Audio-Visual Speech Processing: A Case Study of Biologically-Inspired Information Fusion

R. I. Damper
Information: Signals, Images, Systems Research Group
School of Electronics and Computer Science
University of Southampton, UK

Biologically-Inspired Information Fusion Workshop, University of Surrey
22/23 August 2006
Structure

1. Introduction
2. AV speech perception by humans
3. AV speech recognition by machine
4. Discussion of issues
Speech is not solely an acoustic/auditory phenomenon.

Some (but not all) movement of the articulators is visible . . .

. . . plus related movements of, e.g., cheeks, nostrils, etc.

(Also some complementary “body language” information.)

Visual information especially useful for speech communication under difficult acoustic conditions.

A very well known example is ‘lipreading’ (strictly ‘speech-reading’) by deaf and hard of hearing.

A paradigmatic case of biologically-inspired information fusion!
The classic experimental psychologists’ approach to studying information fusion in human subjects is to set the two sources in opposition…

The McGurk effect… sets visual and auditory information in conflict.

Three possible outcomes, depending upon relative strength:

1. audio information dominates
2. visual information dominates
3. audio and visual information ‘blend’
What do we See/Hear?

Audio /ba/ bilabial
AV /da/ alveolar
visual /ga/ velar

What happens when we present audio /ga/ and visual /ba/?
The McGurk Effect Works . . .

- . . . better for some consonant combinations than for others (McGurk and McDonald, 1976).
- . . . better for consonants than for vowels (Summerfield and McGrath, 1984).
- . . . even if audio and visual information are from speakers of different gender (Green et al., 1991).
- . . . with subjects from different language backgrounds (Massaro et al., 1993).
- . . . with highly schematic face representations and even if subjects are unaware that they are looking at a face (Rosenblum and Saldaña, 1996).
- . . . with young infants (Rosenblum et al., 1997).

The effect is automatic, unconscious, obligatory.
Cross-modal discrepancy effects are not specific to speech (e.g., Shams’ illusion).

Another striking demonstration—where audio and visual information regarding source location are in conflict—is the ventriloquism effect.

Ultimately, such illusions are of less interest for their own sake than for what they can tell us about perception in general.

Over-riding of auditory by visual information (or vice versa) can be explained by classical speed of processing or selective attention theories.

Blending is more difficult . . . one attempt is ‘object identity theory’ (Bedford, 2001).

But this is essentially a gestalt psychology approach—more conceptual than mechanistic.
Towards a Mechanistic Theory?

- Multisensory neurons . . . could provide a common currency for cross-modal fusion (Stein and Meredith, 1993; Stoffregen and Bardy, 2001).

- Massaro (1987, 1998) has developed a fuzzy logical model of perception (FLMP) based on four assumptions:
  1. each information source is evaluated to determine the (continuous) degree to which it specifies various alternatives;
  2. each source is evaluated independently;
  3. sources are then integrated to give an overall (still continuous) degree of support for each alternative;
  4. perception follows strongest degree of support among alternatives.

- Maximum likelihood?
For its first several decades of existence, ASR was resolutely unimodal!

This has been slowly changing (cf., Stork and Hennecke, 1996; Vatikiotis-Bateson et al., 2005).

Fuelled by availability of cheap computer power, cheap CCD cameras, etc., researchers are increasing looking to exploit more and richer input information.

Very little attempt is made to use knowledge (what knowledge?!) of how humans process audio-visual information.

A key issue is early versus late integration.

NB: this has both an architectural and a temporal interpretation!
In early (architectural) integration, audio and visual features are integrated into a single, large feature vector before classification.

**PRO:** allows statistics of audio and visual information to be jointly modelled.

**CON:** poses a possibly severe data modelling problem (‘curse of dimensionality’).
In late integration, audio and visual recognition is performed separately and the results combined.

Combination can be *soft* (scores) or *hard* (decisions).

**PRO:** easier data modelling and training of classifiers, noise resilience, variance reduction by central limit theorem.

**CON:** assumes independence of audio and visual information!
Issues

Psychology versus engineering: should we engineers care how humans do it?

Two outstanding issues in modelling perception and in engineering AV speech processing systems are:

1. Early versus late integration of information (architectural and temporal)
2. Early versus late continuous → discrete conversion of information

What should the fusion rule be in late integration? Trainable or fixed? etc. etc.

Neuroimaging? fMRI studies?

What about sensory fission (aka. parallelism)? A motivation for multi-stream speech processing.