Spectro-Temporal Interactions in Auditory and Auditory-Visual Speech Processing

Ken W. Grant
Walter Reed Army Medical Center, Washington, D.C.

Steven Greenberg
The Speech Institute, Oakland, CA

http://www.wramc.amedd.army.mil/departments/aasc/avlab
grant@tidalwave.net
Acknowledgments

Collaborations:
David Poeppel and Virginie van Wassenhove
Neuroscience and Cognitive Science Program, University of Maryland, College Park, MD

Funding:
NIH Grant: DC 00792-01AI
NSF Grant (subcontract): SBR 9720398 - Learning and Intelligent Systems Initiative of the National Science
Speech Recognition in Noise and Reverberation

- Primary complaint expressed by hearing-impaired and elderly patients
- Important for machine recognition (ASR)
Noisy, Reverberant Speech: Demo
Goal: Improve Speech-to-Noise Ratio

- Signal Processing (e.g., noise reduction algorithms)
- New Technologies (e.g., directional microphones)
- Speechreading and Auditory-Visual Integration
Auditory-Visual vs. Audio Speech Recognition

Roughly 6 dB improvement in S/N; roughly 30% improvement in intelligibility for NH subjects
Auditory-Visual vs. Audio Speech Recognition

Speech-to-Noise Ratio (dB)

Percent Correct Recognition

NH - Auditory Consonants
HI Auditory Consonants
HI-Auditory Sentences
ASR Sentences
(adapted from Kollmeier et al., 2001)
Roughly 6 dB improvement in S/N; roughly 30% improvement in intelligibility for NH subjects.

<table>
<thead>
<tr>
<th>Speech-to-Noise Ratio (dB)</th>
<th>Percent Correct Recognition</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
</tr>
<tr>
<td>10</td>
<td>70</td>
</tr>
<tr>
<td>15</td>
<td>60</td>
</tr>
<tr>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>-5</td>
<td>40</td>
</tr>
<tr>
<td>-10</td>
<td>30</td>
</tr>
<tr>
<td>-15</td>
<td>20</td>
</tr>
</tbody>
</table>

Graph labels:
- NH - Audiovisual Consonants
- NH - Auditory Consonants
- HI - Audiovisual Consonants
- HI Auditory Consonants
- HI-Auditory Sentences
- ASR Sentences
Spectral Interactions

Audio-visual benefit depends on the spectral locus of the acoustic signal

- AV Benefit is determined primarily by redundancy between acoustic and visual information
- Redundancy can be estimated by information transmission
Auditory-Visual Spectral Interactions: Consonants

Speechreading + Speech Envelope Bands

CUNY Sentences

Percent Correct

Prefilter Condition

SA WB 500 Hz 1600 Hz 3150 Hz
Redundancy Hypothesis – Modeling Results

Auditory consonant recognition based on perfect transmission of indicated feature. Responses within each feature category were uniformly distributed.

Predicted AV consonant recognition based on PRE model of integration (Braida, 1991).
Spectral Interactions - Summary

• Speechreading provides information mostly about place-of-articulation
Spectral Interactions - Summary

- Speechreading provides information mostly about place-of-articulation
- Auditory-visual speech recognition is determined primarily by complementary cues between visual and auditory modalities
Spectral Interactions - Summary

• Speechreading provides information mostly about place-of-articulation

• Auditory-visual speech recognition is determined primarily by complementary cues between visual and auditory modalities

• The most intelligible auditory speech signals do not necessarily result in the most intelligible auditory-visual speech signal
Spectral Interactions - Summary

• Speechreading provides information mostly about place-of-articulation

• Auditory-visual speech recognition is determined primarily by complementary cues between visual and auditory modalities

• The most intelligible auditory speech signals do not necessarily result in the most intelligible auditory-visual speech signal

• Acoustic cues for voicing and manner-or articulation are the best supplements to speechreading
Spectral Interactions - Summary

• Speechreading provides information mostly about place-of-articulation

• Auditory-visual speech recognition is determined primarily by complementary cues between visual and auditory modalities

• The most intelligible auditory speech signals do not necessarily result in the most intelligible auditory-visual speech signal

• Acoustic cues for voicing and manner-or articulation are the best supplements to speechreading

• These cues tend to be low frequency
Temporal Window for A and AV Integration
AUDIO-ALONE
EXPERIMENTS
Word Intelligibility - Single and Multiple Slits

Slit Asynchrony Affects Intelligibility

Desynchronizing the slits by more than 25 ms results in a significant decline in intelligibility.

The effect of asynchrony on intelligibility is relatively symmetrical.

Cross-Spectral Temporal Asynchrony Effects

AUDITORY-VISUAL EXPERIMENTS
Auditory-Visual Tasks

**IEEE Sentences**
- Recognition of key words
  - Audio slits 1 + 4
  - Video presented at various temporal asynchronies

**CV Syllables**
- Recognition of McGurk pairs
  - Audio /pa/, /ba/, /ta/, /da/
  - Video /ka/, /ga/, /ta/, /da/
- Synchrony identification and discrimination
  - Yes/No single interval simultaneity judgments
  - Congruent versus incongruent tokens
Auditory-Visual Asynchrony - Paradigm

- Synchronous Audiovisual (Natural Speech Alignment)
- Asynchronous Audiovisual (Video Leads Audio)
- Asynchronous Audiovisual (Audio Leads Video)

Audio Onset re: Natural Speech Alignment (ms)
Cross-Modality Temporal Asynchrony Effects: Sentences

IEEE Word Recognition (%)

Audio Delay (ms)

Audio Alone - Bands 1 + 4

Speechreading Alone
McGurk Synchrony Paradigm

Original /ka/

Dubbed /pa/

Time (ms)
Temporal Integration in the McGurk Effect

Response Probability

Audio Delay (ms)

/pa/ /ta/ /ka/

[Graph showing response probability over audio delay (ms) for /pa/, /ta/, and /ka/ sounds]
Synchrony Identification - Natural vs. McGurk AV Tokens

Probability of Simultaneity vs. Audio Delay (ms)

- AVd
- A VT
- AVp
- AVg

The graph illustrates the probability of simultaneity for different AV tokens across various audio delays. The x-axis represents audio delay in milliseconds, ranging from -500 to 500 ms, while the y-axis shows the probability of simultaneity, ranging from 0 to 1.
Spectro-Temporal Integration: Summary

**Within Modality (Cross- Spectral Auditory Integration)**
- TWI is symmetrical
- TWI roughly 20-40 ms (phoneme?)

**Across Modality (Cross-Modal AV Integration)**
- TWI is highly asymmetrical favoring visual leads
- TWI is roughly 160-250 ms (syllable?)
- TWI for Incongruent CV's (McGurk Stimuli) is not as wide as TWI for natural congruent CV's