Audio-Visual Speech Processing

COM 4110 / COM 6070

Jon Barker

j.barker@dcs.shef.ac.uk

http://www.dcs.shef.ac.uk/~jon

Department of Computer Science

University of Sheffield

General Information

Lectures: Weeks 1-5, 7-11

- Thursday, 9:00-9:50, St George, LT04

Assessment:

- A single practical assignment (70% of the marks).
  - Details will be made available in week 4.
- A short written question (30% of the marks).
  - Available week 9.

Web page:
http://www.dcs.shef.ac.uk/∼jon/campus_only/teaching/COM4110/

Resources

Books:

- Stork, D. G. & Hennecke, M. E. (Eds.) (1996). *Speechreading by Humans and Machines: Models, Systems and Applications*. Springer-Verlag,
- Dodd, B. & Campbell, R. (Eds.) (1987). *Hearing by Eye: The Psychology of Lip-Reading*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- R. Campbell, B. Dodd & D. Burnham (Eds.) (1998). *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual Speech*. Psychology Press Ltd. New York.

Papers:

- Most lectures will focus on a few key papers.
- These papers can all be downloaded from the COM4110 web page.

Resources

Web Links: Some good starting points.

- **Haskins 'Talking Heads' Web Site**
  An overview of audio-visual speech processing research.
  http://www.haskins.yale.edu/haskins/heads.html
- **UCSC Perceptual Science Lab Speechreading Links**
  Links to automatic speechreading research sites.
  http://mambo.ucsc.edu/psl/lipr.html
- **AVISA** - Homepage of the Auditory-Visual Speech Association.
  http://www.uws.edu.au/marcs/AVISA/AVISA.htm

Lecture 1: Introduction to Audio-Visual Processing

Objectives

- To motivate the study of audio-visual processing.

Topics

- What is audio-visual speech processing.
- An introduction to audio-visual research themes and applications.
- Reasons for the sudden growth in interest in AV speech research.

Reading

- *Issues in audiovisual spoken language processing (when, where and how?)*, Bernstein, Burnham and Schwartz, Proc ICSLP 2002

Audio-Visual Processing Technologies

AV processing technologies aim to exploit visual information to improve the **robustness** and **naturalness** of audio-based processing in adverse acoustic conditions.

Example contexts:

- **Audio-Visual Speech Recognition** (MMSP99, ICASSP00, ICME00)
  Combine visual speech (visemes) with audio speech (phones) recognition.
- **Audio-Visual Speaker Recognition** (AVSP99, MMSP99, ICME00)
  Combine audio signatures of a person with visual signatures (face-id).
- **Audio-Visual Speaker Segmentation** (RIAO00)
  Combine visual scene change with audio-based speaker change.
- **Audio-Visual Speech-Event Detection** (ICASSP00)
  Combine visual speech onset cues with audio-based speech energy.
- **Audio-Visual Speech Synthesis** (ICME00)

Audio-Visual Speech Processing Applications

Some example applications:

- Audio-visual **speaker recognition** for security applications.

  Face recognition and voice recognition have limited reliability. Audio-based speaker recognition can be fooled by impersonation. Systems which identify a person based on both how they sound and how they look may have a much higher reliability.

- Audio-visual **speaker tracking** for interactive robots.

  Visually tracking a moving speaker is prone to error due to visual occlusion. This problem can be solved using sound localisation.

- Audio-visual **speech synthesis** for computer games and CGI characters.

  High quality facial synthesis of speaking characters needs a good deal of knowledge about the correspondences between audio and visual aspects of speech. Even minor mismatches between the visual and acoustic cues are readily apparent and give the animations an unnatural appearance.

Audio-Visual Speech Processing Applications

Further example applications:

- **Audio-driven lip-morphing** for cross-language film dubbing.

  When films are dubbed into foreign languages the asynchrony between the lip movements and the speech can be very distracting. By manipulating the lip images, the lips of the speaker can be transformed to match the acoustics of the new language while maintaining a realistic visual appearance.

- Audio-visual **speech enhancement**.

  Visual speech cues can be used to help segregate acoustic speech signals from background noise. Such techniques can be a useful preprocessing stage for automatic speech recognition, or can be employed to clean speech signal before transmission across telecommunication networks, or they can be used in conjunction with hearing aids such as cochlear implants.

- Audio-visual **speech recognition** for deployment in noisy environments.

  Automatic speech recognition can benefit from visual speech signals at a number of levels (we will focus on this application in the later lectures).

Recent History of AV Speech Processing Research

- **1995** - NATO Advanced Study Institute meeting, 'Speechreading by Man and Machine: Models, Systems and Applications'
- **1996** - Auditory-Visual Speech Processing symposium at ICSLP'96 in Philadelphia.
- **1997** - 1st AVSP conference, AVSP '97 satellite conference of Eurospeech '97. The AVSP conference has become an annual event.
- **1998** - Founding of the AVISA (Auditory-Visual Speech Association) as an ISCA special interest group, to promote interest and activity in auditory-visual speech processing.
- **2000** - John Hopkins, CSLP Workshop on Audio-Visual Speech Recognition

Growing number of audio-visual ASR papers appearing at major speech processing conferences. ICSLP, Eurospeech, ICASSP.

Growing number of AVSP research groups springing up.

Audio-Visual Automatic Speech Recognition: Motivation

- Lack of ASR robustness to noise / mismatched training-testing conditions despite decades of research.
- Humans speechread to increase intelligibility of noisy speech.
- Humans involuntarily fuse audio and visual information in speech recognition.
- Hearing impaired people can speechread extremely well.
- Visual and audio information are partially complementary:

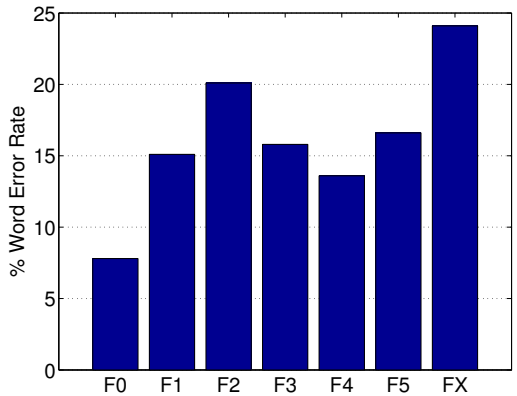
  i.e. Acoustically confusable phonemes are often visually distinct.
- Falling cost of capture, storage and realtime processing of visual information.

Automatic Speech Recognition: State of the Art

Woodland et al, 99

Results of the HTK Broadcast news transcription system.

- **F0** Prepared
- **F1** Conversation
- **F2** Telephone
- **F3** Background music
- **F4** Background noise
- **F5** Non-native speakers
- **FX** All other - e.g. combinations



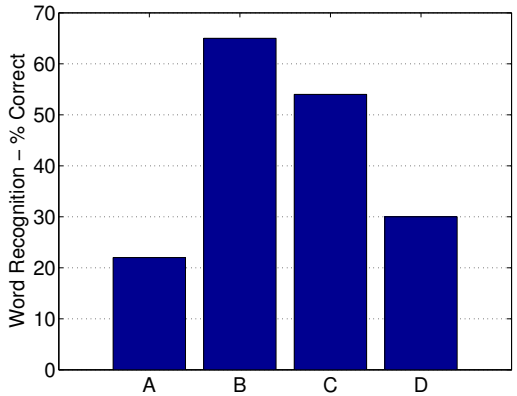
These were among the best results at the 1998 DARPA Broadcast news evaluations.

Human Audio-Visual Speech Recognition

Summerfield, 79

Human transcription of audio-visual speech at SNR < 0 dB

- **A** - Acoustic Only
- **B** - Acoustic + full video
- **C** - Acoustic + lip region
- **D** - Acoustic + 4 lip-points



Visual-Only ASR Performance

Digit/Alpha Recognition

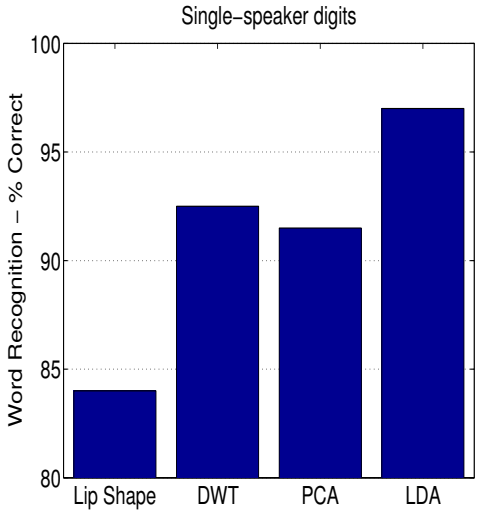
Potamianos et al. (1998)

- Single-speaker digits = 97 %
- Single-speaker alphas = 71 %
- 50-speaker alphas = 37 %

Phoneme Classification

Potamianos et al. (2000)

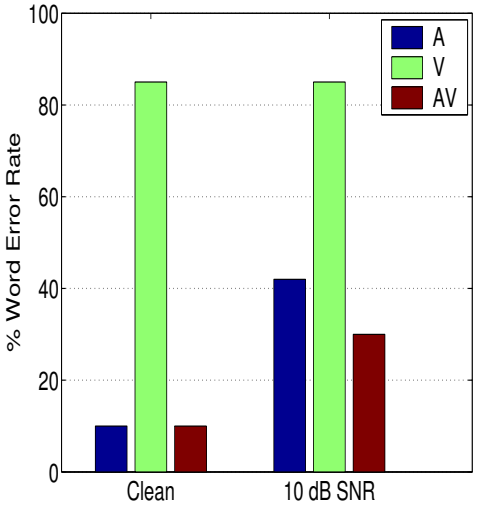
- 162-speakers, LVCSR = 37.3 % phone recognition



AV Large Vocab Continuous Speech Recognition

Large Vocab ASR *IBM (2000)*

- 5.5 hrs train, 162 speakers
- 1.5 hrs test, 162 speakers
- **Features**:
  - **Audio**: Cepstra+LDA+MLLT
  - **Video**: DWT+LDA+MLLT
  - **AV**: (Cepstra,DWT)+LDA+MLLT
- **Noise**:
  - speech noise at 10 dB SNR



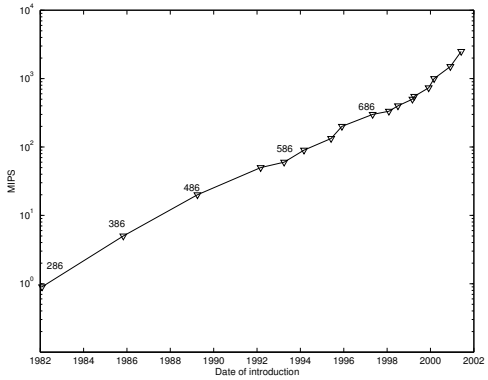
Falling Cost of AV Research + Applications

There are three main cost considerations:

- **Data capture**.
  - i.e. video cameras, a/d converts.
- **Data storage.**
- **Data processing.**
  - May be either realtime processing or offline processing depending on the application.

Exponential Increase in Processor Speed

The exponential rise in processing power of Intel processors:



The processing power of Intel processors has roughly doubled every 2 years since 1980, while the cost of the processors remains roughly constant.

c.f. data rate for 50Hz 600x800 24-bit RGB video = 72 Mbytes/s

Falling Cost of Storing Video Data

Many of the AV applications that have been mentioned require processing very large amounts of data.

This data may either be needed during the development of the system (e.g. training speech recognisers), or maybe central to the application (e.g. video indexing and retrieval).

Prohibitive storage costs have previously made the development, or deployment, of many AV processing applications commercially unattractive.

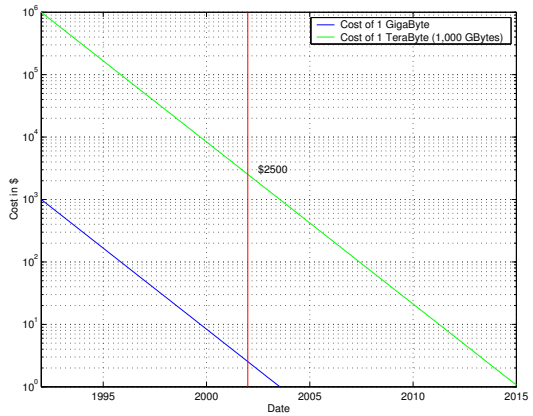
However, a dramatic fall in the cost of storing video data in a convenient form has removed this financial barrier.

The cost of storing video data has fallen due to two factors:

- **Falling cost of storage media.**
- **Increased sophistication of video compression.**

Falling Cost of Disk Storage

The cost of storing 1 terabyte of data on hard disk has roughly halved every year throughout the 1990's.



This trend appears set to continue for at least the next 10 years.

Video Compression

New video compression algorithms (e.g. MPEG) radically reduce storage requirements.

MPEG2 compression can compress video data by a factor of 30 to 40 without noticeable loss in video quality.

Example 1:

Consider uncompressed high quality digital TV data (CCIR-601 resolution 4:2:2 serial digital video)

- Requires 20 MB of storage per second. So a 2 hour film requires about 144 GB.
- With MPEG2 compression a 2 hour video + audio + subtitles can be stored on a single 4.7 GB DVD.

Example: Cost of Storing an AV Speech Corpus

Example 2:

Consider a 100 hour high quality AV speech corpus.

Uncompressed this may require around 7 terabytes of storage.

- 10 years ago this would cost around $7,000,000!
- Today storing the uncompressed data would cost less than $10,000.
- With MPEG2 compression it can be stored in <240Gb. 250 GB disk drives can be bought for less than £70!

Video processing that can be performed in real time on a modern computer, would take around 2 days per hour of data on an equivalently priced machine 10 years ago.

Audio-Visual ASR Challenges

- **Suitable databases are only just emerging:**
  - AT&T (*Potamianos, 1997*): 1.5 hr, 50 subjects, connected letters.
  - Surrey (*Pigeon, 1997*): M2VTS (37 subjects, connected digits).
  - IBM (*Basu et al., 1999*): LVCSR AV database (50 hrs, >260 subjects).
  - Clemson (*Patterson, 2002*): CUAVE (connected digits).
  - Sheffield, GRID (2005): Small vocabulary, 36 subjects, 20 hours
- **Design of the visual front end.**
  - Face and lip/mouth tracking.
  - Visual features suitable for speechreading.
- **How best to fuse the audio and visual information?**
  - Feature fusion or decision fusion?
  - Integration level: state, phone or word?

Publicly Available Audio-Visual Speech Databases

A survey of 'Natural Interaction Multimodality' data resources published Feb, 2002 (*http://isle.nis.sdu.dk/reports/wp8/*).

Unfortunately, there are still no publicly available database suitable for large vocabulary speaker independent AV ASR.

Introduction: Course Overview

- **Lecture 1:** Introduction.
- **Lecture 2:** The perceptual basis for AV speech processing.
- **Lecture 3:** Image Processing using MATLAB.
- **Lectures 4 & 5:** Face detection and tracking.
- **Lecture 6 & 7:** Visual feature parameterisation.
- **Lecture 8 & 9:** Audio and visual feature integration.
- **Lecture 10:** Computational Auditory-Visual Scene Analysis (CAVSA).

2: The Perceptual Basis for AV Speech Processing

The lecture will examine how the brain exploits visual information in the perception of speech.

- Speechreading cues,
- Point light studies,
- Visemes versus phonemes.

How does the brain fuse auditory and visual information? Several key experiments will be considered:

- The McGurk effect
- The ventriloquist illusion
- AV point light studies

3: Image Processing using MATLAB

This lecture provides an introduction to the MATLAB numeric computation tool. The lecture provides sufficient MATLAB knowledge to tackle the course's practical assignment.

- Digital representation of images.
- A short introduction to MATLAB.
- Basic image processing.
- Details of the assessed practical assignment.

4&5: Face Detection

These lectures will present an examination of the various methods that may be employed to detect faces in arbitrary scenes. Including:

- Fisher faces,
- Eigen-faces,
- Neural networks

The challenges in dealing with lighting and pose variation will be discussed.

6&7: Visual Feature Parameterisation

These lectures will describe the three different type of visual feature that are employed in audio-visual speech processing:

- Low level video pixel based (such as image transform features).
- High level model-based features (lip shape and appearance modelling).
- Hybrid features, based on the combination of high and low level features.

The processing techniques that are employed to extract such features will be studied.

8&9: Audio and Visual Feature Integration

Successfully combining the information present in the audio and visual channels is paramount to improving robustness and performance of audio-only speech recognition.

These lectures will discuss a number of issues relevant to the "fusion" problem:

- Feature fusion versus decision fusion.
- Modelling audio-visual asynchrony.
- Source reliability estimation.

10: Computational Auditory-Visual Scene Analysis (CAVSA)

Auditory Scene Analysis (ASA)

- In most acoustic environments, we hear a mixture of sounds.
- How can we attend to one voice in a cocktail party, or attend to the violins in an orchestral recording?
- Listeners perform an auditory scene analysis:
  - The mixture of sounds is decomposed into a collection of sensory elements.
  - Elements that are likely to have arisen from the same source are grouped, forming a structure (stream), which can be interpreted by higher centres.

Over the last 10 years much work has been done to try and build computational models of ASA (Computational ASA, **CASA**). Very recently new work has emerged which has attempted to broaden the scope of ASA to encompass **visual** cues (**CAVSA**).

The final lecture will examine this work and the emerging research problems it poses.

References

- Basu et al. (1999) Audio-visual large vocabulary continuous speech recognition in the broadcast domain. In *Proc. IEEE 3rd Workshop on Multimedia Signal Processing*, pages 475–481.
- Patterson et al. (2002) CUAVE: A new audio-visual database for multimodal human-computer interface research, In *Proc. Int. Conf. on Acoustics, Speech and Signal Proc.*, Orlando.
- Pigeon and Vandendorpe (1997) The M2VTS Multimodal face database (Release 1.00), In *Proc. First International Conference on Audio- and Video-based Biometric Person Authentification*, pages 403–409, Crans Montana, Switzerland.
- Potamianos, Cosatto, Graf and Roe (1997) Speaker independent bimodal database for bimodal ASR, In *Proc. European Tutorial Workshop Audio-Visual Speech Processing*, pages 65–68, Rhodes.
- Potamianos, Graf and Cosatto (1998) An image transform approach for HMM based automatic lipreading, In *Proc. International Conference on Image Processing*, volume III, pages 173–177.
- Potamianos et al. (2000) A cascade image transform for speaker independent automatic speechreading. In *Proc. International Conference on Multimedia and Expo*, volume II, pages 1097–1100, New York
- Summerfield (1979) Use of visual information for phonetic perception. *Phonetica, 36*, 314–331.
- Woodland et al. (1999) The 1998 HTK broadcast news transcription system. In *Proc. DARPA Broadcast News Workshop*, Virginia.