FULL FACE AUDIO-VISUAL SPEECH RECOGNITION

Benjamin X. Hall, John Shawe-Taylor and Alan Johnston

OVERVIEW

Automatic Speech Recognition:

-Process of turning acoustic speech into words

OVERVIEW

Automatic Speech Recognition:

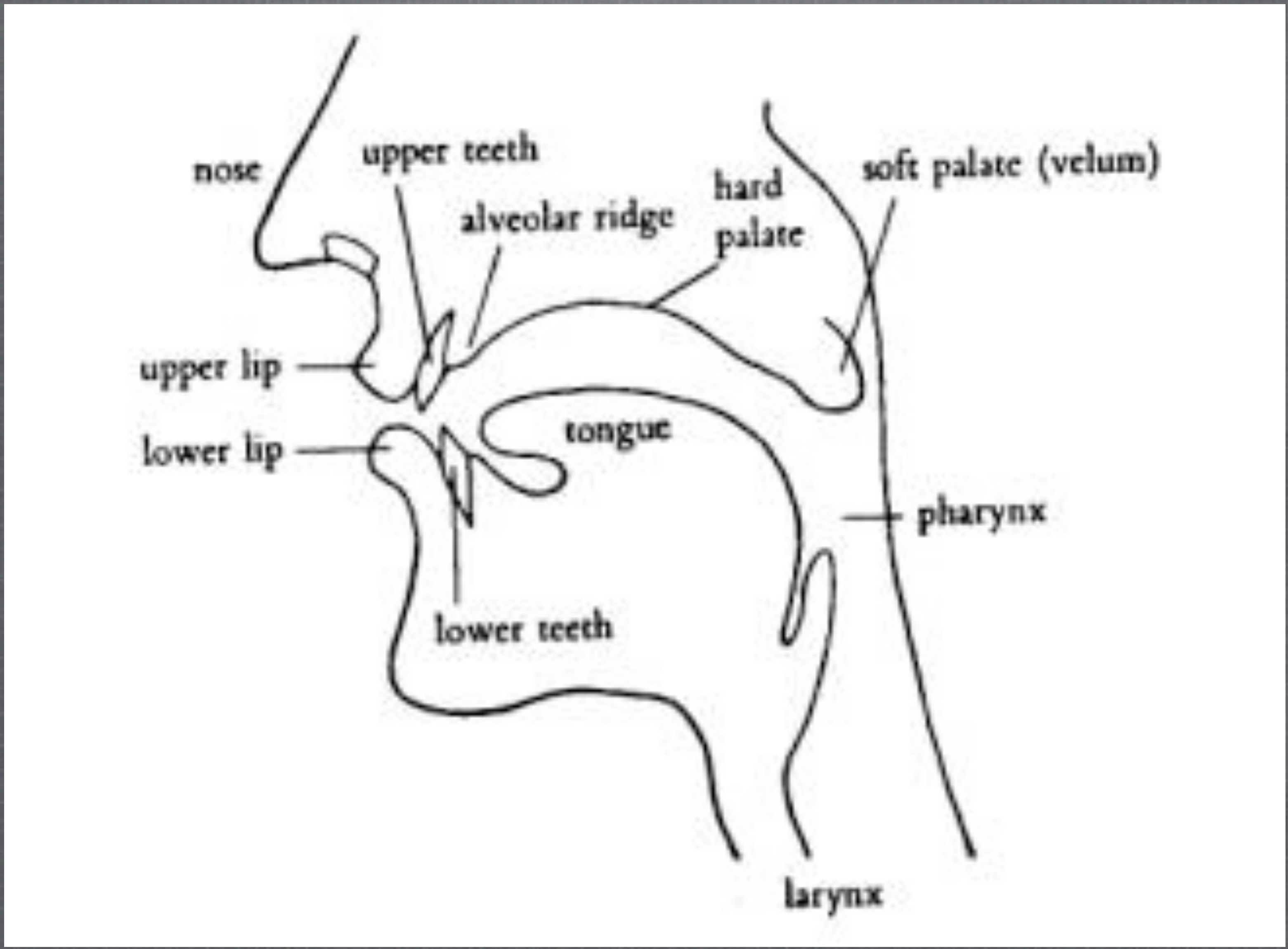
-Process of turning acoustic speech into words

Matured Technology

-HMMs
-Commercialised
-Plateaued

Siri, Android Voice Search, Dragon

VISUAL MODELS



OVERVIEW

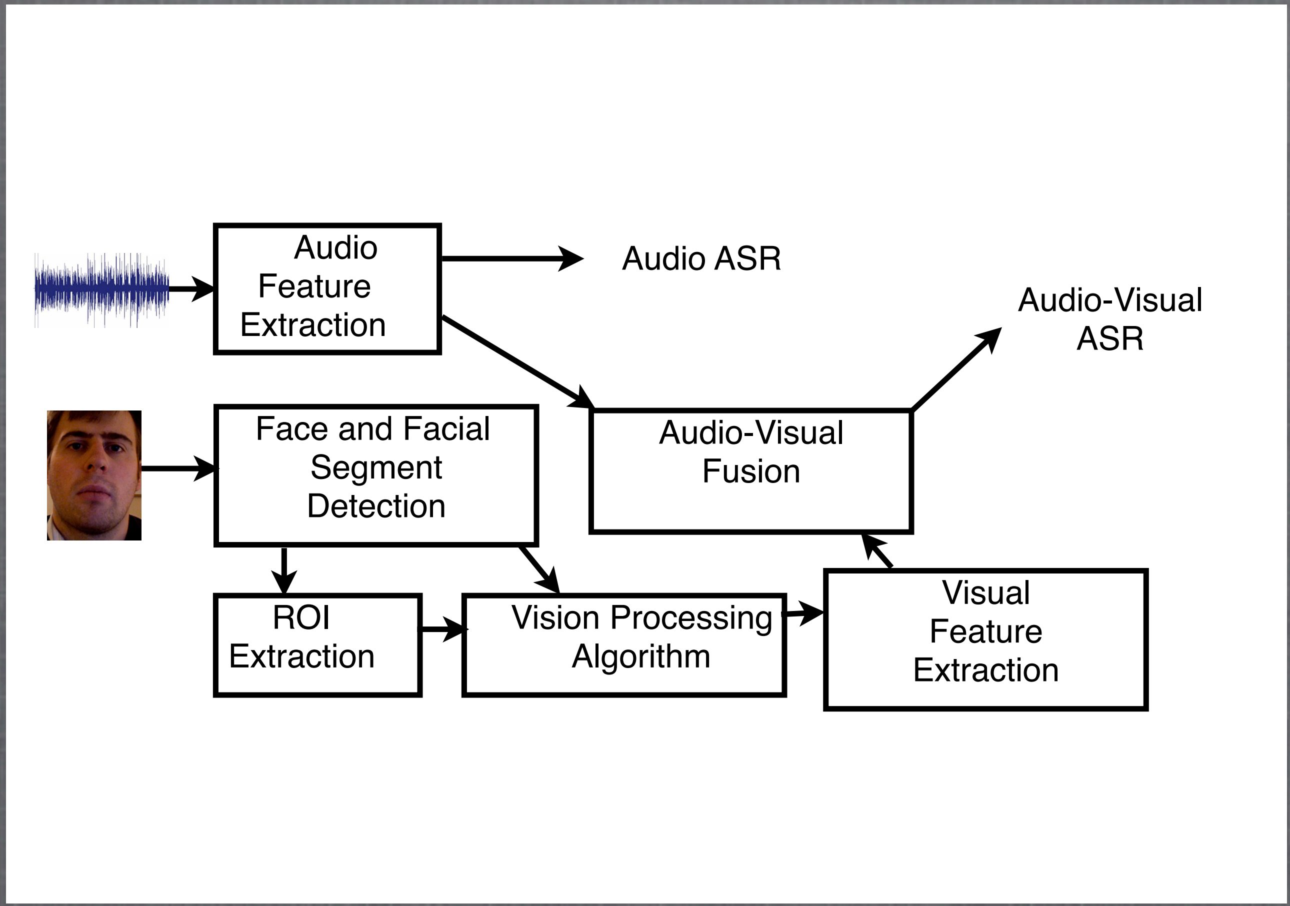
Automatic Speech Recognition:

- -HMMs
- -Commercialised
- -Plateaued

Visual Automatic Speech Recognition:

- -Inclusion of Visual Information
- -Fused with audio

VISUAL MODELS



VISUAL ALGORITHMS

Two broad categories :

-Shape based models

-Lip models

VISUAL ALGORITHMS

Two broad categories :
- -Shape based models
  - -Lip models
- -Appearance based models
  - -DCT type II

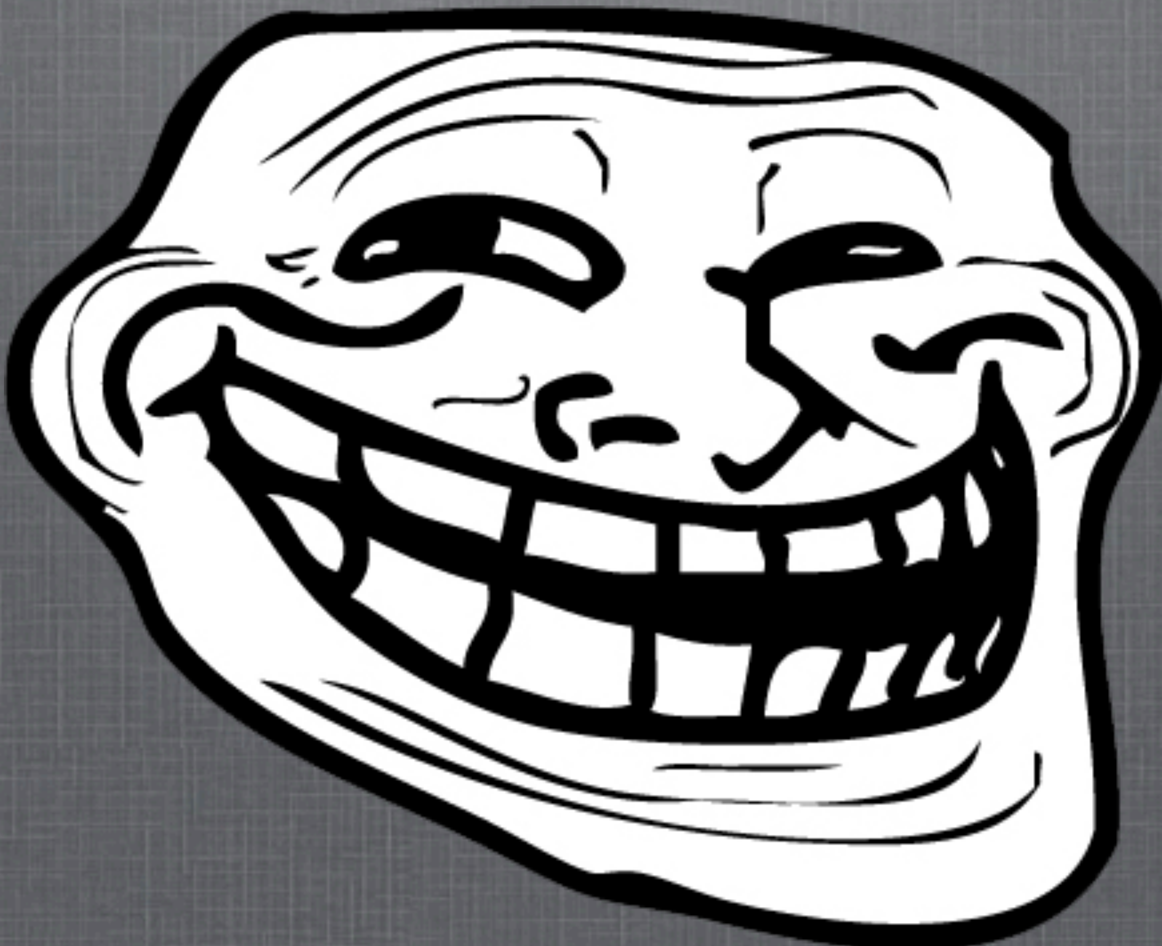
$$X_k = \sum_{n=0}^{N-1} x_n \cos[\frac{\pi}{N}(n+\frac{1}{2})k] \qquad k = 0,\ldots,N-1$$

EMPLOYED TROPES

Framing

Intense Visual Focus

PROBLEMS?



PROBLEMS?

Intolerant to visual occlusions

PROBLEMS?

Intolerant to visual occlusions

Implementation:
- Webcams
- Fixed Focus CCDs

No optical zoom

DISSONANCE

Contrast to human understanding

DISSONANCE

Contrast to human understanding

Jordan and Sergeant demonstrated Visual Speech understanding is exhibited at distances too great for teeth, tongue and mouth positions to be clearly definable.

DISSONANCE

Contrast to human understanding

Jordan and Sergeant demonstrated McGurk effects

are exhibited at distances too great for teeth, tongue and mouth positions to be clearly definable.

Preminger et al. selectively masked aspects of the face during speech production and observed visual speech understanding

BERISHA'S WORK

BERISHA ET. AL

SOLUTIONS?

Multi Channel Gradient Model

Derived from investigation into ratio-conditioning problem

$$I(x+dx, t+dt) = I(x,t) + \frac{\delta I(x,t)}{\delta x}dx + \frac{\delta I(x,t)}{\delta t}dt + O(dx^2, dt^2)$$

SOLUTIONS?

$$\frac{\delta I(x,t)}{\delta t}$$

$$\frac{\delta I(x,t)}{\delta x}$$

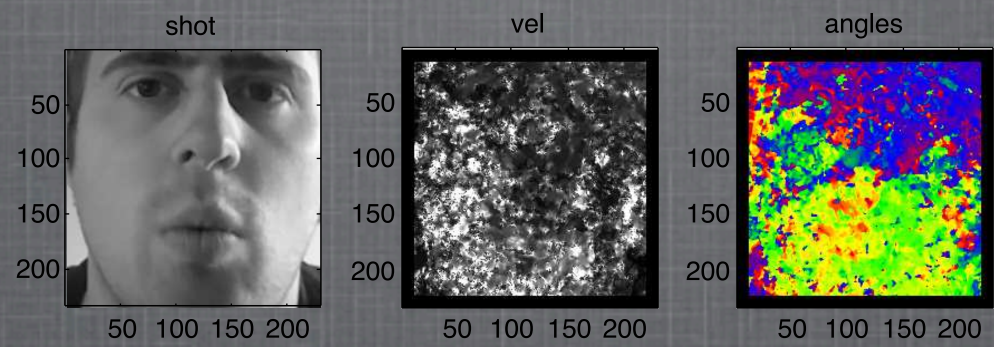
$$(D_t \ I, \ D_{tx} \ I \ ,..., \ D_{t(n-1)x} \ I \ )$$

$$(D_x \ I, D_{xx} \ I \ , ..., Dn_x \ I \ )$$

$(v'X - T)$ which requires $v' = (XT/XX)$

$$v' = \frac{\sum_n \frac{\delta^n}{\delta x^n} I \frac{\delta^{n-1}}{\delta x^{n-1}} \frac{\delta}{\delta t} I}{\sum_n \frac{\delta^n}{\delta x^n} I \frac{\delta^n}{\delta x^n} I}$$

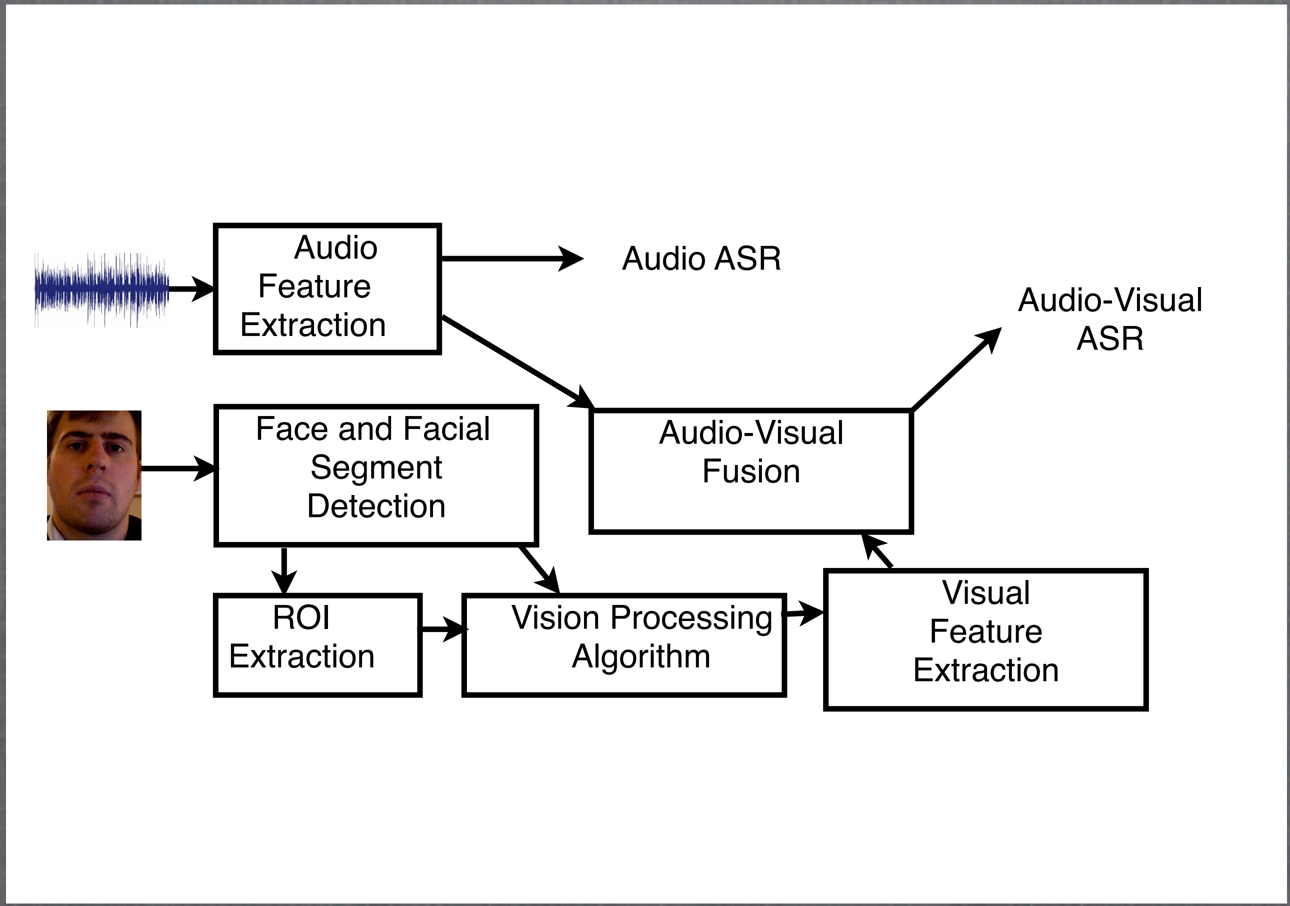
VIDEOS



VIDEOS

videos of MCGM

VISUAL MODELS



OCCLUSIONS

FORMULATION

Sound waveform is turned into MFCCs

From MCGM:
Angular Information and Speed are mapped onto velocity
Linear PCA

Late Fusion HMMs to classify

TESTING

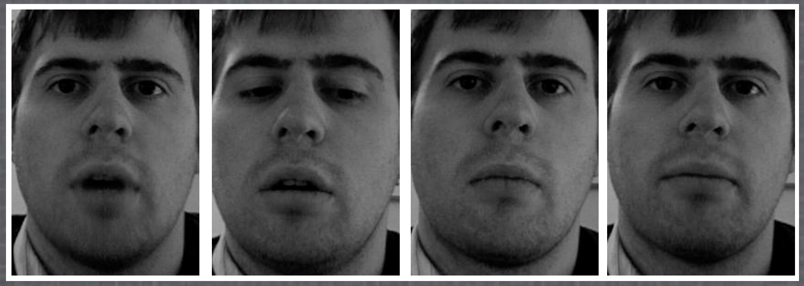
Testing is against DCT Type-II
Using similar pipeline as before

Required because new database

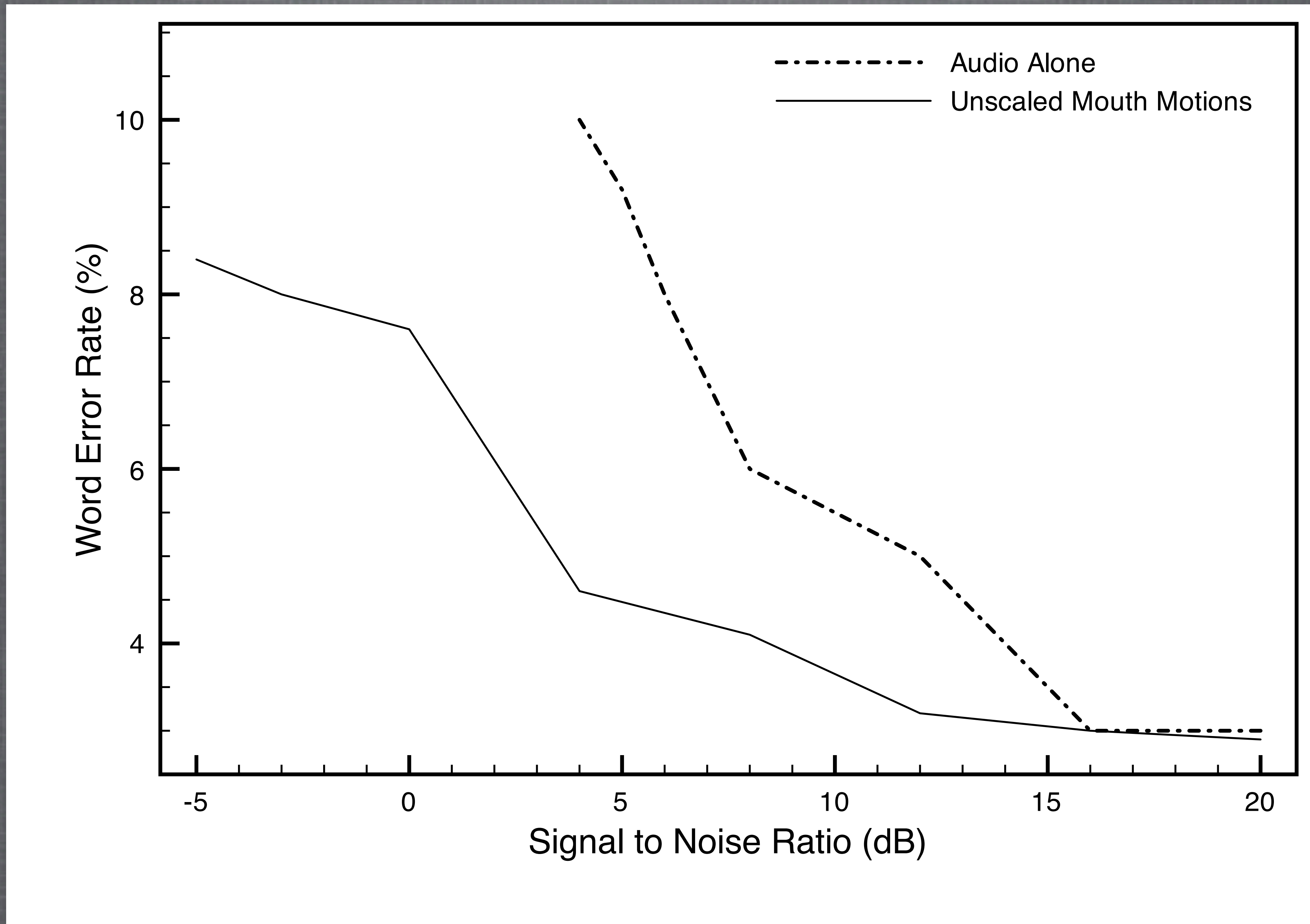
DATABASE

None of the existing ones did what we needed

Single Speaker, Full Face, Simple Words

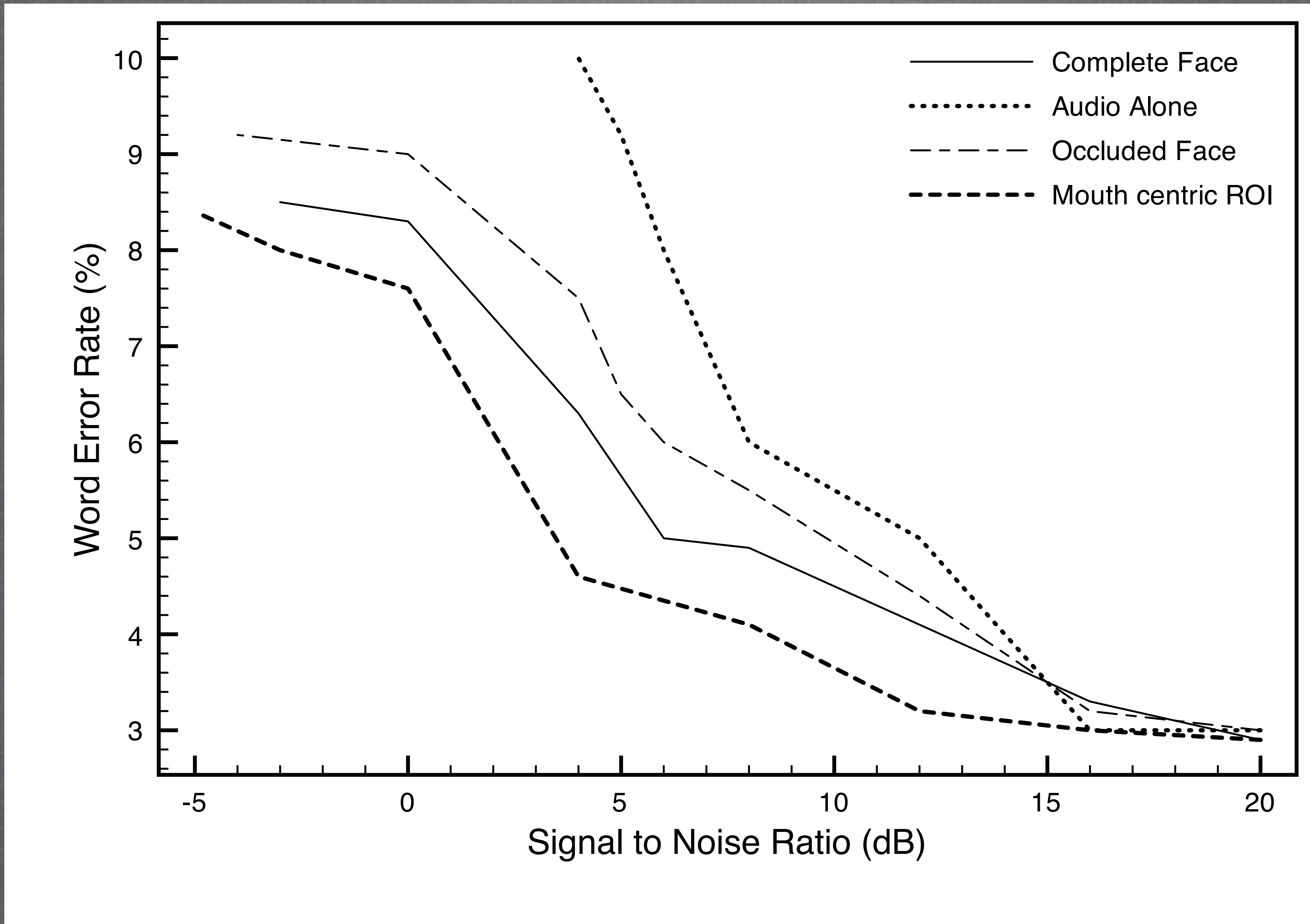


RESULTS



Audio Alone
Unscaled Mouth Motions
Word Error Rate (%)
10
8
6
4
-5
0
5
10
15
20
Signal to Noise Ratio (dB)

RESULTS



Complete Face
Audio Alone
Occluded Face
Mouth centric ROI
Word Error Rate (%)
10
9
8
7
6
5
4
3
-5
0
5
10
15
20
Signal to Noise Ratio (dB)

CONCLUSIONS

Possible to extract information from the surrounding face

Not as good as DCT type-II, in optimal conditions

FUTURE WORK

Expand database, currently only single speaker

Improved feature selection, at the moment very basic

REFERENCES

J E Preminger, H B Lin, M Payen, and H Levitt. Selective visual masking in speechreading. *Journal of Speech, Language and Hearing Research*, 41(3):564–575, 1998.

T.R. Jordan and P. C. Sergeant. Effects on visual and audiovisual speech recognition. *Lang. Speech*, 43:107–124, 2000.

F. Berisha, A.Johnston, and P. McOwen. *Facial Mimicry*. PhD thesis, University College London, 2006.

Thanks to ESPRC, UCL, UCL CS, CoMPLEX, PASCAL,

MORE?

MOUTH SCALING