Audio-Visual Scene Analysis:
A brief survey of interesting audio-visual interactions.

Jon Barker
Speech and Hearing Research
University of Sheffield

Audio-visual interactions

• Two areas of study:
  – The influence of visual input on auditory perception.
  – The influence of acoustic input on visual perception.

• Some questions:
  – Conditions necessary for AV interaction?
  – At what stage of processing do interactions occur?
  – Mechanisms?

Overview

• Primitive Audio-Visual interactions
  – Visual cues influencing auditory perception
  – Acoustic cues influencing visual perception
• Multimodal speech perception
• Auditory Scene Analysis + Vision

Ventriloquist illusion

Bertelson and Radeau (1981)

The speech audio is mislocated: it is heard to be coming from Position 2

Synchronised lip movements from Position 2

Speech audio coming from Position 1

Ventriloquist illusion

• Often presented as a speech illusion – but it occurs with very simple stimuli
  – A visual flash will cause a synchronized acoustic beep to be mislocalised.
• Occurs even when subjects don’t perceive visual stimulus.
  – experiments with subjects with unilateral visual neglect (Bertelson et al., 2000).
  – Seems to imply that auditory and visual information are integrated at a pre-perceptual stage of processing.
• Under the right conditions the localization of the visual stimulus is also affected by the position of the acoustic stimulus...
Auditory Driving
Gebhard and Mowbray (1959)
- Early demonstration of influence of sound on visual perception.
  - Subjects are presented with a flickering light and a fluttering sound.
  - Flicker rate held constant, flutter rate increased or decreased.
  - Subjects perceive changes in the (constant) flicker rate.
- The reverse V->A effect was not observed.
  - i.e. changing the visual flicker rate does not alter perception of the audio flutter rate.
- More detailed recent work showing auditory influence on visual temporal rate perception (e.g. Recanzone, 03)

Intersensory Temporal Locking
Frendrich and Corballis (2001)
- Subjects are asked to judge either where the circle is when the flash occurs or where it is when the click occurs.
- Click and flash judged to be part of the same event so they share a single temporal reference.
  - Error in judgement when attending flash
  - Error in judgement when attending click
- Does this auditory-induced temporal relocation of visual events generalise to not repetitive stimuli?

Intersensory Temporal Locking
Frendrich and Corballis (2001)
- Click and flash judged to be part of the same event so they share a single temporal reference.
  - Error in judgement when attending flash
  - Error in judgement when attending click
- Does this auditory-induced temporal relocation of visual events generalise to not repetitive stimuli?

Temporal Ventriloquism
Morein-Zamir et al. (2003)
- Employed a Temporal Order Judgement (TOJ) task.
  - Either top LED or bottom LED comes on first.
  - Sounds played before and after lights.
- Playing 2 beeps consistently induces subjects to mistakenly perceive 2 flashes.
  - Playing one sound between lights coming on has no effect.
- Not symmetrical - Showing two flashes does not make subjects hear a single beep as two beeps.

- A single flash is accompanied by a varying number of auditory beeps.
- Subjects asked to report how many flashes are seen.
  - Playing two sounds between lights coming on makes task harder.
  - Playing one sound between lights coming on has no effect.
- Not symmetrical - Showing two flashes does not make subjects hear a single beep as two beeps.
Audio-Visual Integration

- Modalities contribute to percept according to their relative reliability of measurements. Generally:
  - Auditory: Reliable spatial location.
  - Vision: Reliability of temporal location.
- Auditory: Unreliable spatial location. Reliable temporal location.
- Not a new idea; c.f. modality appropriateness hypothesis - Welch and Warren (1986).
- Judgement of location of blurred visual stimuli is influenced by sound.
- Successful prediction of modality dominance by simple models based on maximum likelihood principle.
- Important implications for A-V models:
  - When combining A and V estimates, the estimates alone are not enough, we also need a "confidence measure" i.e. an estimate of the variance as well as the mean.

Integration

Visual percept

Auditory percept

Perceived acoustic position

Increasing separation

Beyond a certain separation the interpretation switches from that of a single object to that of two separate objects

Audio-Visual Integration

Overview

- Primitive Audio-Visual interactions
  - Visual cues influencing auditory perception.
  - Acoustic cues influencing visual perception.
- Multimodal speech perception.
- Auditory Scene Analysis + Vision.

Audio-Visual Integration

Preamble – Is speech special?

- Speech perception exhibits many unusual properties, including many interesting cross-modal effects.
- Two schools of thought:
  - Speech engages a "special mode of perception", e.g. Remez et al. (1994)
  - Speech is not processed in a qualitatively different way from nonspeech. "speech perception is continuous with the perception of other natural events", e.g. Bregman (1990), Rosenthal (2005)

Audio-Visual Speech Processing

- Sumby and Pollack (1954),
  - 
  untrained listeners routinely employ visual information when listening to speech in adverse conditions.
  - two syllable words + stationary noise at varying SNR.
  - visual cues equivalent to improving SNR by ~15 dB.
Audio-Visual Speech Intelligibility

- Summerfield (1979)
  - Where are the visual cues?
  - Visual information concentrated in lip region, but also cues elsewhere.

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

Audio-Visual Speech

- Early idea – primacy of auditory information.
  - Visual information used as a 'back up' system.
  - Employed when there is reason to doubt the acoustic signal.
- This is not the case
  - Visual information can influence AV speech perception even when acoustic signal seems unambiguous.
  - AV integration can occur even when the visual cues contradict seemingly clear acoustic cues.

The McGurk Effect


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

The listener hears a d or th.
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).
- Visual information can be reduced to a few moving points (Rosenblum & Saldaña, 1996).
- It works when observers touch (rather than look-at) the face (Fowler & Dekle, 1991).
- Effect persists even if listeners are instructed to focus attention on the acoustic signal.

AV speech detection studies

Grant and Seitz (2000), Grant (2001)

- Compare detection threshold for speech in audio only and audio-visual conditions.
- Found visual speech reduces the detection threshold by 2-3 dB.
- Effect disappears if the Audio-Visual synchrony is artificially disturbed.

Fuzzy Logical Model of Perception

Massaro (1987, 1998)

- Treats A and V as independent channels
- Describes how A and V integrate, not whether they will integrate.

AV speech detection

- Strong correlation between dynamics of lip area and of energy envelope of the F2 region of the spectrum.
AV source separation


- Identification of voiced versus unvoiced plosives. [gy gu dy du] vs [ty tu ky ku]
- Speech babble added at ~ -9 dB SNR
- Identical video stimulus dubbed onto all acoustic signals.

Using acoustic cues alone, subjects can’t reliably discriminate between g and k.

Visual stimulus aids stream separation, so “g” is heard more reliably

Visual stimuli is totally ambiguous

Overview

- Primitive Audio-Visual interactions
  - Visual cues influencing auditory perception.
  - Acoustic cues influencing visual perception.
- Multimodal speech perception.
- Auditory Scene Analysis + Vision.

Ventriloquist illusion with two talkers – Driver (1996)

Condition 1:
Visual and audio target cues at same position
Ventriloquist illusion with two talkers – Driver (1996)

Does the illusory separation increase intelligibility?

Condition 2: Identical audio but visual cue displaced.

Listeners were able to better recognise the target words when the target and masker appeared to be coming from different locations.

Integrating auditory scene analysis and vision speech information

Two simultaneous speech sources at Position 1

Spectro-temporal source 'glimpses'

Target lip features (originate at Position 2)

AV target = Pos 2

Audio masker = Pos 1

Can separate sources by attending to Pos 1 or Pos 2

Integrating auditory scene analysis and vision speech information

Two simultaneous speech sources at Position 1

Spectro-temporal source 'glimpses'

Target lip features (originate at Position 1)

AV target = Pos 1

Cannot separate sources by attending to a position

Supplying visual cues for masker

Condition 1: Masker lips present

Does viewing the masker lips increase intelligibility of the target speech?

Condition 2: Masker lips occluded
Supplying visual cues for masker

Small but significant increase in intelligibility when masker's lips are visible.

Integrating auditory scene analysis and vision speech information

Two simultaneous speech sources at Position 1

Spectro-temporal source 'glimpses'

Masker lip features (originate at Position 1)

Audio target = Pos 1

AV masker = Pos 1

Can separate target by focusing on Audio glimpses?

Summary

- Audio and visual information are fused into a single percept if there is sufficient evidence of a common cause.
  - Common spatial location
  - Common temporal dynamics
  - Fit to models
- Modalities contribute to the common percept according to the relative reliability of their measurements.
  - Visual system provides reliable spatial information
  - Auditory system provides reliable temporal information
- Auditory and visual processing probably interact at several different processing stages, e.g. consider speech.
  - Visual information may be employed in sound source separation
    - i.e. release from energetic masking
  - Visual information may guide auditory attention
    - i.e. release from informational masking
- Visual phonetic information is combined with acoustic phonetic information to better identify the speech sound being uttered.
- The importance of such mechanisms for processing non-speech events remains unclear.

References

- Bernstein et al. (1994) Psychophysiology 31, 130-140
- Driver (1996) Nature 381(6577), 66--68
- McGeoch and MacDonald (1976) Nature 264, 746--748
- Rosenblum and Saldana (1996) J. Experimental Psychology 22(2), 318-331