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

Lecture 2: Audio-Visual Speech Perception

Objectives

- To examine how Humans process audio-visual speech.

Topics

- An overview of speech production.
- Identifying speechreading cues.
- Audio-visual fusion and the McGurk Effect.

Reading

- *Speech recognition and sensory integration*, Massaro & Stork, Am. Sci., 1998, 86:236-244. [http://www.amsci.org/amsci/articles/98articles/massaro.html].
- *The perceptual basis for audiovisual integration*, Rosenblum, Proc ICSLP, 2002

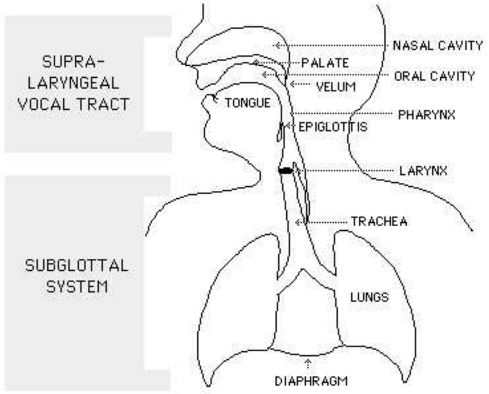
Lecture 2: Audio-Visual Speech Perception

Overview

- Speech Production.
- Visual Speech Perception.
  - ♦ Visemes.
  - ♦ Point-Light Studies.
  - ♦ Dynamic versus Static Features.
- Audio-Visual Speech Perception.
  - ♦ Combining the Audio and Visual Modalities.
  - ♦ The Ventriloquist Illusion.
  - ♦ The McGurk Effect.

Speech Production

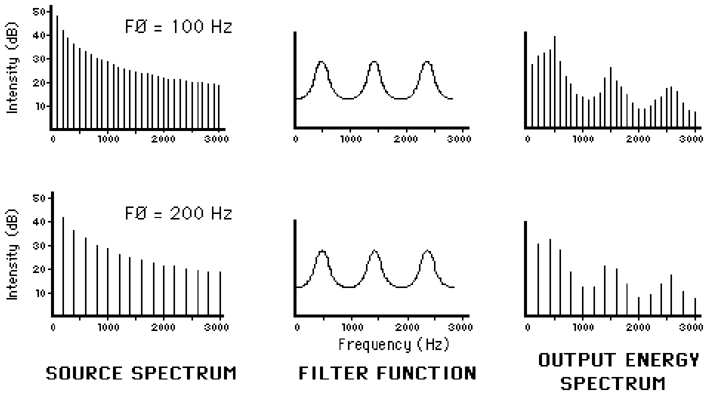
Speech production can be viewed as a **source/filter model**:



- A sound **source** is produced by subglottal system
- The source is **filtered** by supra-laryngeal vocal tract.

Speech Production - The Source/Filter Model

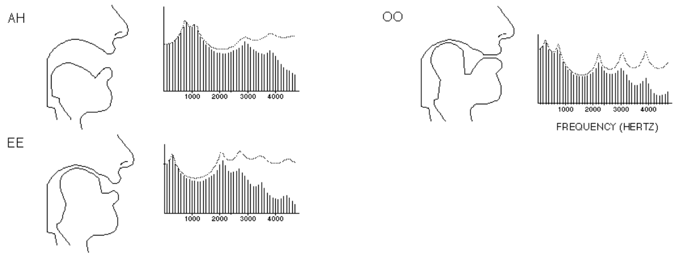
The figures below shows the effect the vocal tract filtering has on the source **frequency spectrum** of a segment of **voiced speech**.



The vocal tract **amplifies** some frequency regions and **attenuates** others.

Speech Production - Vowels

Vowels are determined by the frequency of the **vocal tract resonances (‘formants’)**.



The **frequency of the resonances** depends on the shape of the vocal tract, and hence the position of the **tongue** and the shape of the **mouth opening**.

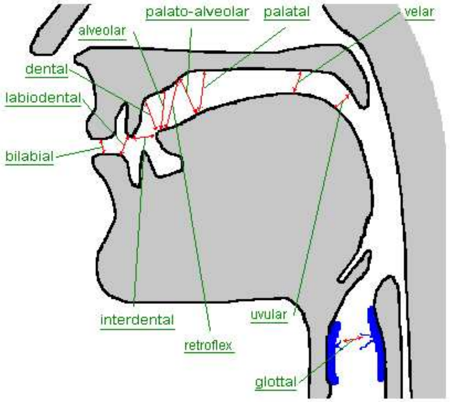
- Formant frequencies **decrease** with **lip rounding**, & **increase** with **spreading**.
- A **mouth constriction** lowers the first formant and raises the second formant.

So while the visible shape of the speaker’s lips does not uniquely determine the vowel, it does hold information about the vowel identity.

Speech Production - Consonants

Consonants are formed by creating **constrictions** in the vocal tract.

The identity of the consonant will depend on the where the constriction occurs (***‘the place of articulation’***).



The following places of articulation may be clearly visible:

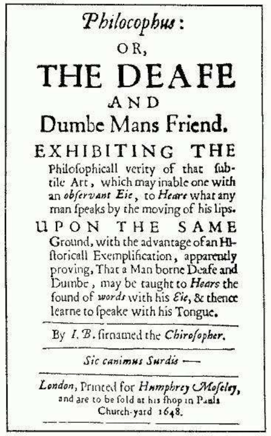
- **Bilabial** (**p** and **b**): The lips are momentarily closed.
- **Labiodental** (**f** and **v**): The upper teeth close against the lower lip.
- **Interdental** (**th**): The tongue is pushed between the teeth.

Audio-Visual Speech Perception: Lipreading

The study of lipreading has a long history.

“Exhibiting the philosophical verity of that subtle art, which may enable one with an observant eye, to hear what any man speaks by the moving of his lips.”

John Bulwer, “Philocophus” (1648)



Philocophus:
OR,
THE DEAFE
AND
Dumbe Mans Friend.
EXHIBITING THE
Philoſophicall verity of that ſubtile Art, which may inable one with an *obſervant Eie*, to *Heare* what any man ſpeaks by the moving of his lips.
UPON THE SAME
Ground, with the advantage of an Hiſtoricall Exemplification, apparently proving, That a Man borne Deafe and Dumbe, may be taught to *Hears* the ſound of *words* with his *Eie*, & thence learne to ſpeake with his Tongue.

By *I. B.* ſirnamed the *Chiroſopher*.

Sic canimus Surdis —

London, Printed for *Humphrey Moſeley*,
and are to be ſold at his ſhop in Pauls
Church-yard 1648.

Physician John Bulwer (fl. 1644-1662) is known for being the first Englishman to develop a method for communicating with the deaf and dumb.

Audio-Visual Speech Intelligibility

The importance of visual features has been quantified as far back as 1954:

Sumby and Pollack “Visual contribution to speech intelligibility in noise.” (1954)

Sumby and Pollack demonstrated that **untrained listeners** routinely employ visual information when trying to understand **speech in adverse conditions**.

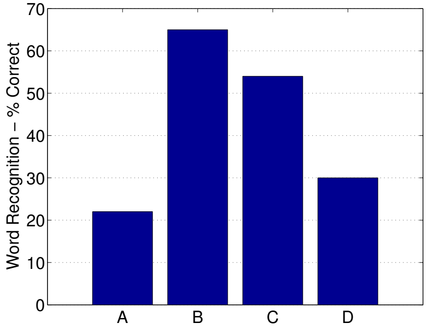
Their experiments showed that the addition of visual cues can increase **speech intelligibility** as much as **removing 15 dB of noise**.

It appears lipreading occurs at a **subconscious level**. It is something that is **learnt at a very early age**.

Audio-Visual Speech Intelligibility

More recent work has tried to determine the **precise nature** of the visual information that is being employed.

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



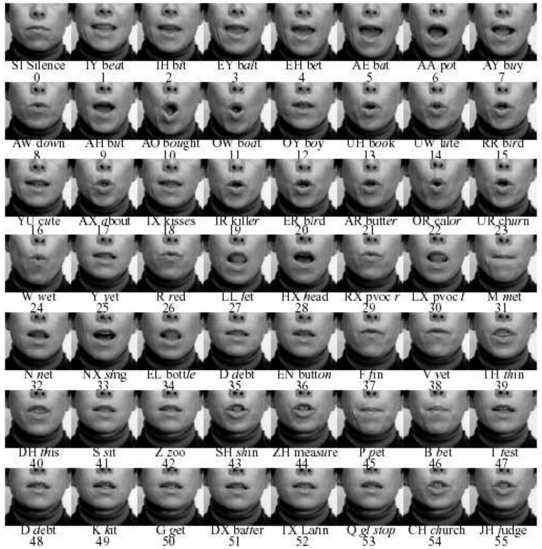
Summerfield (1979) studied the intelligibility of noisy audio-visual speech (SNR <0dB) under a number of **different visual conditions**.

There is useful information **outside the lip region**. Even four moving lip points can provide a significant increase in intelligibility.

Phoneme Mouth Shapes

Using large AV speech corpora and image processing techniques it is possible to determine the **average mouth shape** occurring at the midpoint of each phoneme.

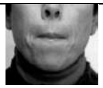
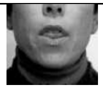
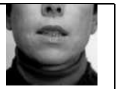
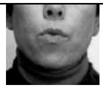
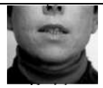
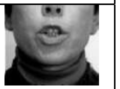
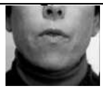
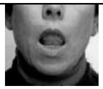
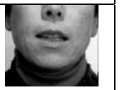
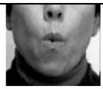
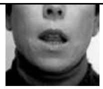
Some pairs of phonemes appear very **similar** (e.g. **p** vs. **b**), others pairs are easily **distinguishable** (e.g. **l** vs **r**).



Keith Waters and Thomas M. Levergood, DEC Tech Report CRL 93/4, Image reproduced by permission of the Cambridge Research Lab of Digital Equipment Corporation.

Visemes Versus Phonemes

Through perceptual studies phonemes can be arranged into **viseme** groups such that those in the same group are **visually indistinguishable**. Below is the phoneme to viseme mapping determined by *Summerfield (1987)* for **English consonants**:

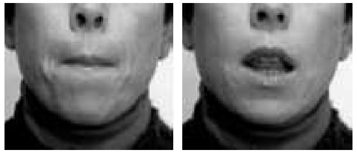
phonemes	viseme	phonemes	viseme	phonemes	viseme
p, b, m		th		s, z	
r		d,t		sh, zh	
f, v		l		y	
w		n, k, g			

A similar phoneme to viseme mapping can be made for vowels.

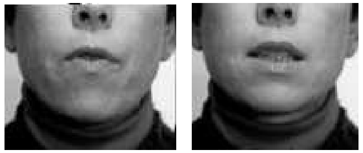
Audio-Visual Complementarity

Visemes provide information that **complements** the phonetic stream and therefore **reduces confusability**.

- Example 1: ***m*** (left) and ***n*** (right) are acoustically confusable but are visually distinct: in ***m*** lips close at onset, where as in ***n*** they do not.



- Example 2: The unvoiced fricatives ***f*** (left) and ***s*** (right) are also acoustically confusable but are likewise visually distinct.



What are the Most Important Visual Features?

If we want to build an audio-visual speech recognition system then we need to know **which visual features are most relevant** to the recognition task.

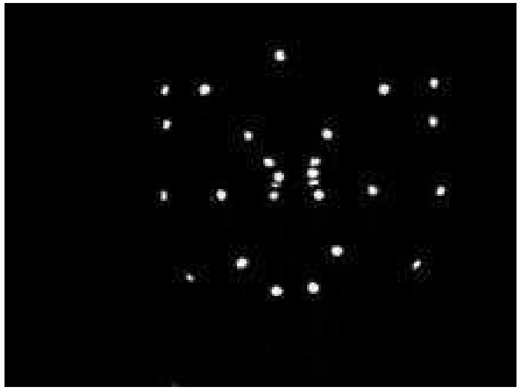
If it just lip motion? Is tongue visibility important? If there useful information in cheek movements?

An obvious approach to this question is to **analyse the response of Human speech recognition performance** under conditions where the visual information available is carefully controlled.

The amount of visual information can be controlled by using the '**point-light technique**' (largely pioneered by Summerfield).

The Point-light Technique

What is this? Random dots? A butterfly?



When the dots are moving, most people are able to correctly identify the image as the lower half of a face.

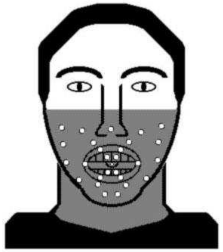
[http://www.faculty.ucr.edu/ rosenblu/lab-index.html]

The Point-light Technique

The point-light technique is very simple:

- **Florescent dots** are attached to the face of the speaker.
- **Lighting conditions** are arranged so that only the dots can be seen.

Here is the pattern used in the previous video sequence:

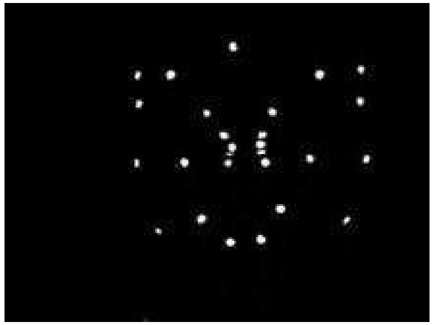


The point-light technique can help us isolate the specific information used in visual speech perception.

A Demonstration of the Point-light Technique

The video shows the utterance "*The tree fell on the house*" spoken three times, each time with an **increasing number of points present**:

1. Lips only,
2. Lips, teeth and tongue-tip,
3. Lips, teeth, tongue-tip, cheeks, jaw, nose and forehead.



Studies using this technique have shown that point-lights on the **lips**, **teeth** and **tongue-tip** are most useful for enhancing noisy speech.

Rosenblum, L.D. et al. (1996).*J. Sp. & Hear. Research 39(6), 1159-70.*

Higher-level Visual Cues

Note, that although lip and tongue tip movement are most directly correlated with the phonetic stream, Humans can make use of more general **high-level visual cues** to aid in speech understanding:

- **Head movements** - e.g. nodding in agreement, shaking,
- **Gestures** - hand gestures support speech,
- **Eye-brow movements** - reflect the speaker's intonation,
- **Facial expression**.

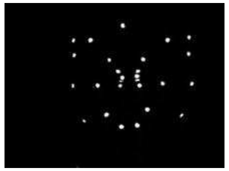
The use of such information to aid automatic speech recognition/understanding is a research area that has only very recently been addressed.

Dynamic Features Versus Static Features

What is more important for speech reading,

- the **static** configuration of the facial features,
- or the **dynamic** relative motion of the features?

This question is still openly debated.



One possibility is that **static** information is more useful for **vowel** recognition, and **dynamic** information is more reliable for **consonant** recognition.

We will return to this point in later lectures when considering the design of automatic AV speech recognition systems.

Audio-Visual Feature Integration

Where and how are the audio and visual modalities combined?

Old assumption: '**Primacy of auditory information**' - i.e. visual information is a 'backup' when we can't hear properly.

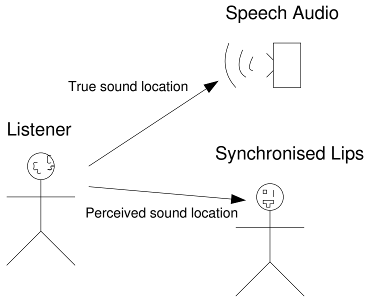
Over the last 40 years psychologists have come to realise that audio-visual integration is far **more subtle** than this.

Two major challenges to the 'Primacy of auditory information' assumption:

- **The Ventriloquist Illusion**
- **The McGurk Effect**

The Ventriloquist Illusion

The **ventriloquist illusion**, arises when people mislocate sounds towards their **apparent visual source**.

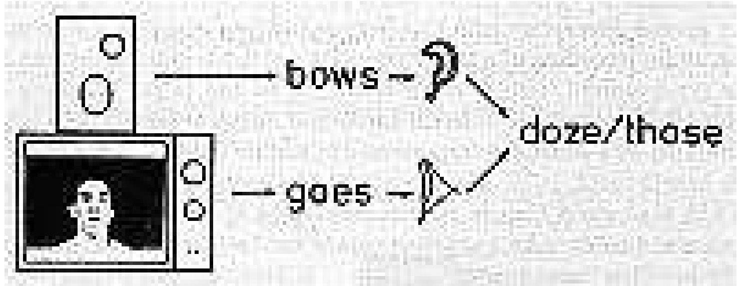


This applies powerfully for speech sounds and ***matching* lip movements**, e.g. on TV actors' voices appear to emanate from the screen rather than the loudspeakers, but the illusion breaks down for badly dubbed speech.

Witkin, H. A., Wapner, S. & Leventhal, *J. exp. Psychol. 43, 58-67 (1952) Bertelson, P and Radeau, M,* Perception and Pychophysics 19(6), 531-535 (1976)

The McGurk Effect

The **McGurk Effect** was first described by Harry McGurk and John MacDonald, 1976.



In the original set-up a *visual* **b** is synchronized with an *acoustic* **g**.

The listener *hears* a **d** or **th**.

McGurk and MacDonald "Hearing lips and seeing voices", Nature 264, 746-748 (1976).

The McGurk Effect: Demonstration

KUHL, P., McGurk Effect, Department of Speech and Hearing Sciences, University of Washington

The McGurk Effect: Demonstration

Depending on the audio and visual stimuli **one of the three following effects** may occur:

- The **visual** stimulus may **override** the **audio** stimulus,
- The **audio** stimulus may **override** the **visual** stimulus,
- The two **stimuli combine** to form a new phoneme.

Audio	+ Visual	-> Heard
BA	+ BA	-> BA
BA	+ VA	-> VA
BA	+ THA	-> THA
BA	+ GA	-> DA

Generality of the McGurk Effect

How general is the McGurk effect?

- It works on perceivers with **all language backgrounds** (e.g. Massaro et al. 1993)
- It works on **young infants** (Rosenblum, Schmuckler, & Johnson, 1997).
- It works when observers are **unaware** that they are looking at a face (Rosenblum & Saldaña, 1996).
- It works when observers **touch** (rather than look-at) the face (Fowler & Dekle, 1991).
- It works **less well with vowels** than consonants (Summerfield & McGrath, 1984).

The McGurk Effect - Interpretation

What does the effect say about speech perception?

- **Visual articulatory information is integrated** into our perception of speech **automatically and unconsciously**.
- The syllable that we perceive depends on the **strength** of both the auditory and visual information.
- Integration of the discrepant audiovisual speech syllables is **effortless and mandatory**.
- Our speech function makes use of all types of relevant information, **regardless of the modality**.

There is some evidence that the brain treats visual speech information *as if it **is** auditory speech*.

The McGurk Effect as a Tool

The McGurk Effect provides an excellent **tool** for studying audio-visual data fusion.

The experimental procedure is a follows:

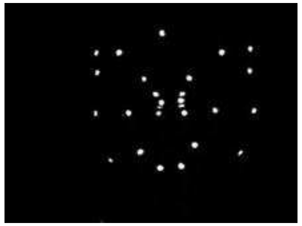
1. We start with the standard McGurk AV stimulus,
2. We manipulate the audio and/or visual components,
3. We measure the strength of the resulting McGurk effect.

Hence, we can **isolate the characteristics** of the visual speech that are **important for AV integration**.

For example, we can use the McGurk effect to re-examine the differences between **dynamic** and **static** visual features.

The McGurk Effect and Point-Light Stimuli

Can the McGurk effect be achieved using point-light video sequences?



- **Dynamic** point-light video **can** induce the McGurk effect.
- However, **static** images **do not** induce a McGurk effect.

Such results suggest that AV integration relies on **correlated dynamics** of the video and audio information.

Rosenblum, L.D. and Saldaña, H.M. (1996). *J. Exp. Psy. 22(2), 318-31*

Summary

- Studies of Human **AV speech perception** are a good place to start when considering the design of an Automatic AV speech recogniser.
- The face provides strong **visual clues** to both consonant identity (lip and tongue motion) and vowel identity (mouth shape).
- We can form phonemes into groups that are visually indistinguishable - **visemes**. Acoustically confusable phonemes are often visually distinct.
- **Point-light studies** can be used to try and locate the most important visual features. Points on the lips, tongue tip and teeth are all important.
- The brain *subconsciously* fuses the visual and acoustic speech information - This is demonstrated by **the McGurk Effect**.
- McGurk effect studies suggest this **AV fusion** relies on the *correlated dynamics* - i.e. the synchrony - of the visual and acoustic feature stream.

Lecture 3 Preview: Image Processing with MATLAB

The next lecture will examine how visual data may be represented digitally.

The lecture will provide an introduction to MATLAB, a 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.

References

- Bertelson and Radeau (1976) Ventriloquism, sensory interaction, and response bias: Remarks on the paper by Choe, Welch, Gilford, and Juola. *Perception and Psychophysics*, 19(6), 531–535.
- Fowler and Dekle (1991) Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology*, 17, 816–828.
- McGurk and MacDonald (1976) Hearing lips and seeing voices. *Nature* 264:746-748, 1976.
- Massaro et al. (1993) Bimodal speech perception: An examination across languages *Journal of Phonetics*, 21, 445–478.
- Massaro and Stork (1998) Speech Recognition and Sensory Integration, In *American Scientist*, 1998, 86:236-244. [http://www.amsci.org/amsci/articles/98articles/massaro.html].
- Rosenblum (2002) The perceptual basis for audiovisual integration, Rosenblum, In *Proc. International Conference on Speech and Language Processing*, 2002.
- Rosenblum, L.D. and Saldaña, H.M. (1996). An audiovisual test of kinematic primitives for visual speech perception *Journal of Experimental Psychology. 22(2), 318-331.*
- *Rosenblum, Schmuckler and Johnson (1997) The McGurk effect in infants. Perception and Psychophysics, 59(3), 347–357.*
- *Rosenblum, Schmuckler and Johnson (1996) Visual kinematic information for embellishing speech in noise. Journal of Speech and Hearing Research 39(6), 1159–1170.*

References - continued

- Sekiyama and Tokhura (1993) Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, 21, 427–444.
- W. H. Sumby and I. Pollack (1954) Visual contribution to speech intelligibility in noise. *JASA*, 26:212-215, 1954.
- Summerfield, A. Q. (1979). Use of visual information in phonetic perception. *Phonetica*, 36, 314-331.
- Summerfield (1987) Some preliminaries to a comprehensive account of audio-visual speech perception. In Barbara Dodd and Ruth Campbell, editors, *Hearing by Eye: The Psychology of Lip Reading*, 3–51, Lawrence Earlbaum Associates.
- Summerfield and McGrath (1984) Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology*, 36, 51–74.
- Witkin, Wapner and Leventhal (1952) *Journal. of Experimental Psychology*, 43, 58–67.

Acknowledgements

- The point-light figures and the point-light video sequences have been taken from Lawrence Rosenblum's excellent 'AudioVisual Speech Web-Lab' [http://www.faculty.ucr.edu/ rosenblu/lab-index.html]