Features for Audio-Visual Speech Recognition

Doctoral thesis by Iain Matthews,
University of East Anglia reviewed
by Tobias Andersen,
Lab. of Computational Eng., HUT

Audiovisual speech primer

- Why is visual speech interesting?
  - It improves the ”SNR” of speech
  - It helps distinguish speakers in a multi-speaker environment
  - The best speech decoder – the human brain – always try to use it

Audiovisual speech primer

Review overview

- Visual speech recognition
  - low level
    - Area
    - Height and width
    - Area histogram
    - Scale histogram
  - high level
    - Active shape model (ASP)
    - Active appearance model (AAP)
- HMM
- AV integration

Low level speech recognition

- Sieves
  - Erosion/dilation
  - Opening/closing
  - Granules and scales

Erosion

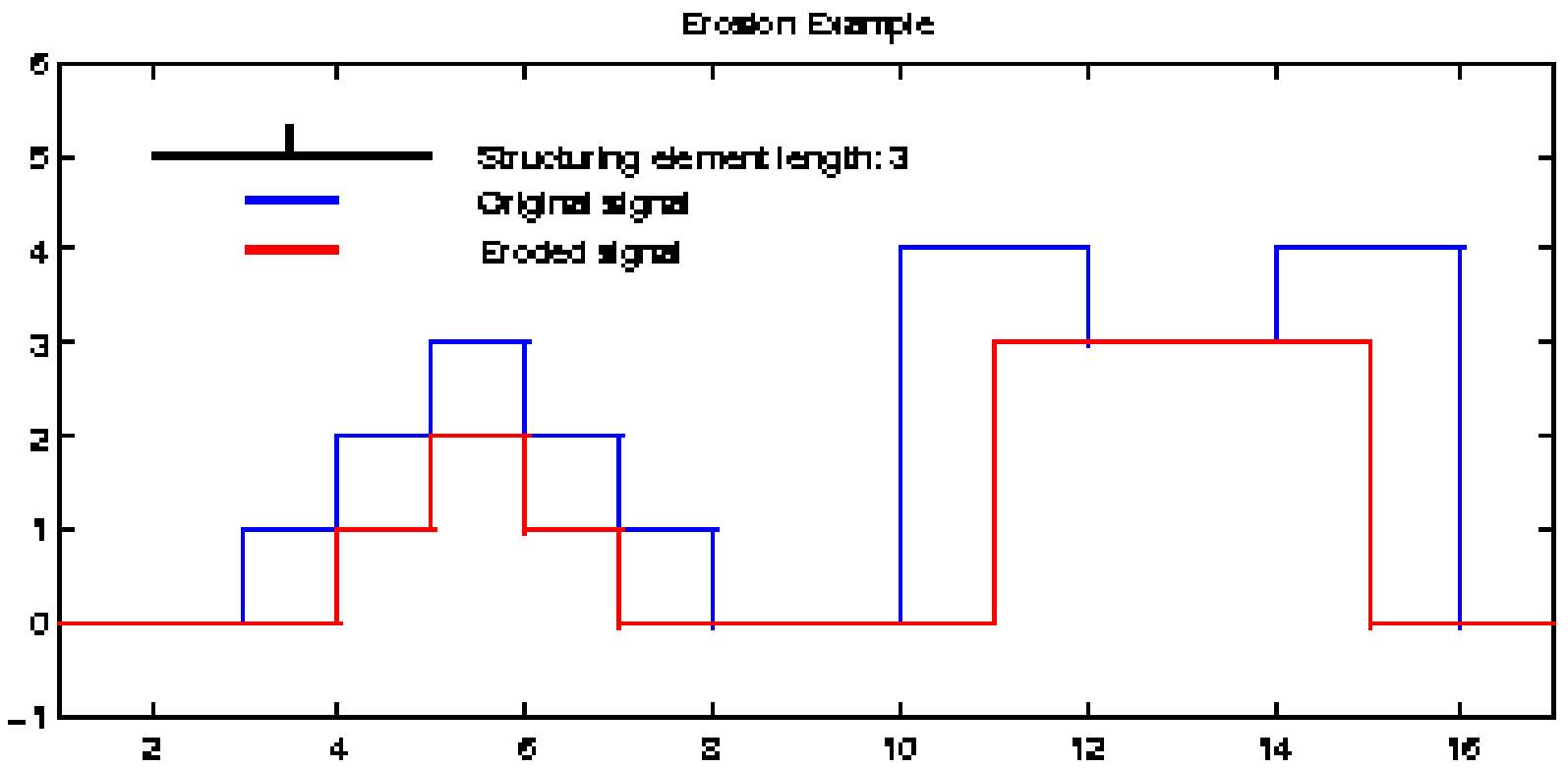


Figure 4.2: Example of greyscale erosion.

Dilation

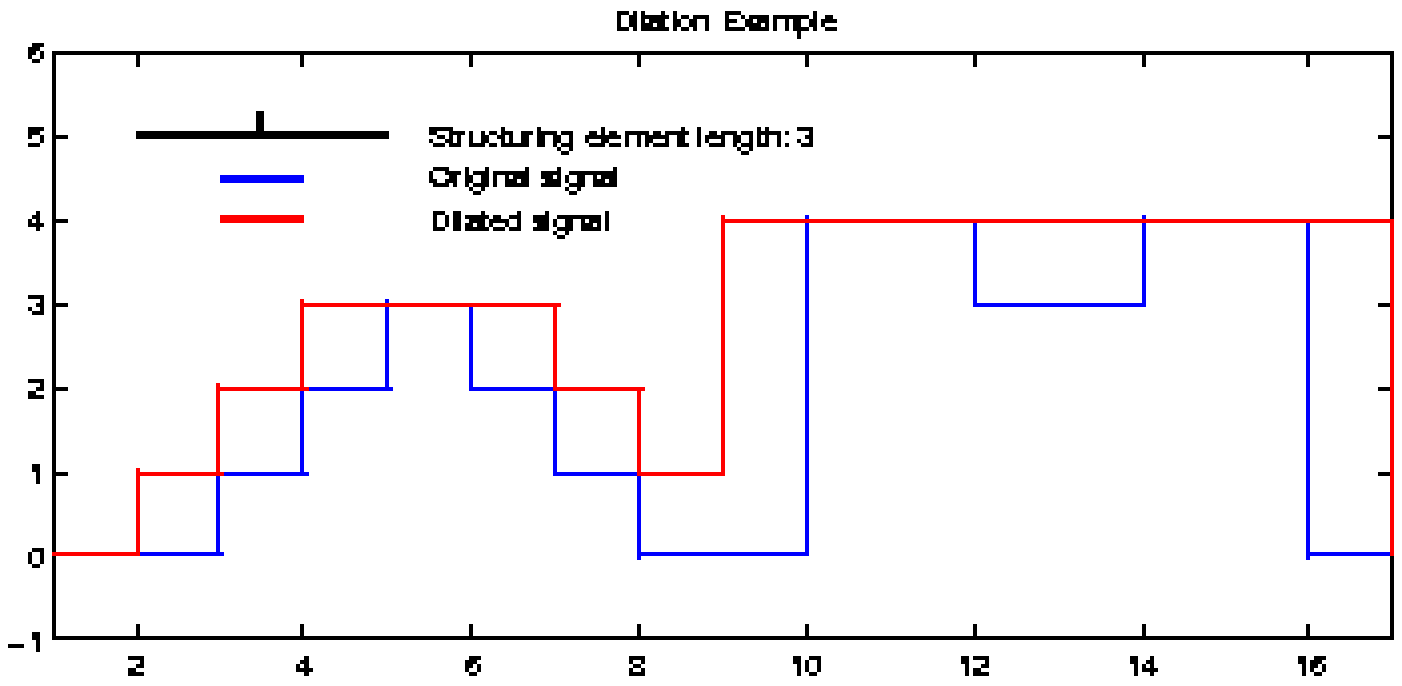


Figure 4.3: Example of greyscale dilation.

Opening and closing

- Opening
  - Erosion followed by dilation
  - Removes positive extrema
- Closing
  - Dilation followed by erosion
  - Removes negative extrema

Recursive median filter

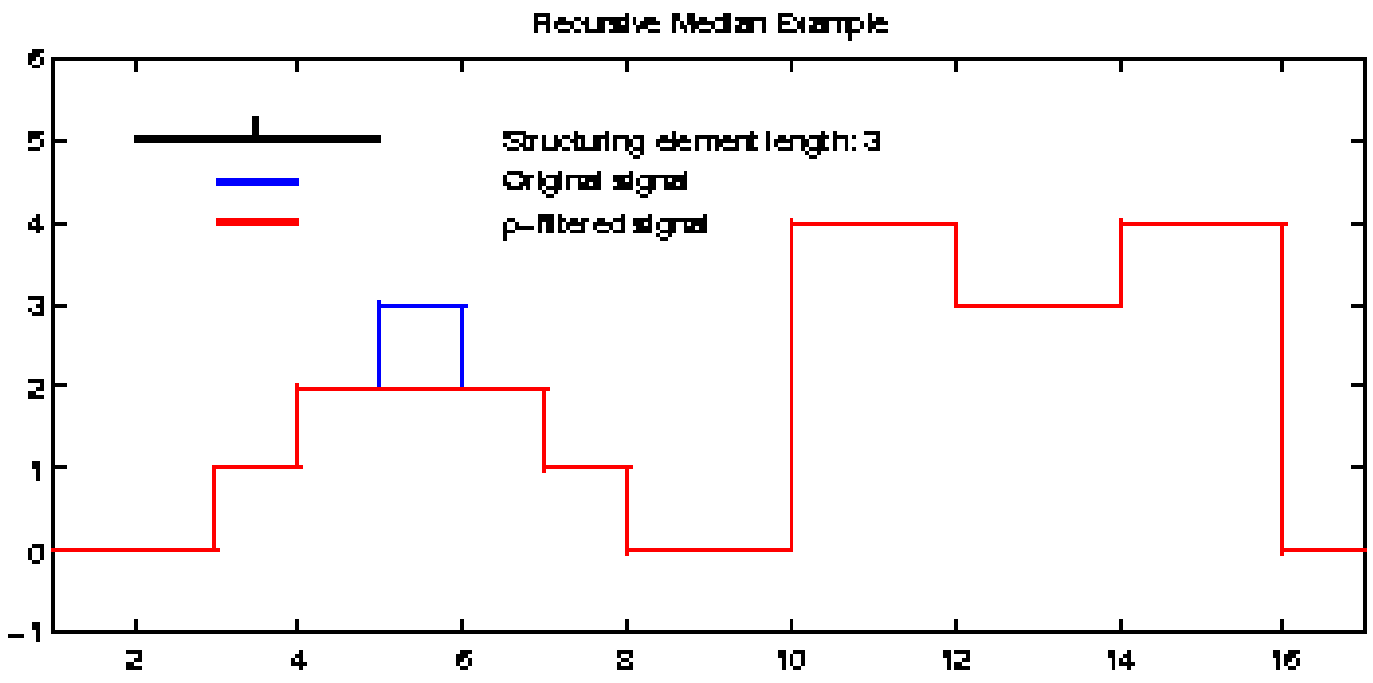
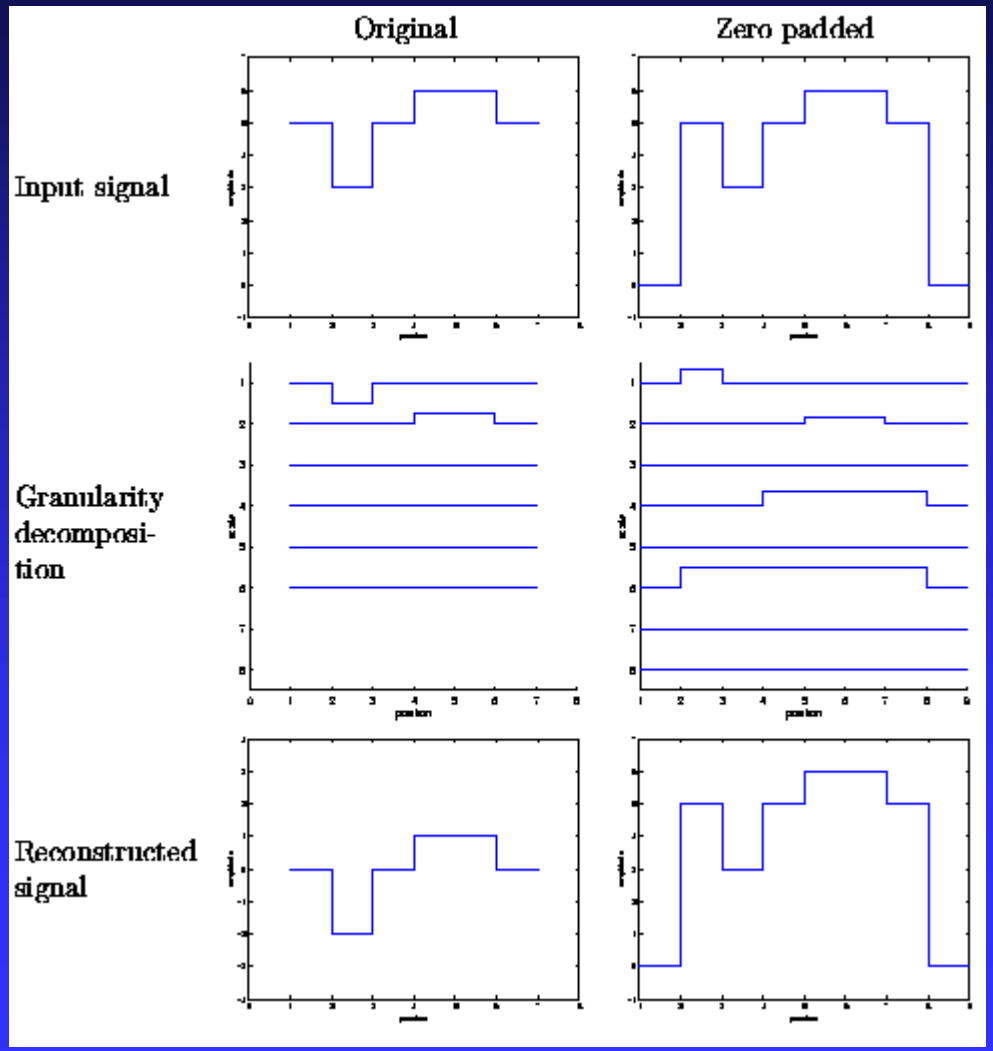


Figure 4.9: Example of a recursive median filter width three

Preservation of baseling



Sieve decomposition

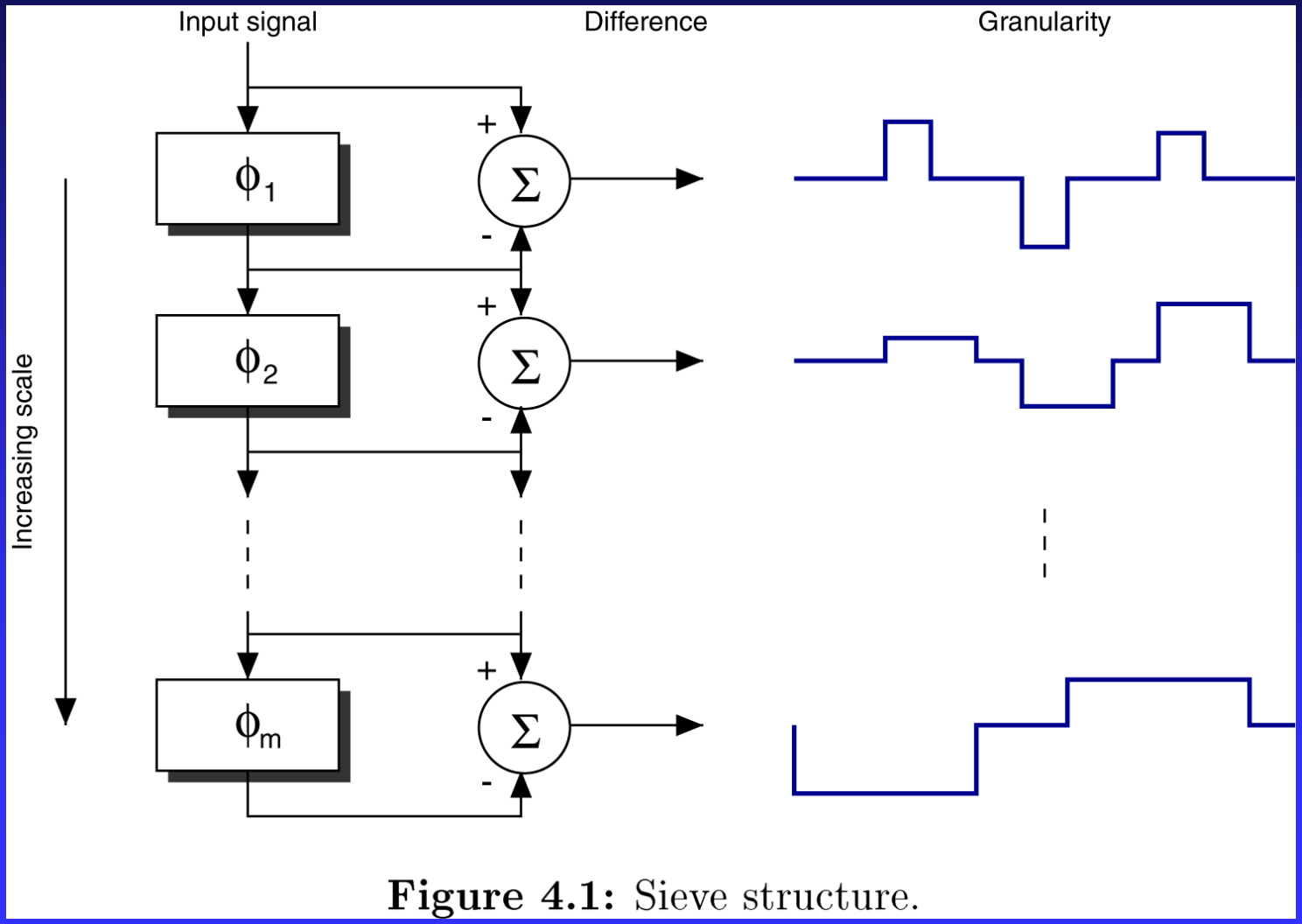
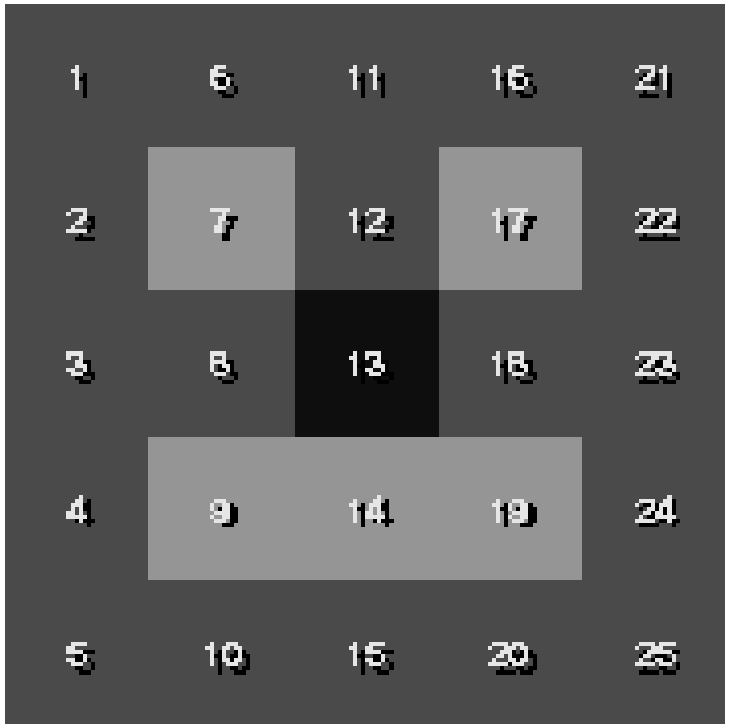


Figure 4.1: Sieve structure.

Sieve decomposition

- The size of the scaling element determines the scale
- At each scale, *granules* may or may not be extracted
- Scales are binned into *channels* to lower computational load

2D sieve decomposition



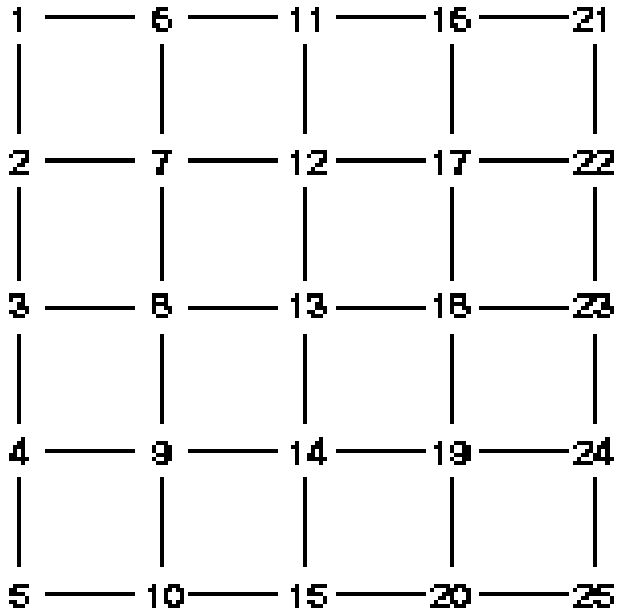
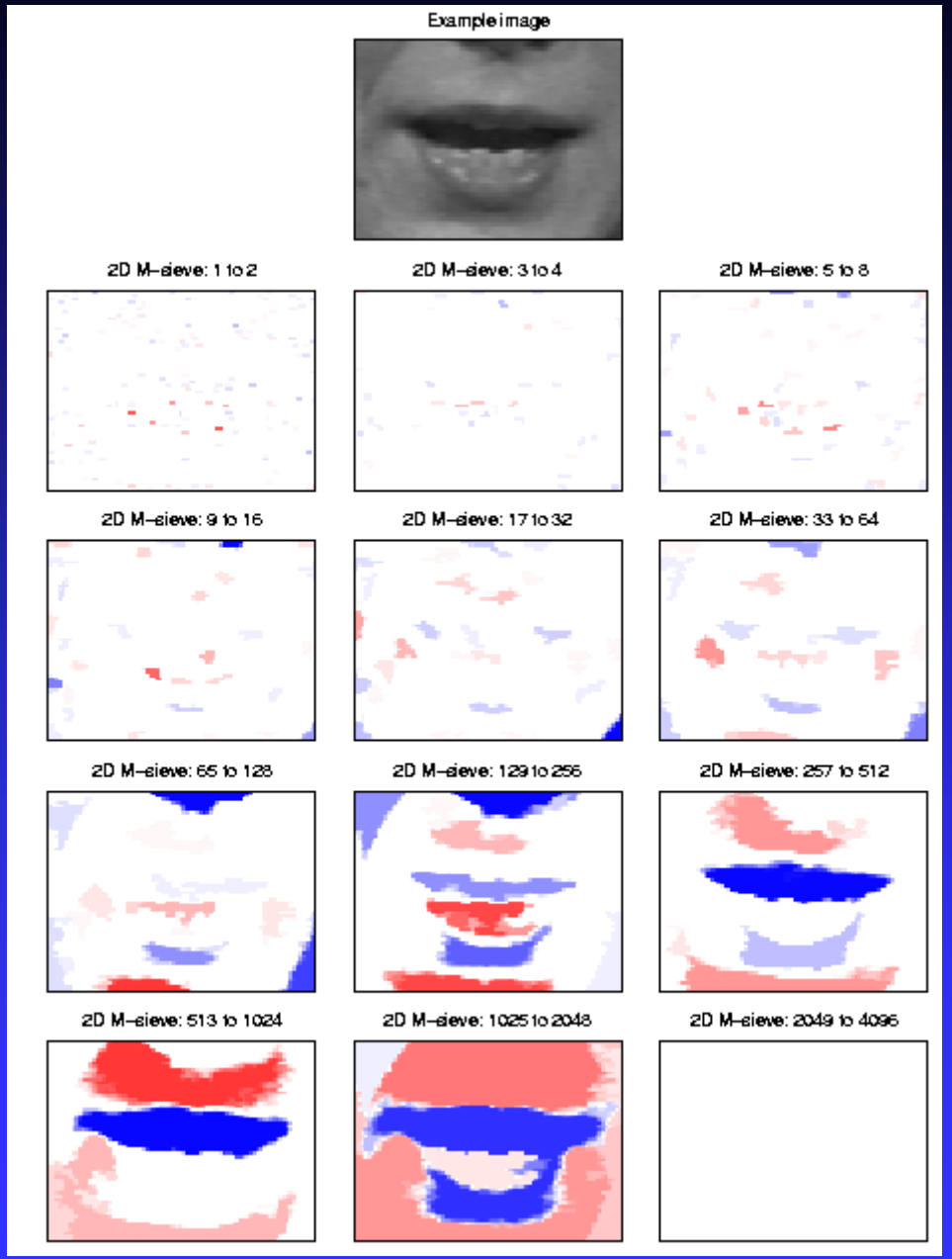


Figure 4.8: Image represented as a four-connected graph

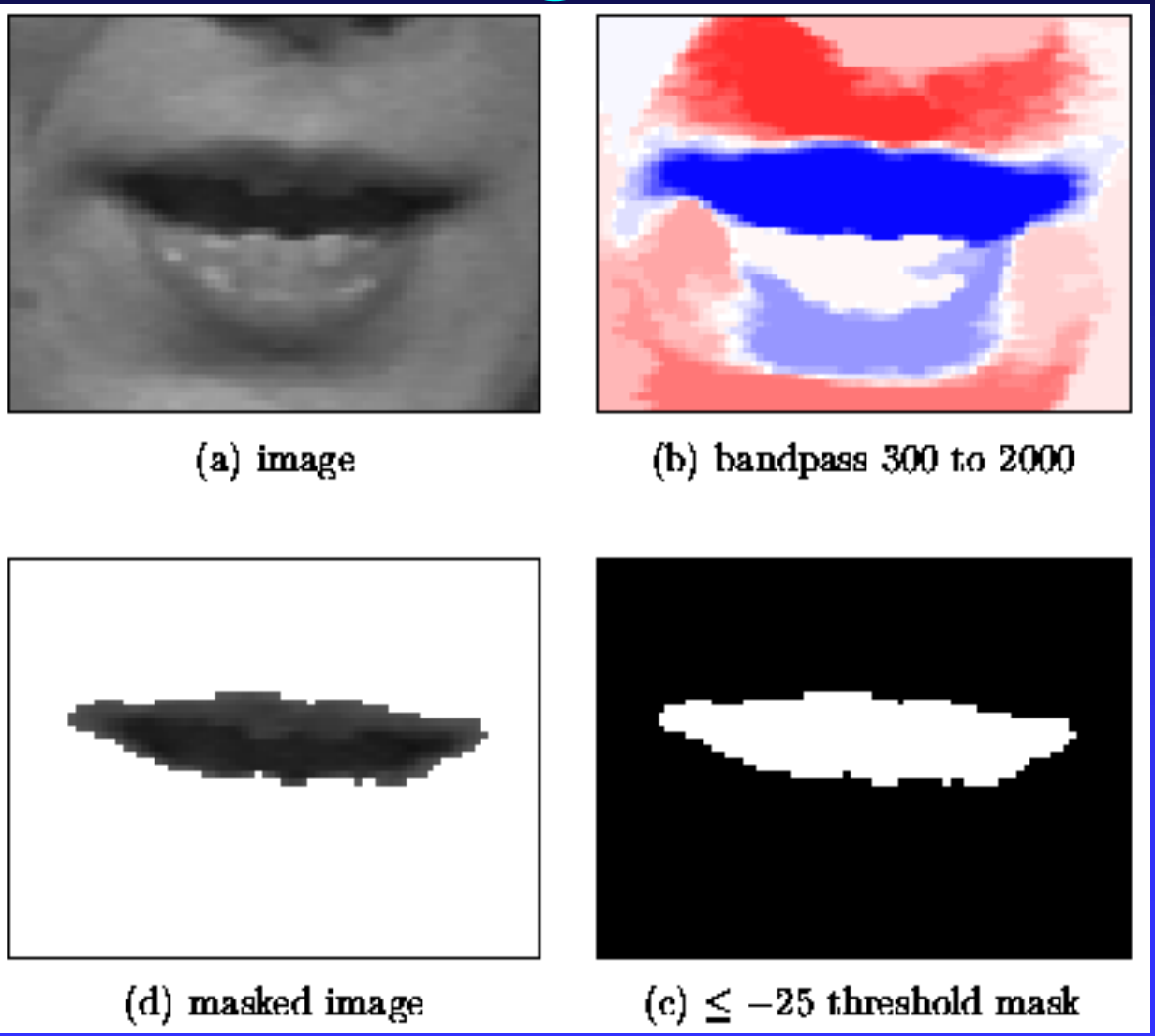
Area tracking



Area tracking

- Idea: Some scale/channel will find the area of the mouth opening
- Problem: what's the right scale?
- Solution: 300-2000

Area tracking



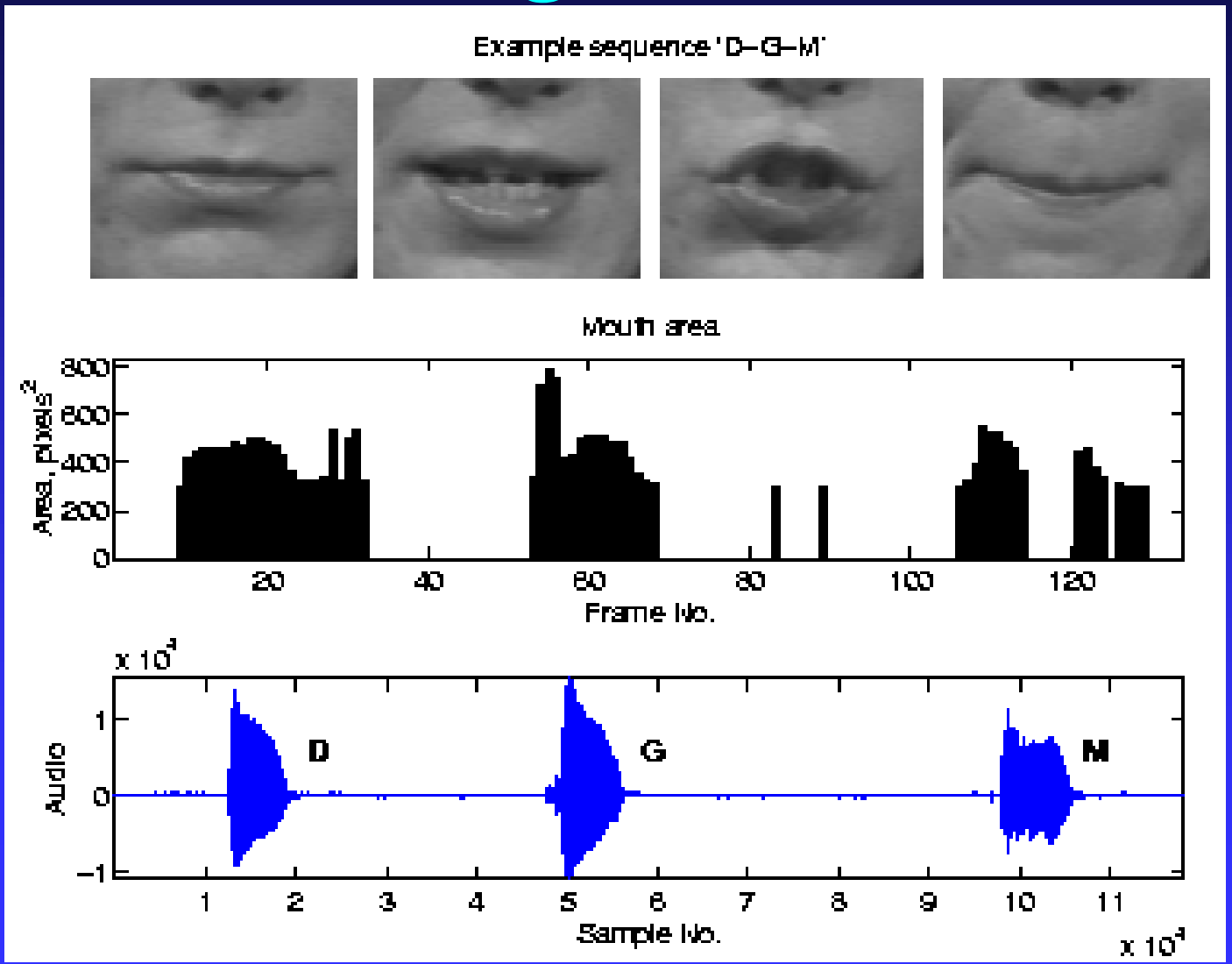
(a) image

(b) bandpass 300 to 2000

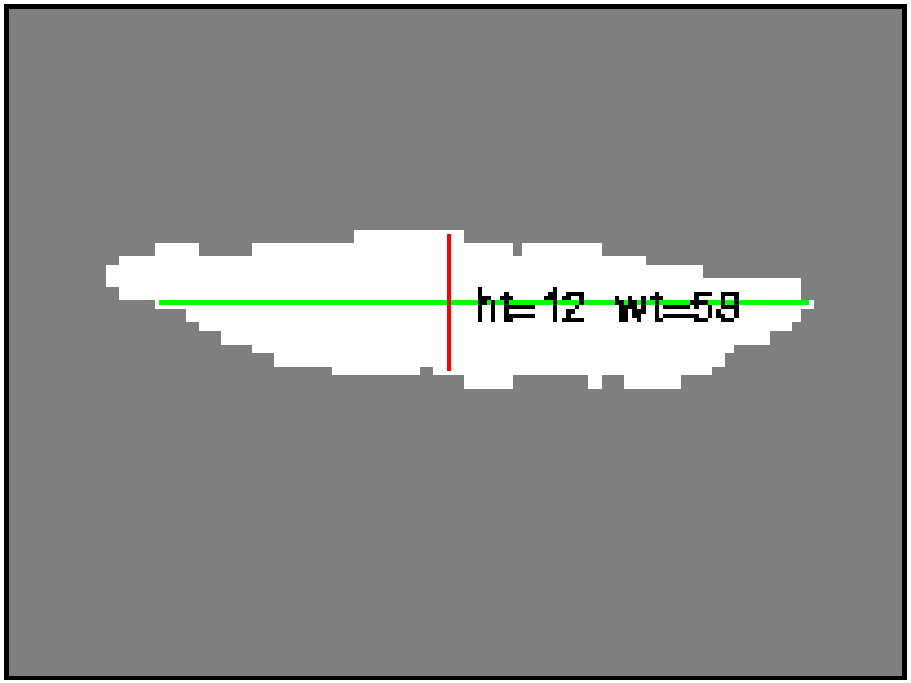
(d) masked image

(c) ≤ -25 threshold mask

Area tracking

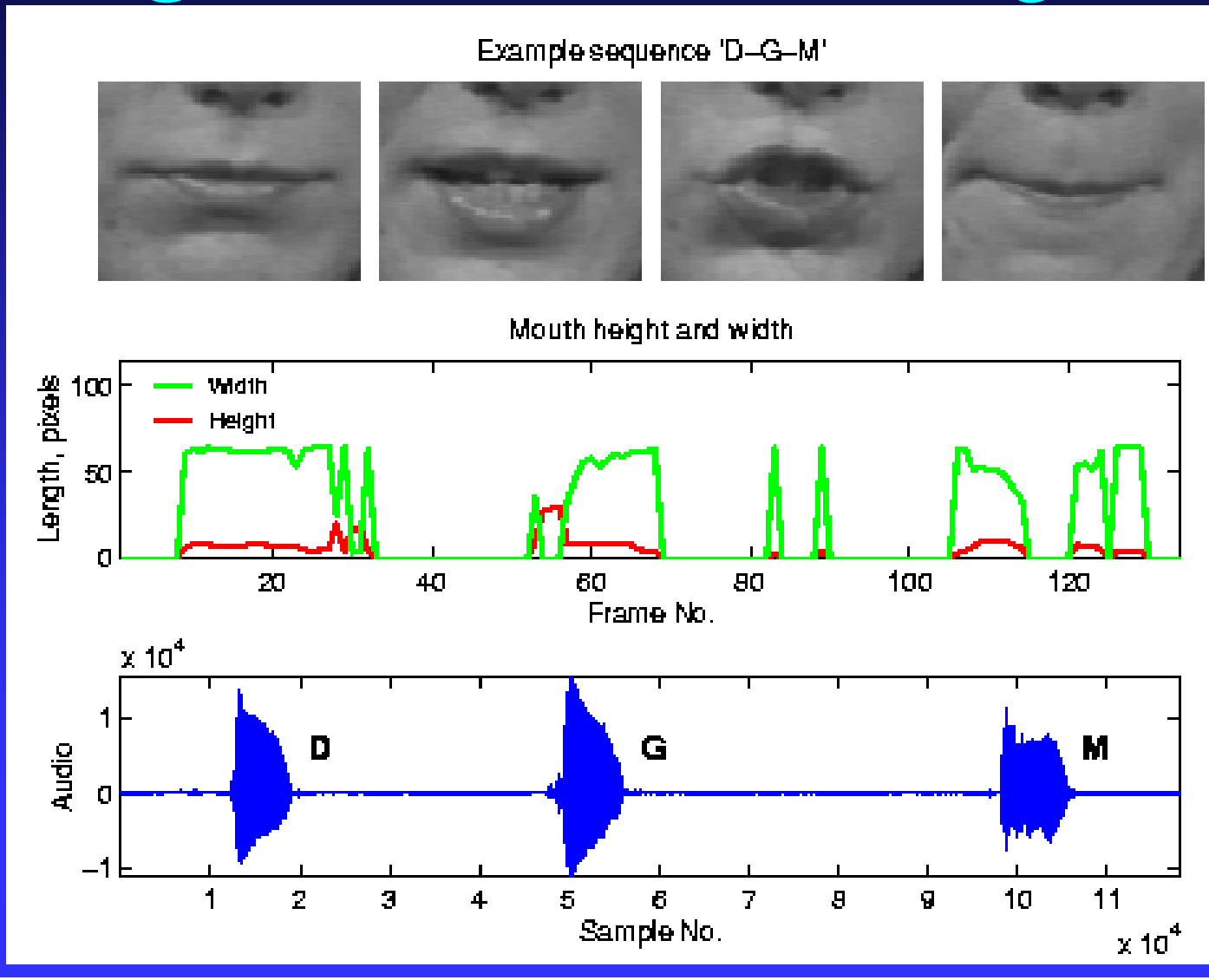


Height and width tracking



(c) measured height and width

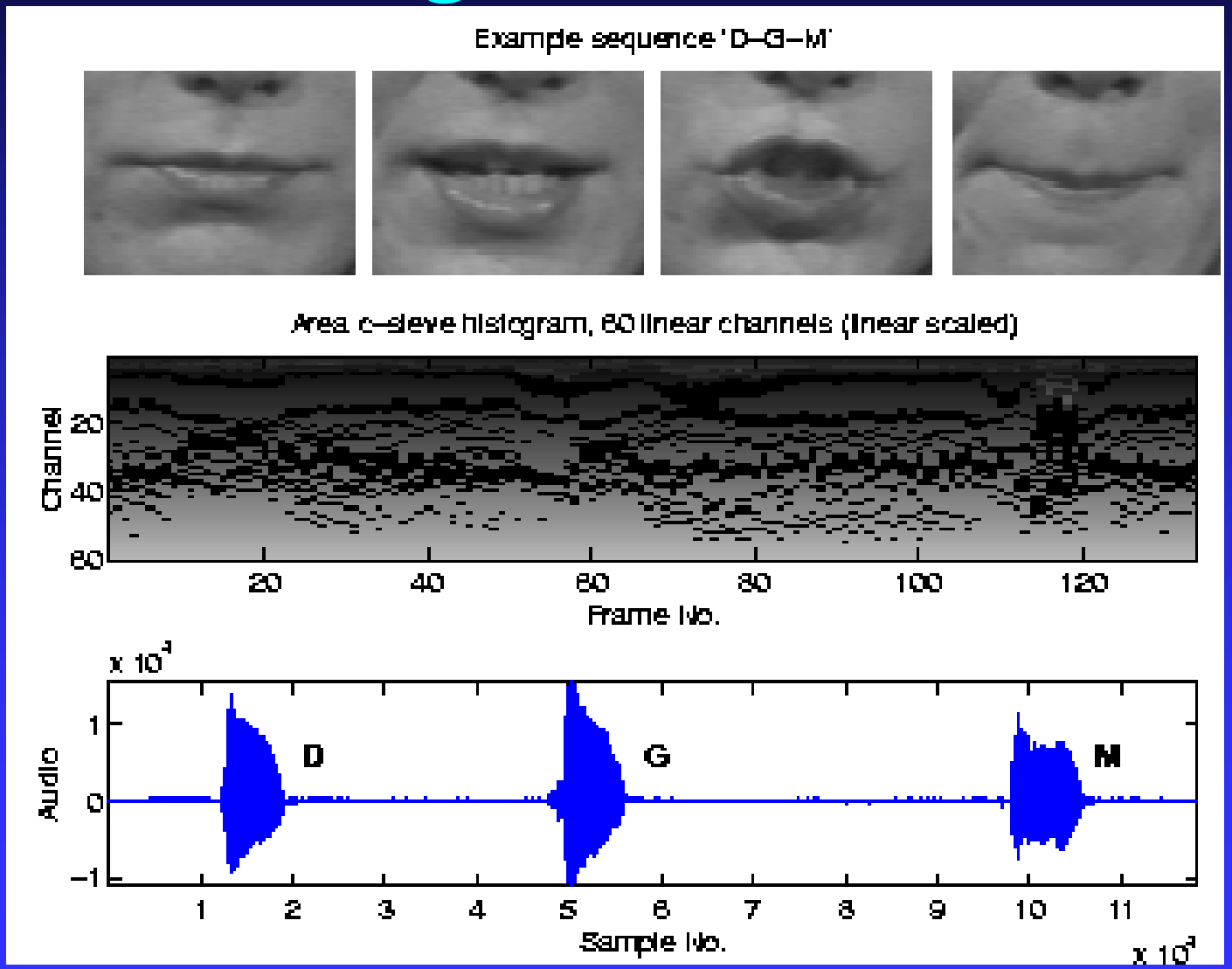
Height and width tracking



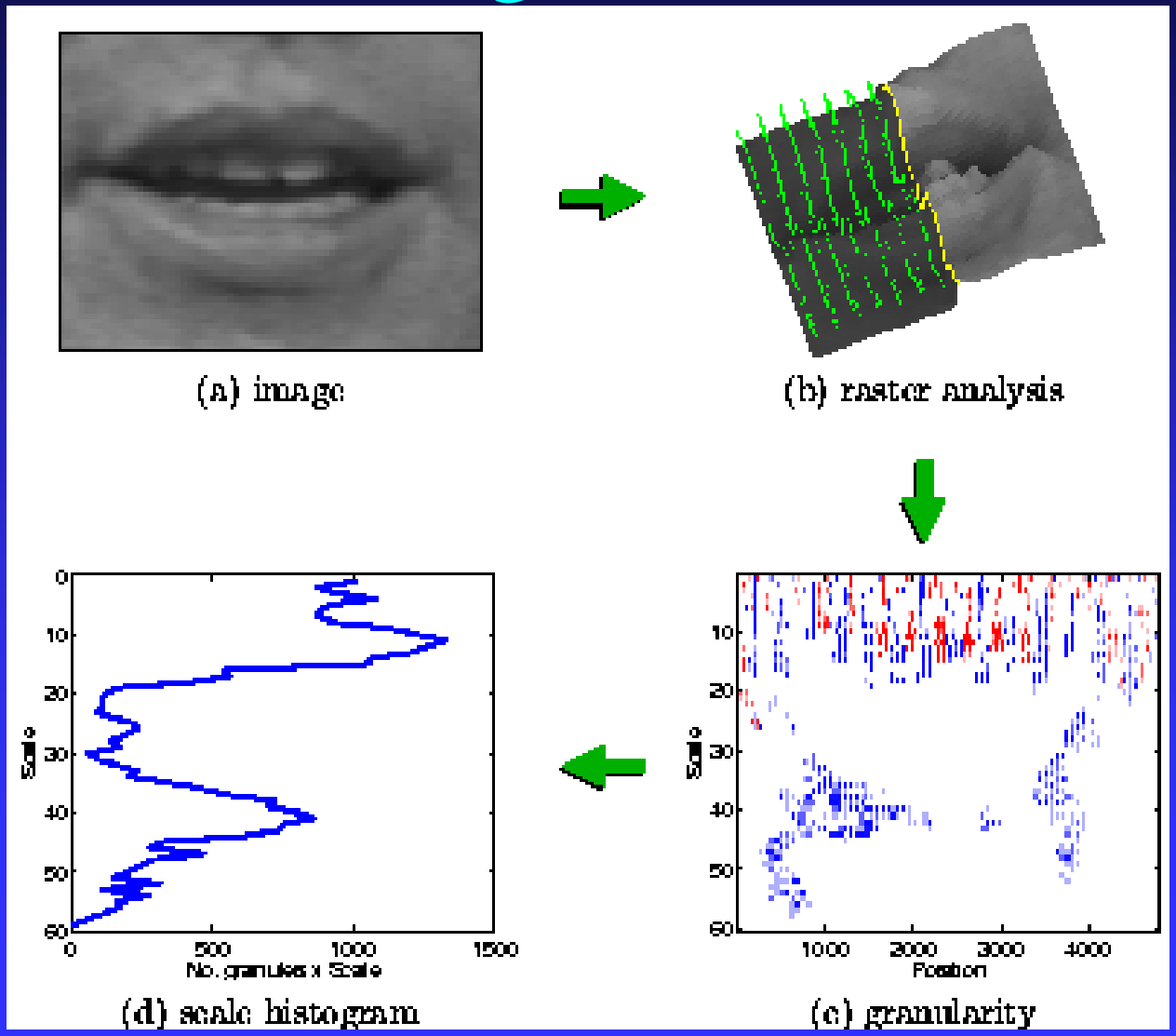
Area histogram

- To overcome the arbitrary choice of channel
- Choose all channels

Area histogram



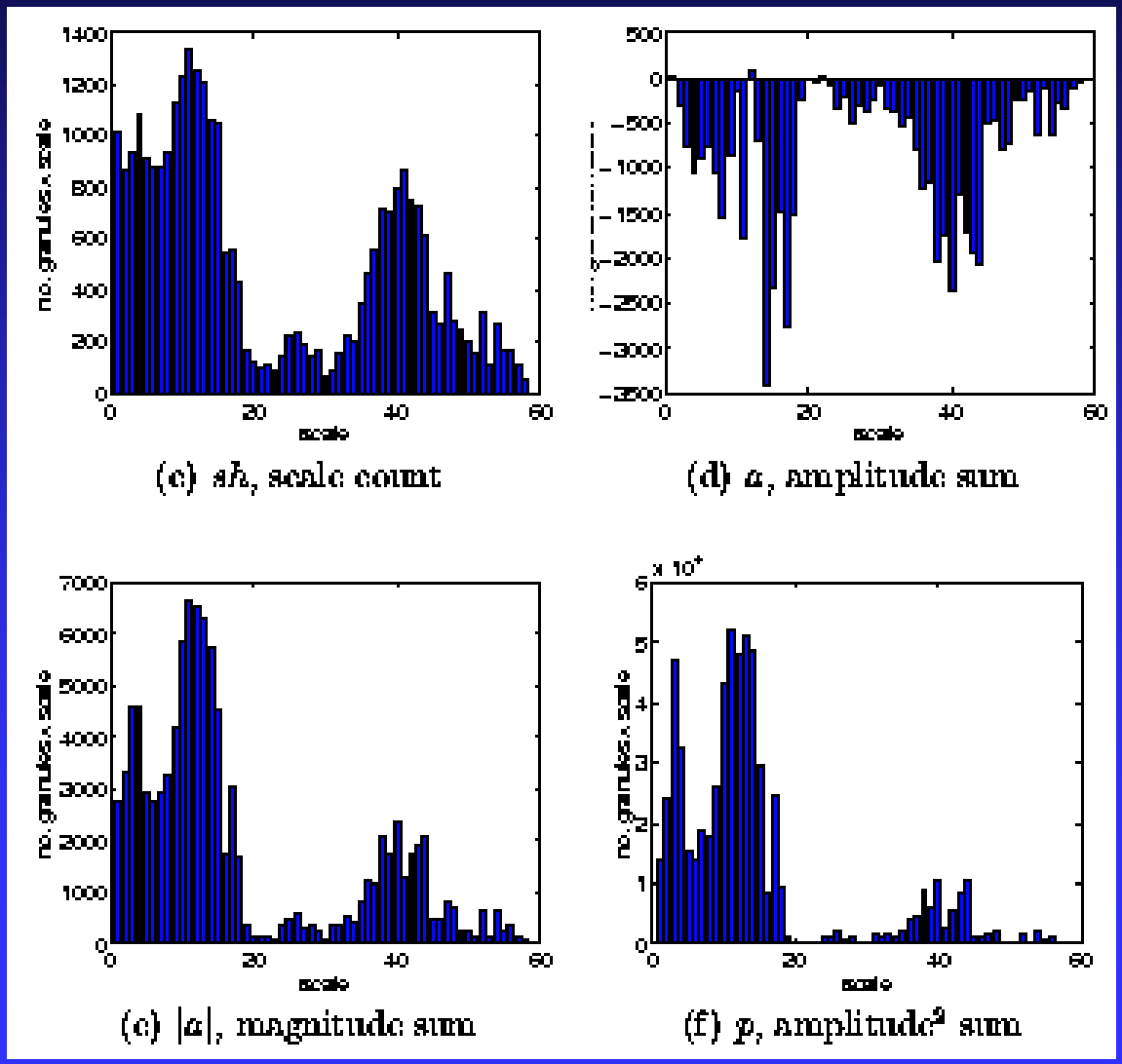
Scale histogram



PCA

- To further reduce the number of channels
- Problem: Covariance or correlation?
- Problem: How many PCs?

Scale histogram



The Full investigation

Attribute	Settings		
Sieve type	median, m	opening, o	closing, c
DC Baseline	preserve	ignore	
Interpolate	N, 25Hz	I, 50Hz	
PCA components	10	20	
PCA type	covariance	correlation	
HMM states	5	7	9
Gaussian modes	1	3	

Key	Type	Details
sh	scale count	count of granules at each scale
a	amplitude sum	sum of amplitudes at each scale
$\|a\|$	magnitude sum	sum of absolute amplitude at each scale
a^2	power sum	sum of squared amplitude at each scale

Table 7.8: Key to scale histogram types.

The Full investigation

- Yet another parameter: Scaling - Multiplication of count by scale
- 2304 combinations of 9 parameters
- Tested on two databases

Evaluation of low level methods

- Area tracking and height/width tracking gave less than 10% correct identification across talkers
- Area histogram was better, but still not good. Less than 20% correct identification across talkers
- All three methods not further discussed

Evaluation of low level methods

- Consistent results (two databases):
  - Covariance better than correlation for PCA
  - Opening performs consistently poorer than median or closing

Evaluation of low level methods

- Inconsistent results (two databases):
- Three Gaussian modes better than one for one training set – vice versa for another
- Preserving the DC is better for one set – poorer for another
- Linear scaling helps for one database – not for another. Conclusion: Better do it.
- 20 PCs better than 10 for one database – vice versa for another

Evaluation of low level methods

- Robustness of results:
  - Linear scaling affects correlation PCA indicating that the results are ”statistically fragile”

High level methods

- Active shape model (ASP)
- Active appearance model (AAP)

Active shape model (ASP)

- Based on a Point Distribution Model
- Landmark and secondary points placed manually on training set
- Spline fitted to points
- Secondary points repositioned evenly along spline

Active shape model (ASP)

- Elimination of pose variation by minimization of

$$E = (\mathbf{x}_1 - M(s,\theta)[\mathbf{x}_2] - \mathbf{t})^T \mathbf{W} (\mathbf{x}_1 - M(s,\theta)[\mathbf{x}_2] - \mathbf{t})$$

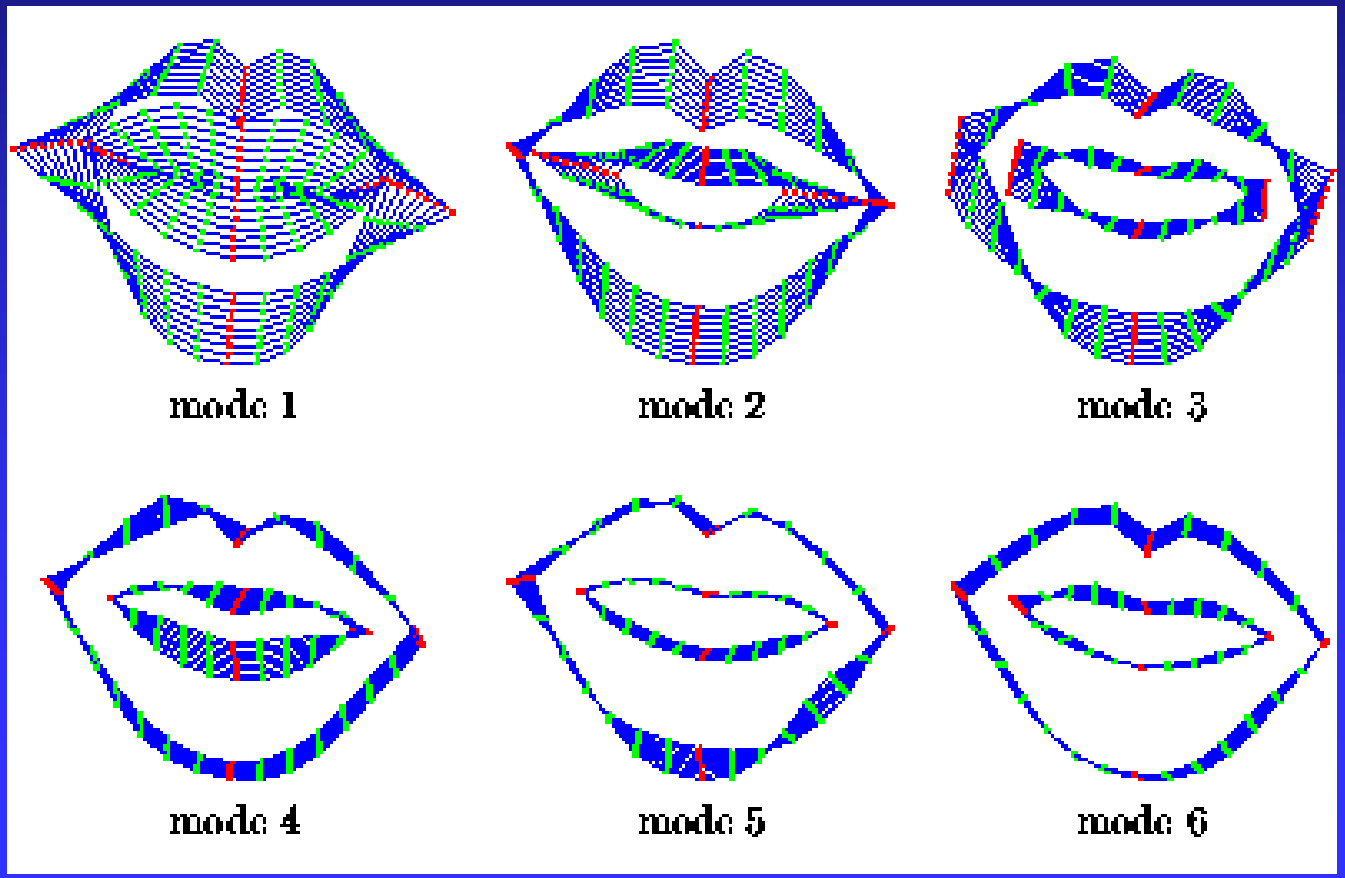
$$M(s,\theta) \begin{bmatrix} x_{jk} \\ y_{jk} \end{bmatrix} = \begin{pmatrix} (s\cos\theta)x_{jk} - (s\sin\theta)y_{jk} \\ (s\sin\theta)x_{jk} + (s\cos\theta)y_{jk} \end{pmatrix}$$

$$\mathbf{t} = (t_{x1}, t_{y1}, \ldots, t_{xN}, t_{yN})$$

- Weights highest for stable points

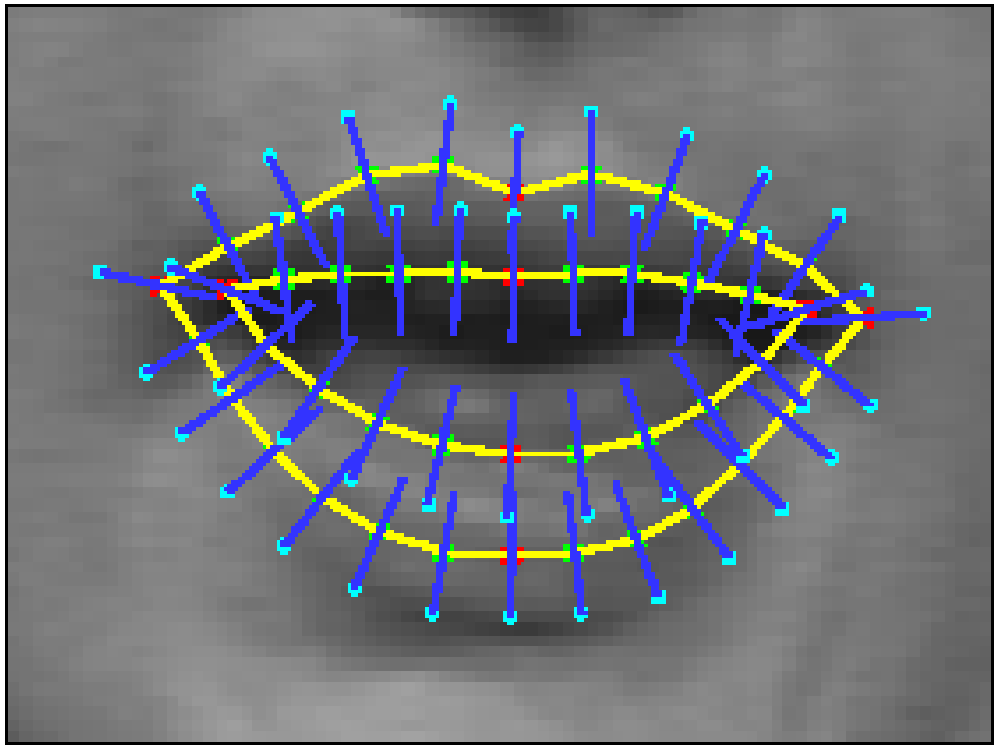
Active shape model (ASP)

- 95% PCA filtering

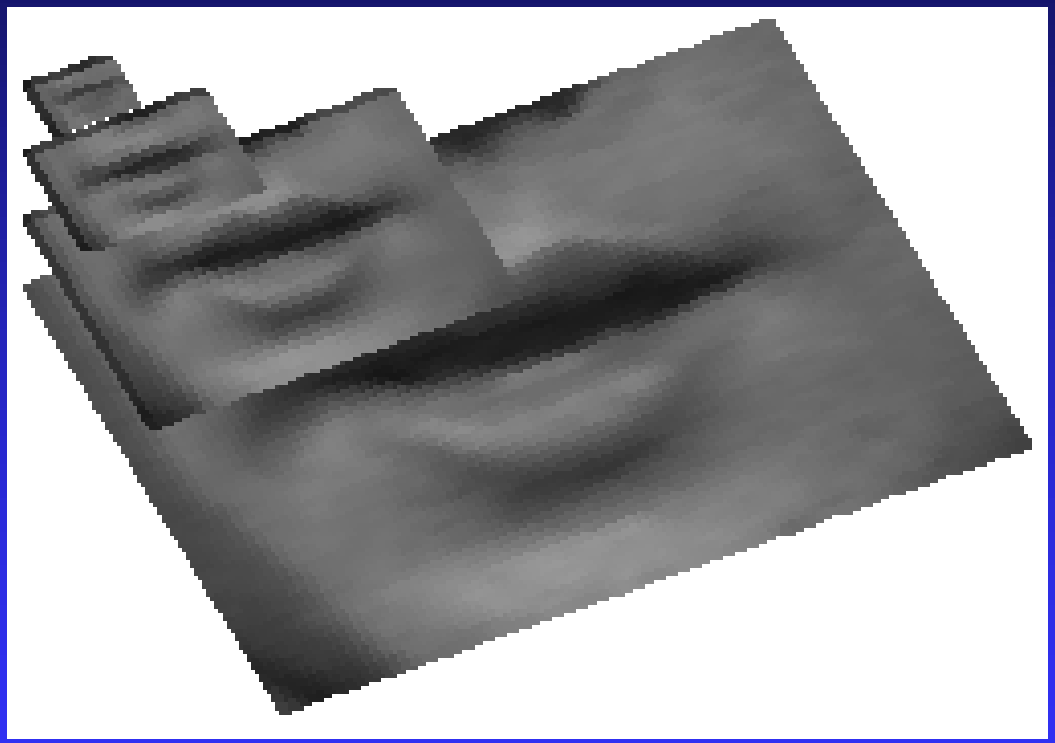


Active shape model (ASP)

- Grey Level Profile Distribution Model (GLPD)



Active shape model (ASP)



- Multi-resolution Fitting
- GLDM starts at coarse scale – moves to finer
- May yield better fit

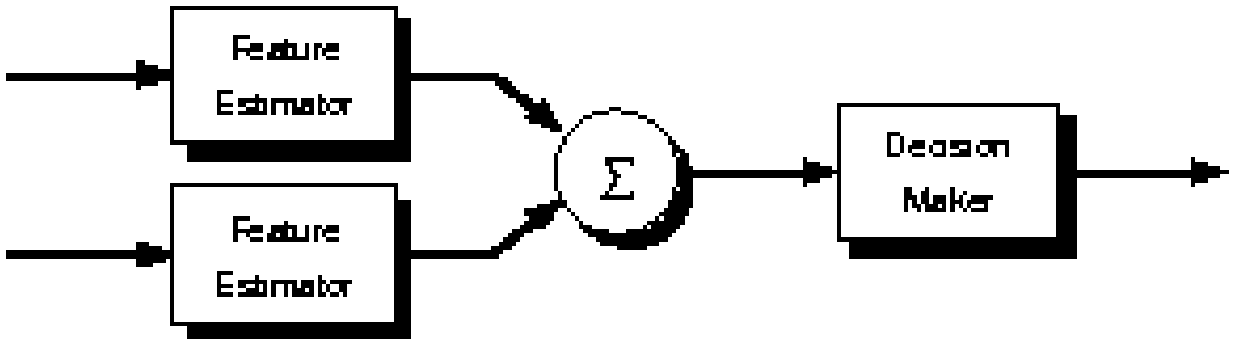
Active Appearance model (AMP)

- Combines the ASM with a greylevel appearnce model (GAM)
- GAM:
  - Warp images to mean shape
  - Pixel values within mean shape forms the GAM

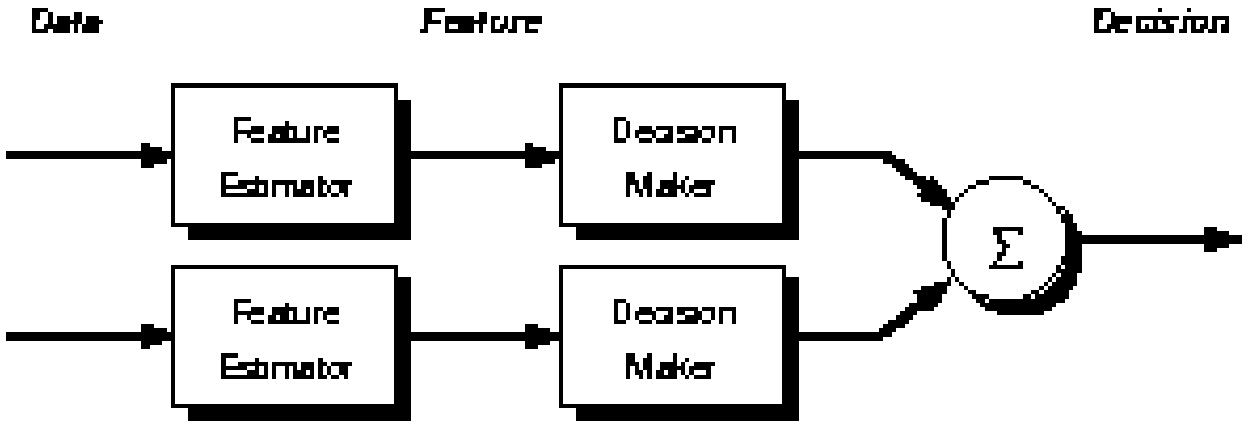
Evaluation of high level methods

- ASM poorer than MSA
- Multi-resolution better
- Longer Greylevel Profiles better
- Initialising optimisation at mean shape better than at previous shape – difficult to break out of local minima

Audiovisual integration



(b) Feature fusion (*early integration*).



(c) Decision fusion (*late integration*).

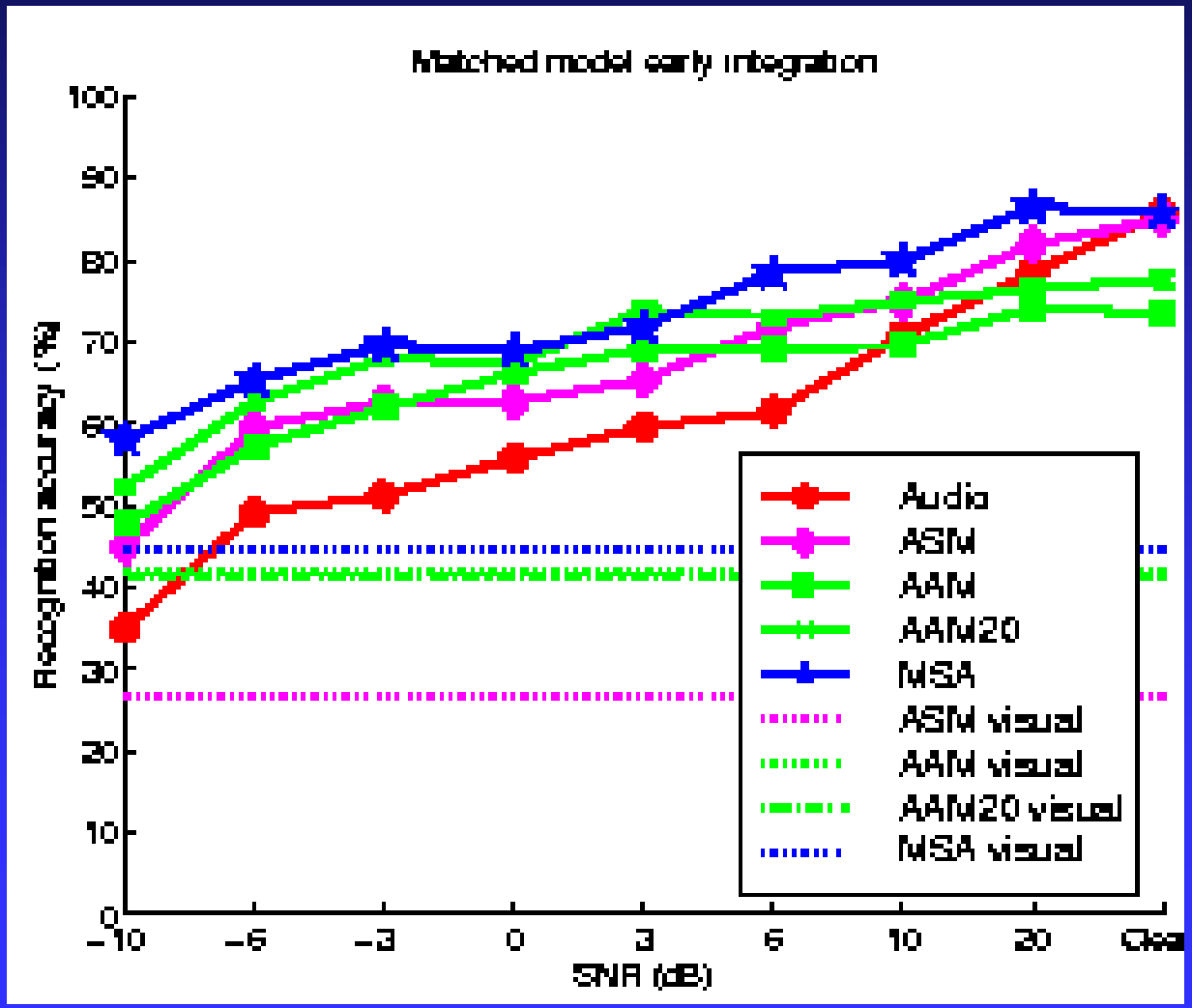
Early audiovisual integration

- SNR variation
- Matched condition
  - Trained at each SNR – tested at each SNR
  - Best possible performance
- Unmatched condition
  - Trained at highest SNR and tested at all SNRs

Early audiovisual integration

- Visual methods tested: MSA, ASM & AAM
- Measure of succes: improvement in performance *compared to audio only*
- Conclusion
  - AAM best with degraded audio, but *lower* than audio only at highest SNR
  - MSA and ASM gains only (and very little) whith more HMM Gaussian modes than possible with audio only (due to less training data).
- Overall: Poor performance

Early audiovisual integration

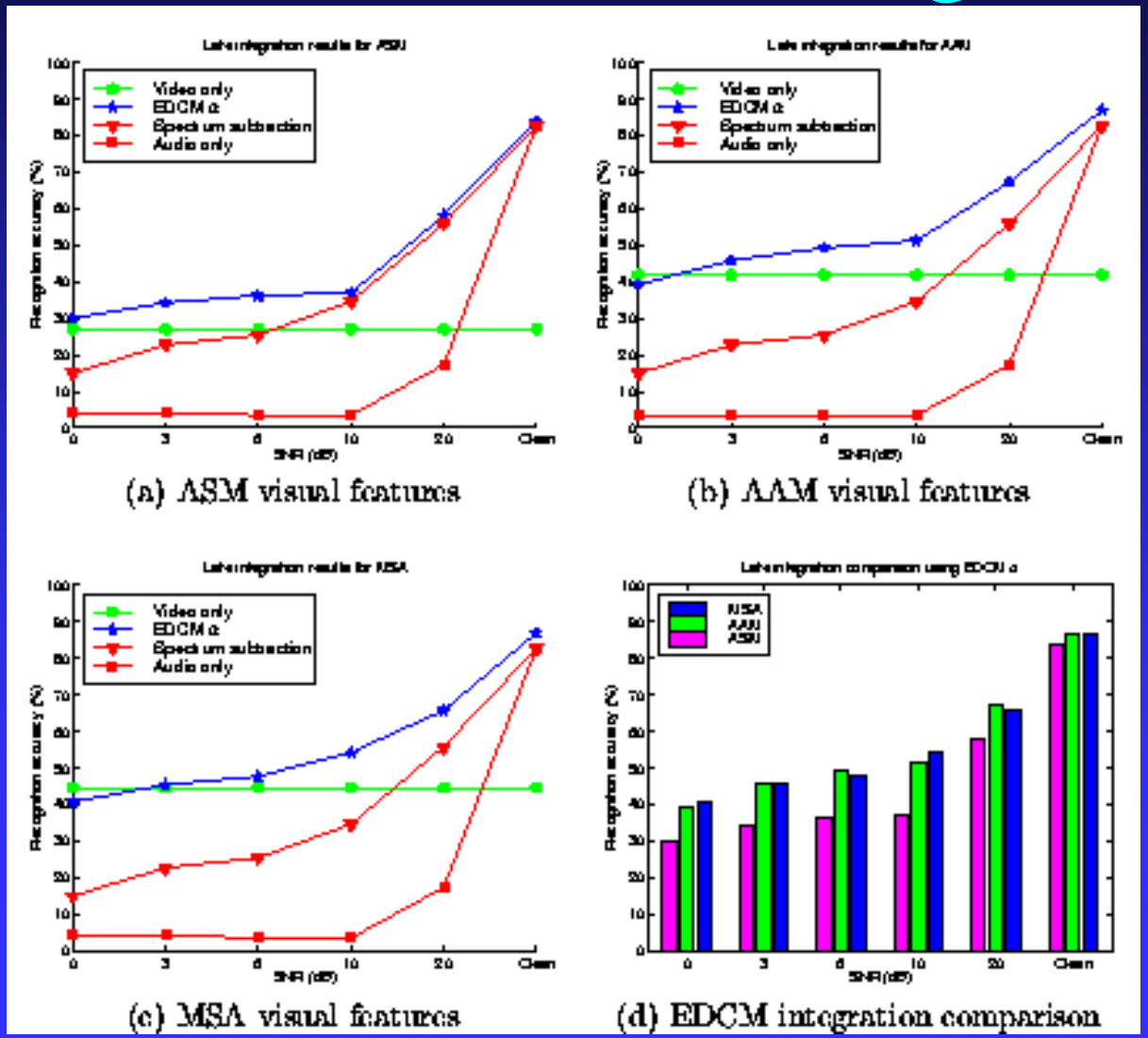


Late audiovisual integration

- Integration rule

$$w^* = \underset{i \in 1,2,\ldots,V}{\arg\max} \{\alpha L(w_i|A) + (1-\alpha)L(w_i|V)\}$$

Late audiovisual integration



(a) ASM visual features

(b) AAM visual features

(c) MSA visual features

(d) EDCM integration comparison

Homework

- Derive Bayes' rule for two conditions
  - P(percept | Audio,Visual)=?
- Compare Matthew's late integration rule with that of Fritsch et al.'s (Katagiri, ch. 7.4)
  - Is any (or both) statistically optimal (Bayesian)? Justify.
  - How does each respond to incongruent (wrongly dubbed) audiovisual speech?
- The McGurk effect was demonstrated during this talk. It is an illusion occuring from incongruent audiovisual speech. It may seem mysterious why the brain would allow such an illusion. Give a reason for why this might not be a mystery.