National Centre for Scientific Research "Demokritos"

Institute of Informatics & Telecommunications

Audio-Visual Automatic Speech Recognition & Related Bimodal Technologies: A Review of the State-of-the-Art & Open Problems

Gerasimos Potamianos

Research Director, Institute of Informatics & Telecommunications,
National Center for Scientific Research (NCSR) "Demokritos"
Athens, Greece

http://www.iit.demokritos.gr/~gpotam

Some words about NCSR "Demokritos"

- Largest Greek gvmnt funded **research center**.
- Located in Athens, Greece.
- Founded in the late **50's**.
- Consists of **8 research institutes** – very diverse.
- Bird's eye view ➔

Institute of Informatics & Telecommunications

- About **100** permanent & collaborating **staff**.
- Over **20** national & EU **projects** currently running.
- One significant concentration is on **computational intelligence systems**.
  - *Text, video, audio processing; knowledge engineering; machine learning*.

Some current EU projects at IIT

- **INDIGO** ➔ *Interaction with Personality and Dialogue Enabled Robots* – single person-robot HCI – cultural heritage domain, anthropomorphic robot.
- **CASAM** ➔ *Computer-Aided Semantic Annotation of Multimedia* – aggregate human and machine knowledge with the ultimate target of minimizing human involvement in the annotation of multimedia content.
- **PRONTO** ➔ *Event Recognition for Intelligent Resource Management* – real-time, knowledge-led support for decision-makers in sectors characterised by large volumes of multi-source, multi-format data.
- **PASCAL2** ➔ *Pattern Analysis, Statistical Modeling and Computational Learning* – NoE.
- **IMPACT** ➔ *Improving Access to Text* – innovative tools to enhance the capabilities of OCR engines and the accessibility of digitised text and lay down the foundations for the mass-digitisation programs
- **SYNC3** ➔ *Synergistic Content Creation & Communication* – intelligent framework for making more accessible the vast quantity of user comments on news issues – connect blogosphere & traditional media sources.
- **AVISPIRE** ➔ *Audio-Visual Speech Processing for Interaction in Realistic Environments* – starting now (FP7–PEOPLE–RG). AV speech processing in broadcast news and meeting domains.

Overview of Presentation

1. Introduction:

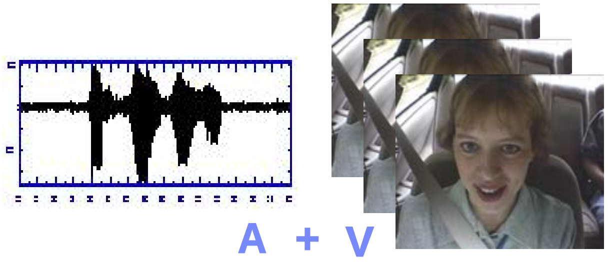
- Motivation.
- Audio-visual speech technologies.
- Potential applications.

2. Audio-visual speech components with emphasis on ASR:

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

3. Other audio-visual speech technologies:

- Speech synchrony.
- Speech enhancement.
- Speech inversion.
- Speaker recognition.
- Speech synthesis.

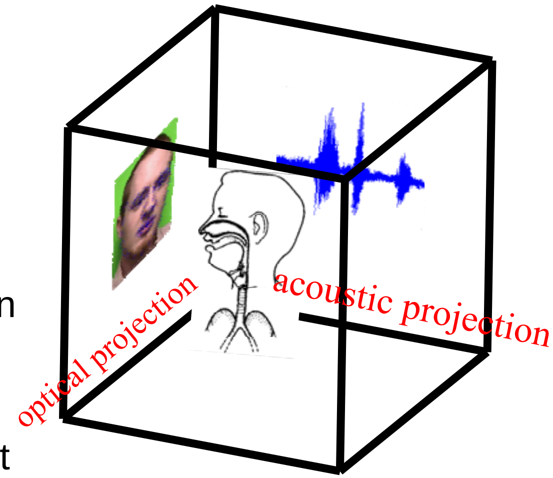


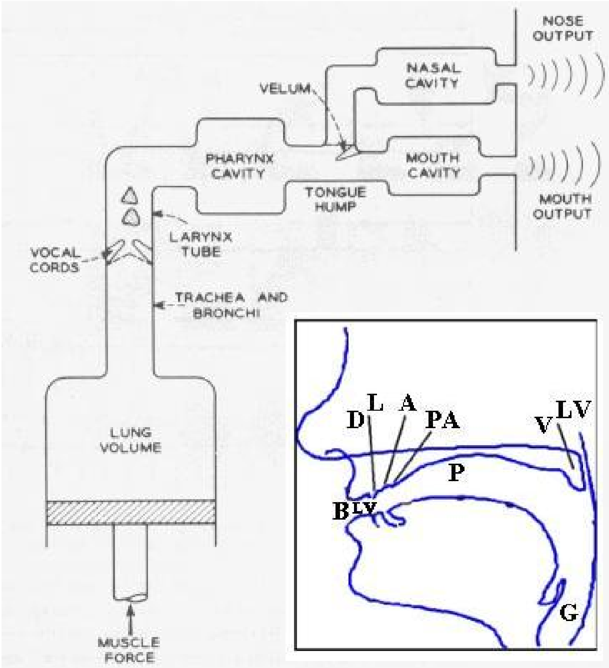
4. Concluding Remarks.

- Summary.
- Acknowledgements.

Motivation – Bimodality of Speech (I)

- **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/.



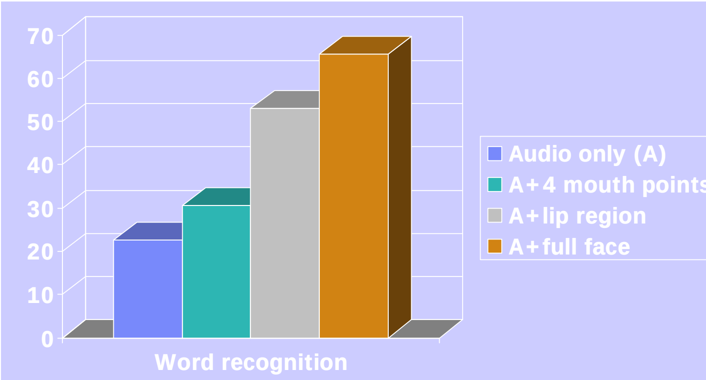


Schematic representation of speech production (J.L. Flanagan, *Speech Analysis, Synthesis, and Perception,* 2nd ed., Springer-Verlag, New York, 1972.)

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

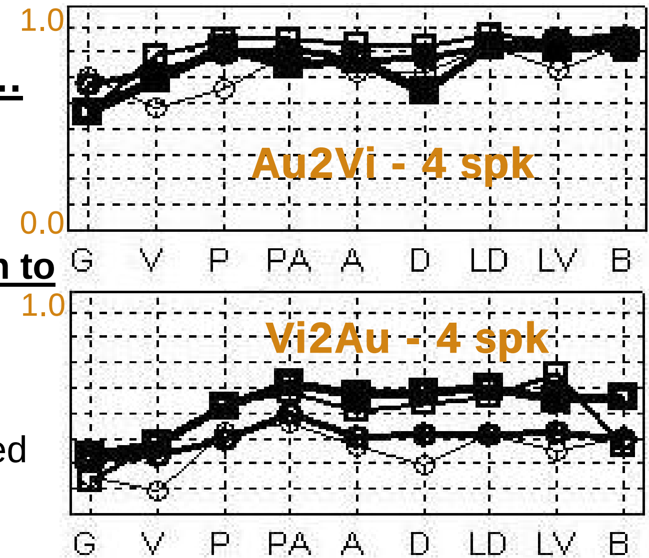
    E.g., human speech perception experiment by Summerfield (1979): Noisy recognition at low SNR.
  - We integrate audio and visual stimuli, as demonstrated by the McGurk effect (McGurk and McDonald, 1976).

    Audio /ba/ + Visual /ga/ → AV /da/
  - **Hearing impaired** people lip-read.

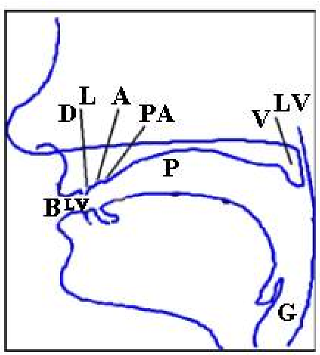


Motivation – Bimodality of Speech (II)

- **Although the visual speech information is less than audio …**
  - **Visemes:** Visually distinguishable classes of phonemes: **6-20**, significantly less than the number of phonemes.
- … 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).



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

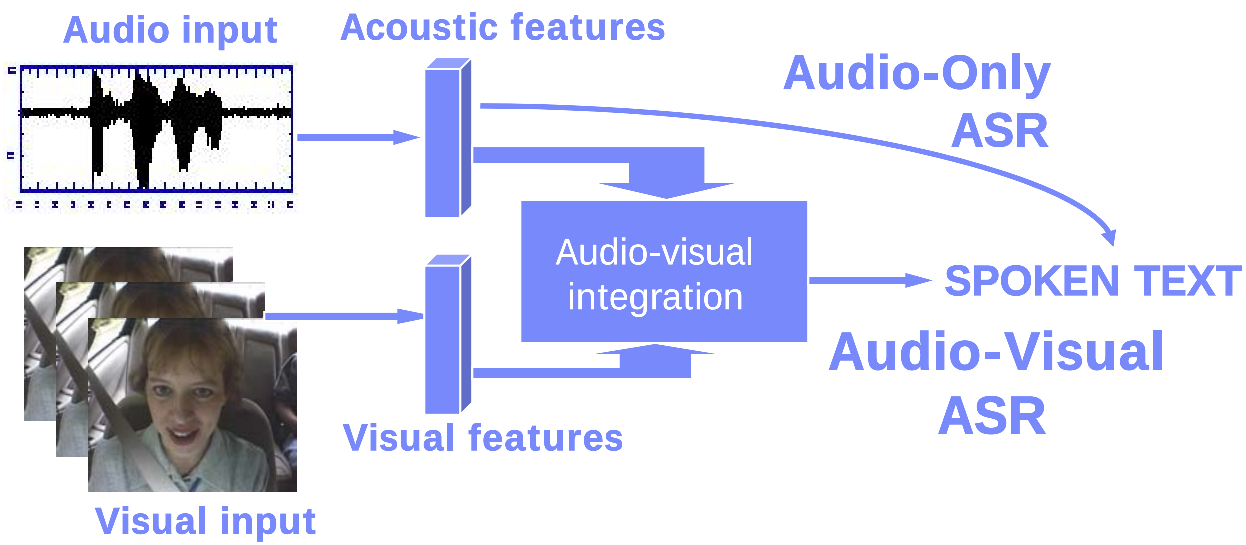
AP	: Approximant	/r, w, y/
LA	: Lateral	/l/
N	: Nasal	/m, n/
PL	: Plosive	/b, d, g, k, p, t/
F	: Fricative	/f, h, s, v, z, θ, ð, ʃ, ʒ/
AF	: Affricate	/tʃ, dʒ/

  - Given the above, and the fact that **noise in the audio and visual channels is in most cases uncorrelated**, this leads to interest in AV speech processing as a means to improve **robustness**.

Audio-Visual Speech Technologies (I)

The following speech technologies can benefit from the visual modality:

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

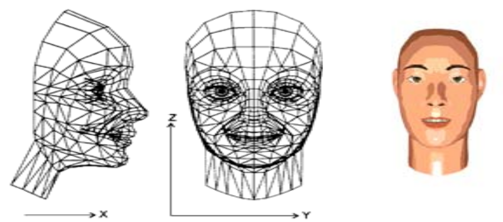


- **Automatic speaker *identification / verification.***

AV Speech Technologies (II)

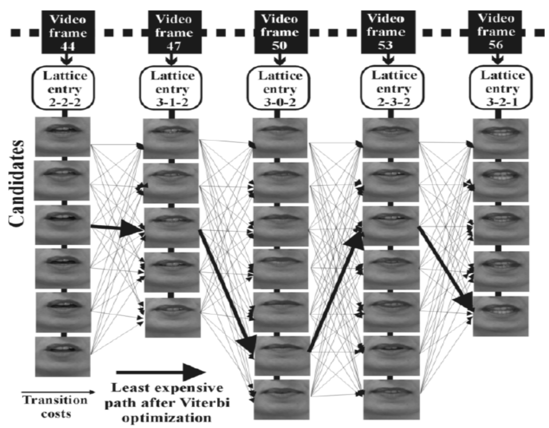
- Speaker localization / speech activity & synchrony detection / speech separation.
- Speech synthesis:

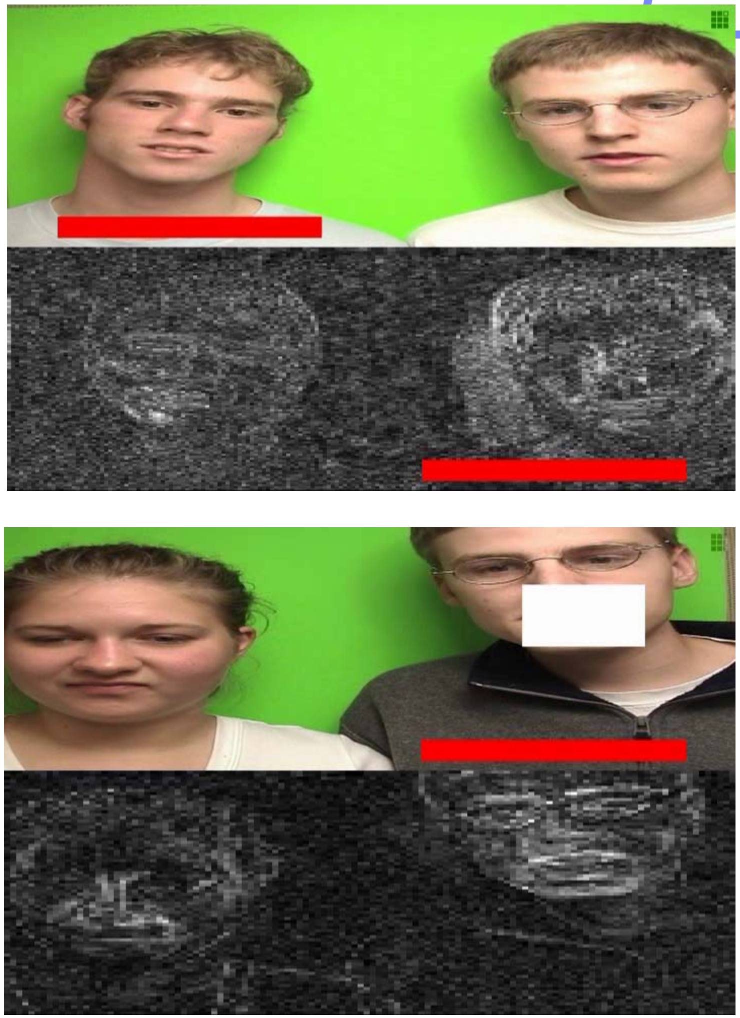
  Model based:



  Sample based:

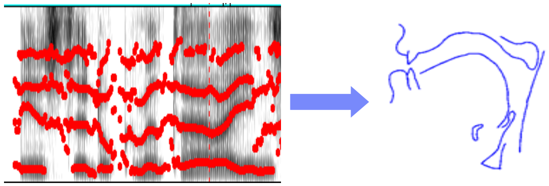
Viterbi search for best mouth sequence (Cosatto et al. 2000).

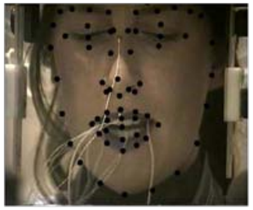




Audio-visual synchrony and tracking (Nock, Iyengar, and Neti, 2000).

- Speech inversion:





Katsamanis et al. 2007

Potential of AV Speech Research & Current State

- Clearly, in scenarios where robustness is an issue and cameras / video is available.
  - **Automobiles.**
  - **Broadcast News.**
  - **Ambient intelligence** environments / **smart rooms**
  - Networks of cameras and microphones in *offices*, *homes*, etc.
  - Advanced **handhelds**.
- Unfortunately, many of these environments represent **significant challenges** to the visual modality as well.
- Coupled with the few resources (data, groups) working on the problem, this has created **significant lag** compared to the progress in acoustic speech processing.
- **Basic approaches** to the problems in the field have followed in the footsteps of traditional acoustic speech research. This has yielded novel algorithms, significant research work, and prototype demo systems.

1. Introduction:

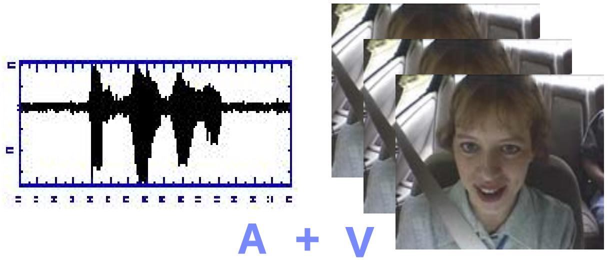
- Motivation.
- Audio-visual speech technologies.
- Potential applications.

2. Audio-visual speech components with emphasis on ASR:

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

3. Other audio-visual speech technologies:

- Speech synchrony.
- Speech enhancement.
- Speech inversion.
- Speaker recognition.
- Speech synthesis.



4. Concluding Remarks.

- Summary.
- Acknowledgements.

National Centre for Scientific Research "Demokritos"

Audio-Visual Databases (I)

- Mostly aiming at **small-vocabulary ASR** tasks.
- Recorded under **ideal AV conditions** – small number of subjects.
- Most commonly used database: **CUAVE**, 36 subjects, connected digits (Paterson et al., 2002).
- Mostly in **English**, but also in Japanese, German, French, …

Audio-Visual Databases (II)

- Large databases have been collected at IBM Research at various environments – both for **LVCSR** and **small-vocabulary** tasks.
  - *Studio, office, automobile, broadcast news, headset* (up to 300 subjects per set).
- Another large database is **AVTIMIT** (MIT):
  - 223 speakers, TIMIT SX sentences.
  - Ideal conditions.

STUDIO

OFFICE

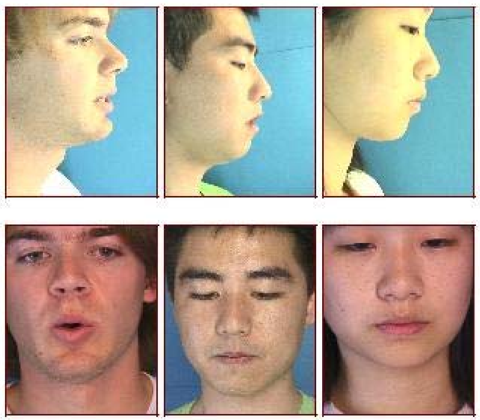
AUTOMOBILE

BROADCAST

- Interesting also multi-sensory databases in the **car environment**:
  - **AVICAR** – 86 subjects (digits, alphas, sentences), 4 channel video recording, 8 channel audio recording.
  - **Aurora 2J, 3J – AV** ➔multiple cameras (infrared channel as well), in-car, Japanese (~100 subjects, Japanese digits).
  - **UTDrive**.

Audio-Visual Databases (III)

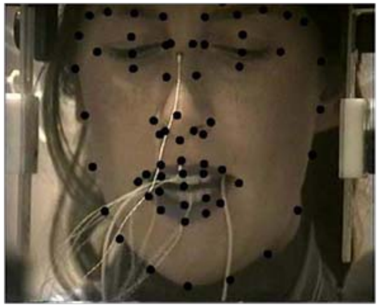
- **Multi-view databases** have been collected by a few groups, e.g. IBM Research, CMU, University of Karlsruhe, etc.



CMU

IBM

- A few databases are also available for some **other tasks** than AVASR, e.g:
  - **XM2VTS**, **VidTIMIT** ➔ speaker recognition.
  - **AVGrid** ➔ speech separation.
  - **MOCHA** ➔ speech inversion.

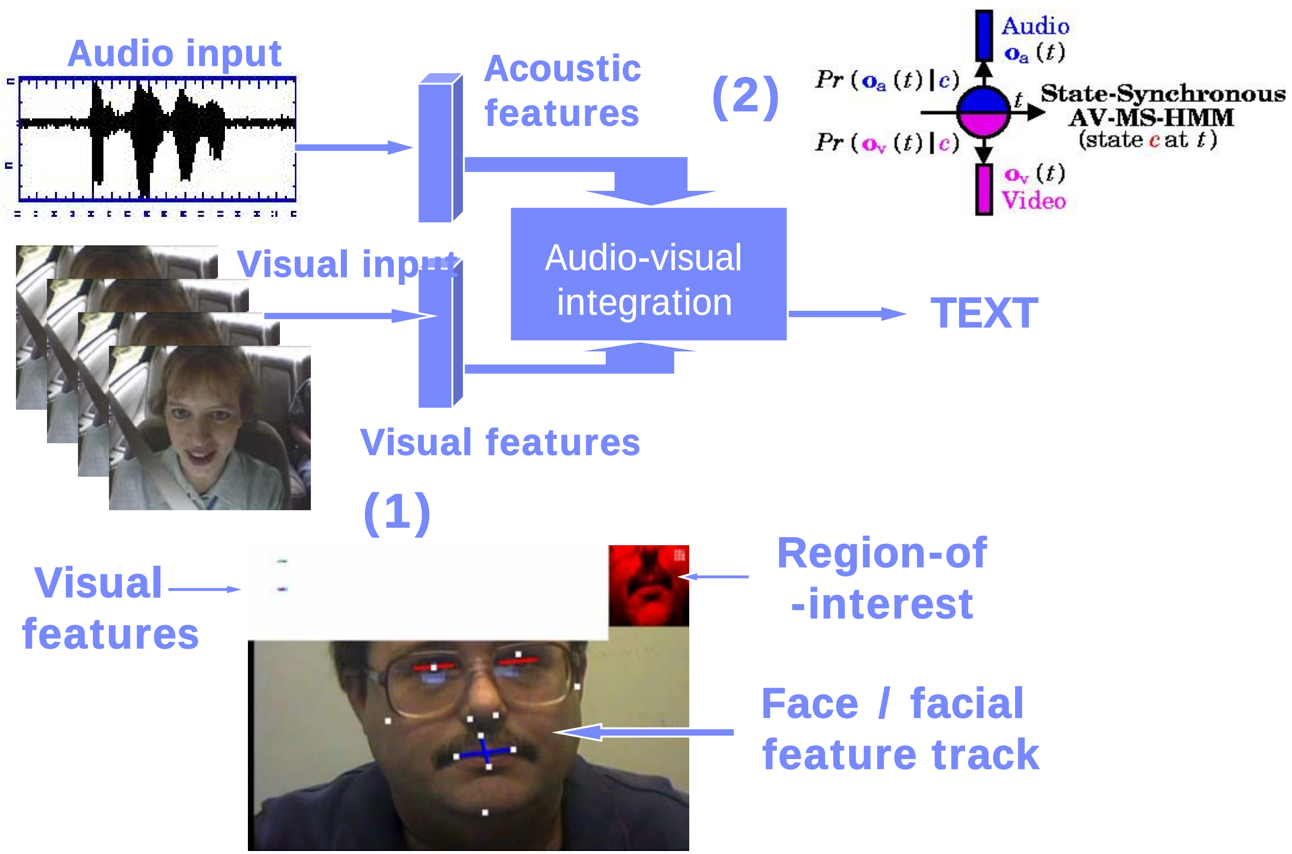


UEdin

National Centre for Scientific Research "Demokritos"

Components of AV Speech Processing

- All AV speech technologies share **two main components**:
  1. **Visual Front End:** Visual channel processing / visual speech representation.
  2. **Fusion:** Audio-visual information “integration” / combination.



AV Speech Processing Components in the Case of ASR

1. Introduction:

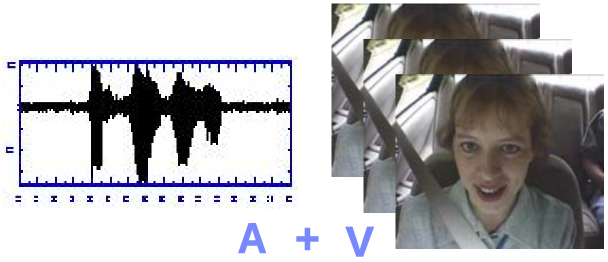
- Motivation.
- Audio-visual speech technologies.
- Potential applications.

2. Audio-visual speech components with emphasis on ASR:

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

3. Other audio-visual speech technologies:

- Speech synchrony.
- Speech enhancement.
- Speech inversion.
- Speaker recognition.
- Speech synthesis.



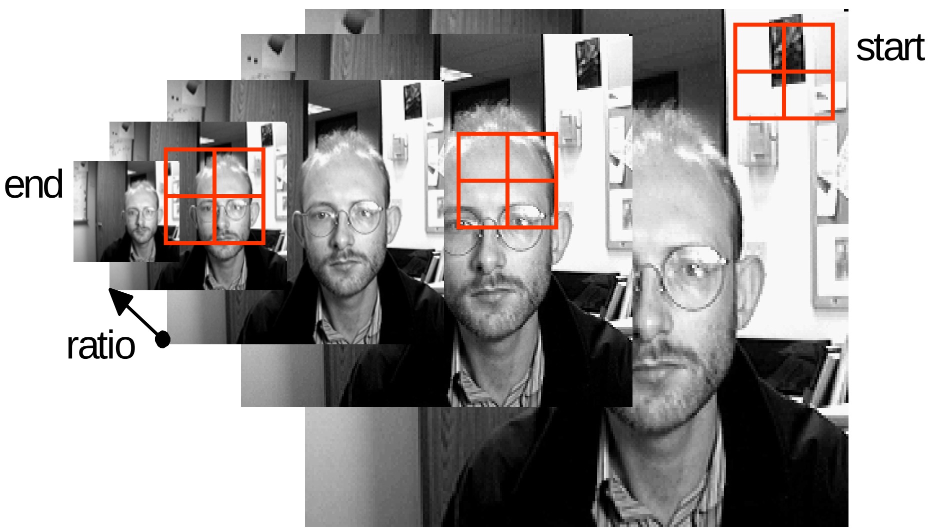
4. Concluding Remarks.

- Summary.
- Acknowledgements.

Face Detection (I)

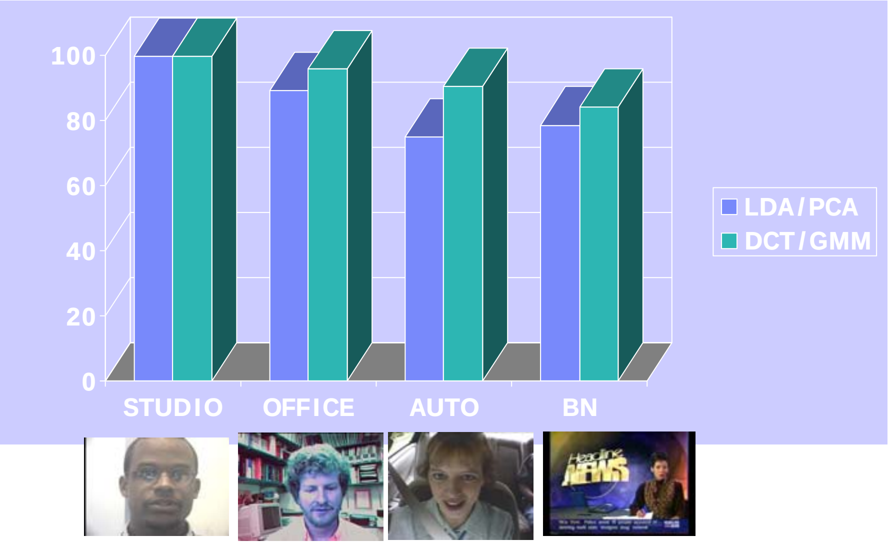
Statistical, appearance based face detection approach, based on "strong classifiers".

- **2-class** classification (into faces / non-faces).
- "**Face template**" (e.g., 11x11 pixel rectangle) ordered into vectors **x** (compressed if desired).
- A **trainable** scheme "scores"/**classifies** **x** into the 2 classes.
- **Pyramidal search** (over locations, scales, orientations) provides face **candidates x**.
- Use your favorite **classifier** (LDA, GMM, NN, SVM, …), favorite **representation** (PCA, DCT), …



Results (in face detection accuracy, %).

More realistic domain → difficulties appear …

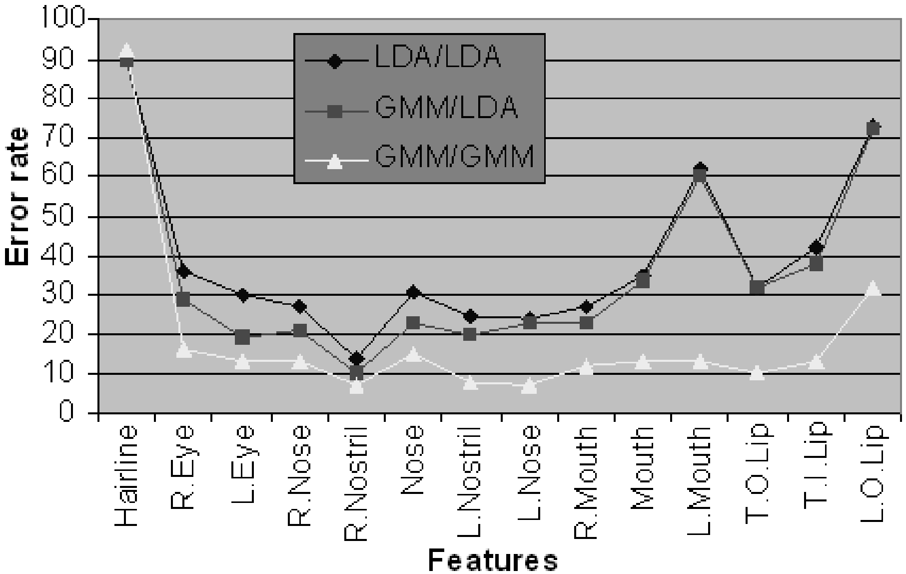


National Centre for Scientific Research "Demokritos"

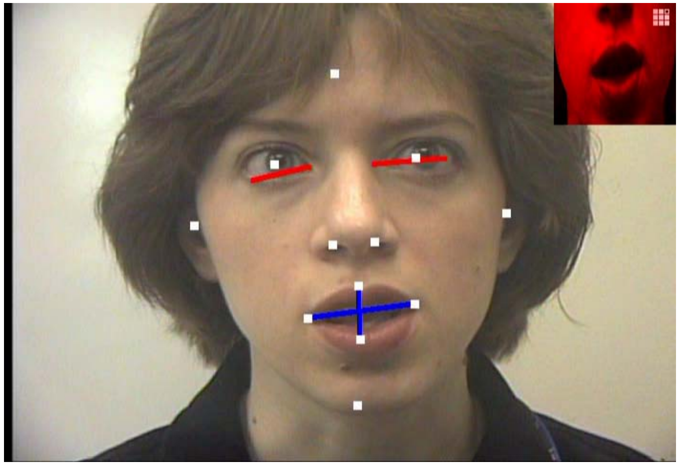
Faces → Facial Features → Region of Interest

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

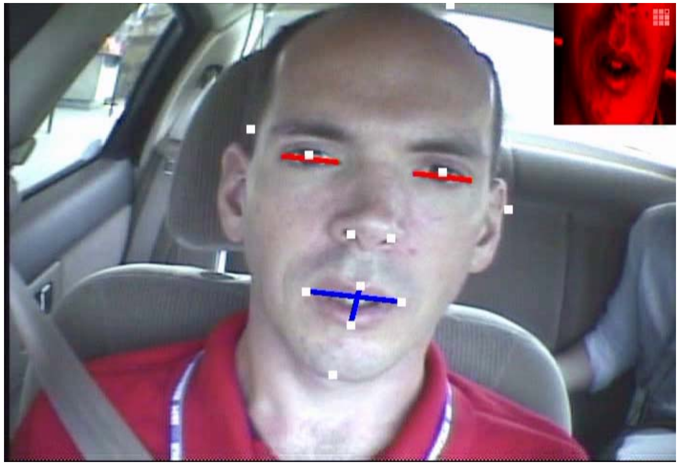
- Similar to face detection. Score *individual* facial feature *templates* by LDA, DFFS, GMMs, etc.



Facial-feature extraction performance



STUDIO



AUTOMOBILE

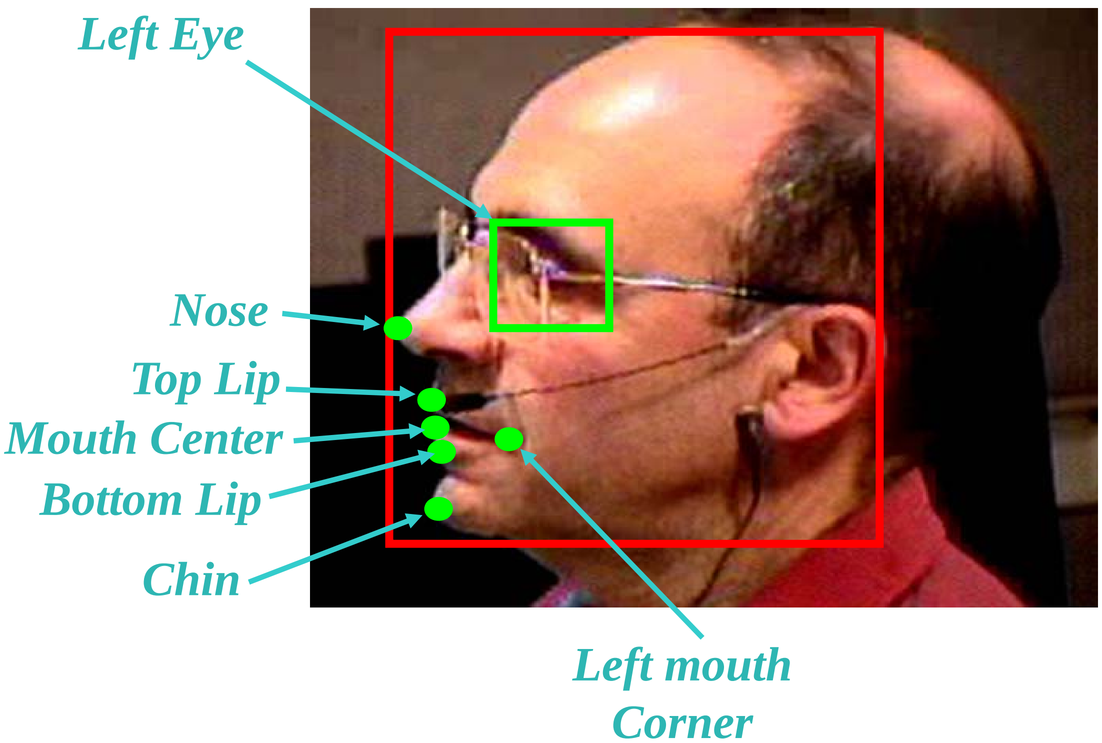
Region-of-interest (ROI):

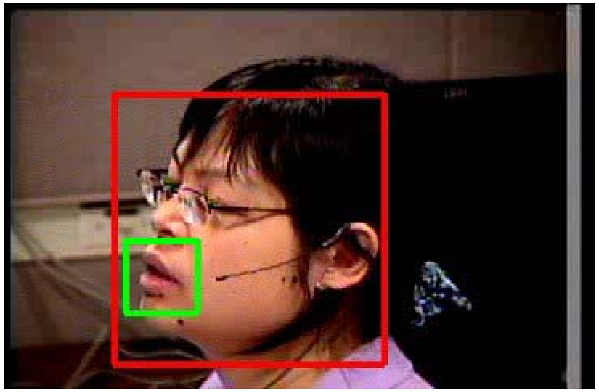
- Assumed to contain “all” visual speech information.
- Typically, a rectangle containing mouth + lower face.
- Appropriately normalized.

Face Detection – ROI Extraction (II)

- … or use cascade of weak classifiers (AdaBoost):
  - Face detection (red box).
  - Seven facial features (green).
  - ROI extraction is based on 3 most reliable facial features.

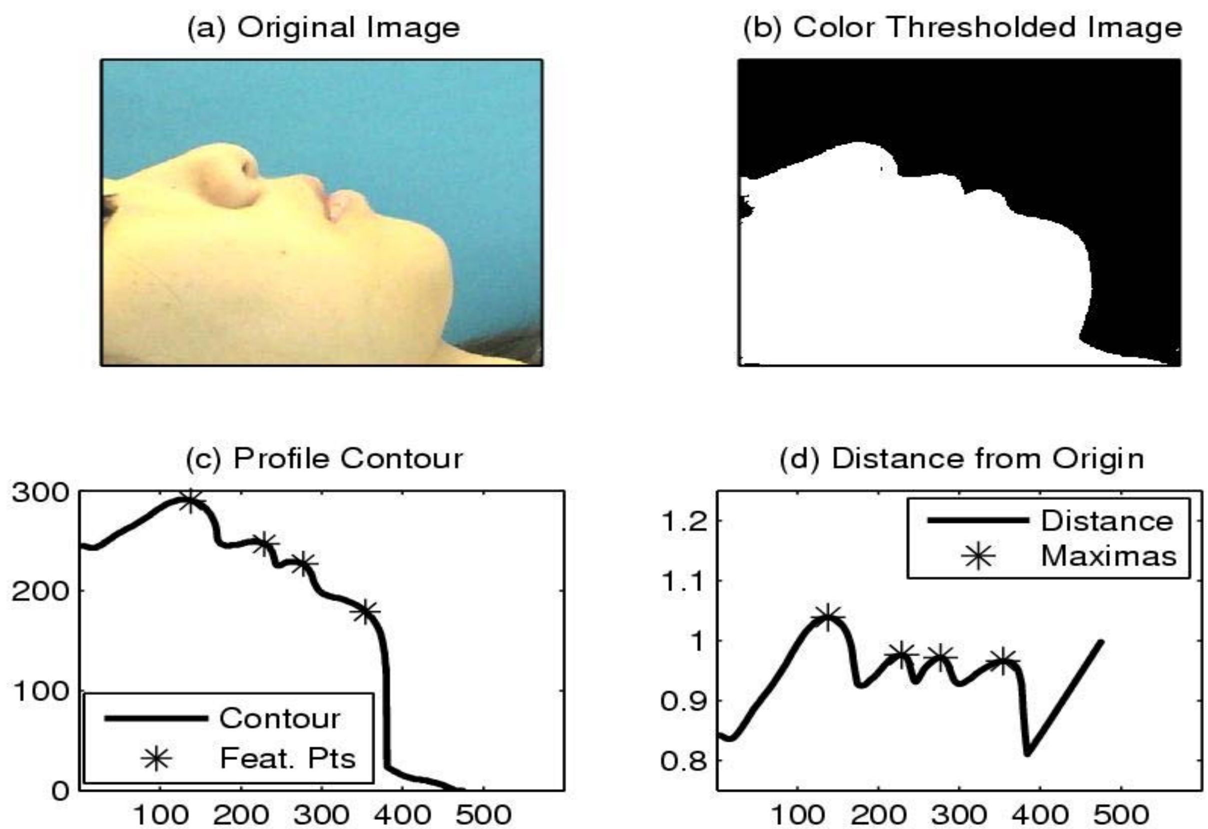
Facial Feature	Acc. (%)
Left Eye	87%
Nose	81%
Top Mouth	79%
Center Mouth	81%
Lower Mouth	73%
Left Mouth	87%
Chin	63%





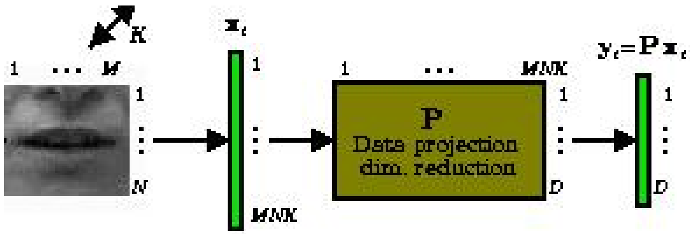
Face Detection – ROI Extraction (III)

- … or use image processing techniques such as:
  - Motion estimation.
  - Color processing.
  - Image segmentation.
  - Face geometry heuristics.



Example from Kumar et al., 2007

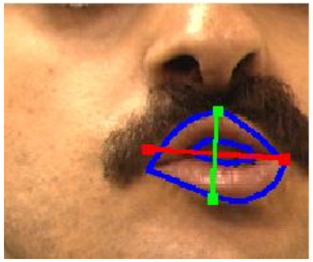
Region-of-Interest → Visual Features

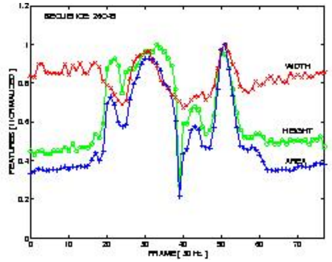


Three types of approaches to feature extraction:

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.



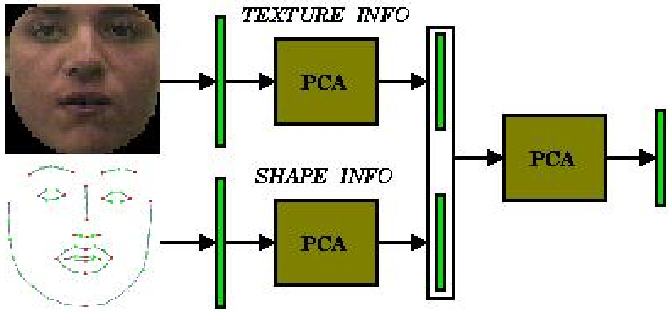


Lip- and face-contour (shape) based:

- Height, width, area of mouth.
- Moments, Fourier descriptors.
- Model based (statistical or geometrical).

Joint shape and appearance features:

- Active appearance models.



Extraction is typically followed by feature post-processing:

- Intra-frame + inter frame **LDA/MLLT** for better within and across frame discrimination.
- … or inclusion of first and second order **derivatives**.
- Feature **normalization** (FMN).
- **Up-sampling** for synchronization to audio feature extraction rate (25, 30 → 100 Hz).

Appearance Based Feature Selection

- Among appearance-type visual features, DCT coefficients are typically used – for example extracted from *64 x 64* pixel ROI.
- This gives rise to large number of features. How to ***select the appropriate*** ones?

Appoaches:

– ***Energy based*** → Select high energy coefficients (baseline approach).

– ***LDA*** → high input dimensionality, stability problems.

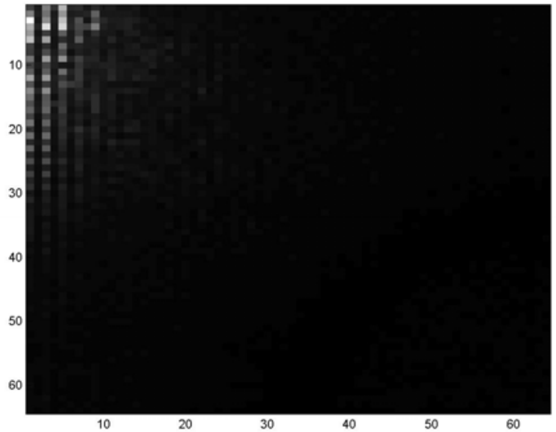
– ***Variance*** → somewhat worse performance than energy based schemes.

– ***Mutual information (MI)*** → promising scheme, but computational problems.

  – Select DCT features x that ***maximize MI wrt speech classes c***.

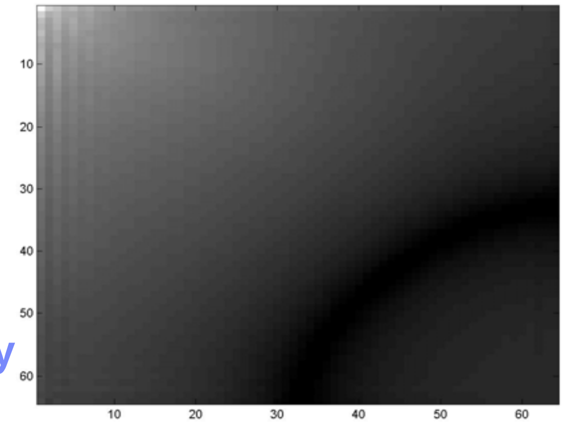
– Disregard even-column features (due to mouth ***symmetry***) and use above schemes.

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



MI

Energy

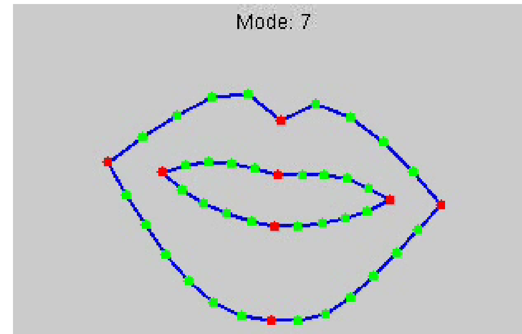


Visual Features – Shape Based Approach

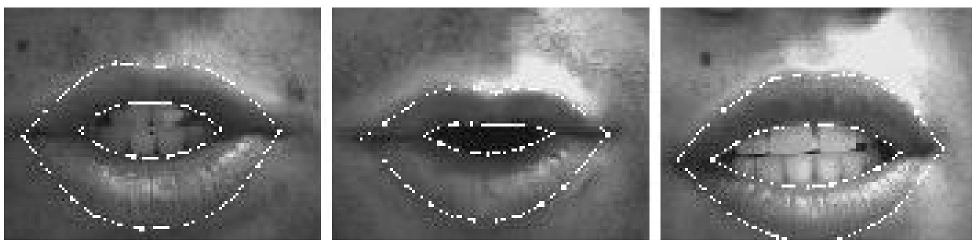
Shape based features represent speech information using lip contour information.

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.



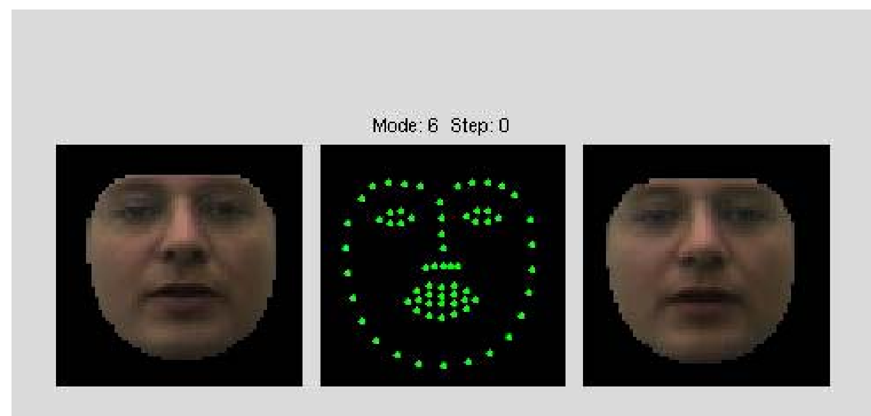
ASM modes trained on IBM data



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

Feature Comparisons

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

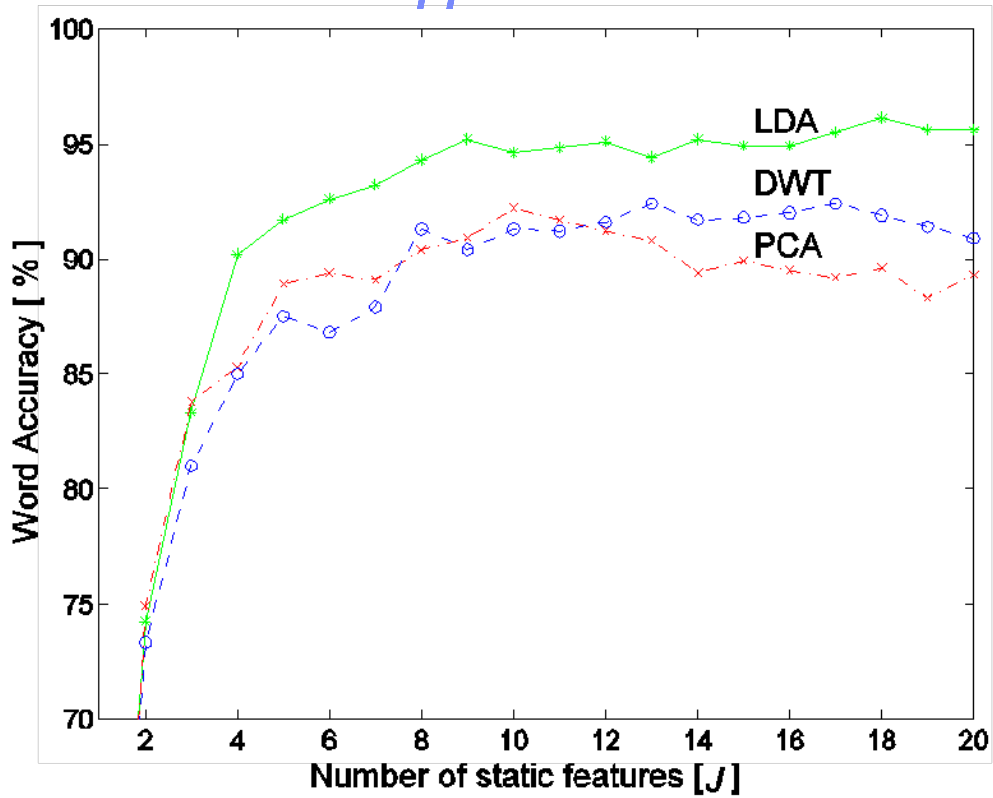
- ***Appearance-*** are better than ***shape-***based features:

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**

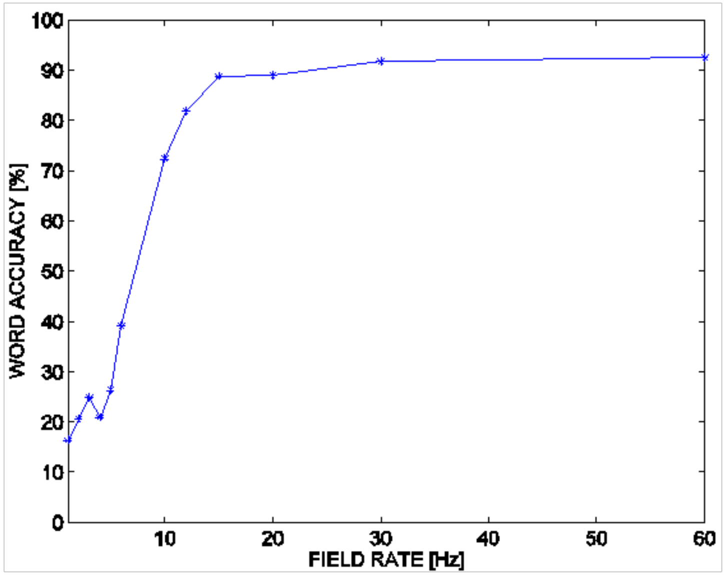
Feature type	%, Word accuracy
Lip-contour based	**83.9**
Appearance (LDA)	**97.0**

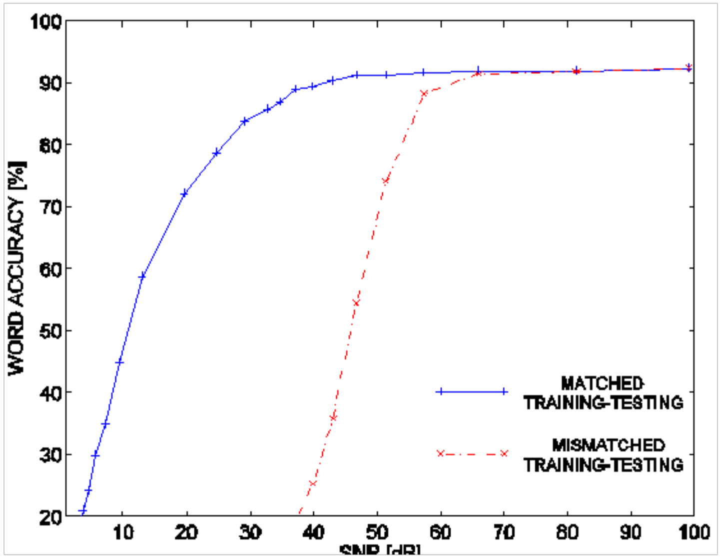
- Comparisons of various ***appearance***-based features

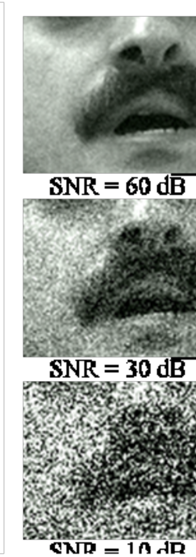


Challenges in Non-Ideal Data

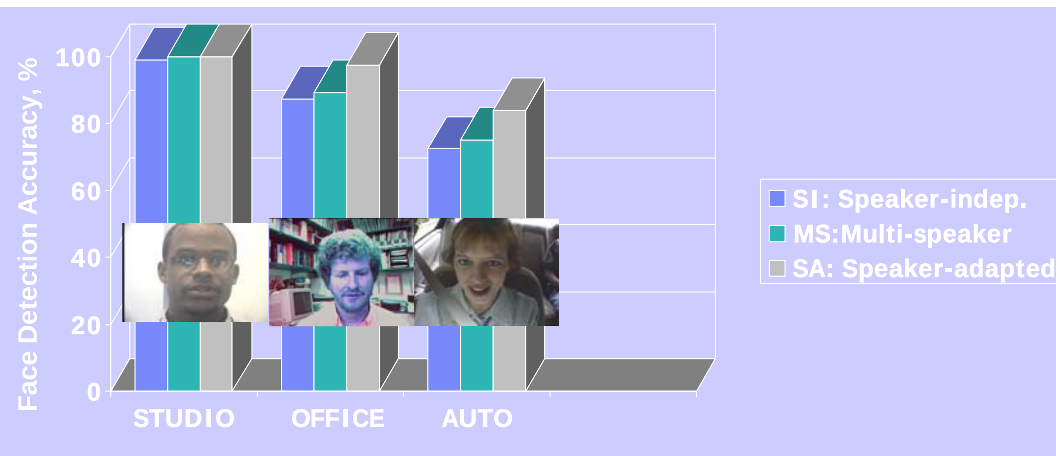
- **Frame rate decimation:** Limit of acceptable video rate for automatic speechreading is **15 Hz.**
- **Video noise:** Robustness to noise only in a matched training/testing scenario.

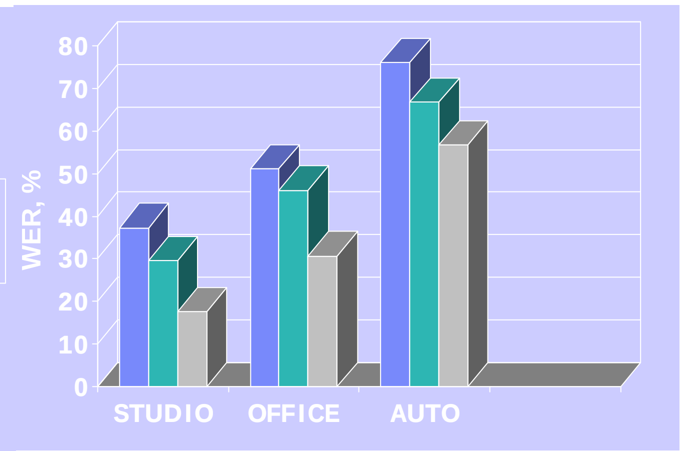






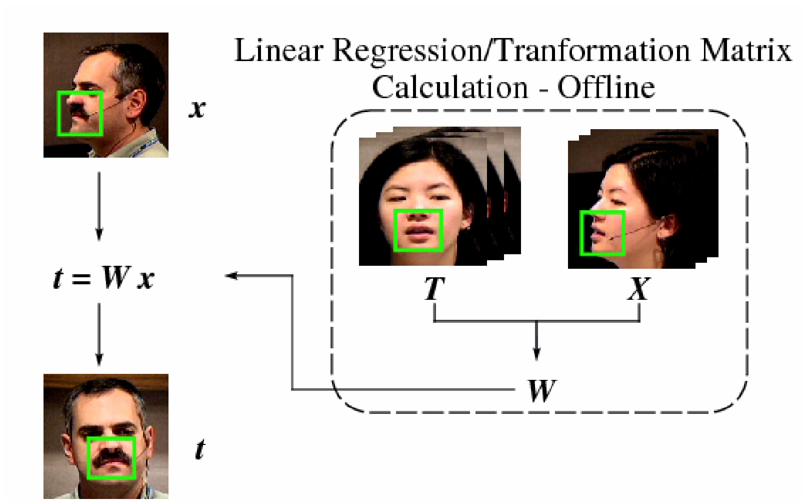
- **Challenging visual domains:** Face detection accuracy decreases → Word error rate increases.

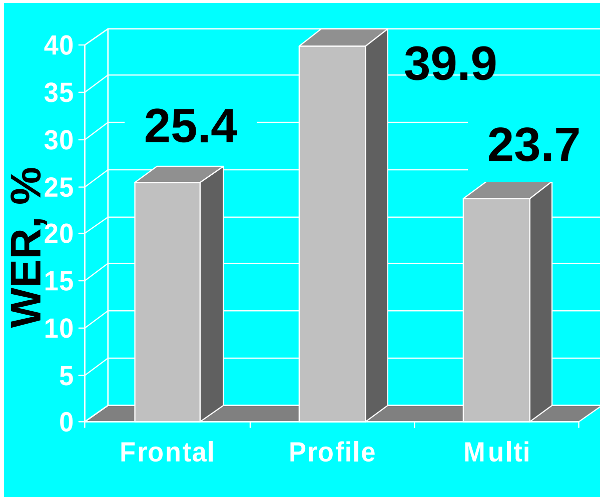




Challenges due to Environment Variation

- It should be noted that appearance based features are susceptible to overtraining to particulars of environment, speaker, pose, etc…
- Robustness to such variability is an issue.
- Similar in nature to problems in speech…
- One example is **head-pose variation**. How to go about statistical modeling?
  - Use ***pose-specific*** visual speech *models*.
  - Throw all pose data into the same "cooking pot" – "***single speech model fits all***".
  - Do this, but at some "***pose-normalized***" space.
- For the latter, one can estimate a **linear regression** matrix, ***W***, from *undesirable* pose-space ***X*** (profile) to *desirable* pose-space ***T*** (frontal).





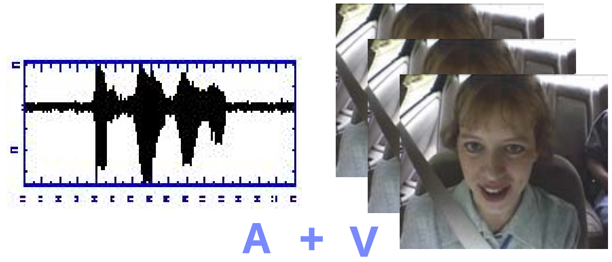
Difference in ASR performance between frontal and profile views.

1. Introduction:
- Motivation.
- Audio-visual speech technologies.
- Potential applications.

2. Audio-visual speech components with emphasis on ASR:
- Data resources.
- Visual feature representation for speech applications.
- **Audio-visual combination (fusion).**

3. Other audio-visual speech technologies:
- Speech synchrony.
- Speech enhancement.
- Speech inversion.
- Speaker recognition.
- Speech synthesis.

4. Concluding Remarks.
- Summary.
- Acknowledgements.



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

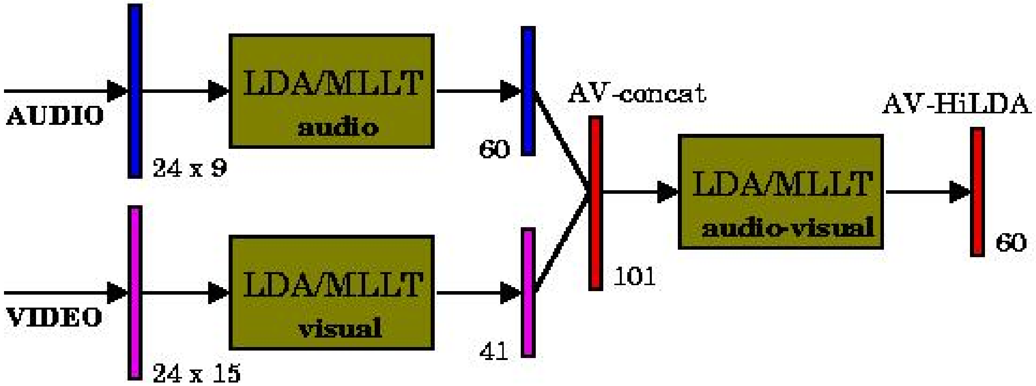
- **Main points in audio-visual fusion for ASR:**
  - Type of fusion:
    - ✓ Combine audio and visual info at the feature level (**feature fusion**).
    - ✓ Combine audio and visual classifier scores (**decision fusion**).
    - ✓ Could envision a combination of both approaches (**hybrid fusion**).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

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.**



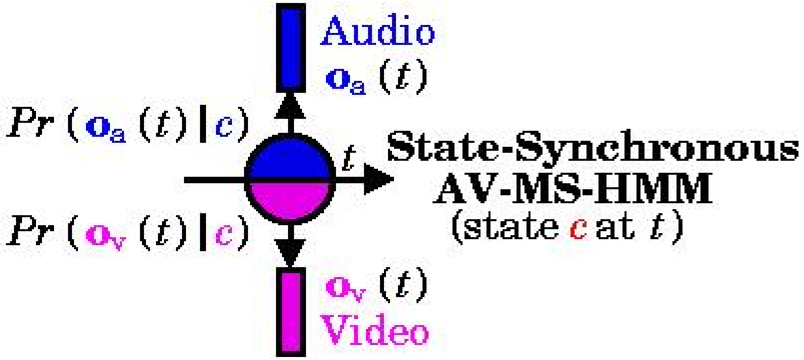
Decision Fusion (I)

- **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**:

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

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



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

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

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

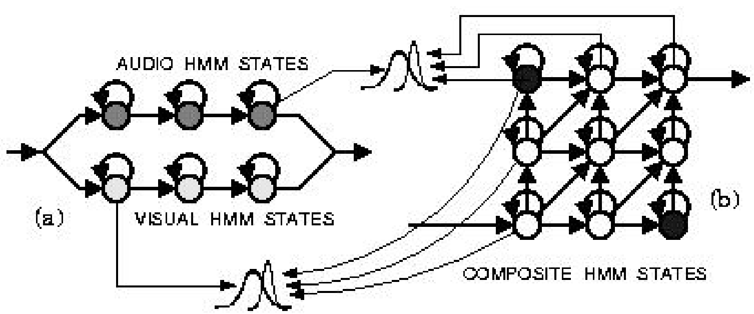
Intermediate Integration

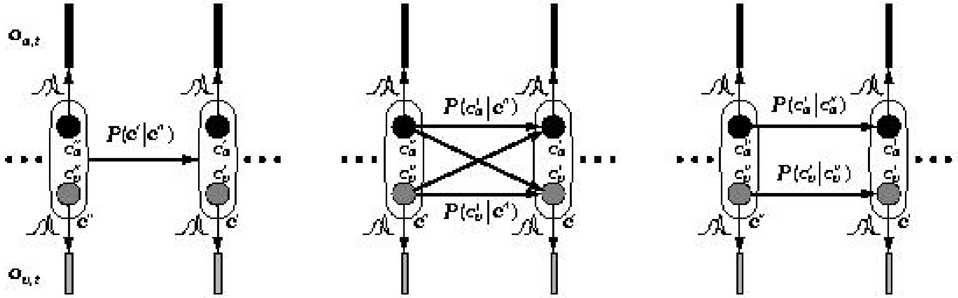
- **Intermediate integration** combines stream scores at a **coarser** unit level than HMM states, such as **phones**. This allows state-asynchrony within each phone.

- Integration model is equivalent to the **product HMM** (Varga and Moore, 1990).
  - Product HMM has "**composite**" (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*|x|*C*| for a 2-stream model.
  - Class-conditional observation probabilities can follow the MS-HMM paradigm, i.e.:

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

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





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: $[t_{i,j,s}^{\text{start}}, t_{i,j,s}^{\text{end}}]$
  - Then rescore:

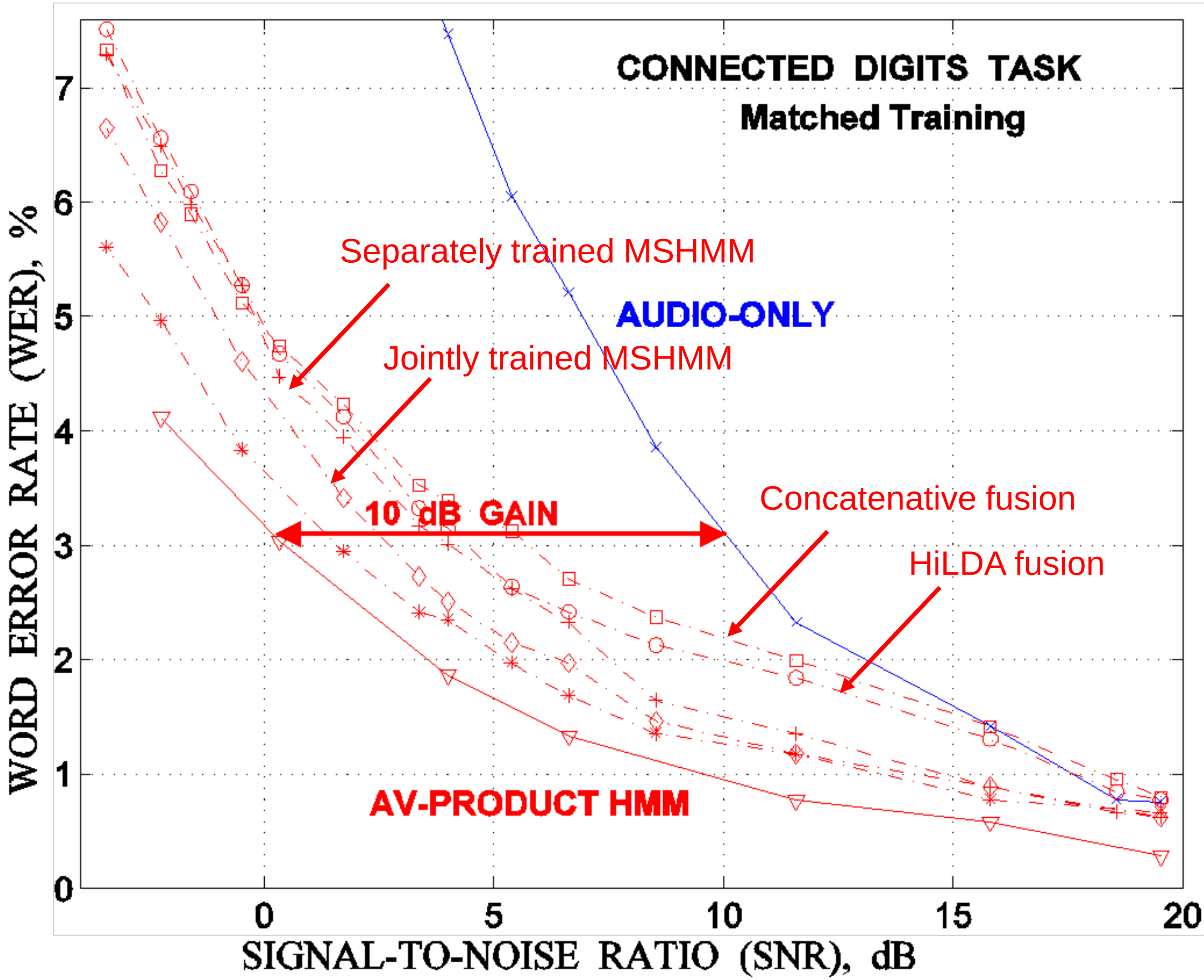
$$\Pr[\mathbf{h}_i] \propto \Pr_{\text{LM}}(\mathbf{h}_i)^{\lambda_{\text{LM}}} \prod_{s \in S} \prod_{j=1}^{N_i} \Pr(\mathbf{o}_{s,t}, t \in [t_{i,j,s}^{\text{start}}, t_{i,j,s}^{\text{end}}] \mid c_{i,j})^{\lambda_{s,c_{i,j}}}$$

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

AVASR: Fusion Results (I)

- 50-subjects, **connected-digits** database in ideal environment.
- **Product HMM fusion** is superior to state-synchronous fusion.
- Effective SNR gain: **10 dB** SNR.

[Potamianos et al., 2003]



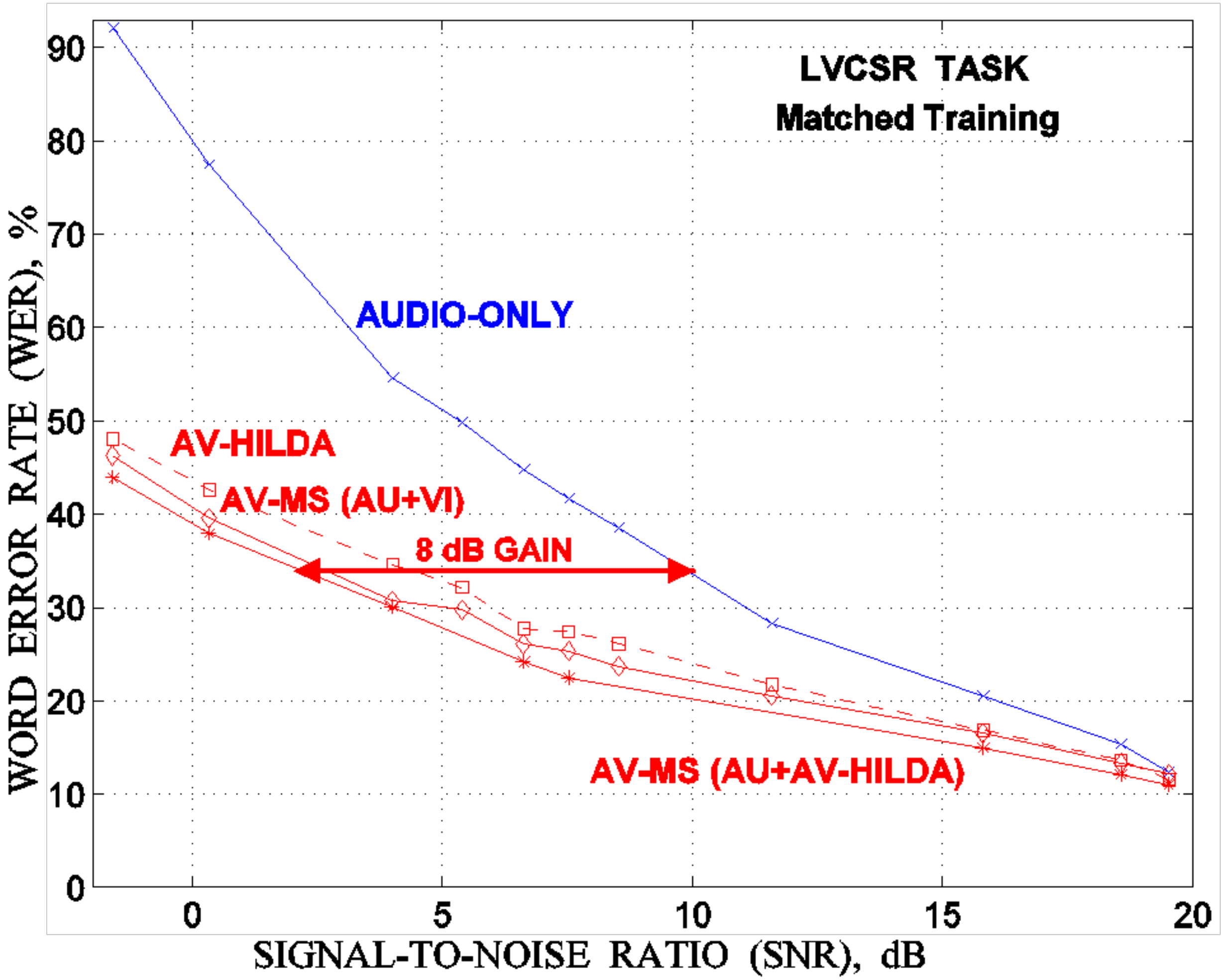
AVASR: Fusion Results (II)

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.

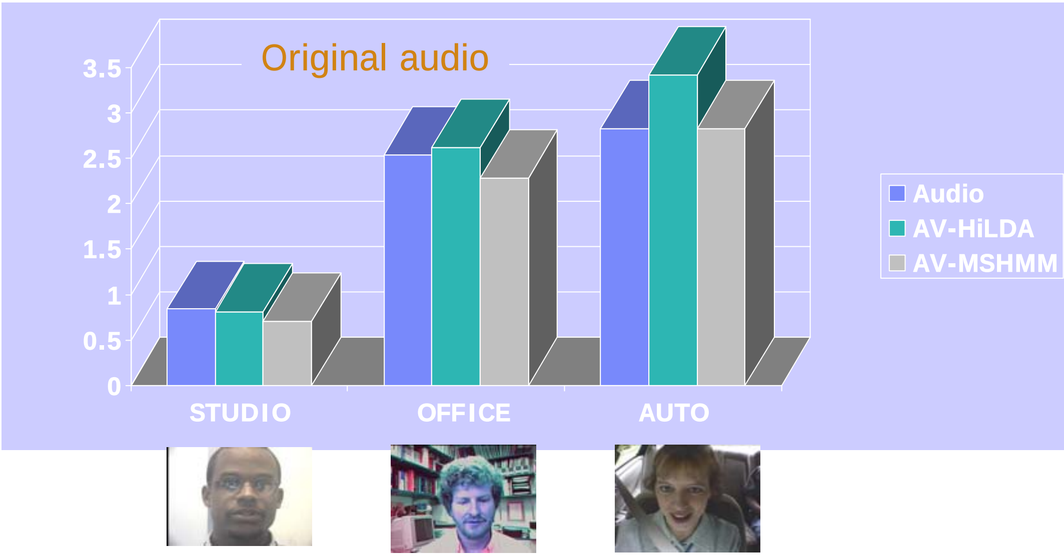
[Potamianos et al., 2003]

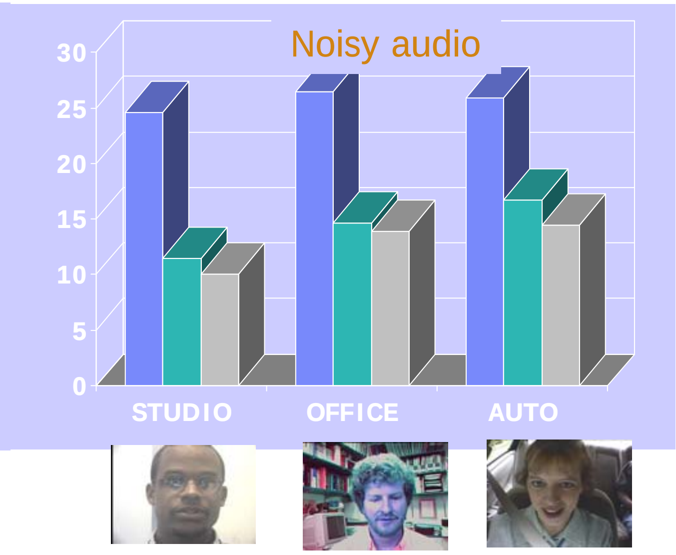


AVASR Results

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, %.





[Potamianos et al. 2003]

Stream Reliability Modeling for Fusion

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

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

- 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.

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

Fusion – Global Stream Weighting

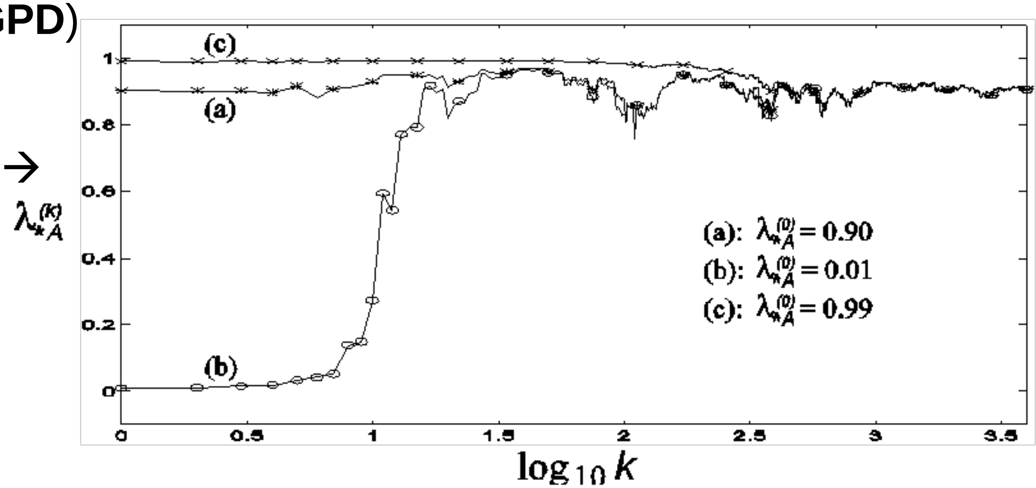
- Stream weights **cannot** be obtained by **maximum-likelihood** estimation, as:

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

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)

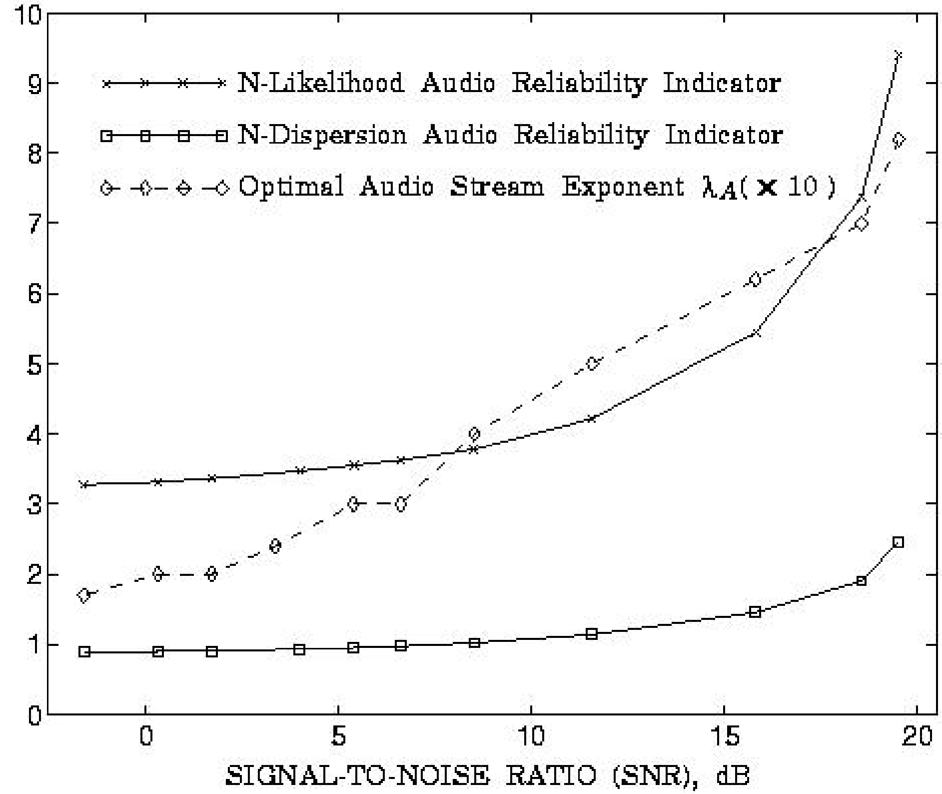


Fusion – Adaptive Stream Weighting

- 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:

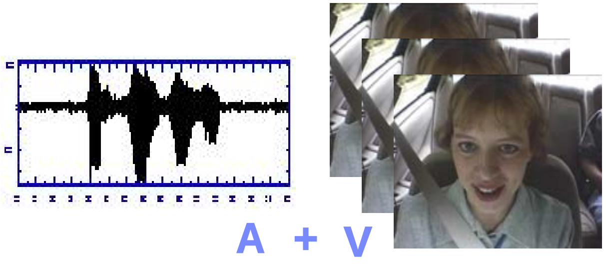
$$\lambda_{A,t} = [1 + \exp(-\sum_{i=1}^{4} w_i \, d_i)]^{-1}$$

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

1. Introduction:
- Motivation.
- Audio-visual speech technologies.
- Potential applications.

2. Audio-visual speech components with emphasis on ASR:
- Data resources.
- Visual feature representation for speech applications.
- Audio-visual combination (fusion).

3. Other audio-visual speech technologies:
- **Speech synchrony.**
- **Speech enhancement.**
- **Speech inversion.**
- **Speaker recognition.**
- **Speech synthesis.**



4. Concluding Remarks.
- Summary.
- Acknowledgements.

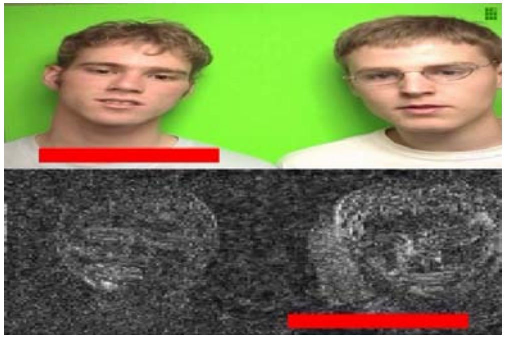
Additional Audio-Visual Speech Technologies

- 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 ***synchrony*** detection.
  - Speech ***enhancement***.
  - Speech ***inversion.***
  - Speaker ***identification / verification***.
  - Speech ***activity detection***.
  - Speech ***synthesis***.

Audio Visual Synchrony Detection (I)

- **Goal** is to detect if audio and visual sources are in sync.

- **Applications:**
  - Biometrics – *spoofing detection*.
  - Improve *speaker diarization*.
  - Speech *source localization*.



- Typical **approaches** in literature employ:
  - **Mutual information** between audio & visual features (Hershey & Movellan, 2000).

$$I(A;V) = \mathbf{E} \log \frac{p(a,v)}{p(a),p(v)} \gtrless \lambda$$

  - **Hypothesis testing**

    Construct two classes:

    - $\mathcal{H}_1$, AV features (***Z***) in sync.
    - $\mathcal{H}_0$, AV features (***Z***) out of sync.

    Log-likelihood Ratio Test (LRT): $LLR = \log \frac{p(Z;\mathcal{H}_1)}{p(Z;\mathcal{H}_0)} \gtrless \lambda$

- Concise **overview** in: Rua et al. 2009; Bredin & Chollet 2007.

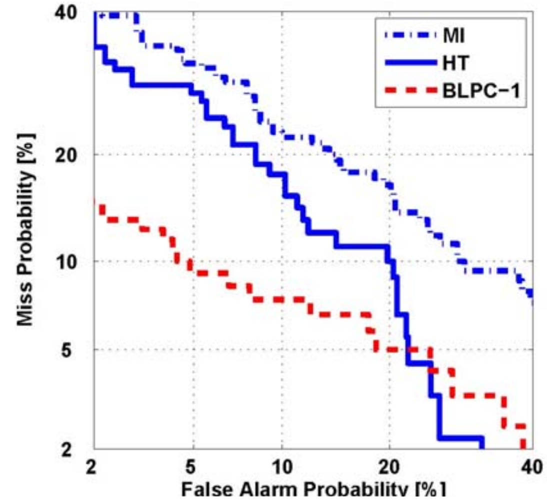
Audio Visual Synchrony Detection (II)

- Above approaches consider AV features to be *statistically independent*.
- An alternative approach has been suggested by *Kumar et al., 2009*, termed **bimodal linear prediction coefficient** (BLPC) approach.
  - Captures the auto-correlation and cross-correlation through meaningful parameters.
  - Jointly models feature evolution in time.
- **Three models** considered:
  - **BLPC-1**: $$a[n] \approx \hat{a}[n] = \sum_{i=1}^{N_a} \alpha[i]a[n-i] + \sum_{j=0}^{N_v} \beta[j]v[n-j]$$
  - **BLPC-2**: $$a[n] \approx \hat{a}[n] = \sum_{i=1}^{N_a} \alpha[i]a[n-i] + \sum_{j=-N_v}^{N_v} \beta[j]v[n-j]$$
  - **BLPC-3**: $$a[n] \approx \hat{a}[n] = \sum_{i=-N_a, i\neq 0}^{N_a} \alpha[i]a[n-i] + \sum_{j=-N_v}^{N_v} \beta[j]v[n-j]$$
- If AV **in sync**, then: $\beta[j] \neq 0, \ \forall\, j$

  if **not**, then: $\beta[j] = 0, \ \forall\, j$
- Coefficients are computed by **MMSE**.
- Method applied on **AV feature pairs** obtained after **canonical correlation analysis** (CCA).

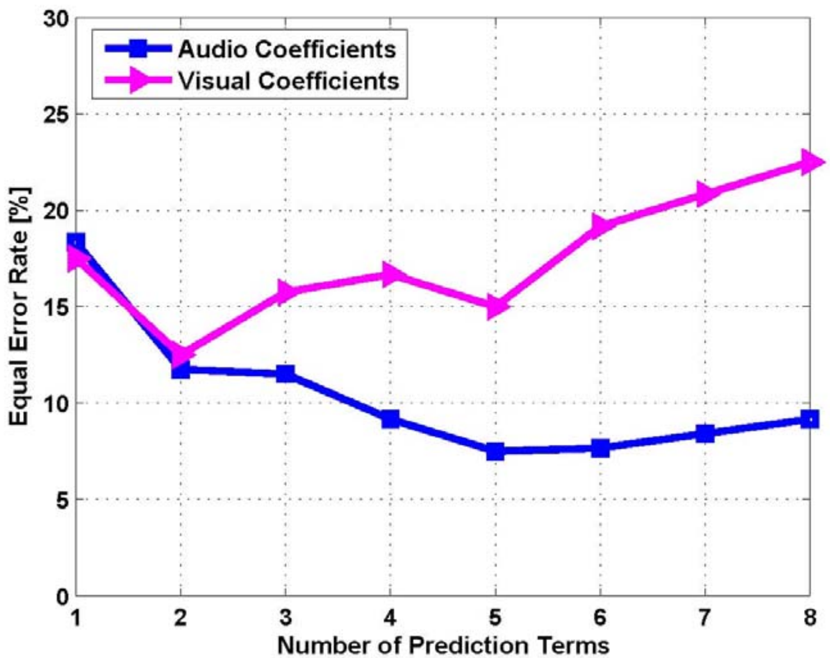
Audio Visual Synchrony Detection (III)

- Some **AV synchrony** detection results on CMU data:

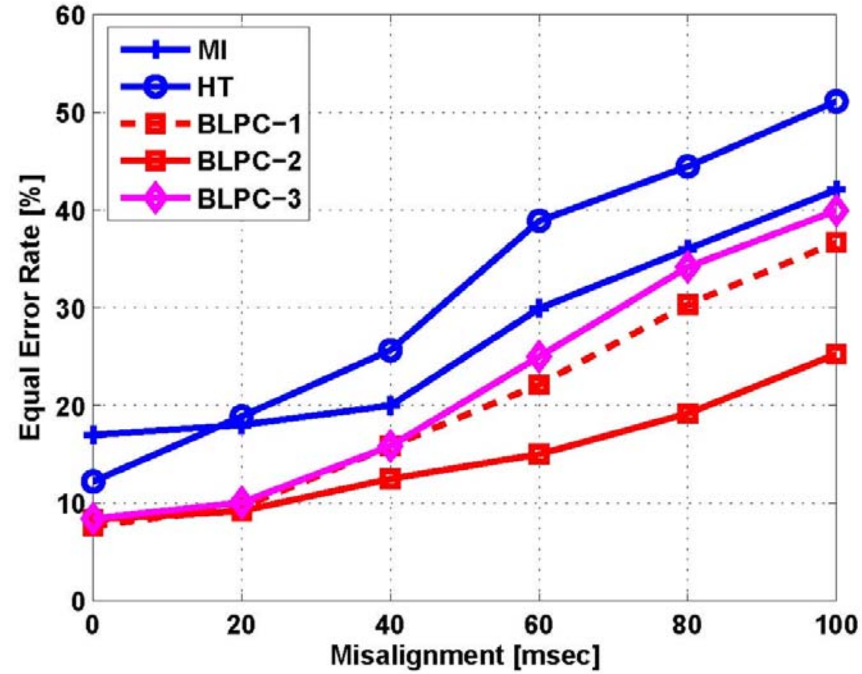


DET curve of various synchrony detection approaches

EER based on audio vs. visual coefficient distance.



Robustness to stream misalignment.



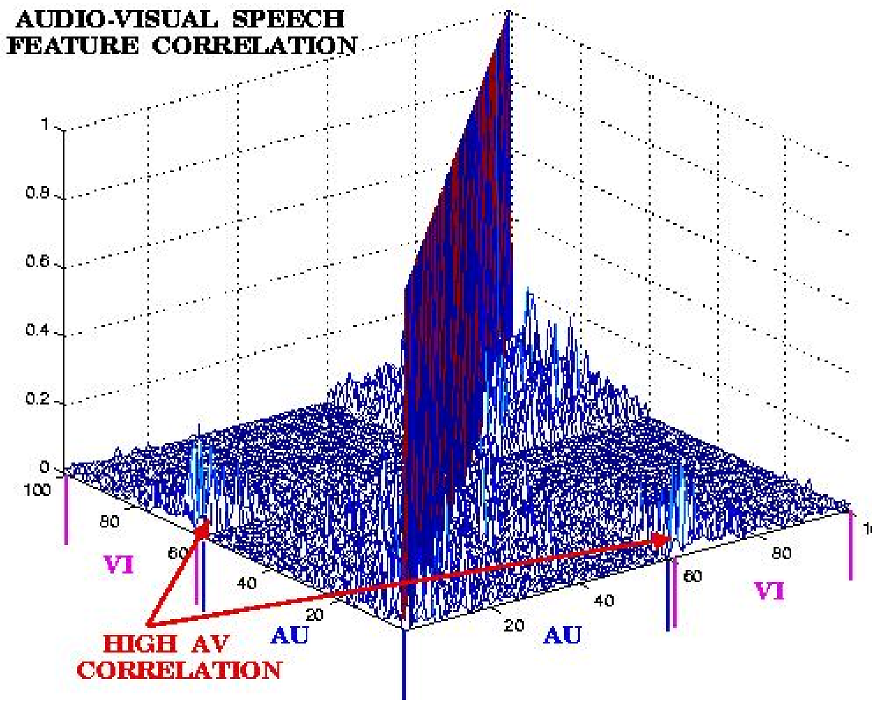
Audio-Visual Speech Enhancement – Overview

- **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.

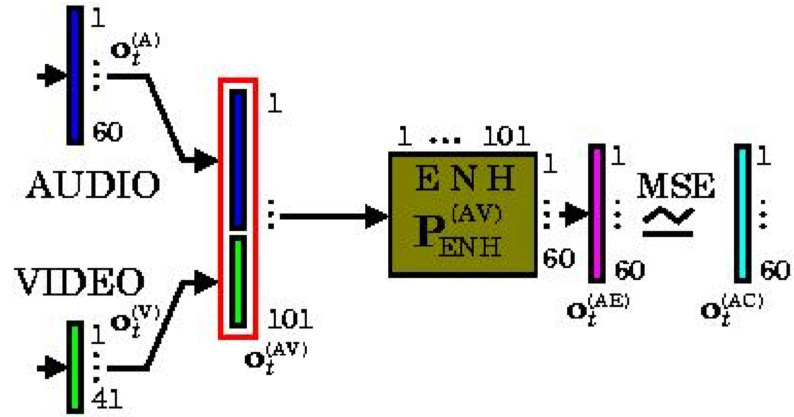
- **Enhancement** can occur in the:
  - **Signal** space (based on **LPC** audio feats.).
  - Audio **feature** space (discussed here).

- **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

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

  - ✓ 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):

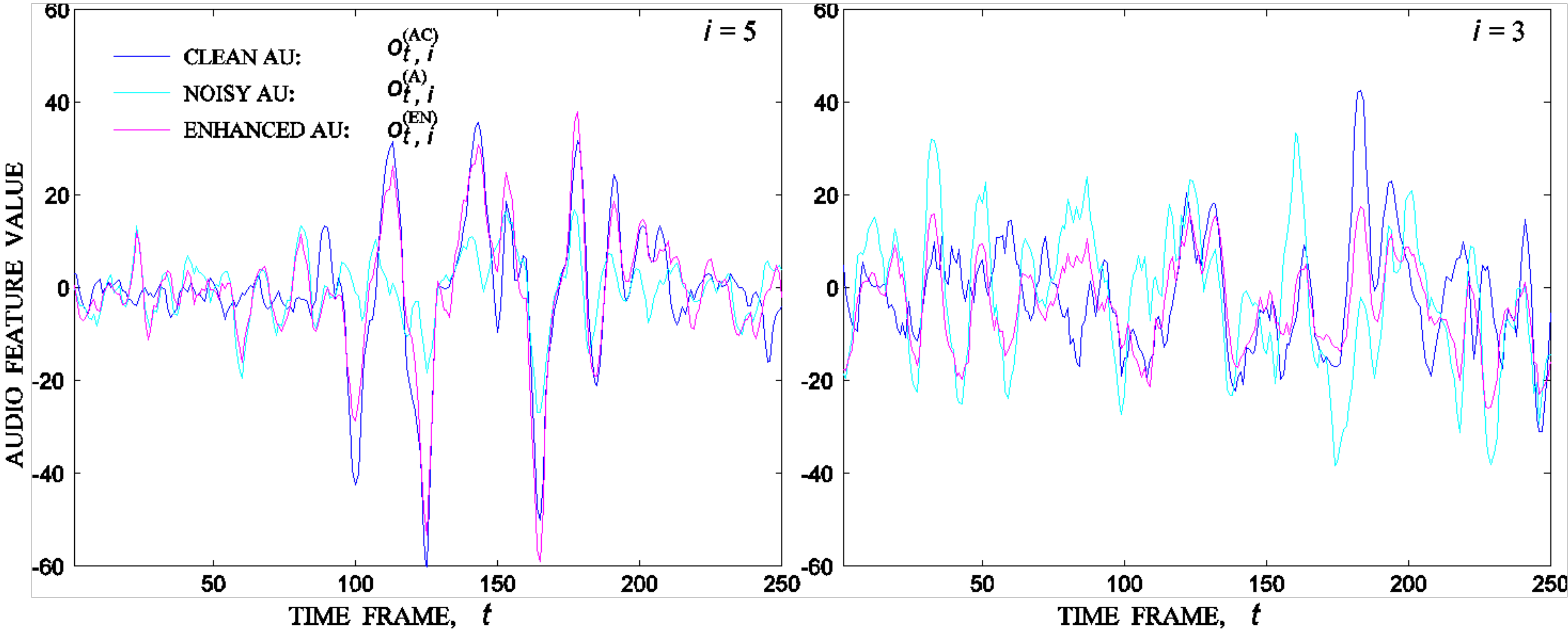
    $$\mathbf{p}_i = \arg\max_{\mathbf{p}} \sum_{t \in T} [\, o_{A,t,i}^{(C)} - < \mathbf{p}, \mathbf{o}_{AV,t} >]^2,\quad i = 1,\ldots,d_A$$

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

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

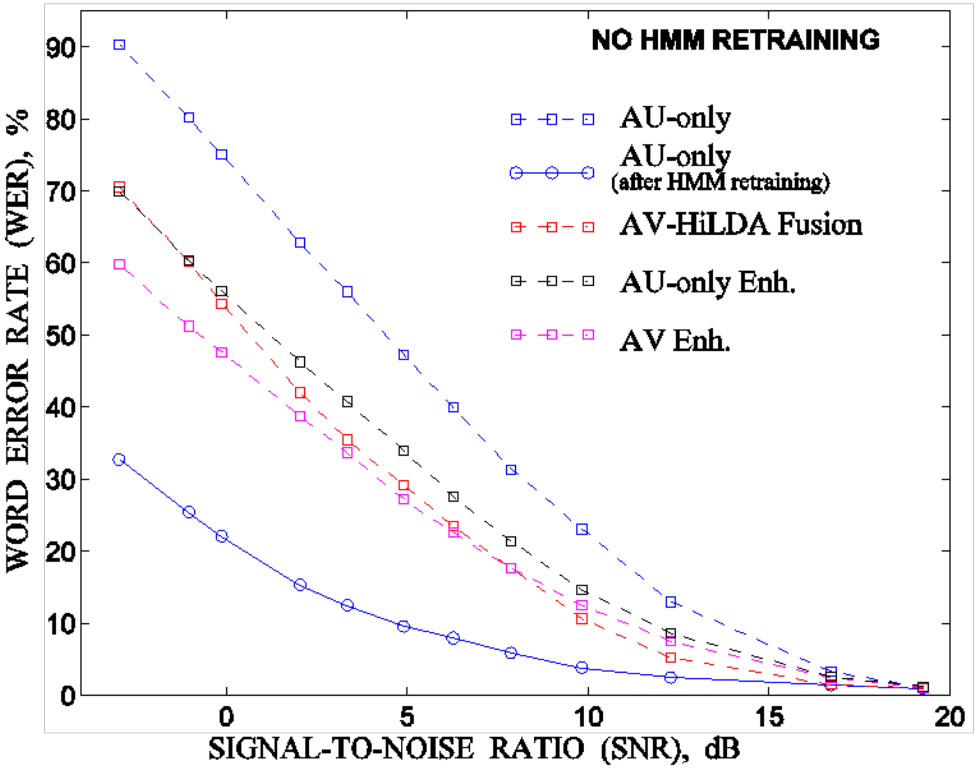
Linear Bimodal Enhancement of Audio (II)

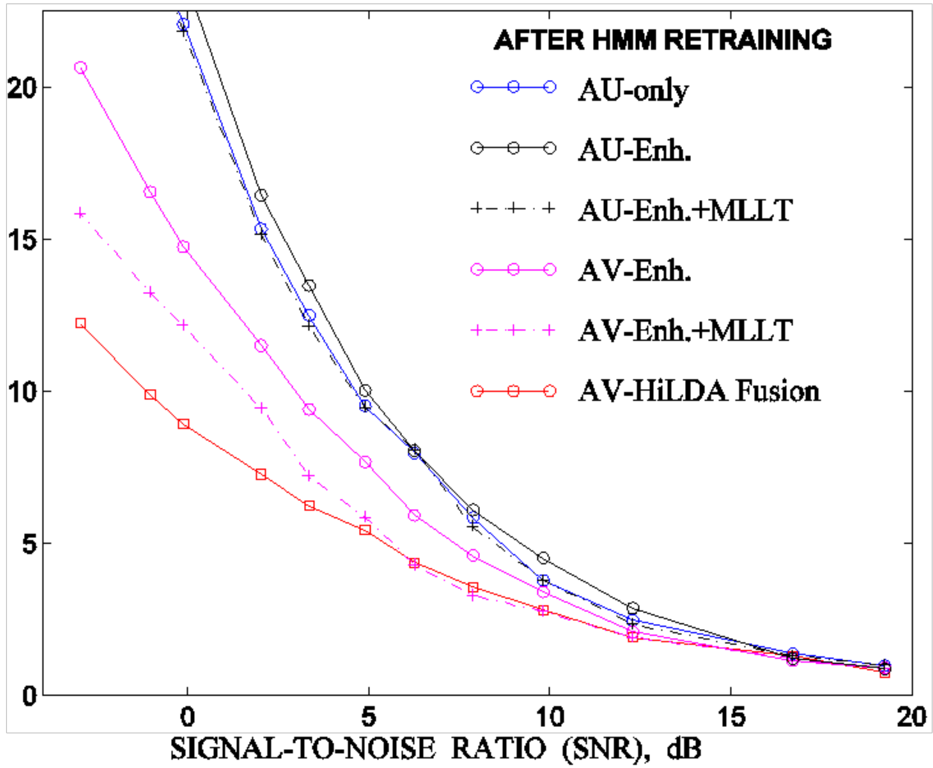
- Examples of **audio feature estimation** using bimodal enhancement (additive speech babble noise at **4 dB SNR**): Not perfect, but better than noisy features, and helps ASR!



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**):

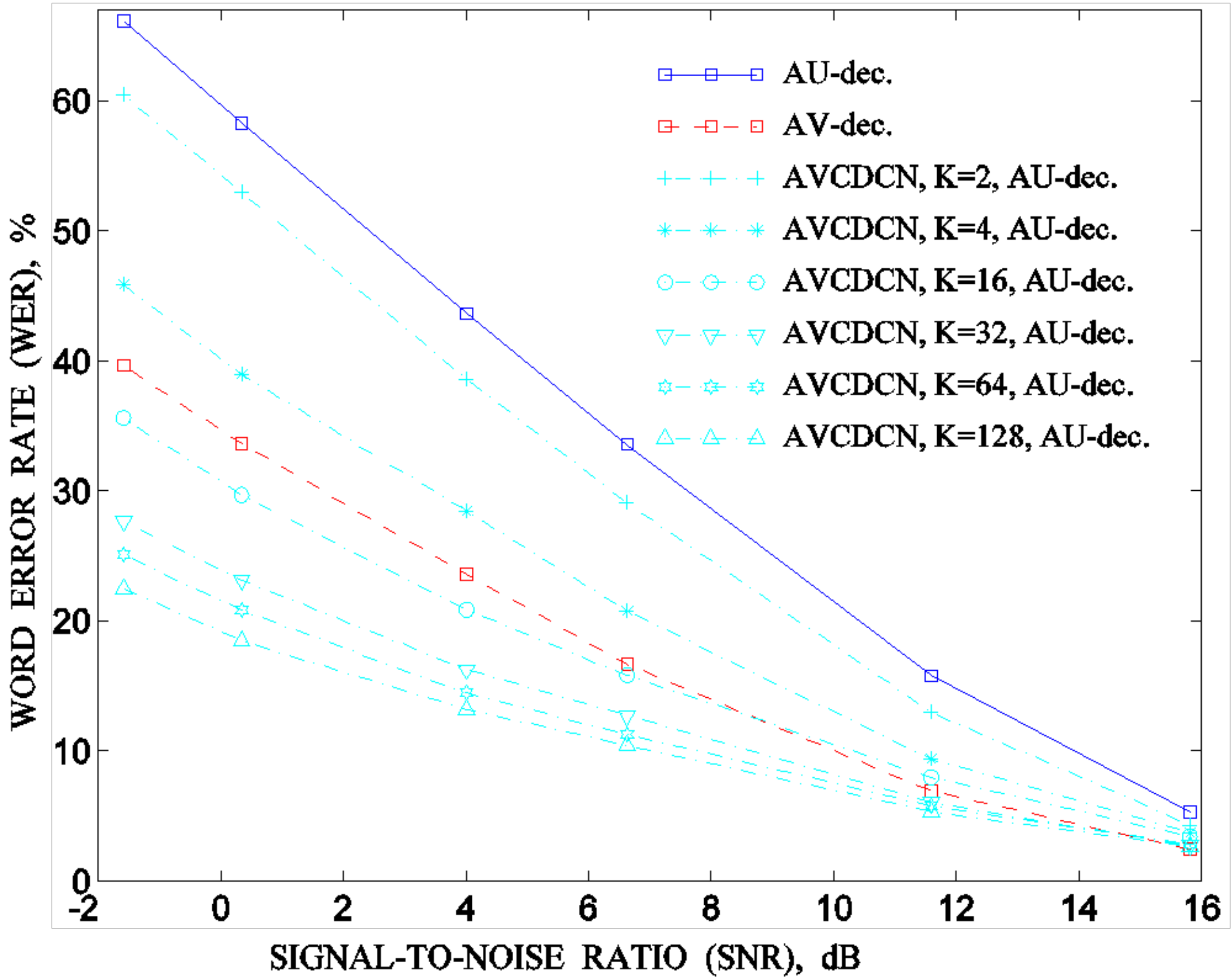
$$\mathbf{o}_{A,t}^{(E)} = \mathbf{o}_{A,t} - \sum_{k=1}^{K} f_{A,k} \Pr(k \mid \mathbf{o}_{AV,t})$$

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

Non-Linear Bimodal Enhancement of Audio (II)

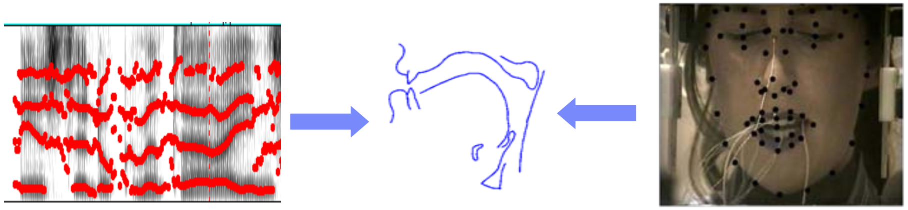
RESULTS:

- **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 Speech Inversion (I)

- Goal is to estimate **vocal tract geometry** and dynamics from observed speech.
- Problem is of interest to speech synthesis & coding, ASR, language tutoring, etc.
- This is an ill-posed inverse problem.
- **Visual channel** can help since some of the articulators are visible.



- Typical approach (Yehia et al., 1998) – observations **y**, articulatory parameters **x**:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \bar{\mathbf{x}}, \Sigma_x) \qquad p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; \bar{\mathbf{y}} + W(\mathbf{x} - \bar{\mathbf{x}}), Q)$$

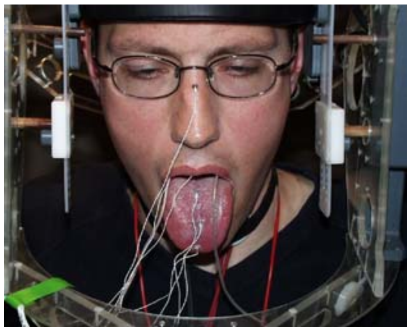
$$\hat{\mathbf{x}} = (\Sigma_x^{-1} + W^T Q^{-1} W)^{-1} (\Sigma_x^{-1}\bar{\mathbf{x}} + W^T Q^{-1}(\mathbf{y} - \bar{\mathbf{y}} + W\bar{\mathbf{x}}))$$

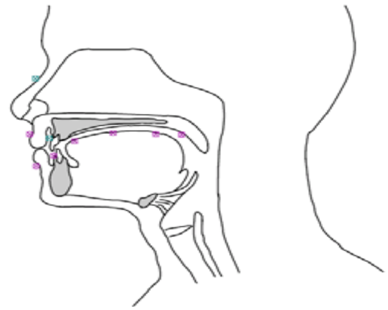
where **W** is estimated with **MSE**.

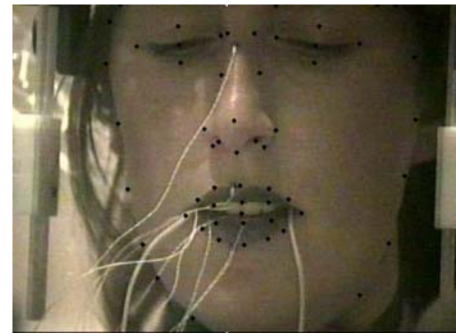
- Better perfomance is achieved with piecewise linear models – **W** depends on HMM sub-phonetic states.
- Smoothing of recovered trajectories is often employed.

Audio-Visual Speech Inversion (II)

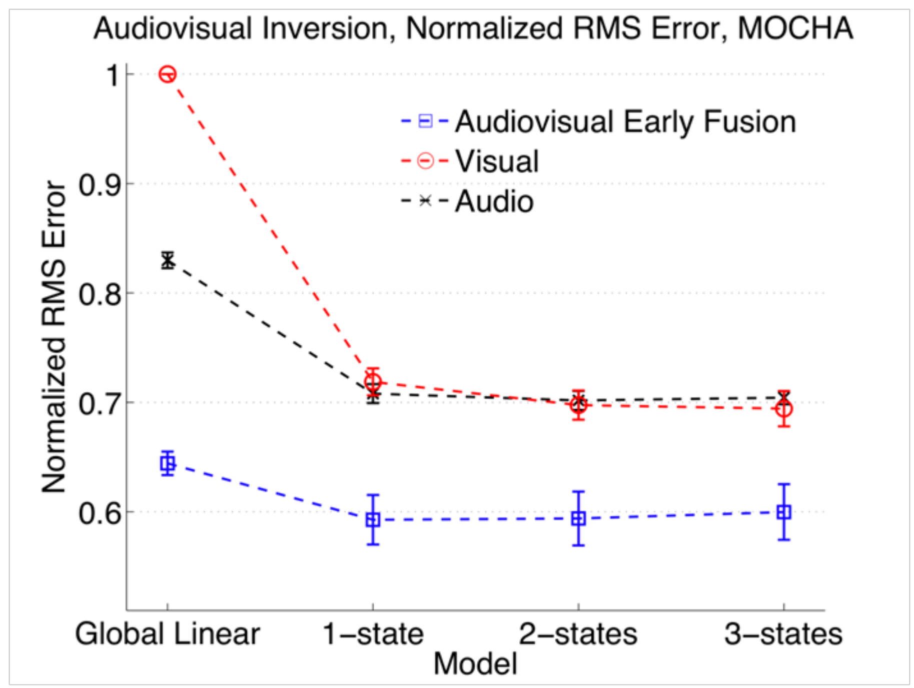
- Experiments on the **MOCHA database** (University of Edinburgh) – 9 EMA coils.







- **Results** (Katsamanis et al., 2009):



Audio-Visual Speaker Recognition – Brief Overview

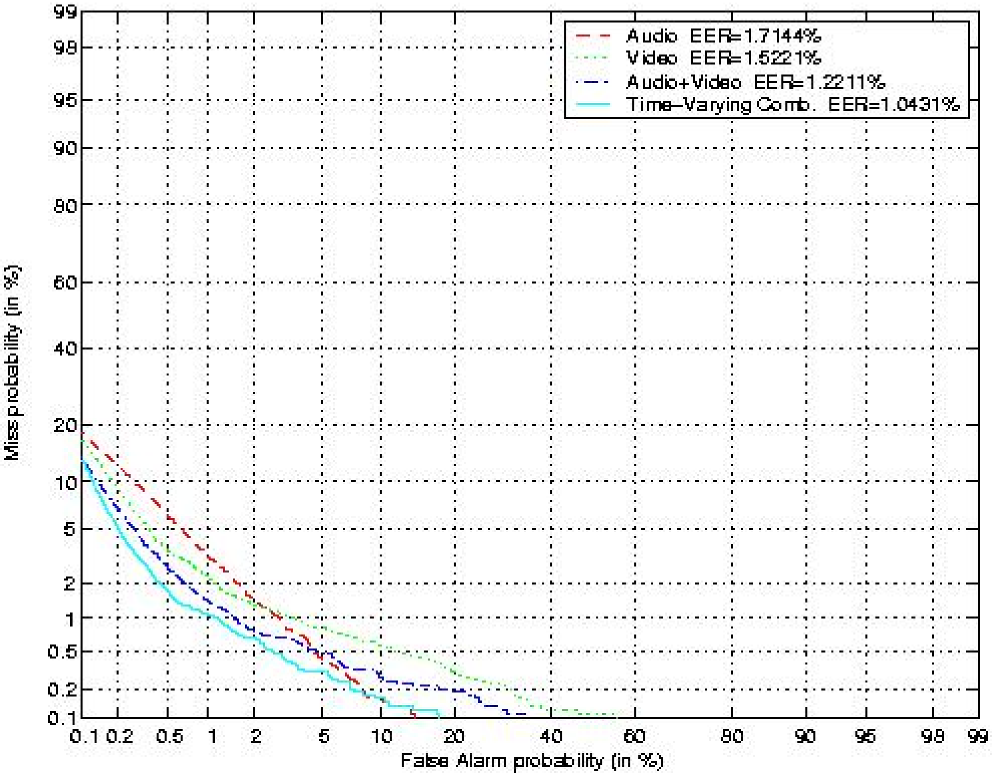
In case of **bimodal data**, the following ***3 information streams*** **can be utilized**:

- 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** %

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

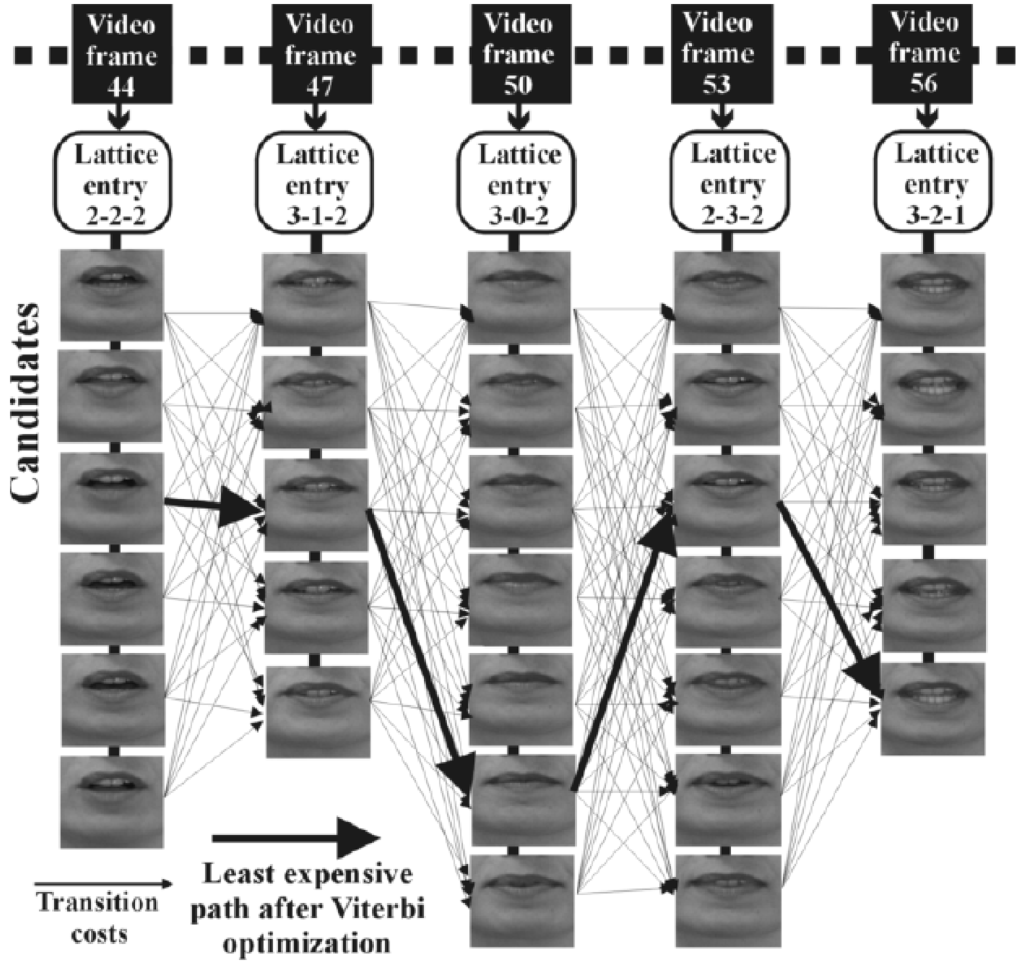
    models
  - Gained popularity due to MPEG-4 facial animation standard
- ***Image or Video-Based***
  - Segments of 2D videos of a speaker are
    - ✓ Acquired
    - ✓ Processed
    - ✓ Concatenated

Boundaries are blurry

Audio-Visual Speech Synthesis (III) – Concatenative Approach

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

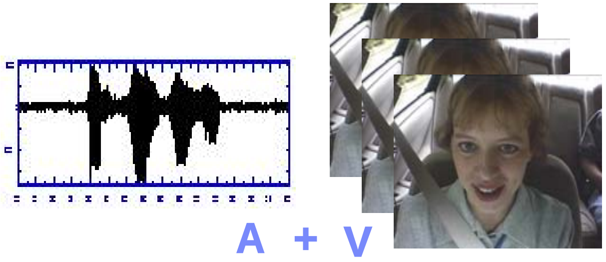
- Analysis of database segments (images or video snippets).
  - Extracts shape or appearance features to allow transition cost computation in concatenation.
- Synthesis stage:
  - Uses dynamic programming approach (Viterbi) to find minimum cost path and "stich" together the best possible image/video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



1. **Introduction:**
   - Motivation.
   - Audio-visual speech technologies.
   - Potential applications.

2. **Audio-visual speech components with emphasis on ASR:**
   - Data resources.
   - Visual feature representation for speech applications.
   - Audio-visual combination (fusion).

3. **Other audio-visual speech technologies:**
   - **Speech synchrony.**
   - **Speech enhancement.**
   - **Speech inversion.**
   - **Speaker recognition.**
   - **Speech synthesis.**



4. **Concluding Remarks.**
   - **Summary.**
   - **Acknowledgements.**

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.
- **Discussed additional AV speech applications.**
  - Synchrony detection.
  - Speech enhancement.
  - Speech inversion.
  - Speaker recognition.
  - Speech synthesis.
- **Experimental results** demonstrate several benefit of visual modality to above technologies.

Current Trends / Open Problems

- **Trends:**
  - Interest shifting towards realistic environments (meetings, broadcasts, automobiles), including multi-sensory environments with multi-speaker interaction.
  - Interest extends beyond ASR problem.
  - Database collection efforts by many sites.

- **Challenges:**
  - Pose modeling, compensation; pose invariant appearance visual features.
  - Robust visual feature extraction for unconstrained visual domains; invariance to environments.
  - Feature representation, selection.
  - Fusion functional, reliability modeling, asynchronous integration within / across modalities.
  - Still lagging common benchmarks in the community.

Acknowledgements

- **Former AT&T colleagues:** *Eric **Cosatto**, Hans Peter **Graf**.*
- **Former IBM colleagues:** *Stephen M. **Chu**, Jonathan **Connell**, Sabine **Deligne**, Jing Huang, Giridharan **Iyengar**, Vit **Libal**, Etienne **Marcheret**, Chalapathy **Neti**, Michael **Picheny**, Larry **Sansone**, Andrew **Senior**, Roberto **Sicconi**.*
- **Former IBM interns:** *Ashutosh **Garg,** Roland **Goecke,** Guillaume **Gravier,** Jintao **Jiang,** Kshitiz **Kumar**, Patrick **Lucey**, Patricia **Scanlon**.*
- **Other:** *Petar S. **Aleksic**, Aggelos K. **Katsaggelos**, Iain **Matthews**, Juergen **Luettin**, Petros **Maragos**, George **Papadopoulos**.*
- **Former funding:** ***IBM*** AR program, EU FP6 projects ***CHIL & NETCARITY*** through IBM Research Participation.
- **Current support** by EU FP7-PEOPLE-RG project ***AVISPIRE.***