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
OVERVIEW

Automatic Speech Recognition:
- HMMs
- Commercialised
- Plateaued

Visual Automatic Speech Recognition:
- Inclusion of Visual Information
- Fused with audio
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

\[ X_k = \sum_{n=0}^{N-1} x_n \cos\left[ \frac{\pi}{N} \left( n + \frac{1}{2} \right) k \right] \quad k = 0, \ldots, N - 1 \]
EMPLOYED TROPES

Framing

Intense Visual Focus
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

Thursday, 20 October 11
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?

\[ \frac{\delta I(x, t)}{\delta t} \quad \frac{\delta I(x, t)}{\delta x} \]

\((D_t I, D_{tx} I, \ldots, D_{t(n-1)x} I)\) \quad \((D_x I, D_{xx} I, \ldots, D_{nx} I)\)

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

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

videos of MCGM
VISUAL MODELS

Audio Feature Extraction ➔ Audio ASR

Face and Facial Segment Detection ➔ Audio-Visual Fusion

ROI Extraction ➔ Vision Processing Algorithm ➔ Visual Feature Extraction

Audio-Visual ASR
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

![Graph showing Word Error Rate (%)](image)

**Signal to Noise Ratio (dB)**

- Complete Face
- Audio Alone
- Occluded Face
- Mouth centric ROI

**Thursday, 20 October 11**
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


Thanks to ESPRC, UCL, UCL CS, CoMPLEX, PASCAL,
MORE?
MOUTH SCALING

Raw image data is downscaled to a variable pixelgrid.

The pixelgrid is then rescaled to manageable dimensions, as so to be used in the HMM framework.

Examples of grid sizes:
- 3x6
- 5x8
- 6x11
- 8x15
- 12x22