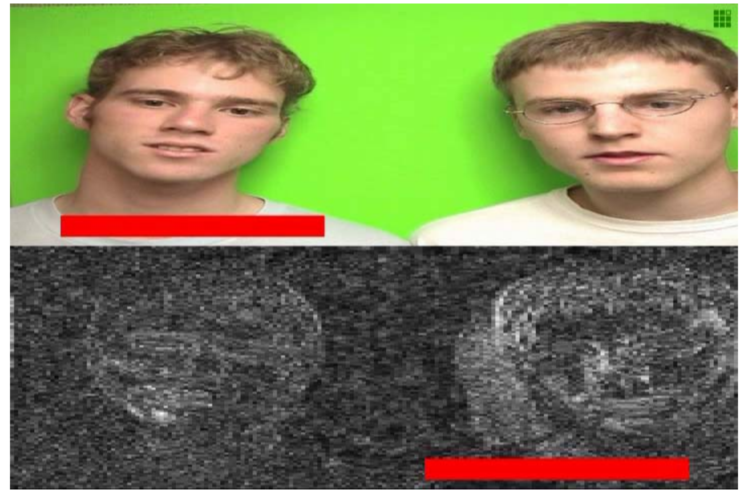
Model based:

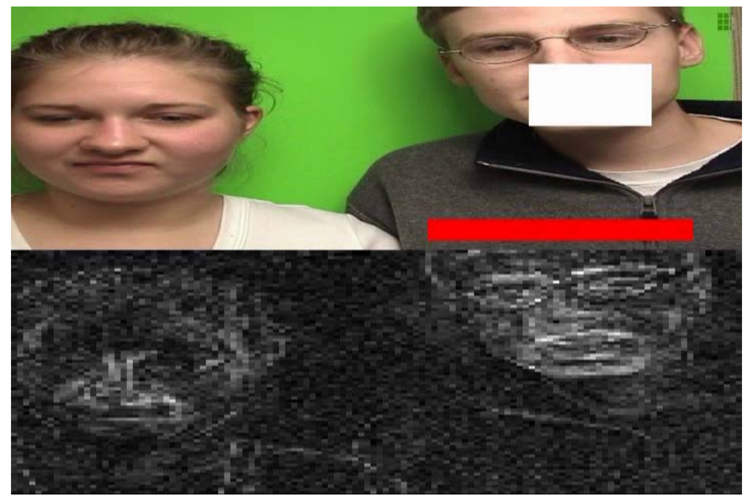


Sample based:



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





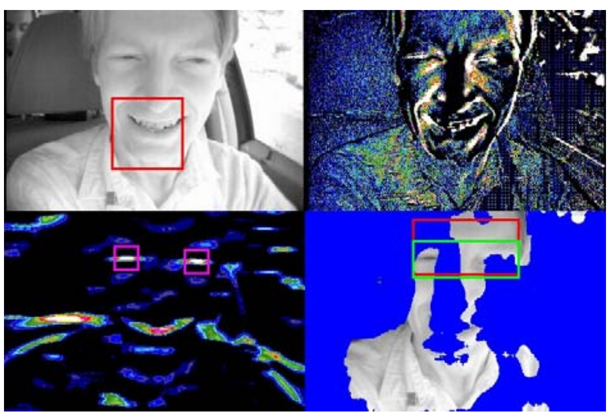
Audio-visual synchrony and tracking (Nock, Iyengar, and Neti, 2000).

Examples of potential application areas of AV work at IBM

- Specially designed audio-visual ***ASR headset***:
  - ***Call centers, trading floors***, etc.

- Audio-visual ***helmet*** for ***ASR***.
  - ***Motorcycles.***
- Visual speech activity detection.
  - ***Automobiles***

    [when is the driver addressing the navigation system?]



Techniques for Audio-Visual Speech Processing

- All above technologies share two main components:
  - Visual processing / representation.
  - Audio-visual fusion.

- In the following, we discuss these two components as relevant to audio-visual ASR (AVASR). In particular, we concentrate on:
  - **Visual processing:**
    - ✓ Face / facial feature detection.
    - ✓ Feature extraction.
    - ✓ Lip-reading results.
  - **Audio-visual integration:**
    - ✓ Feature fusion.
    - ✓ Decision fusion.
    - ✓ Results.

Face Detection – Algorithms

We follow a statistical, appearance based face detection approach:

- **<u>2-class</u>** classification (into faces / non-faces).
- "**<u>Face template</u>**" (e.g., 11x11 pixel rectangle) ordered into vectors **x**.
- A **<u>trainable</u>** scheme "scores"/**<u>classifies</u>** **x** into the 2 classes.
- **<u>Pyramidal search</u>** (over locations, scales, orientations) provides face **<u>candidates x</u>**.
- Can **<u>speed-up</u>** search by using **<u>skin-tone</u>** (based on color information on the *R,G,B* or transformed space), or location/scale (in the case of a video sequence).

Training / scoring (for face "vector" $\mathbf{x}$):

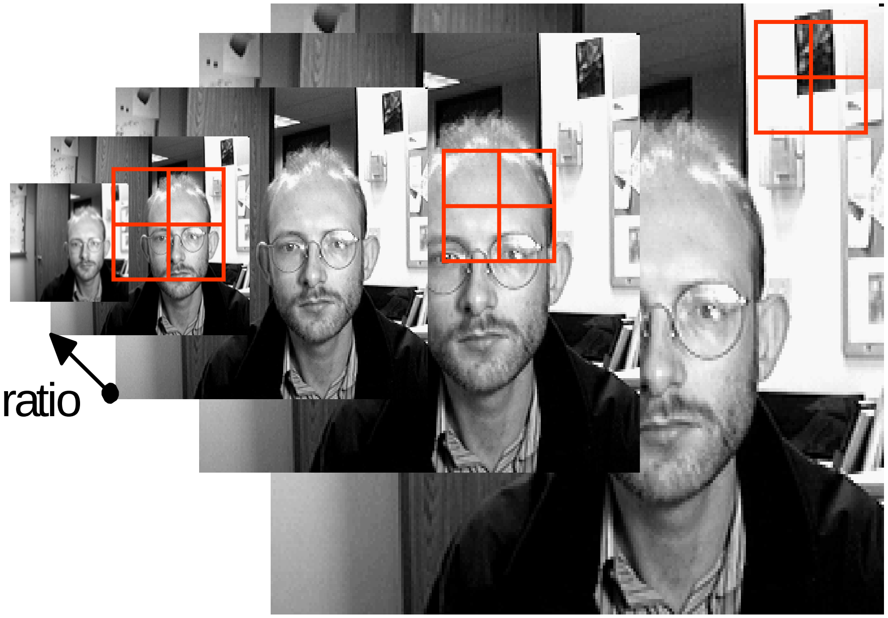
<u>Fisher discriminant</u> detector (LDA):

- One-dimensional projection of 121-dimensional vector $\mathbf{x}$: $y_F = \mathbf{P}_{1 \times 121}\, \mathbf{x}$

<u>Distance from face space</u> (DFFS).

- Obtain **PCA** of training set.
- Projected vectors $\mathbf{y} = \mathbf{P}_{d\mathrm{x}121}\, \mathbf{x}$ have DFFS:

$$\mathrm{DFFS} = \left\| \mathbf{x} - \mathbf{y}\, \mathbf{P}^{\mathrm{T}} \right\|$$



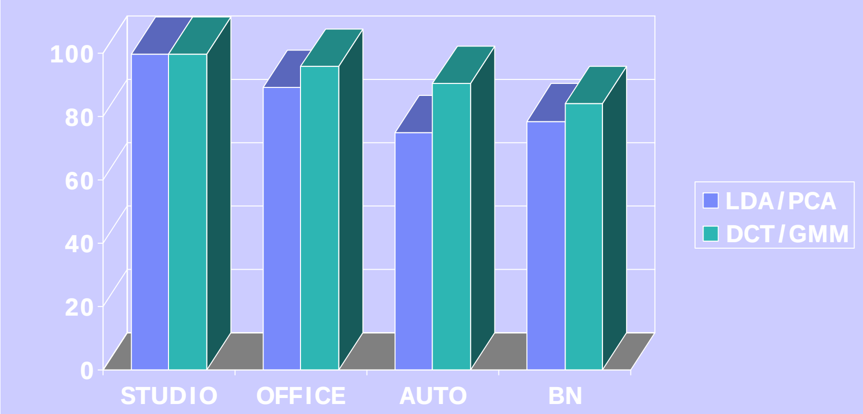
<u>Gaussian mixture</u> classifier (**GMM**):

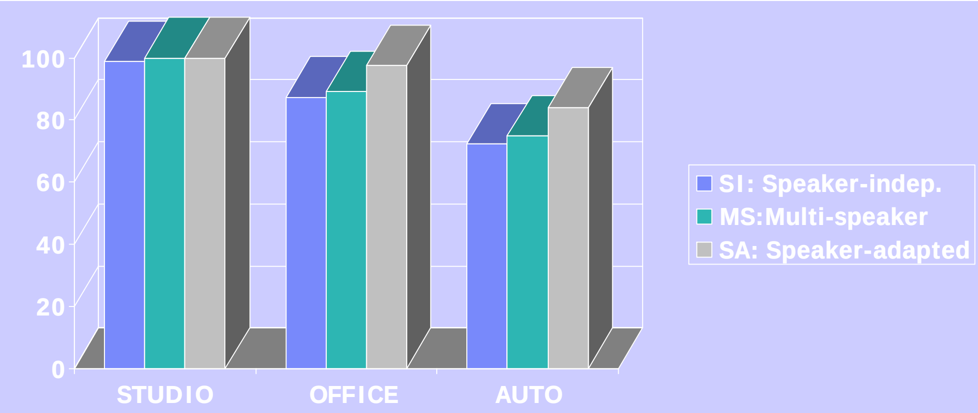
- Vector $\mathbf{y}$ is a PCA or DCT $\mathbf{y} = \mathbf{P}\,\mathbf{x}$.
- Two GMMs model face/non-face: $\Pr(\mathbf{y} \mid c) = \sum_{k=1}^{K_c} w_{k,c} N(\mathbf{y}, \mathbf{m}_{k,c}, \mathbf{s}_{k,c}),\ \ c \in \{f, \overline{f}\}$

Face Detection – Results

- 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:**

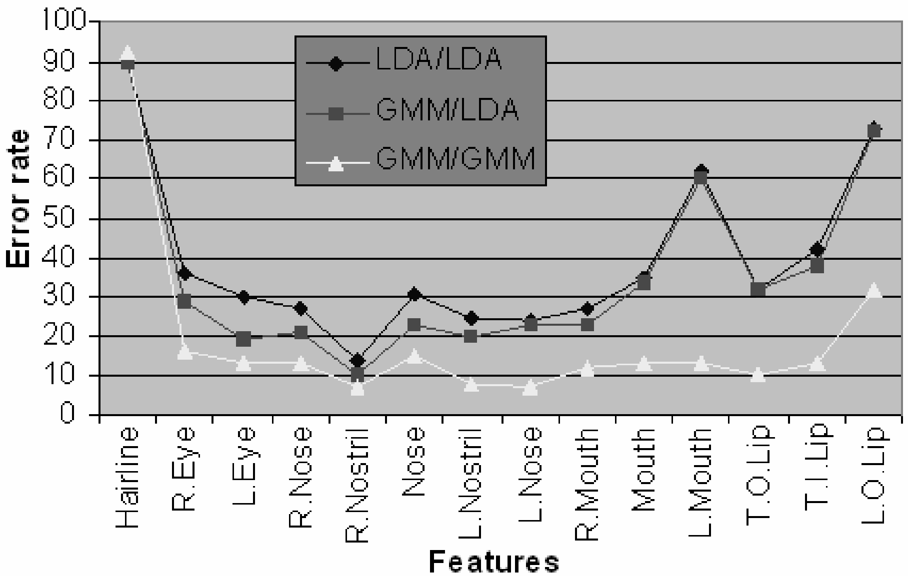




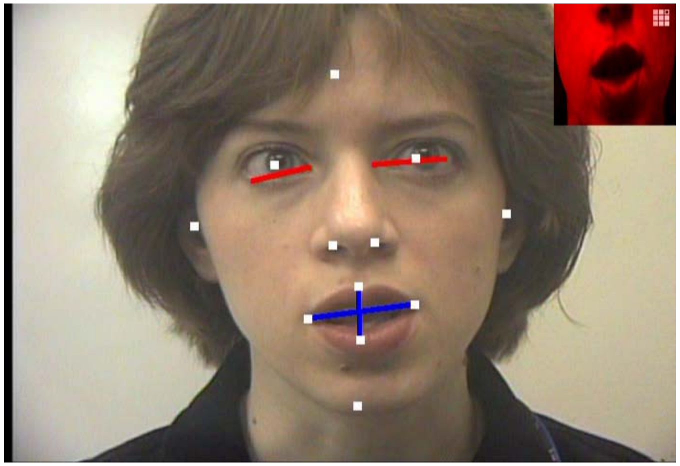
Faces → Facial Features → Region of Interest

From faces to facial features (eyes, mouth, etc):

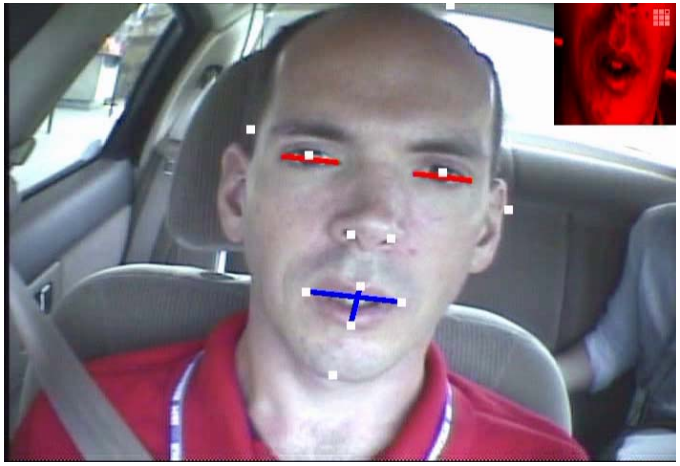
- Similar to face detection. Score *individual* facial feature *templates* by LDA, DFFS, GMMs, etc.



Facial-feature extraction performance



STUDIO



AUTOMOBILE

Region-of-interest (ROI):

- Assumed to contain "all" visual speech information.
- Typically, a rectangle containing mouth + lower face.
- Appropriately normalized.

Region-of-Interest → Visual Features

Three types of / approaches to feature extraction:

Video pixel (appearance) based features:

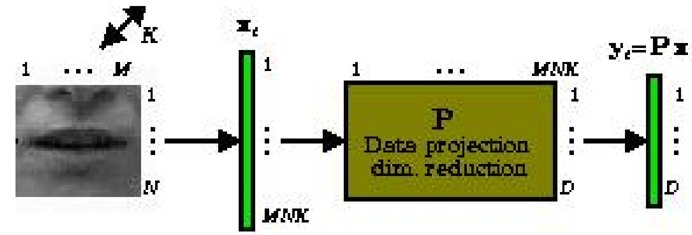
- Lip contours *do not* capture oral cavity information!
- Use compressed representation of mouth ROI instead.
- E.g.: DCT, PCA, DWT, whole ROI.

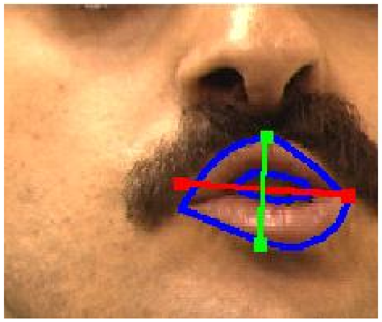
Lip- and face-contour (shape) based:

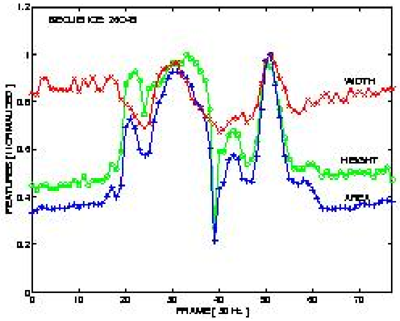
- Height, width, area of mouth.
- Moments, Fourier descriptors.
- Mouth template parameters.

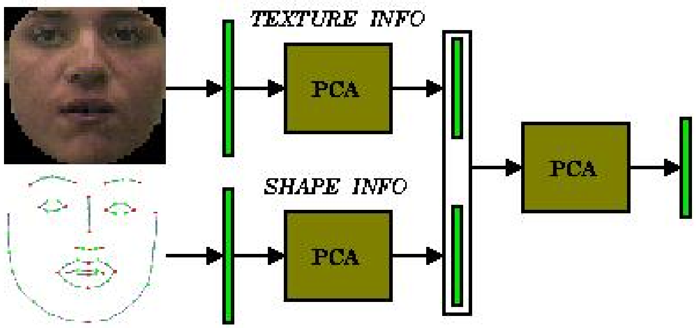
Joint shape and appearance features:

- Active appearance models.







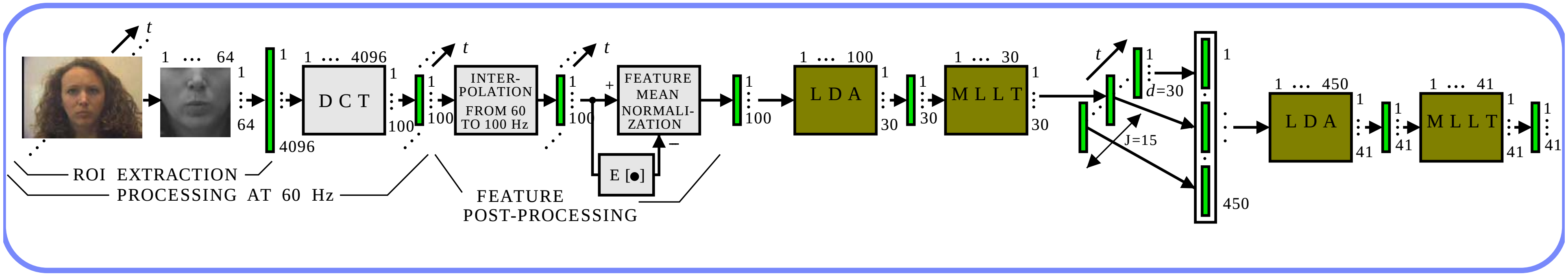


Visual Features – The IBM System

Appearance based approach.

- **Static features:** **100**-dimensional compressed representation of 64 x 64 monochrome ROI using **DCT** *[we'll revisit how to select such features later].*
- **Post-processing:** Intra-frame + inter frame **LDA/MLLT** for better within and across frame discrimination and statistical modeling; **FMN** and **up-sampling.**
- **Final features:** **41**-dimensional at 100 Hz.

Visual front end processing – system diagram:

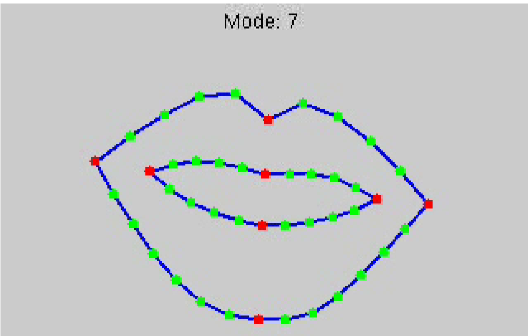


Visual Features – Shape Based Approach

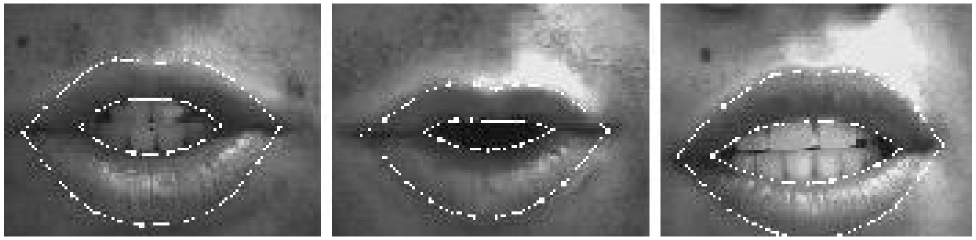
Shape based features represent speech information using lip contour information.

Require “expensive” lip-tracking algorithms, applied within the ROI, using:

- **Snakes** (Kass et al., 1988):
  Elastic curve defined by ***control points***.
- **Deformable templates** (Yuille et al., 1989):
  Geometric model. Typically two or more ***parabolas*** are used.
- **Active shape models** (Cootes, Taylor, Cooper, Graham, 1995):
  A ***PCA*** model of lip contour point coordinates is obtained.

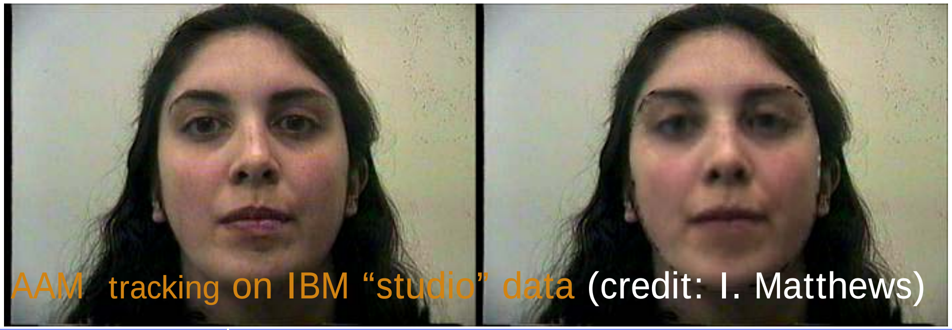


ASM modes trained on IBM data

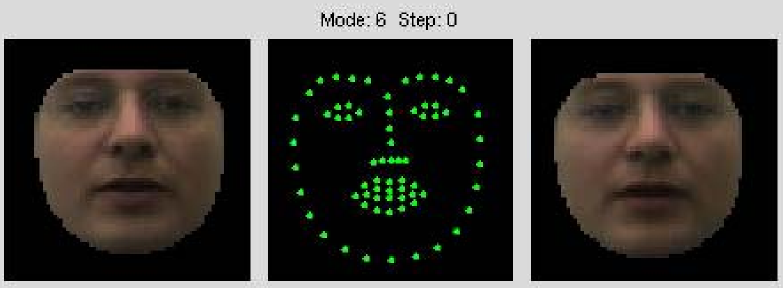


ASM based tracking

- **Active appearance models** (**AAM**s- Cootes et al.,’00):
  In addition to shape, it also builds ***face texture*** PCA.



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



AAM modes trained on IBM data

Visual-Only ASR Results – Feature Comparisons

- Comparisons are based on ***single-subject***, ***connected-digit*** ASR experiments.

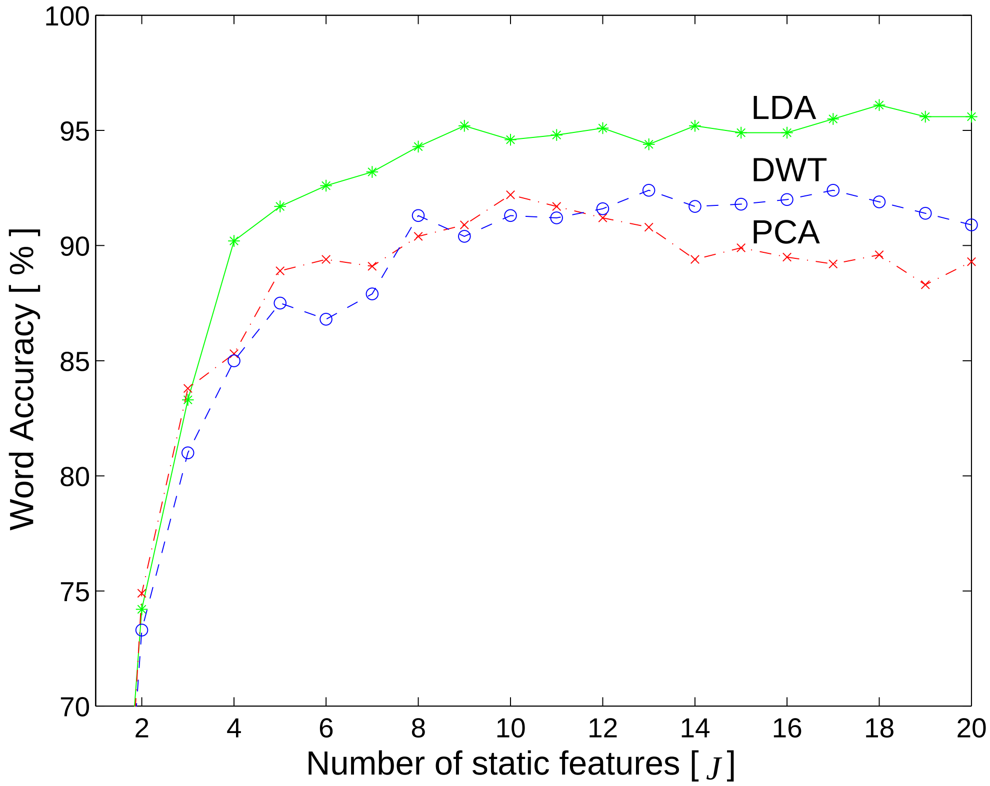
- Appearance- are better than shape-based features:

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**

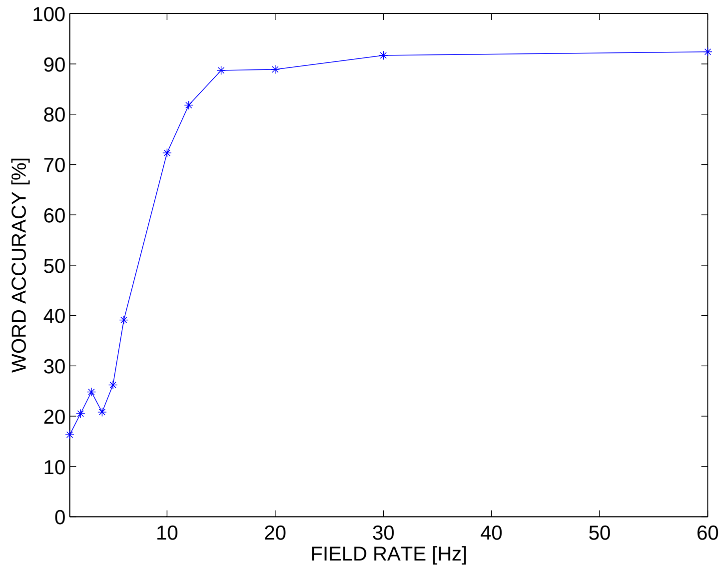
Feature type	%, Word accuracy
Lip-contour based	**83.9**
Appearance (LDA)	**97.0**

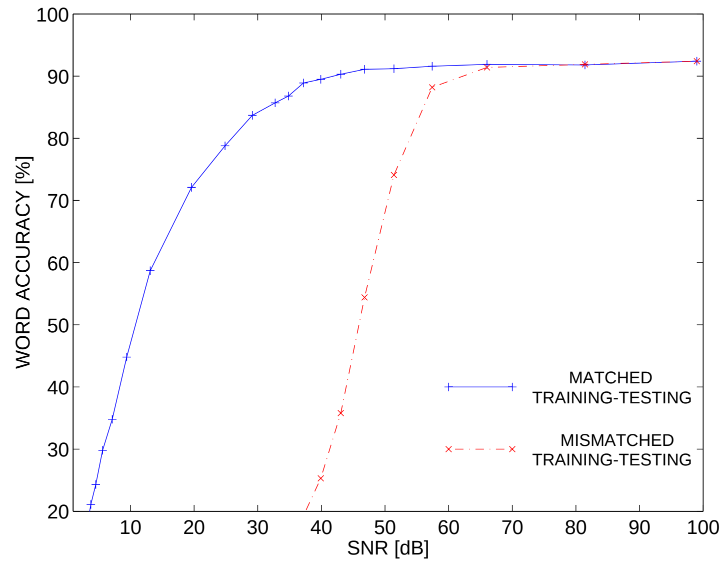
- Comparisons of various appearance-based features:

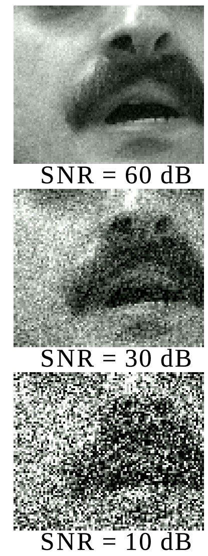


Visual-Only ASR Results – Video Degradation Effects

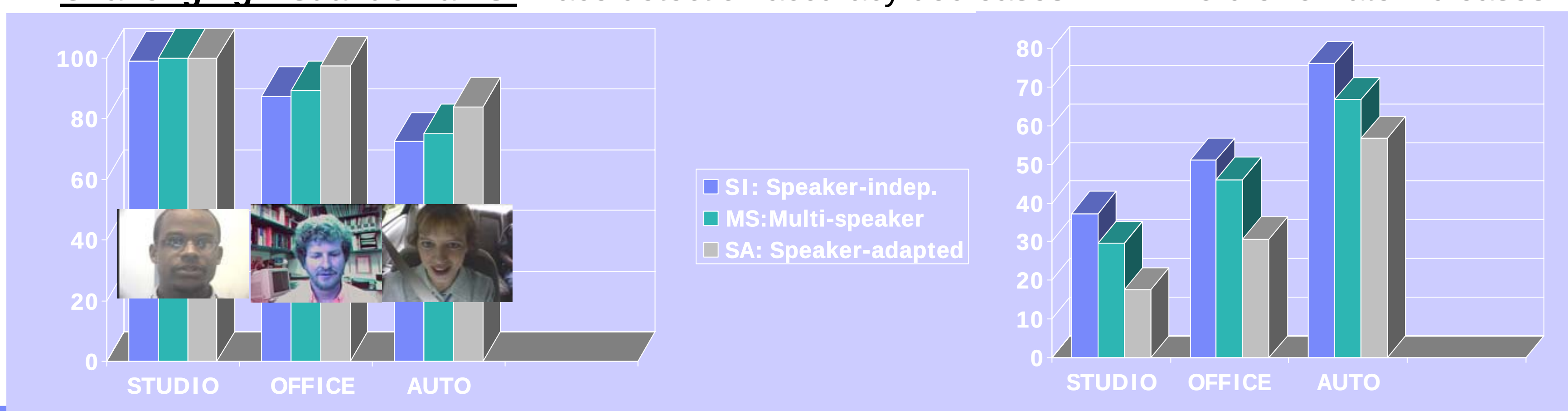
- **<u>Frame rate decimation:</u>** Limit of acceptable video rate for automatic speechreading is 15 Hz.
- **<u>Video noise:</u>** Robustness to noise only in a matched training/testing scenario.







- **<u>Challenging visual domains:</u>** Face detection accuracy decreases → Word error rate increases.



Visual DCT feature selection schemes

- ***Recall:*** Appearance visual features, based on extracted ***64 x 64*** DCT coeffs.

- ***Issue:*** How to ***select the appropriate number*** of visual features?

- ***Appoaches***:
  - ***Energy based***: Select high energy coefficients (baseline approach).
  - ***LDA*** → high input dimensionality, stability problems.
  - ***Variance*** → somewhat worse performance than energy based schemes.
  - ***Mutual information (MI)*** → promising scheme, but computational problems.

- ***MI approach:***
  - Select DCT features x that ***maximize MI wrt speech classes c***.

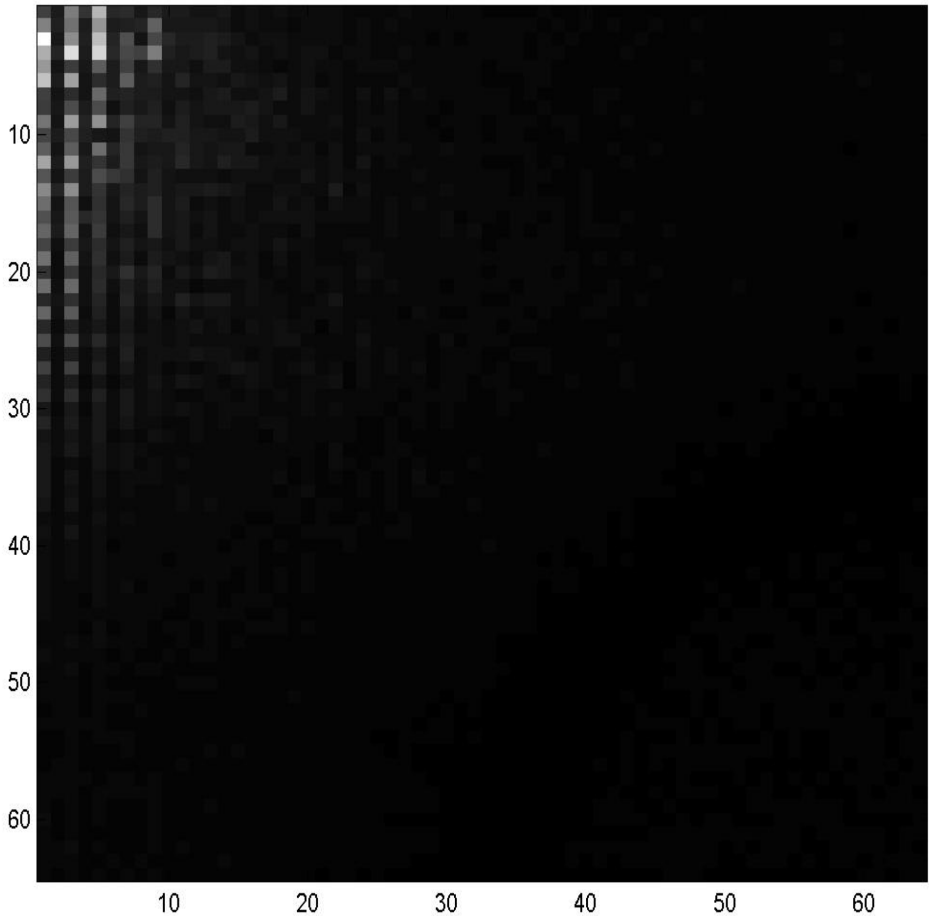
$$I(C;X) = H(C) - H(C|X) = H(X) - H(X|C)$$

$$I(X;C) = -\int_{x\in R} p(x)\log(p(x))dx + \sum_{c=1}^{C} p(c)\int_{x\in R} p(x\mid c)\log(p(x\mid c))dx$$

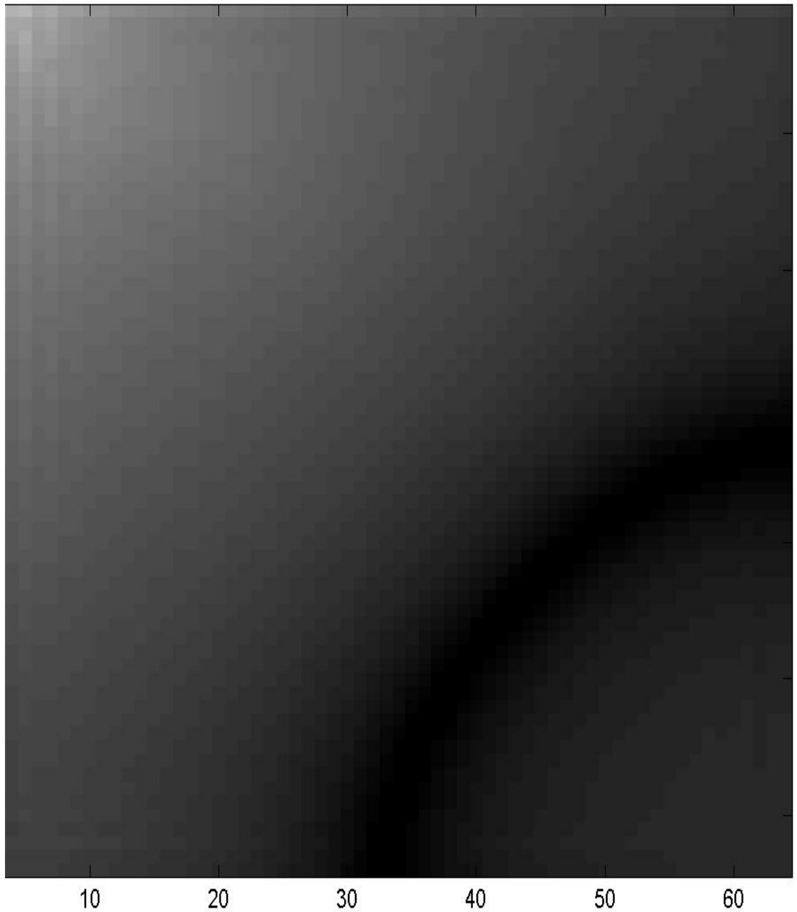
DCT feat. selection – Cont.: Mutual Information vs. Energy (I)

- MI / energy values of 4096 DCT coefficients over training data:

MI



Energy



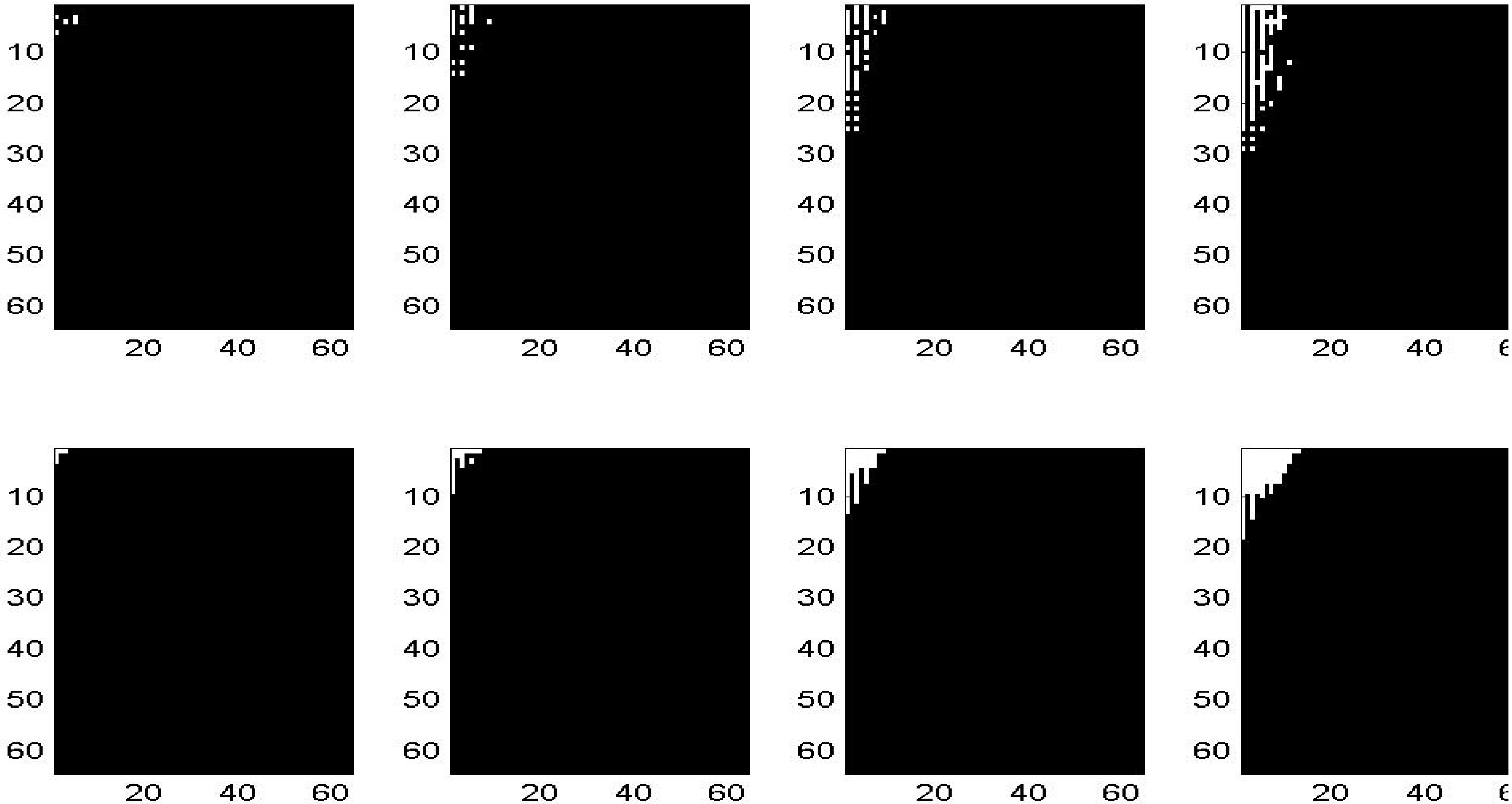
- Clearly, DCT coeffs. at odd columns have very low MI, but still high energy

DCT feat. selection – Cont.: Mutual Information vs. Energy (II)

- Typical MI vs Energy feature selection “masks”:

MI

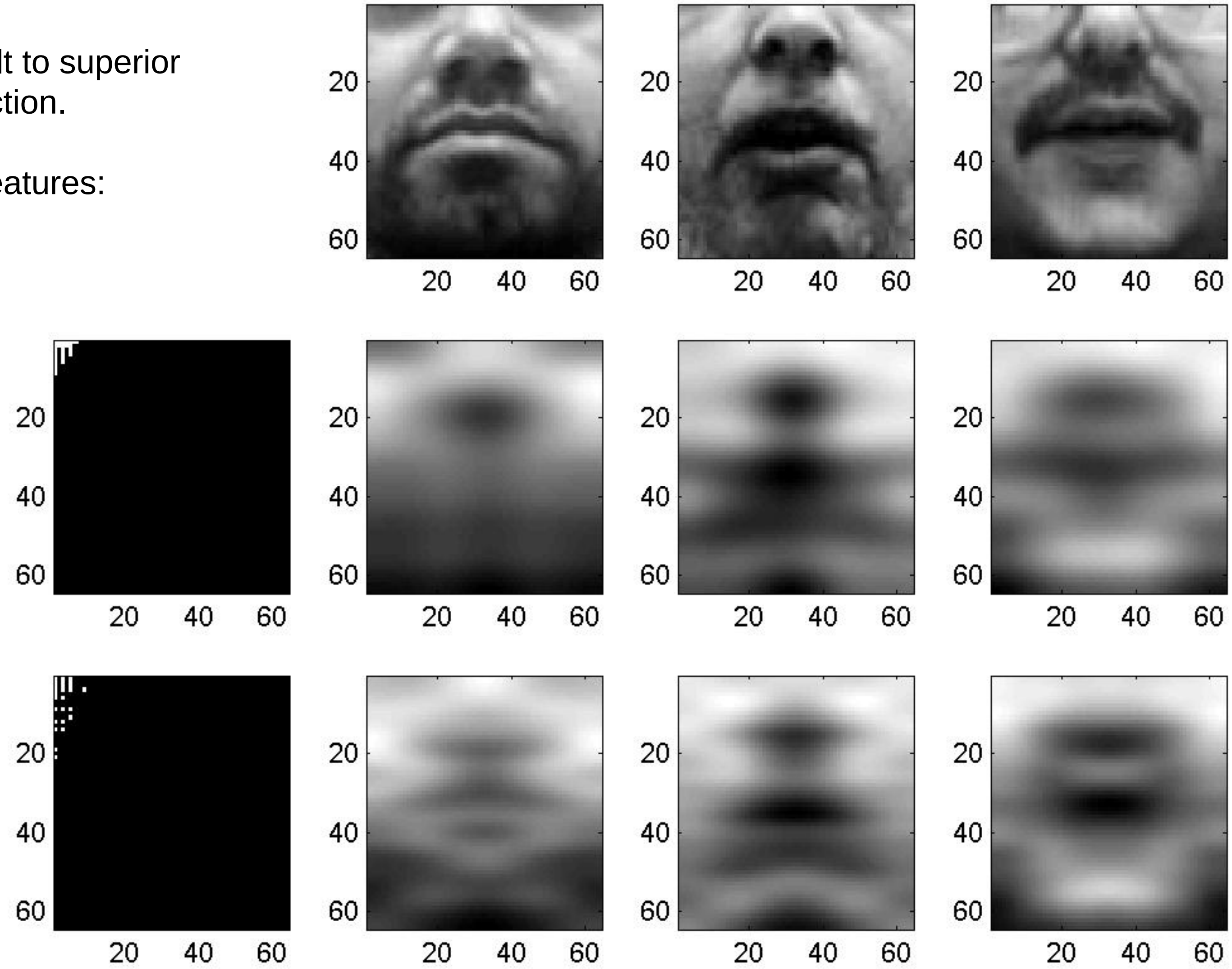
Energy



Larger number of selected features

DCT feat. selection – Cont.: Mutual Information vs. Energy (III)

- MI masks result to superior ROI reconstruction.
- Based on 25 features:

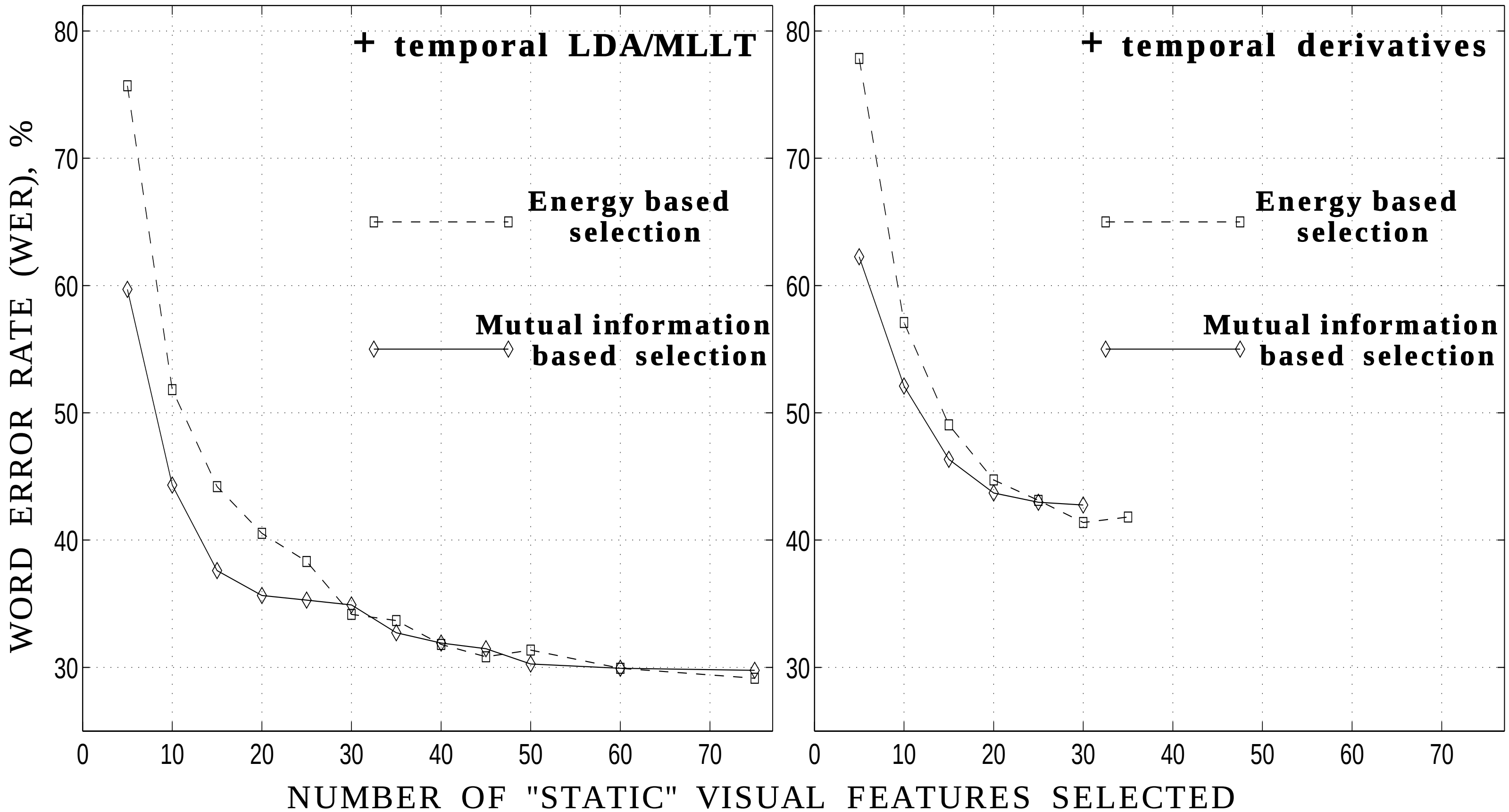


Energy

MI

DCT feat. selection – Cont.: Mutual Information vs. Energy (IV)

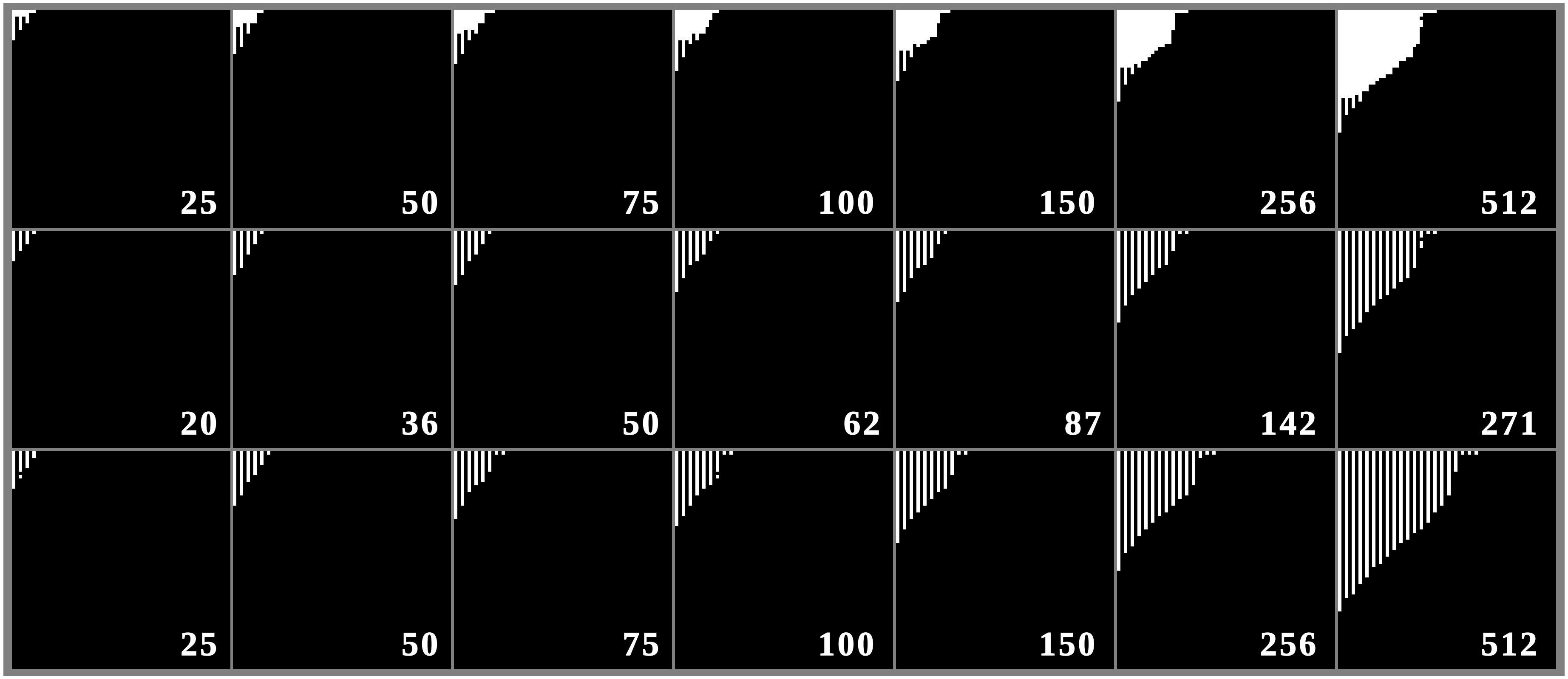
- Unfortunately, MI approach produces better visual features only for low dimensions (< 30).
- Reason: Features are selected independently – due to computational issues.
- Visual only ASR results for connected digits task (studio data):



DCT feat. selection – Cont.: Symmetric Energy Templates (I)

Alternate scheme: Consider even-column DCT coeffs (this is where MI is the highest)!

(a) Baseline energy templates - *Both even + odd DCT components used.*

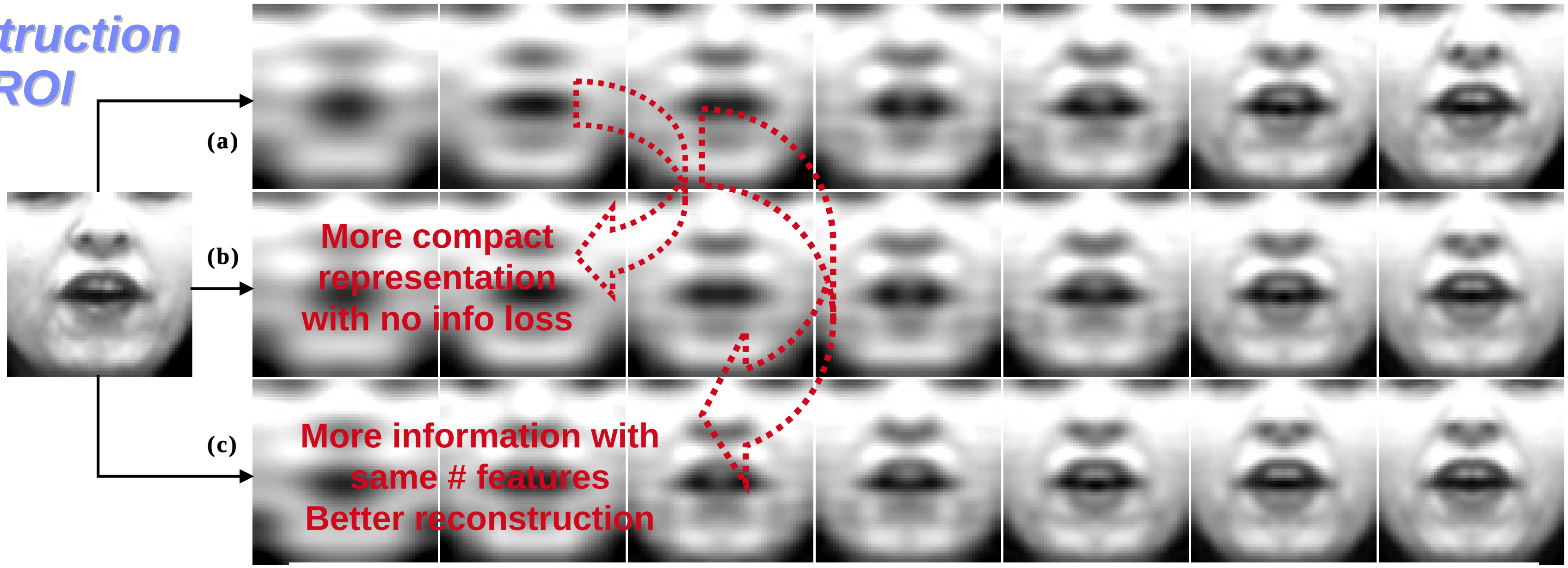


Proposed energy templates

(b) ***Subset*** of (a), with odd DCT components removed → more compact / no loss (?)

(c) ***Same number of elements*** as in (a) → more information (?)

Reconstruction of ROI



More features used

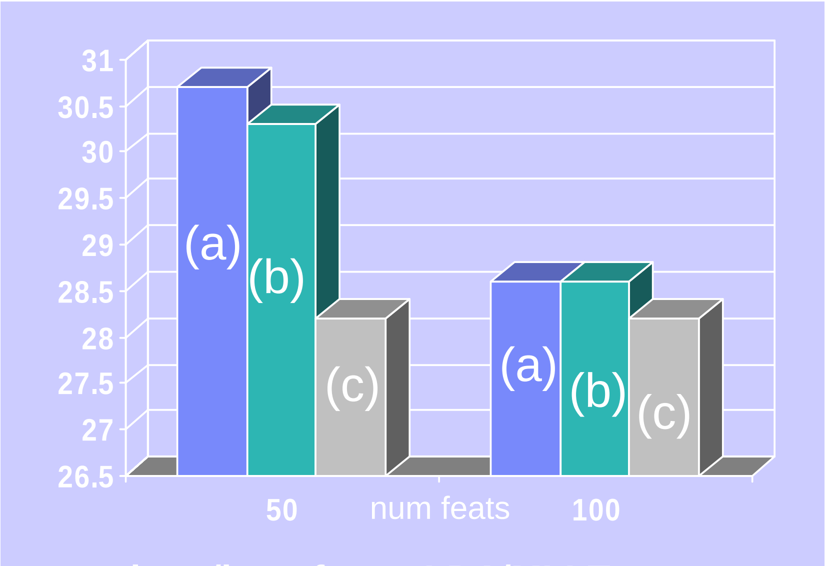
More compact representation with no info loss

+ *DE-ROTATION*

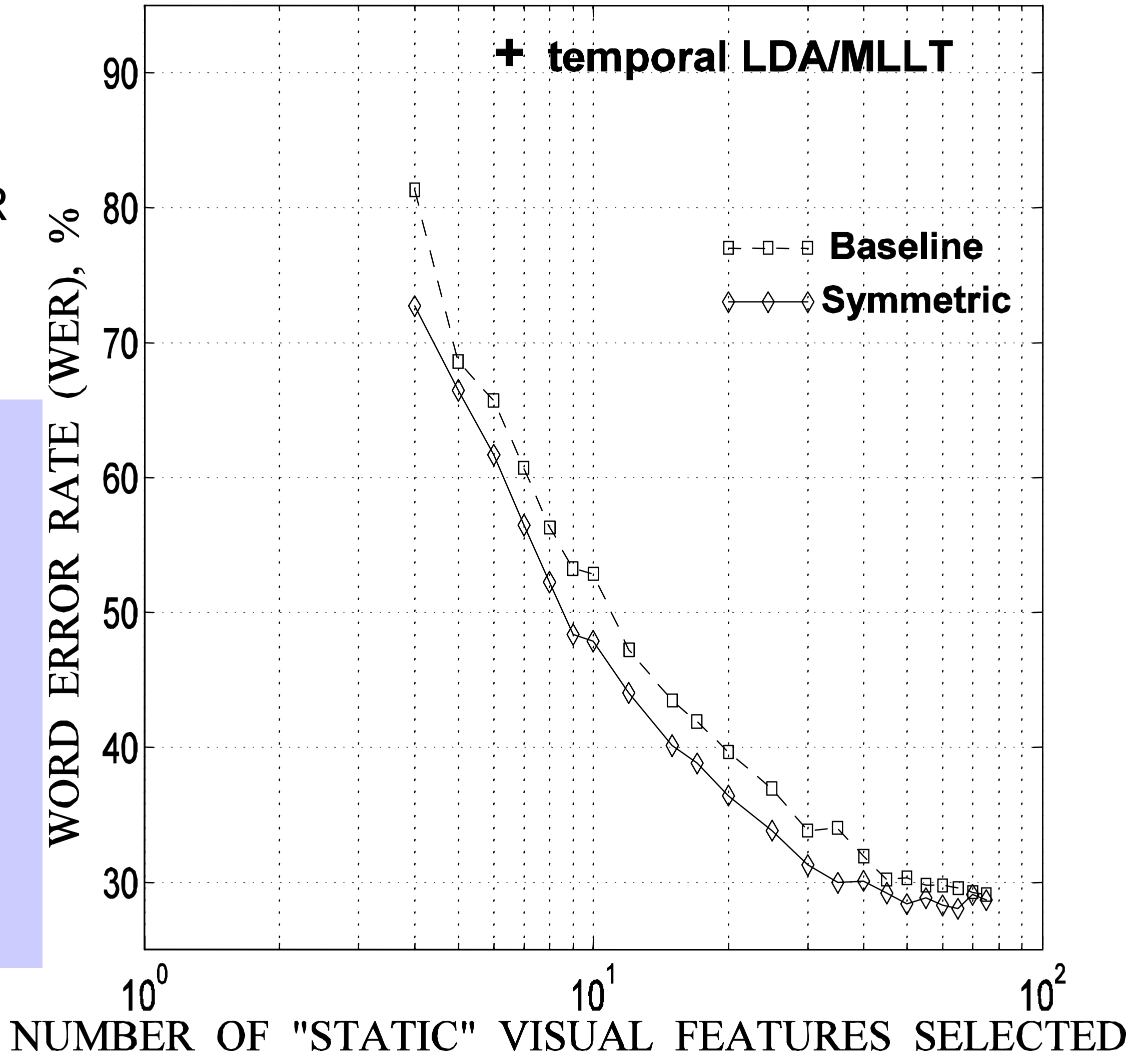
More information with same # features
Better reconstruction

DCT feat. selection – Cont.: Symmetric Energy Templates (III)

Results for “STUDIO” digits ASR



+ intra/inter frame LDA/MLLT



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: $\mathrm{WER}(\mathbf{O}_A, \mathbf{O}_V) \le \min[\mathrm{WER}(\mathbf{O}_A), \mathrm{WER}(\mathbf{O}_V)]$

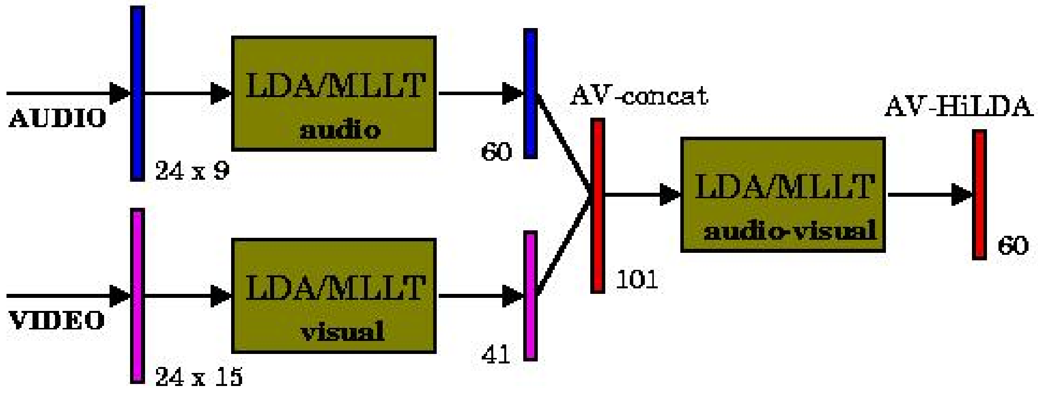
- **Main points in audio-visual fusion for ASR:**
  - **Type** of fusion:
    - ✓ Combine audio and visual info at the feature level (**feature fusion**).
    - ✓ Combine audio and visual classifier scores (**decision fusion**).
    - ✓ Could envision a combination of both approaches (**hybrid fusion**).
  - 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.**

AVASR: Feature Fusion

- **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.

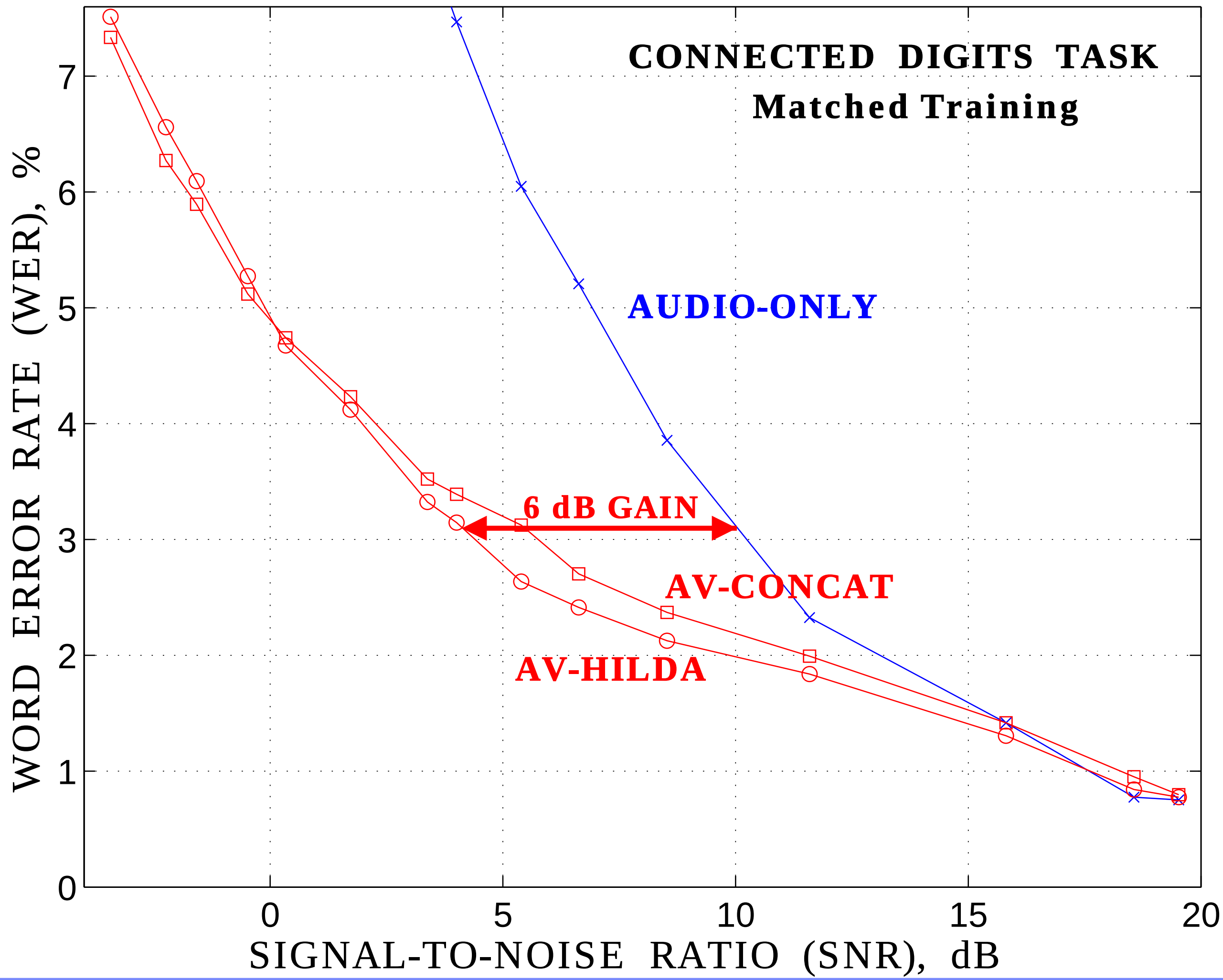
- **Examples:**
  - Feature **concatenation** (also known as **direct identification**).
  - Hierarchical discriminant features: LDA/MLLT on concatenated features (**HiLDA**).
  - **Dominant** and **motor recording** (transformation of one or both feature streams).
  - Bimodal **enhancement** of audio features.

- **HiLDA fusion advantages:**
  - Second LDA learns audio-visual **correlation.**
  - Achieves discriminant **dimensionality reduction.**



AVASR: Feature Fusion Results

- Multiple subjects (50), **connected-digits** (STUDIO dataset).
- Additive babble noise is considered at various SNRs.
- Discriminant feature fusion results in an **effective SNR gain** of **6 dB** SNR.
- Is better than feature concatenation.



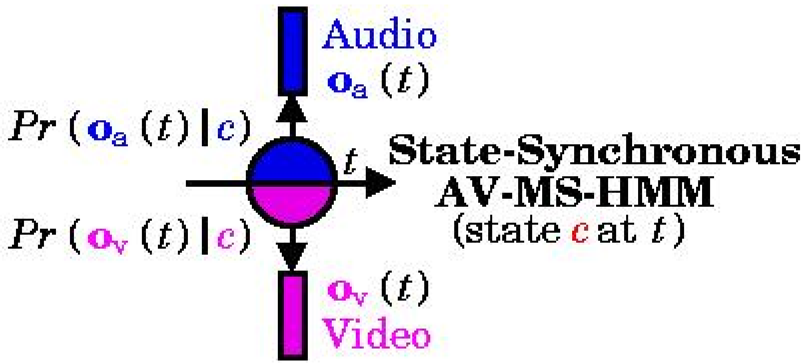
AVASR: Decision Fusion

- **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: $0 \le \lambda_{s,c,t} \le 1; \quad \sum_{s \in \{A,V\}} \lambda_{s,c,t} = 1$
  - MSHMM parameters:

$$\boldsymbol{\theta} = [\boldsymbol{\theta}_A, \boldsymbol{\theta}_V, \boldsymbol{\lambda}], \text{ where:}$$

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

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

AVASR: Decision Fusion – 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):

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

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

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

- Parameters $\boldsymbol{\lambda}$ can be obtained discriminatively – discussed later.

- MS-HMM transition probabilities:

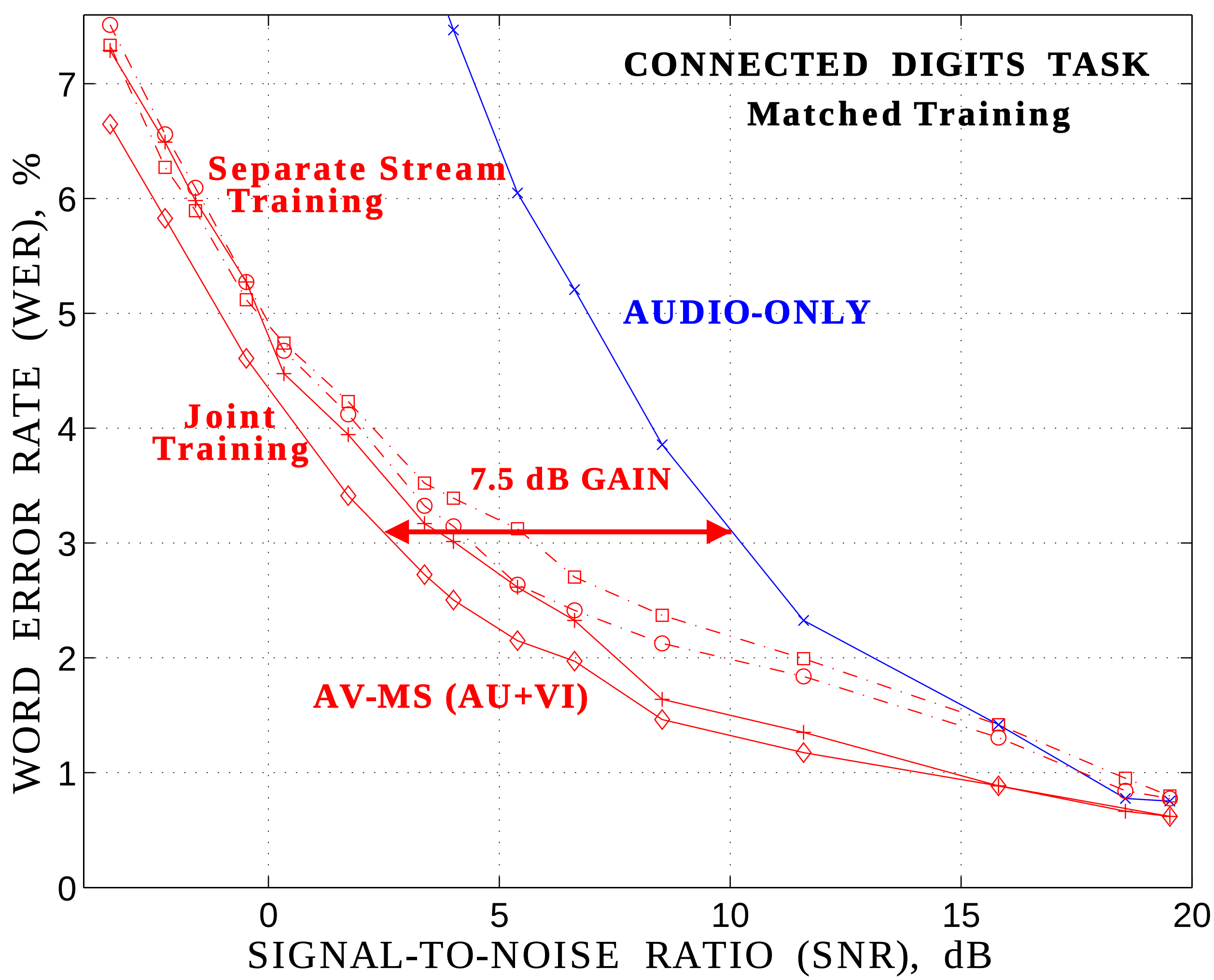
  Scores are dominated by observation likelihoods.

  One can set: $\mathbf{a}_{AV} = \mathbf{a}_A$, or $\mathbf{a}_{AV} = diag(\mathbf{a}_A^{\mathrm{T}} \mathbf{a}_V)$,

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

AVASR: Decision Fusion 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.



AVASR: Asynchrony in Audio-Visual Integration

- So far, we have considered decision fusion with scores computed at **each frame** (HMM state). This paradigm assumes **state-synchrony** of audio and visual observations.

- However:
  - ✓ Audio and visual speech are **asynchronous** – **voice onset time** (VOT).
  - ✓ Bregler et al. (1993) observe stream asynchrony of up to **120 ms** (close to phone duration).
  - ✓ Grant and Greenberg (2001) observe that speech intelligibility does not suffer when visual signal artificially **proceeds** audio by up to **200 ms**.
- Therefore, exploring asynchrony in fusion is of interest.

- In ASR, sequences of classes (HMM states) need to be estimated. Thus, integration of multiple classifiers (audio, visual, HiLDA) does not need to occur at the state level.

- Instead, **asynchronous integration** is possible, by combining scores at the:
  - ✓ *Phone*, syllable, or *word* level (**intermediate integration**). Allows limited, within-unit asynchrony, whereas synchronization is forced at the unit boundaries.
  - ✓ *Utterance* level (**late integration**). Allows complete stream asynchrony, but in practice, it requires a cascade-fusion implementation (non-real-time).

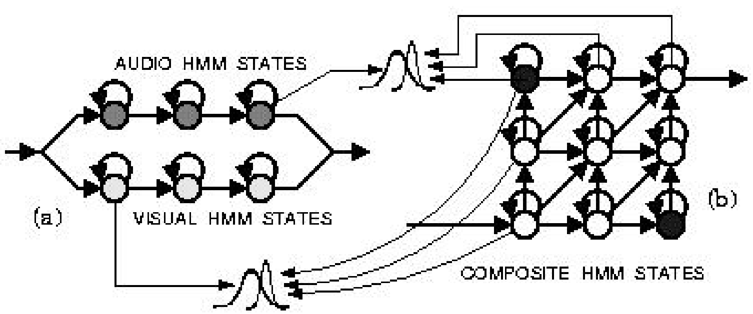
AVASR: Intermediate Integration

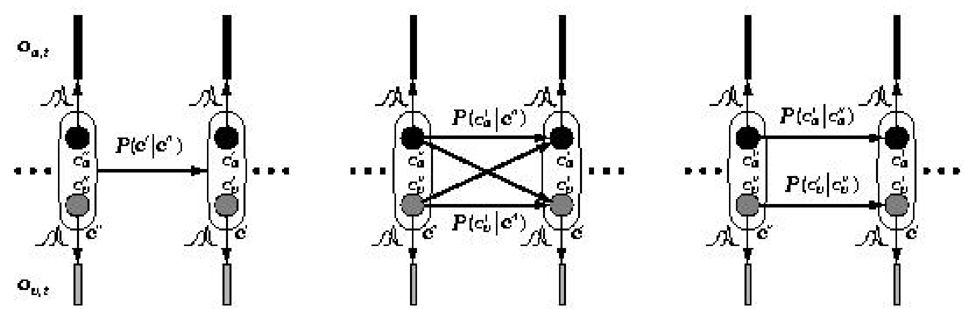
- **Intermediate integration** combines stream scores at a **coarser** unit level than HMM states, such as **phones**. This allows state-asynchrony within each phone.

- Integration model is <u>equivalent to the **product HMM**</u> (Varga and Moore, 1990).
  - Product HMM has "<u>**composite**</u>" (audio-visual) states: $\mathbf{c} = \{ c_s \,,\ s \in S \} ,\ \text{ i.e., } \mathbf{c} \in C^{|S|}$
  - Thus, state space becomes larger, e.g., |*C*|x|*C*| for a 2-stream model.
  - Class-conditional observation probabilities can follow the MS-HMM paradigm, i.e.:

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

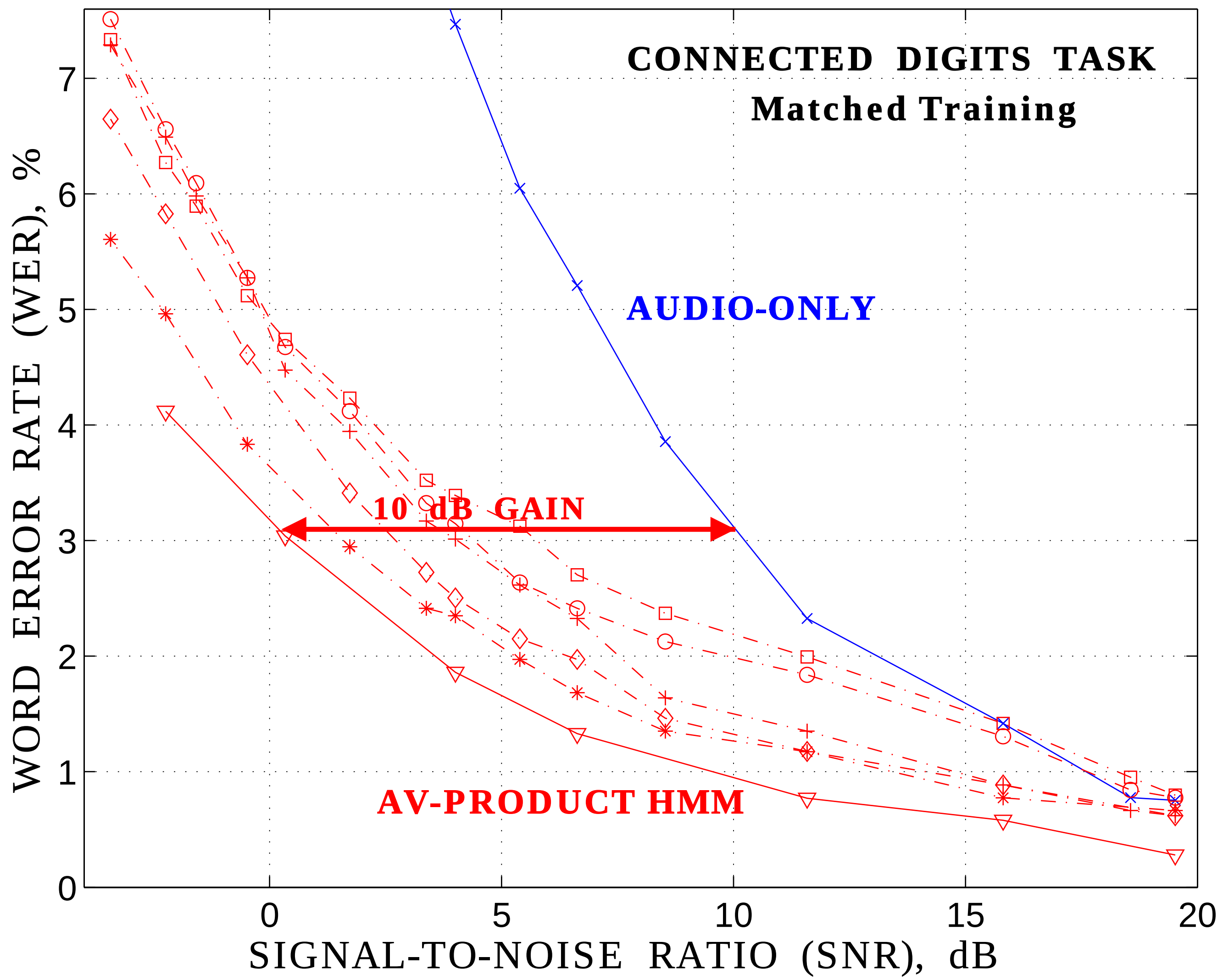
  - If tied, the observation probabilities have **same number** of parameters as state-synchronous MS-HMM.
  - Transition probabilities may be more. Three possible models:





AVASR: Intermediate Integration Results

- Recall the **connected-digit** ASR paradigm.
- **Product HMM fusion** is superior to state-synchronous fusion.
- Effective SNR gain: **10 dB** SNR.



AVASR: Late Integration

- Late integration advantages:
  - ✓ Complete asynchrony between the stream observation sequences.
  - ✓ No need for same data rate between the streams.
- General implementation:
  - ✓ In **cascade** fashion, by rescoring of n-best sentence lists or lattice word-hypotheses.
  - ✓ Thus, real-time implementation is not feasible.
- Typical example: **Discriminative model combination** (**DMC**).
  - ✓ For each utterance, use audio to obtain n-best list: $\{\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n\}$
  - ✓ Force-align each hypothesis phone sequence $\mathbf{h}_i = \{c_{i,1}, c_{i,2}, \ldots, c_{i,N_i}\}$ per modality s into: $[t^{\text{start}}_{i,j,s}, t^{\text{end}}_{i,j,s}]$
  - ✓ Then rescore:

$$\Pr[\mathbf{h}_i] \propto \Pr_{\text{LM}}(\mathbf{h}_i)^{\lambda_{\text{LM}}} \prod_{s \in S} \prod_{j=1}^{N_i} \Pr(\mathbf{o}_{s,t}, t \in [t^{\text{start}}_{i,j,s}, t^{\text{end}}_{i,j,s}] \mid c_{i,j})^{\lambda_{s,c_{i,j}}}$$

  - ✓ All weights are discriminatively trained to minimize WER in a held-out set.

AVASR: Stream Reliability Modeling

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

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

- Stream exponents model reliability (information content) of each stream.
- We can consider:
  - ✓ **<u>Global weights</u>**: 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.

$$\lambda_{s,c,t} \longrightarrow \lambda_{s,c}$$

  - ✓ **<u>Adaptive weights</u>** at a **<u>local</u>** level (**<u>utterance</u>** or **<u>frame</u>**): Assumes that the environment varies locally (more practical). Requires stream reliability estimation at a local level, and mapping of such reliabilities to exponents.

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

AVASR: Global Stream Weighting

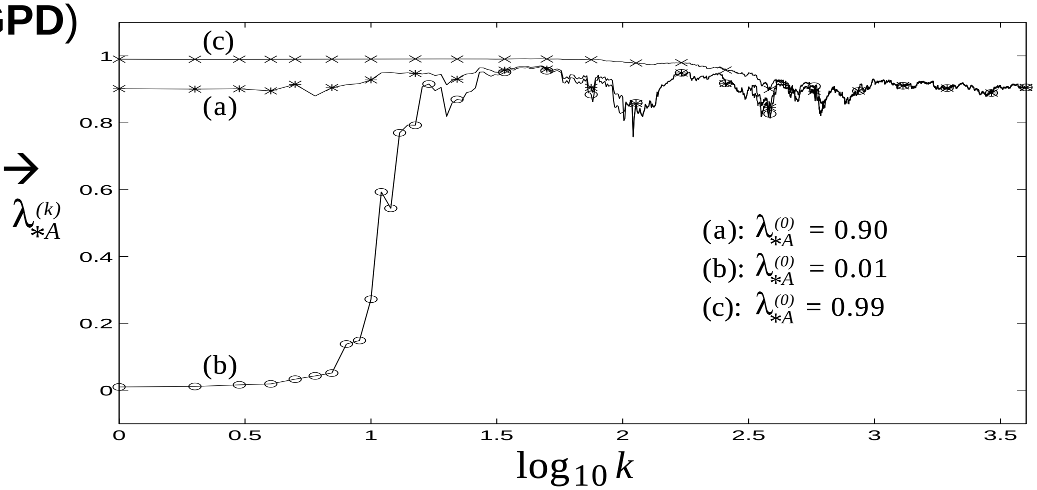
- Stream weights **cannot** be obtained by **maximum-likelihood** estimation, as:

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

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 →

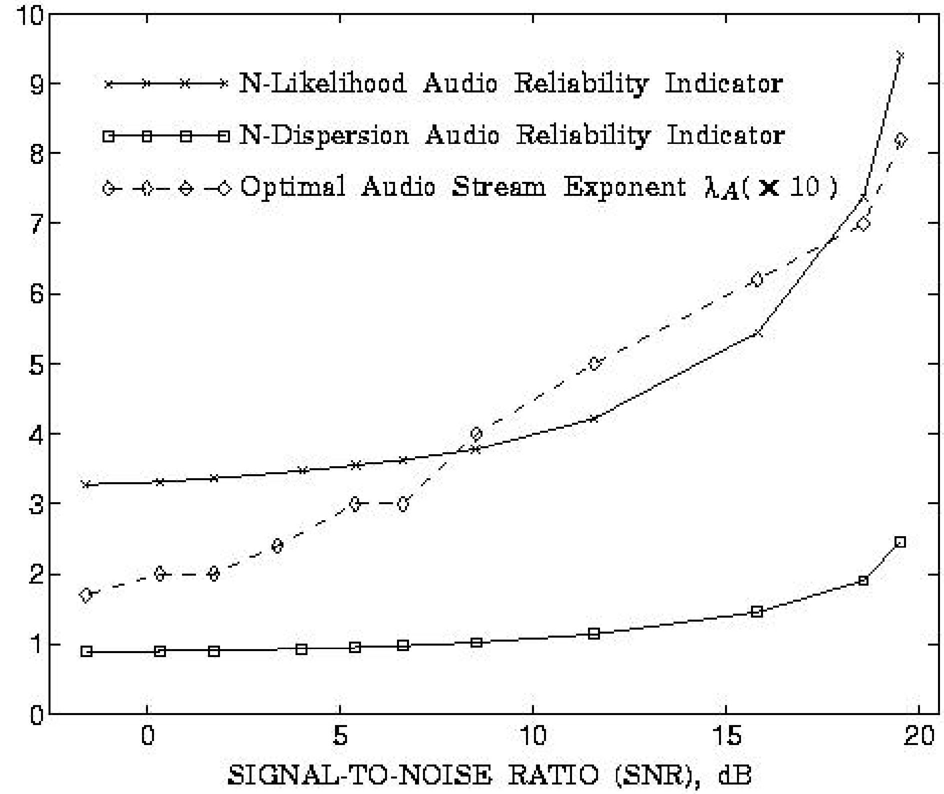
      (GPD based estimation)



AVASR: Adaptive Stream Weighting

- In practice, stream reliability varies **locally**, due to audio and visual input degradations (e.g., noise bursts, face tracking failures, etc.).
- **Adaptive weighting** captures variations, by:
  - **Estimating** environment **reliabilities.**
  - **Mapping** them to stream exponents.
- Stream reliability indicators:
  - **Acoustic** signal based: SNR, voicing index.
  - **Visual** processing: Face tracking confidence.
  - **Classifier** based stream reliability indicators:
    - ✓ 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**:

$$L_{s,t} = \frac{1}{N-1} \sum_{n=2}^{N} \log \frac{\Pr(\mathbf{o}_{s,t} \mid c_{s,t,1})}{\Pr(\mathbf{o}_{s,t} \mid c_{s,t,n})}$$



    - ✓ N-best log-likelihood **dispersion**: $D_{s,t} = \frac{2}{N(N-1)} \sum_{n=2}^{N} \sum_{n'=n+1}^{N} \log \frac{\Pr(\mathbf{o}_{s,t} \mid c_{s,t,n})}{\Pr(\mathbf{o}_{s,t} \mid c_{s,t,n'})}$
- 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).

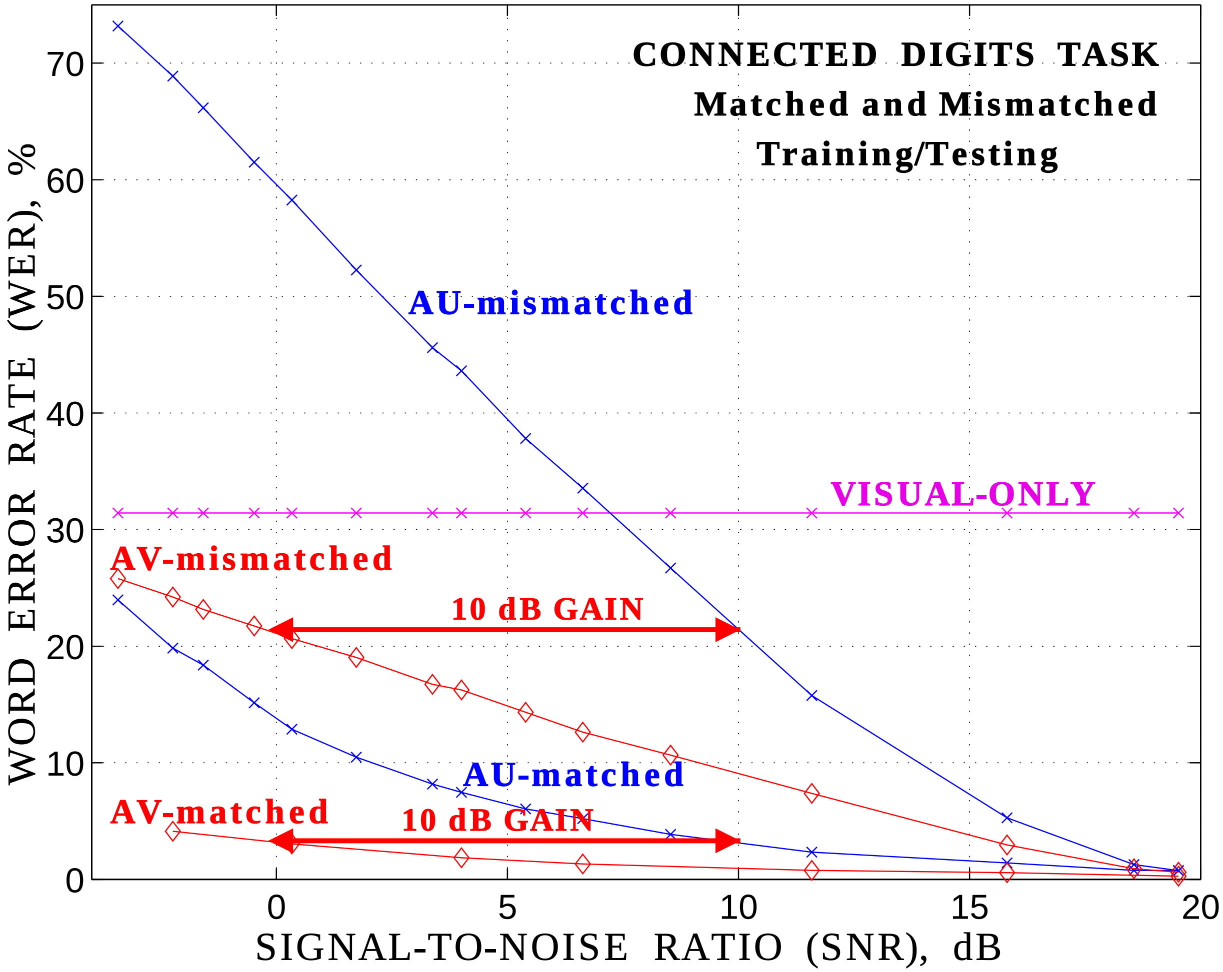
AVASR: Summary of Fusion Results

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

- Multi-speaker training / test.
- 50 subjects, 10 hrs of data.
- Additive noise - many 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**.

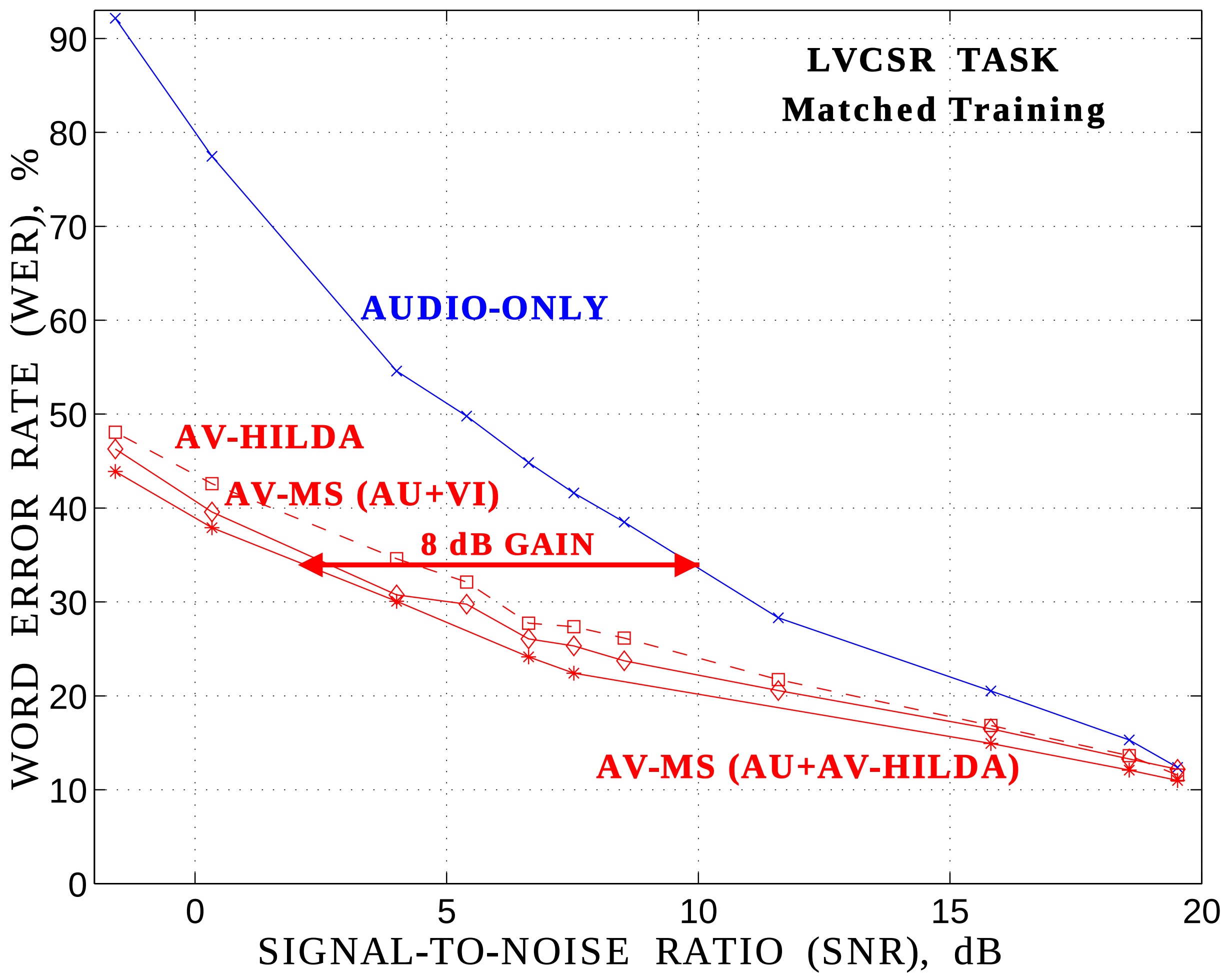


AVASR: Summary of Fusion Results – Cont.

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.

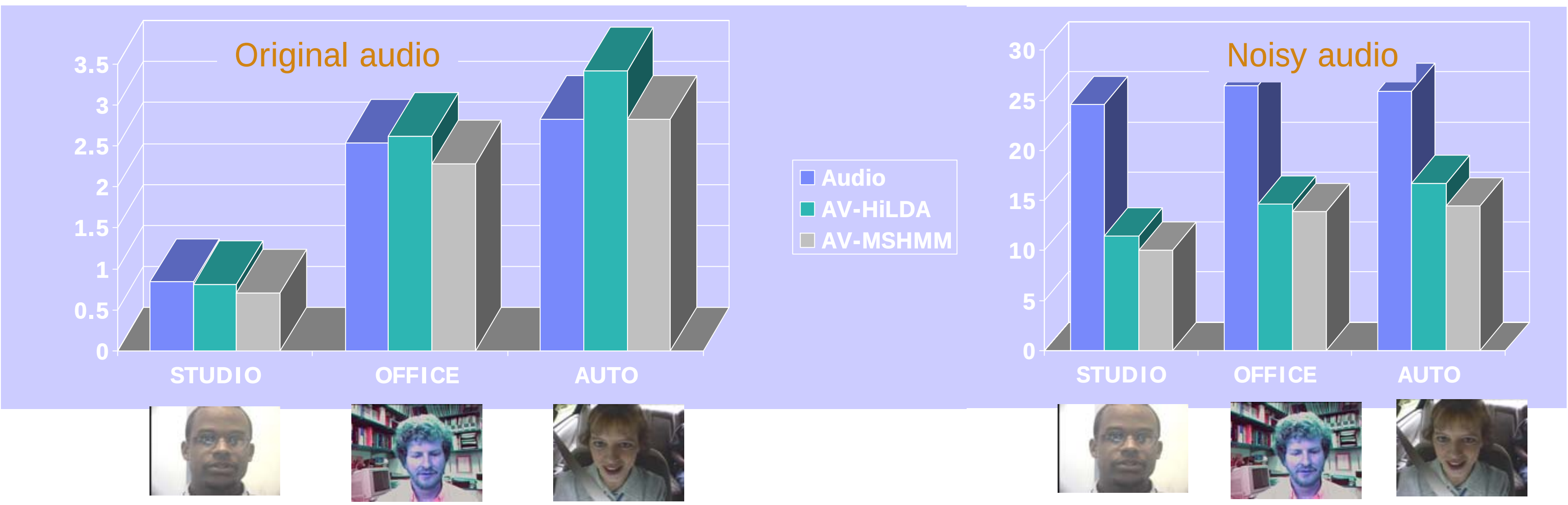
- **8 dB** effective SNR **gain** using hybrid fusion.



AVASR Results – Cont.

AV-ASR in challenging domains:

- 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, %.

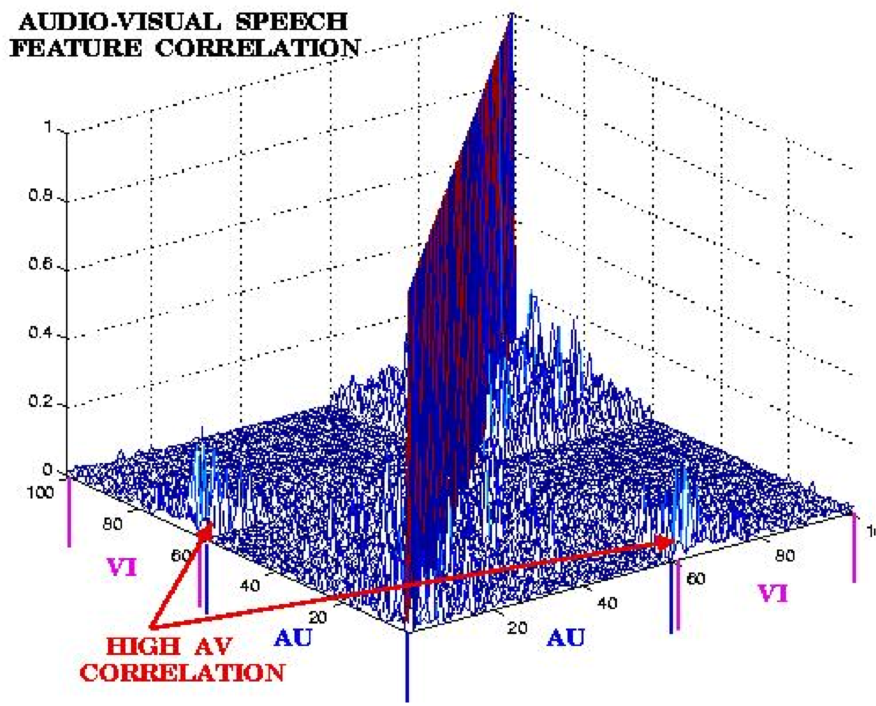


Additional Audio-Visual Speech Technologies

- So far, we have discussed the ***two*** main ***components*** of AV speech processing, as applied to the problem of ***audio-visual ASR***.
- These components are ***shared*** & are relevant to a number of audio-visual speech processing applications, as discussed in the Introduction.
- We briefly discuss a few of them:
  - Speech ***enhancement***.
  - Speaker ***identification / verification***.
  - Speech ***activity detection*** – visual only is discussed.
  - Speech ***synthesis***.

Audio-Visual Speech Enhancement – Brief Overview

- **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.

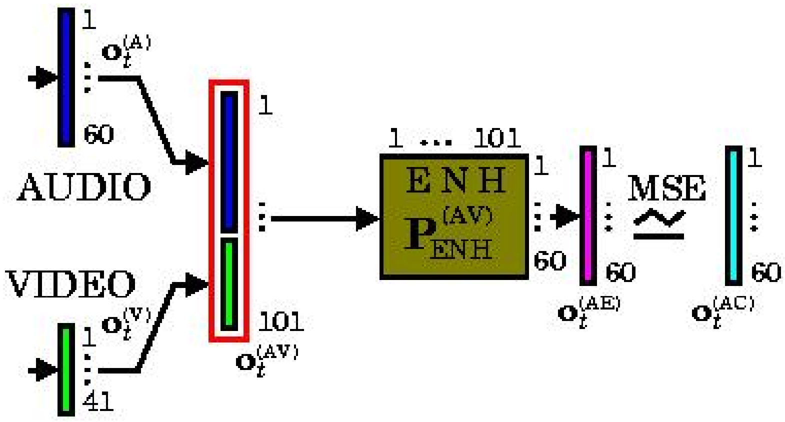
Linear Bimodal Enhancement of Audio (I)

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

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

  - ✓ Seek linear transform $\mathbf{P}$, s.t:

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



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

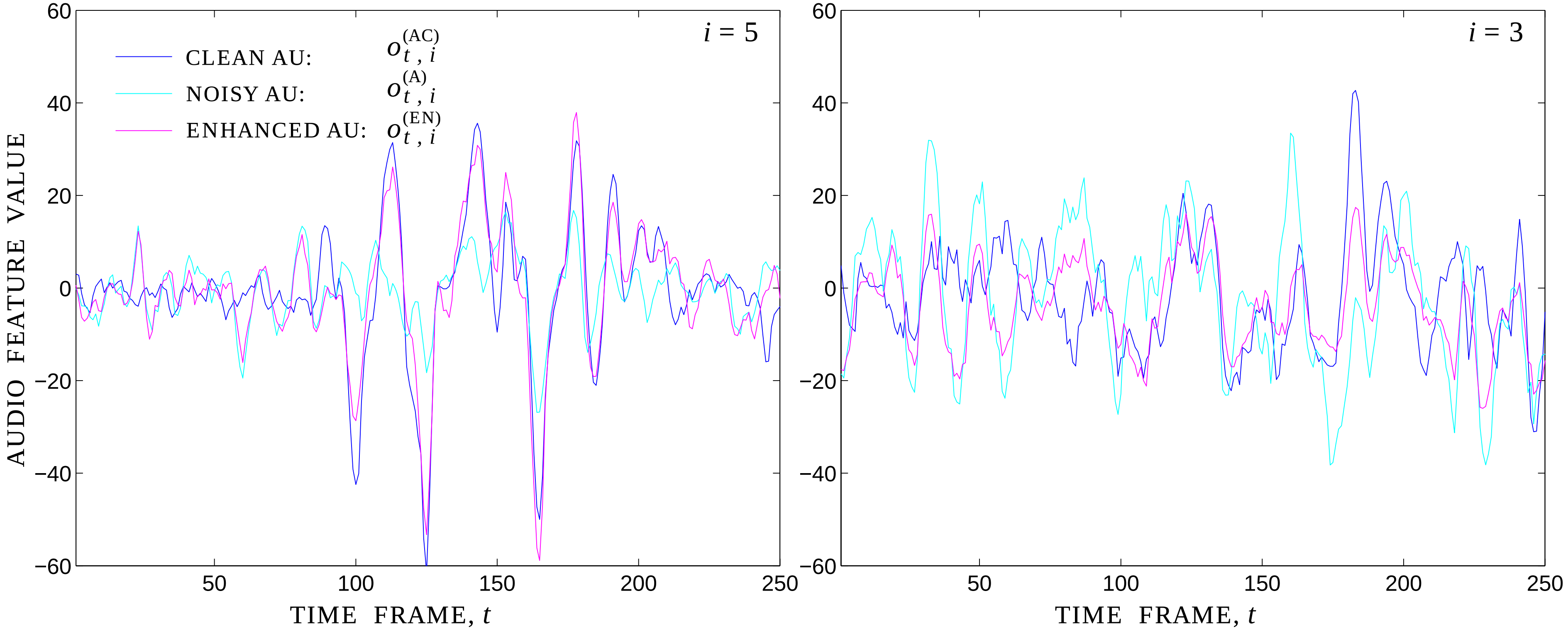
    $$\mathbf{p}_i = \arg\max_{\mathbf{p}} \sum_{t \in T} [\, o^{(C)}_{A,t,i} - <\mathbf{p}, \mathbf{o}_{AV,t}>]^2,\quad i = 1,\ldots,d_A$$

  - ✓ Solved by d_A systems of Yule-Walker equatiions:

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

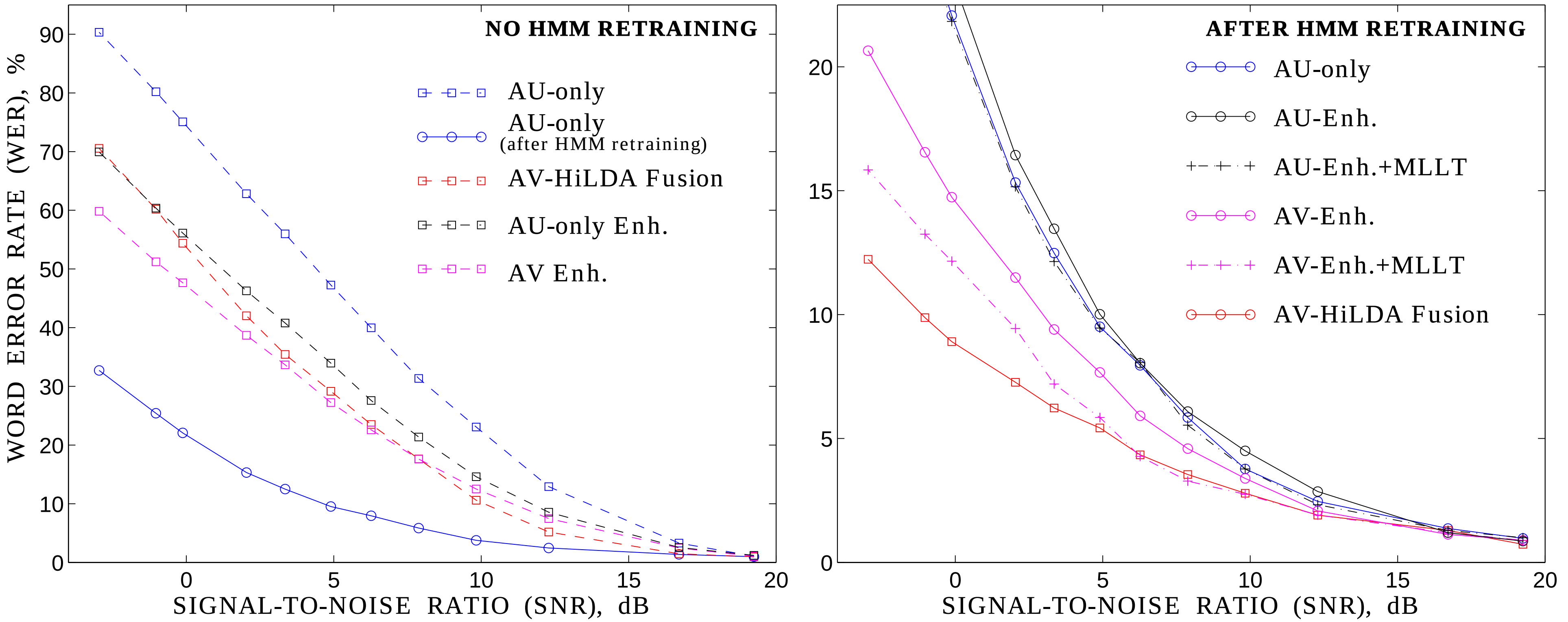
Linear Bimodal Enhancement of Audio (II)

- 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 Enhancement of Audio (III)

- **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.



Non-Linear Bimodal Enhancement of Audio (I)

- **Codebook-dependent cepstral normalization** (**CDCN**):
  - A feature-space technique for robust ASR.
  - Approximates the non-linear effect of noise on clean features by a piece-wise constant function, defined in terms of a "codebook" $\{f_{A,k}\}$:

$$\mathbf{o}_{A,t}^{(E)} = \mathbf{o}_{A,t} - \sum_{k=1}^{K} f_{A,k} \Pr(k \mid \mathbf{o}_{A,t})$$

  - Codebooks are estimated by minimizing MSE over audio data:

$$f_{A,k} = \frac{\sum_{t \in T} (\mathbf{o}_{A,t} - \mathbf{o}_{A,t}^{(C)}) \Pr(k \mid \mathbf{o}_{A,t}^{(C)})}{\sum_{t \in T} \Pr(k \mid \mathbf{o}_{A,t}^{(C)})}$$

- CDCN can be **extended** to use audio-visual data instead (**AV-CDCN**):

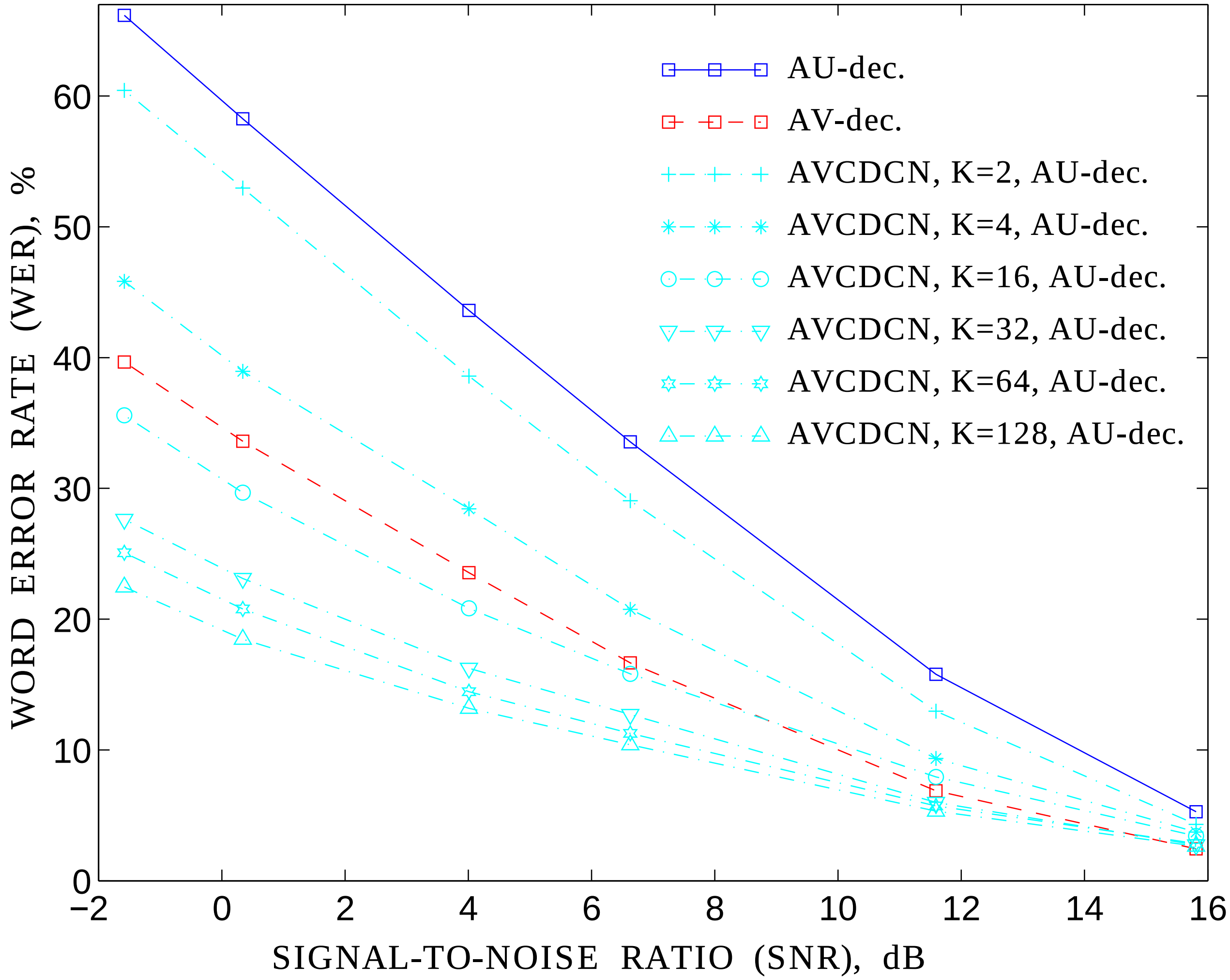
$$\mathbf{o}_{A,t}^{(E)} = \mathbf{o}_{A,t} - \sum_{k=1}^{K} f_{A,k} \Pr(k \mid \mathbf{o}_{AV,t})$$

where **codebook posteriors** $\{\Pr(k|\mathbf{o}_{AV,t})\}_k$ are estimated by **EM** on AV data.

Non-Linear Bimodal Enhancement of Audio (II)

RESULTS (Deligne et al., '02).

- **ASR** performance using AVCDCN vs. audio-only and AV-HiLDA features.
- **Task:** Connected digits, HMMs trained on clean audio.
- Various **codebook sizes** are compared in AVCDCN.
- AVCDCN outperforms feature fusion!



Audio-Visual Speaker Recognition – Brief Overview

In case of **bimodal data**, the following ***3 information streams*** **can be utilized**:

- Sound – **audio** based speaker recognition
- Static video frames – **face** recognition
- Mouth ROI video sequences – **visual** speech based speaker recognition.

Examples **of fusing two or three single-modality speaker-recognition systems:**

Audio + visual-labial (IBM:Chaudhari et al.,03)

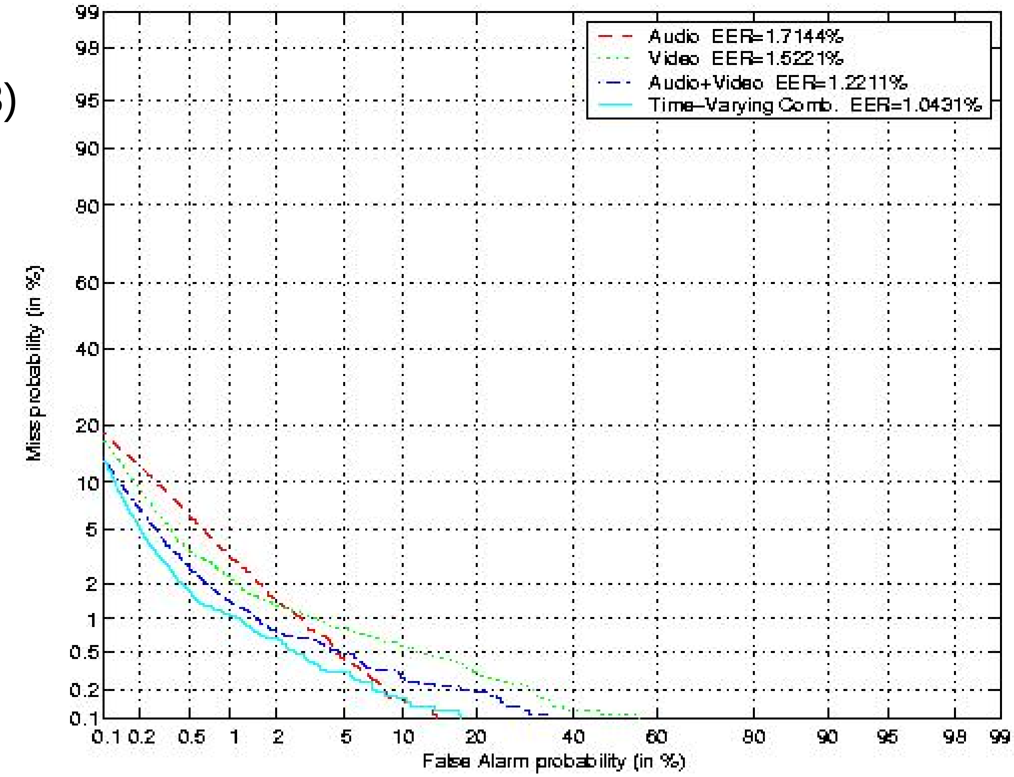
- ID-error: A: 2.01, V: 10.95, AV: **0.40** %
- VER-EER: A:1.71, V: 1.52, AV: **1.04** %

Audio +visual-face (IBM: Maison et al., 99)

- ID-error-*clean*: A: 7.1, F: 36.4, AF: **6.5**
- ID-error-*noisy*: A:49.3, F: 36.4, AF: **25.3** %

Audio + visual + face (Dieckmann et al., 97):

- ID-err: A: 10.4, V: 11.0, F: 18.7, AVF: **7.0** %

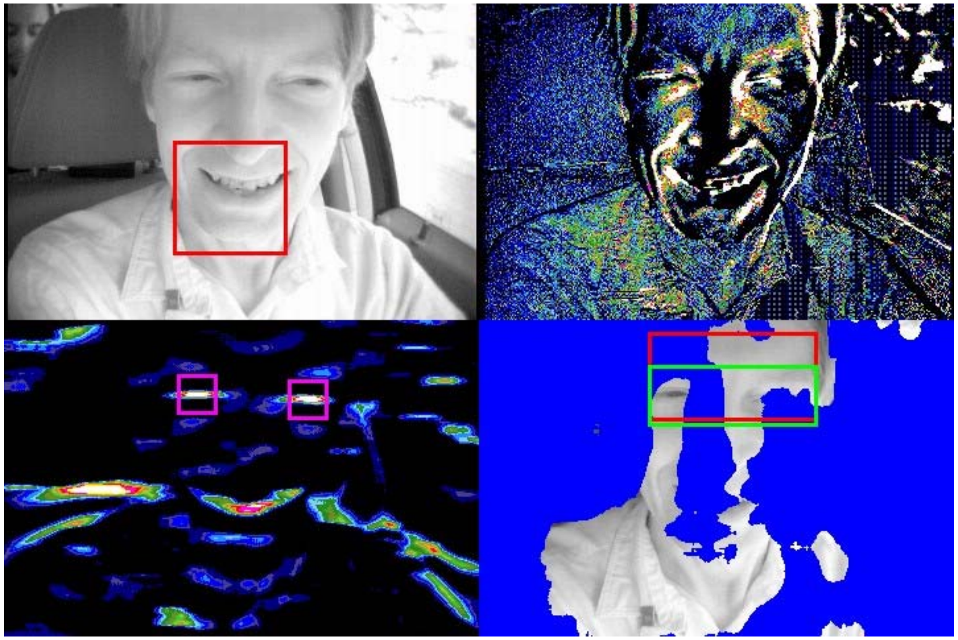


In-Vehicle Visual Speech Activity Detection (I)

- ***WHAT:*** **Speech activity** detection in the automobile, using **visual input** from specially designed sensor, and "**low-cost**" algorithms, aiming at embedded implementation.

- ***The visual sensor:***
  - **Monochrome** (visible + near IR sensitive) equipped with synchronously **flashing IR** LEDs.
  - Allows depth segmentation based on the near objects brightness difference (due to the flashing IR LEDs)

- ***The algorithms:***
  - ✓ Find driver's ***head***.
    Uses synchronous IR **depth finder.**
    Establishes search region for eyes.
  - ✓ Find ***eyes***.
    Uses **matched-filter template** search.
    Establishes search region for mouth.
  - ✓ Find ***mouth***.
    Threshold **patch** based on eyes.
  - ✓ Analyze ***mouth motion***.
    High area **variability** → speech.



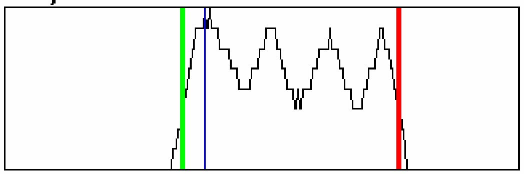
Visual Speech Activity Detection (II) – Alternative Head Finding

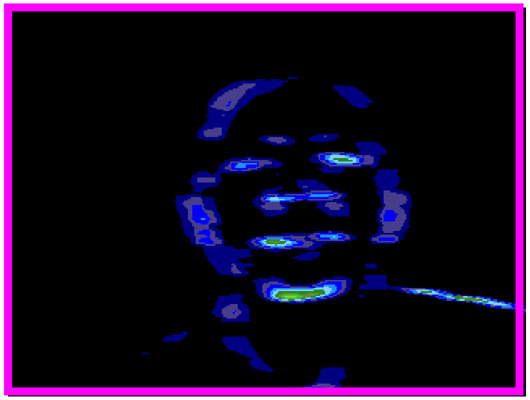
- **Searches for central moving region:**
  Uses just a **standard** camera (B&W).
  No flashing IR lights (can be always on).
  More **tolerance** for frame rate and resolution.

- **Steps:**
  Find **frame difference** over interval.
  Accumulate **motion** evidence.
  Project density to find **head limits**.

- **Complexity:**
  Approximately the same **complexity** as the synchronous IR version (previous slides).

head limits

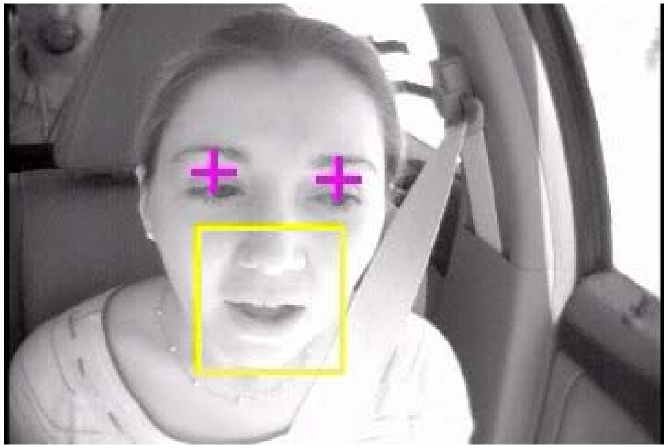




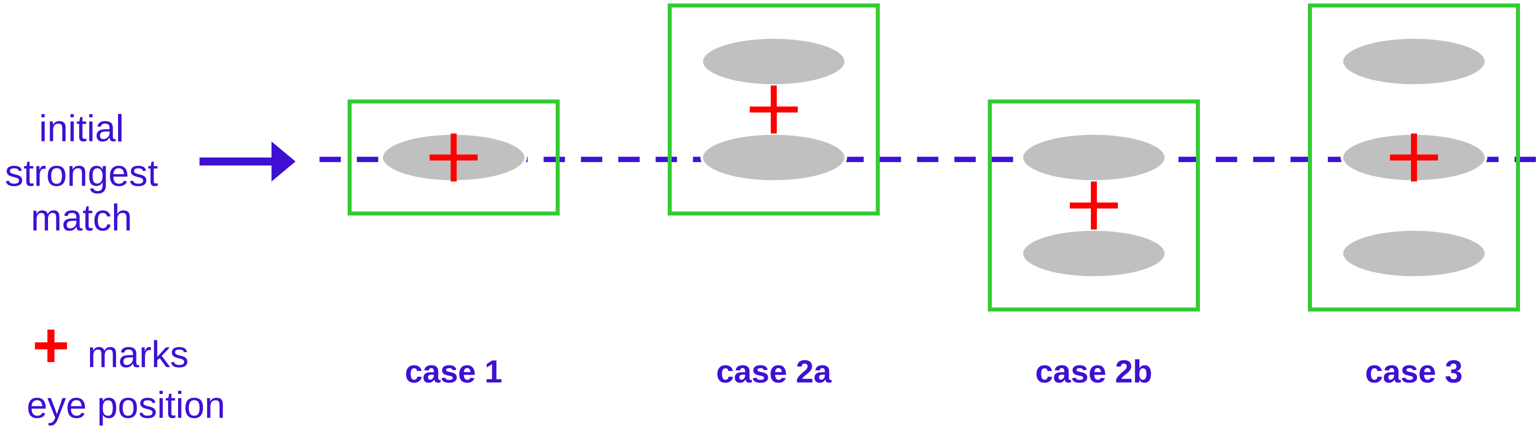
motion history

Visual Speech Activity Detection (III) – Improved Eye Detection

- **Extend** single-template search (per eye) to improve eye detection.
  - Use **multi-bar** model.
  - Accounts for **eyebrows** and **eyeglass frames**.



- **Steps:**
  - Look for **strongest black bar** candidate.
  - Also look above and below.
  - Pick **eye center** based on pattern.



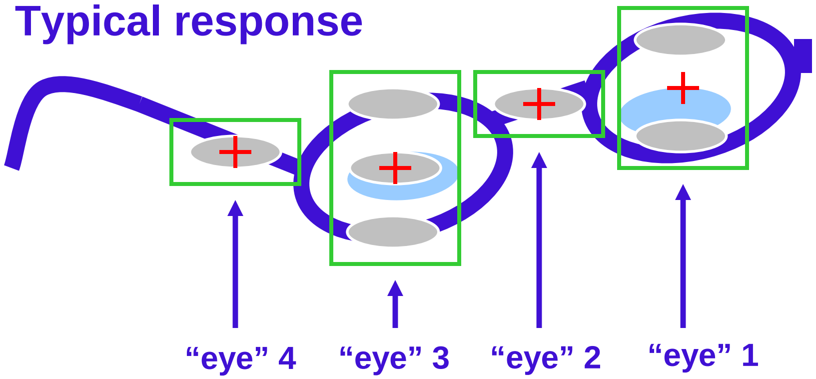
Visual Speech Activity Detection (IV) – Handling Glasses

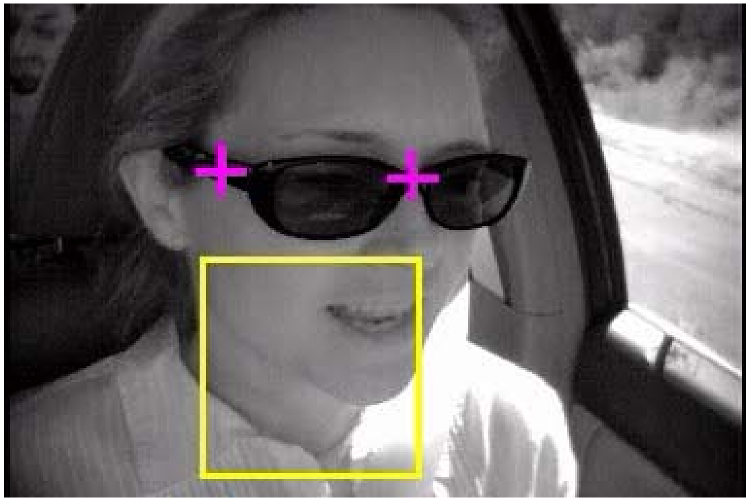
- Detection of eyes in presence of **glasses** needs special handling.
- Our algorithm utilizes a "**four-eye**" model.
  - Glasses look like a chain of virtual eyes.

Algorithmic steps:

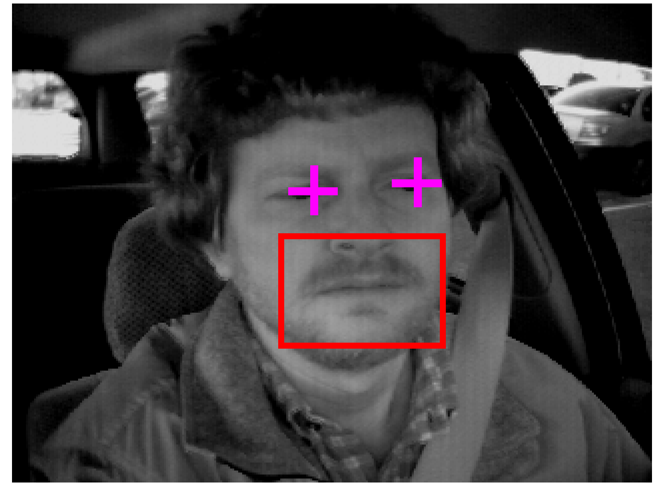
- Find **up to four** multi-bar candidates.
- Choose **one** candidate as an **anchor**.
  - Prefer leftmost or rightmost.
  - Prefer closest to previous eyes.
- Pick mate with the best spacing.
  - Validate separation & tip angle.
  - Pick different anchor if violation.
- Adds negligible processing.

Typical response

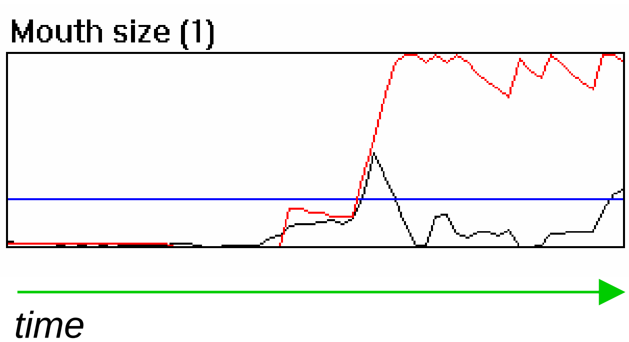


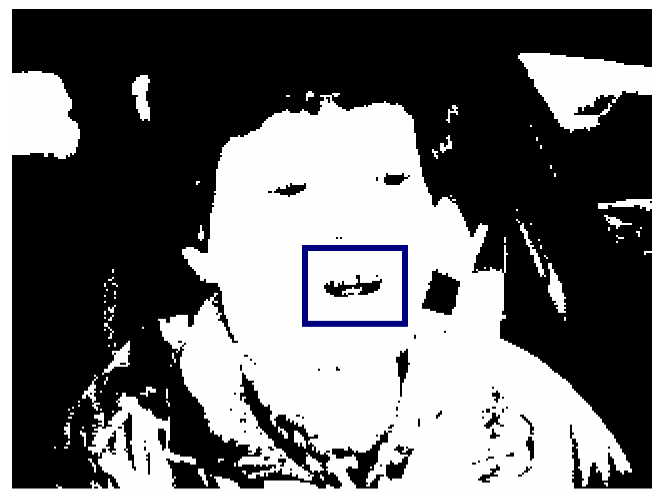


Visual Speech Activity Detection (V) – Mouth Measurement



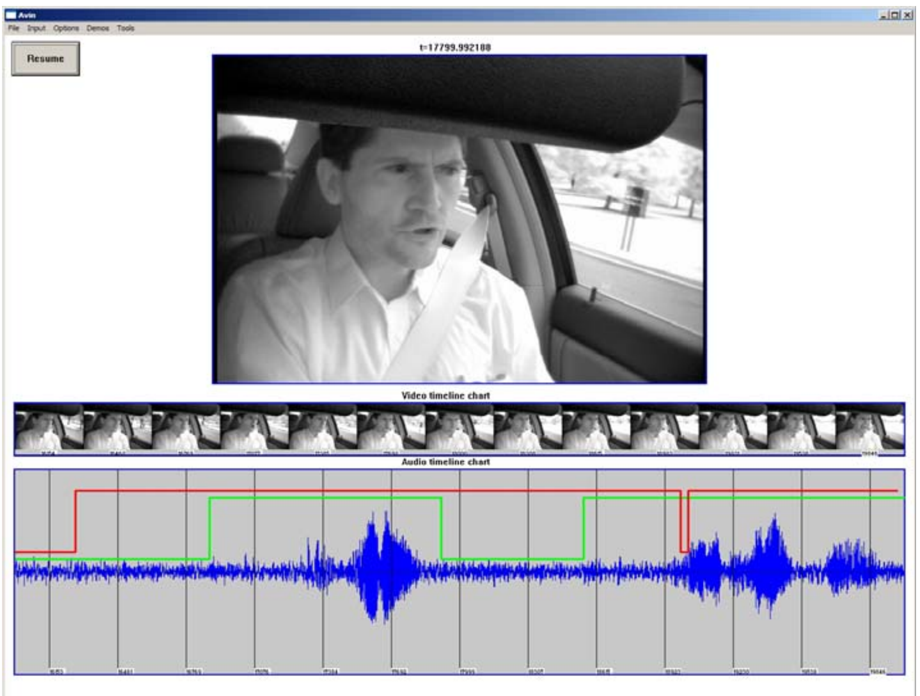
- Basic Steps:
  - ✓ Find likely mouth area based on eyes
  - ✓ Look for and track dark blotch of mouth
  - ✓ Monitor change in size over time
- Refinements:
  - ✓ Uses bar-mask to re-center mouth area
  - ✓ Uses gray scale average for low resolution





Visual Speech Activity Detection (VI) – Data & Evaluation

- **AV Data:**
  - 10 drivers, 10 passengers, 4:45 hours total.
  - Good lighting, head-pose, expression variation.
- **Evaluation:**
  - Ground-truth and evaluation tools developed.
  - Metrics: (SDER, NDER).
  - Results: (17%,19%).



Audio-Visual Speech Synthesis (I)

- The ***goal*** is to automatically generate:
  - Voice and facial animation from arbitrary ***text***; or:
  - Facial animation from arbitrary ***speech***.

- ***Potential applications***:
  - Human communication and perception.
  - Tools for the hearing impaired.
  - Spoken and multimodal agent-based user interfaces.
  - Educational aids.
  - Entertainment (synthetic actors).

- For example:
  - A view of the face can improve intelligibility of both natural and synthetic speech significantly, especially under degraded acoustic conditions.
  - Facial expressions can signal emotion, add emphasis to the speech and support the interaction in dialogue.

Audio-Visual Speech Synthesis (II) - Approaches

- ***Model-Based*** (or knowledge-based)
  - Face is modeled as a 3D object
  - Control parameters deform the 3D structure using
    - ✓ Geometric
    - ✓ Articulatory
    - ✓ Muscular

    models
  - Gained popularity due to MPEG-4 facial animation standard

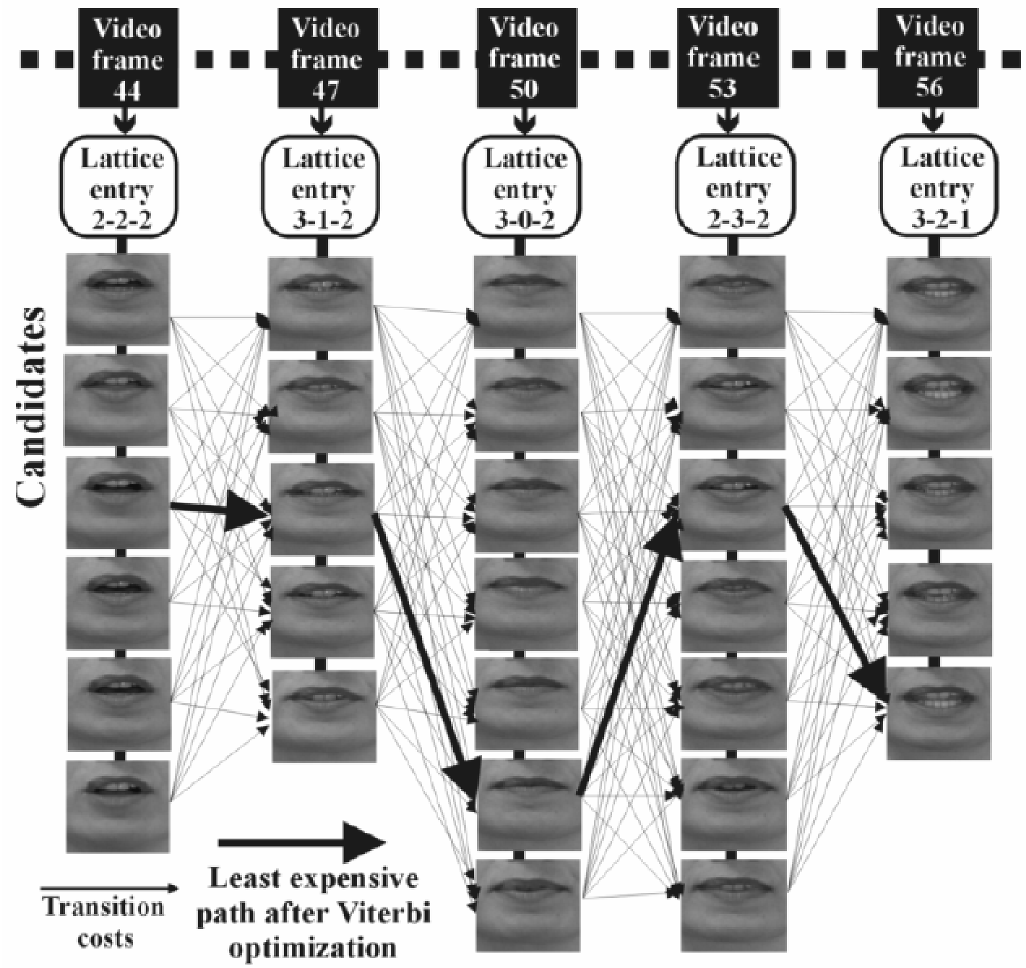
- ***Image or Video-Based***
  - Segments of 2D videos of a speaker are
    - ✓ Acquired
    - ✓ Processed
    - ✓ Concatenated

Boundaries are blurry

Audio-Visual Speech Synthesis (III) – Concatenative Approach

Basic components of this approach are similar to the AV-components discussed earlier.

- Analysis of database segments (images or video snippets).
  - Extracts shape or appearance features to allow transition cost computation in concatenation.
- Synthesis stage:
  - Uses dynamic programming approach (Viterbi) to find minimum cost path and "stich" together the best possible image/video snippets.



Audio-Visual Speech Synthesis (IV) – Speech Driven Animation

- Goal: Synthesize video directly from the acoustic signal.

- Approaches are classified into
  - Symbol based:
    - Audio signal is first translated into an intermediate discrete representation – sequence of phonemes.
  - Regression based.
    - A direct continuous association between acoustic and visual features is sought.

- Both constitute interesting cases of audio-visual fusion; Can be accomplished with various techniques:
  - HMMs (correlation HMMs).
  - Regression.
  - Artificial Neural Networks.

AV Speech Processing – Conclusions

- Discussed the **motivation** & **benefits** of visual information for various speech technologies.
- Audio-visual speech processing requires **visual feature extraction** & **audio-visual fusion**.
- For visual processing, **appearance-based visual features seem preferable**.
  – Achieve better performance.
  – Are computationally inexpensive.
  – Robust to video degradations.
  – Require approximate only face/mouth tracking
- For audio-visual integration, **decision fusion approaches are preferable**:
  – Draws from the classifier combination paradigm.
  – Allows direct modeling of the reliability of each information stream
  – Offers a mechanism to directly model audio-visual asynchrony at various levels.
- **Experimental results** demonstrate the huge benefit of visual modality to ASR.
  – Sizeable gains in clean acoustics.
  – 8-10 dB gains in effective SNR.
- **Discussed additional AV speech applications.**
  – Identification / verification.
  – Speech enhancement.
  – Speech activity detection.
  – Speech synthesis.
- Many **problems remain open**:
  – Pose modeling, compensation; pose invariant appearance visual features.
  – Robust visual feature extraction for unconstrained visual domains.
  – Additional work in decision fusion: Fusion functional, reliability modeling, asynchronous integration.

Acknowledgements

- **IBM colleagues**: *Stephen M.* ***Chu****, Jonathan* ***Connell****, Sabine* ***Deligne****, Giridharan* ***Iyengar****, Vit* ***Libal****, Chalapathy* ***Neti****, Larry* ***Sansone****, Andrew* ***Senior****, Roberto* ***Sicconi***.

- **Work during past IBM internships:**
  - *Ashutosh* ***Garg*** (Google, CA) – AV fusion (frame dependent weights).
  - *Roland* ***Goecke*** (ANU) – AV speech enhancement (linear model).
  - *Guillaume* ***Gravier*** (INRIA/IRISA, FR) – AV fusion (MS/product HMMs).
  - *Jintao* ***Jiang*** (HEI, CA) – Face detection improvements.
  - *Zhenqiu* ***Zhang*** (UIUC, IL) – AAMs/AdaBoost for face detection.
  - *Patrick* ***Lucey*** (QUT) – Multi-view AVASR.
  - *Patricia* ***Scanlon*** (Lucent, IRL) – Visual feature selection.

- **Other:**
  - *Petar S.* ***Aleksic****, Aggelos K.* ***Katsaggelos*** (Northwestern University, IL): ICIP tutorial (AV synthesis).
  - *Iain* ***Matthews*** (CMU, PA), *Juergen* ***Luettin*** (Bosch, GmbH, Germany): Summer 2006 workshop at JHU/CLSP, MD.

