Auditory-Visual Speech Perception Laboratory

- Research Focus:
  - Identify perceptual processes involved in auditory-visual speech perception
  - Determine the abilities of individual patients to carry out these processes successfully
  - Design intervention strategies using signal processing technologies and training techniques to remedy any deficiencies that may be found.
- Primary Funding:
  - NIH Grant: DC 00792-01AI
  - NSF Grant (subcontract): SBR 9720398 - Learning and Intelligent Systems Initiative of the National Science
  - DARPA Grant (subcontract): ONR Award N000140210571
- Staffing:
  - Lab Director: Ken W. Grant
  - Research Associate: Mary T. Cord

Auditory-Visual Speech Perception Laboratory

Collaborations:

Steven Greenberg - The Speech Institute, Oakland, CA

David Poeppel - University of Maryland, College Park, MD

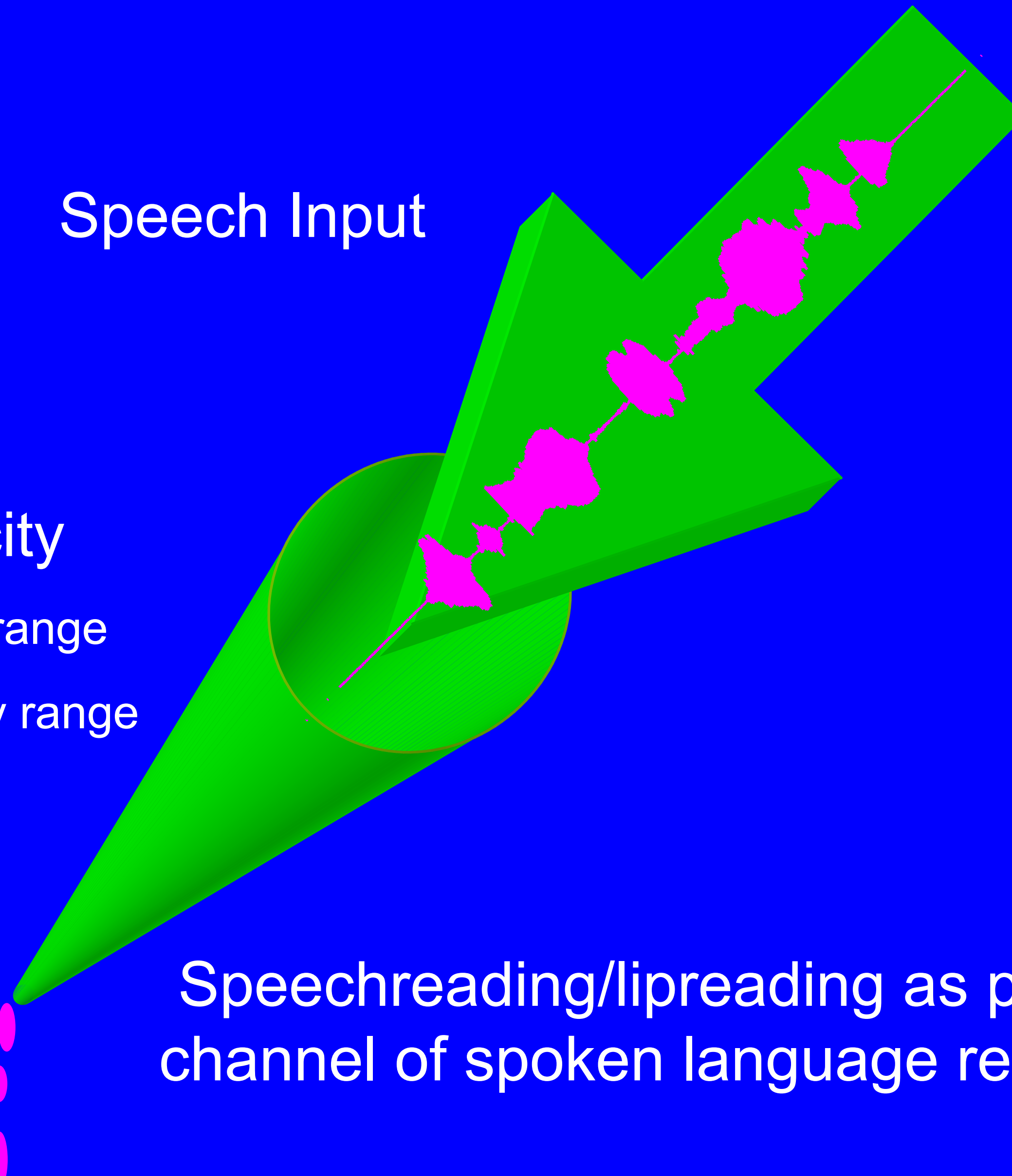
Virginie van Wassenhove - University of Maryland, College Park, MD

Topics

- Auditory Supplements to Speechreading
- Bimodal Comodulation
- Spectro-Temporal Window of Integration

Auditory Supplements to Speechreading

Selective Needs of Severe-to-Profound Hearing Loss

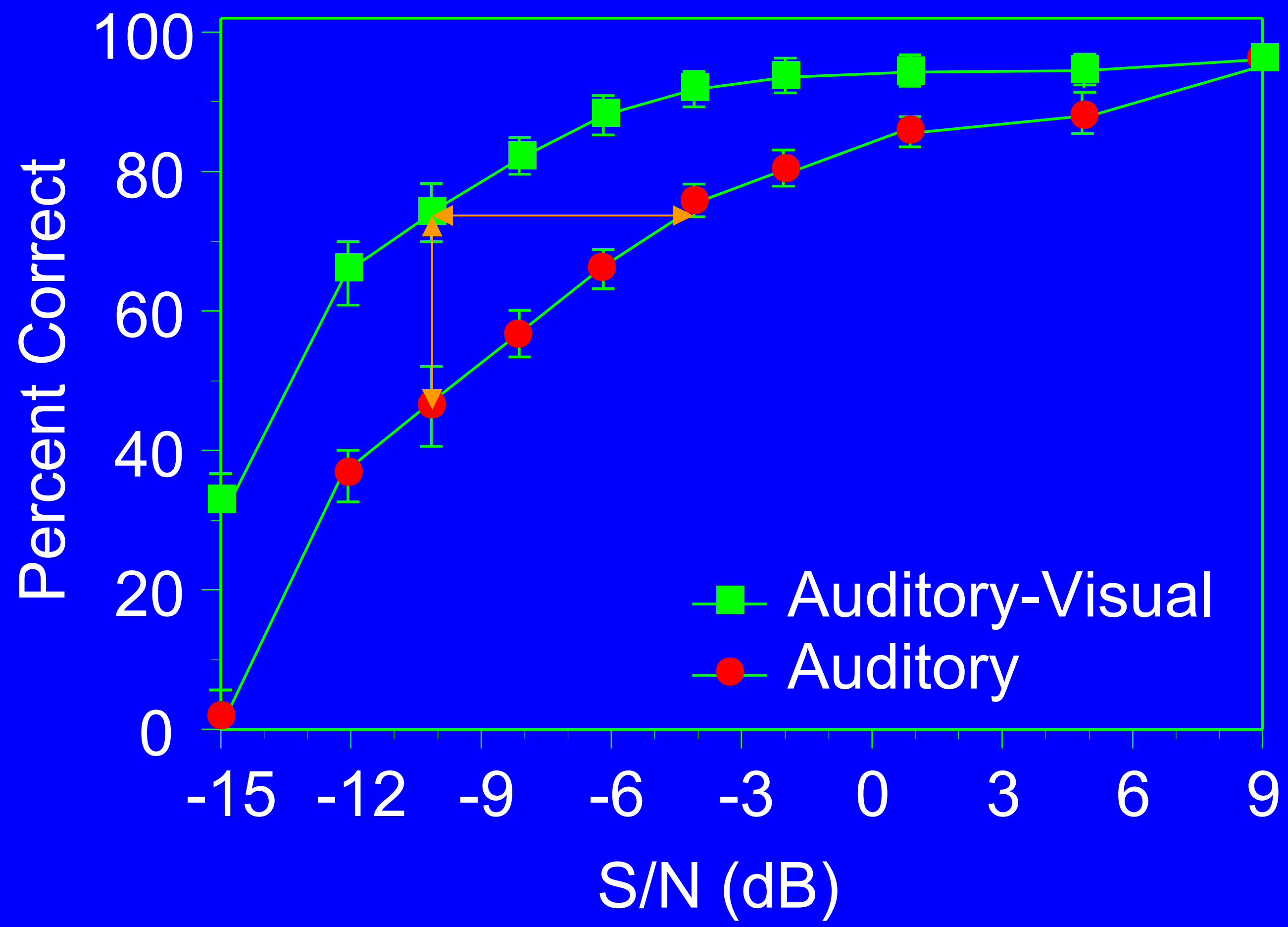


Hearing Capacity

- compressed dynamic range
- compressed frequency range

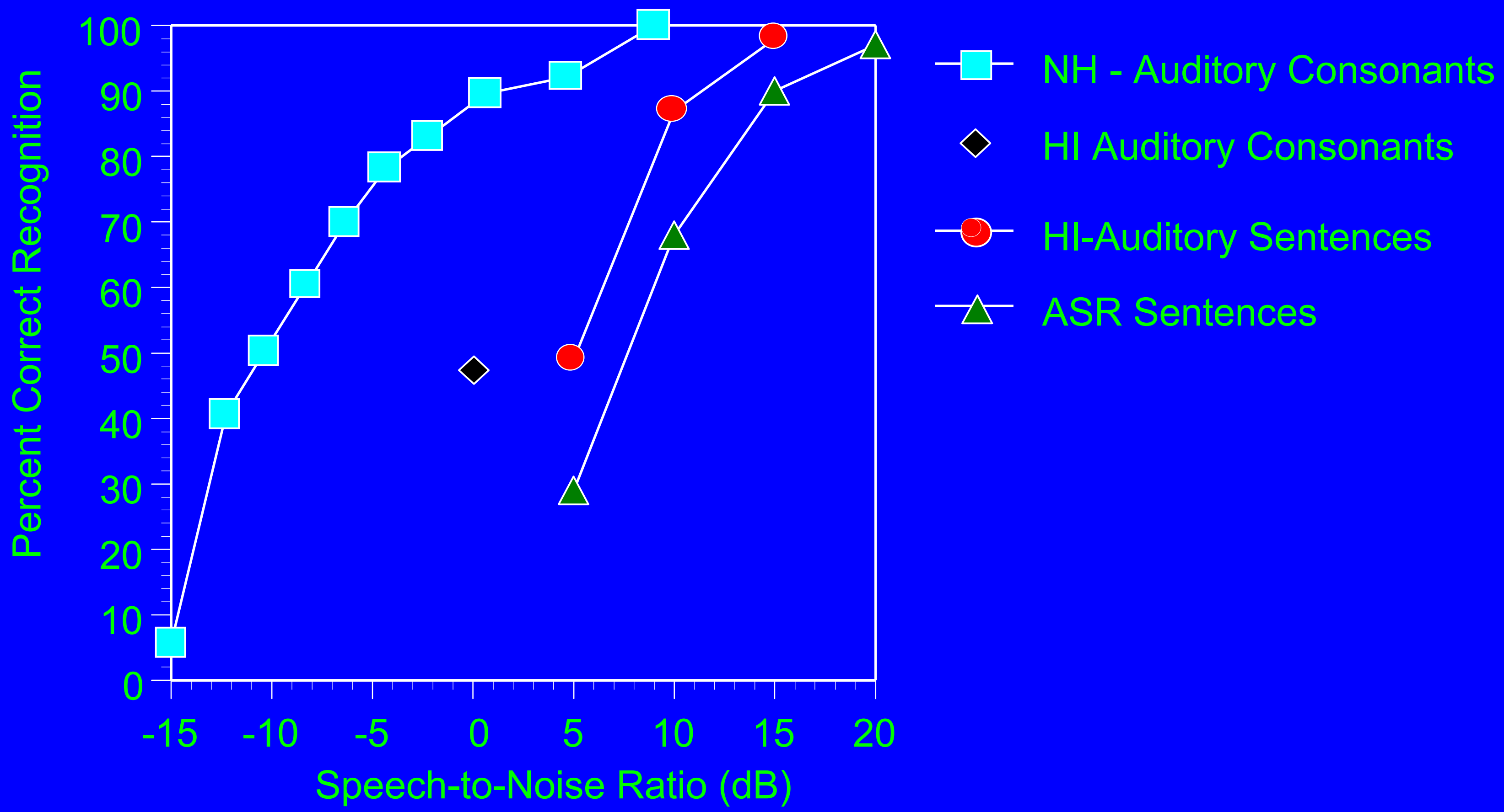
Speechreading/lipreading as primary channel of spoken language reception

Speech Recognition: Sentences

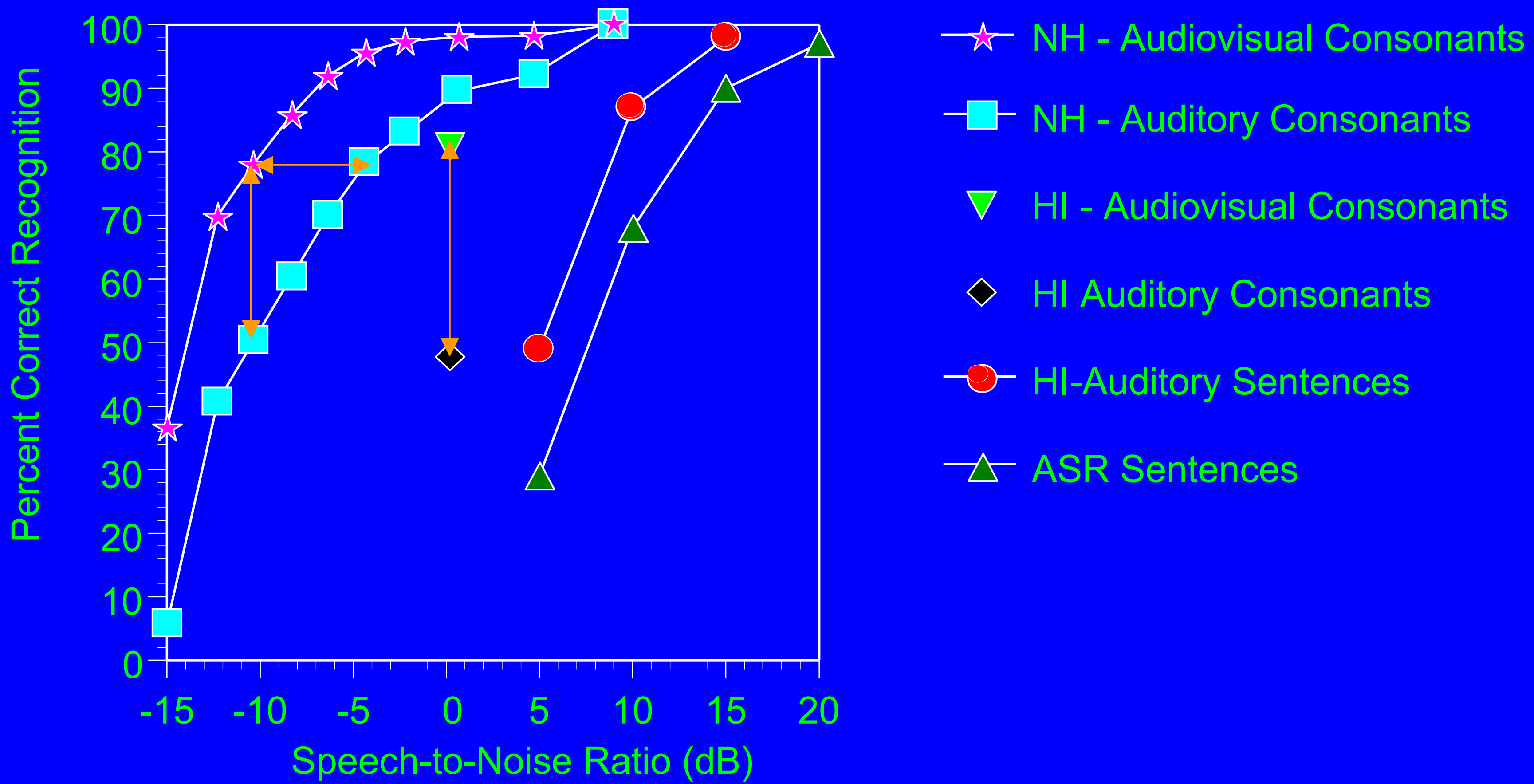


Roughly 6 dB improvement in S/N; roughly 30% improvement in intelligibility for NH subjects

Speech Recognition: Consonants



Auditory-Visual vs. Audio Speech Recognition



Roughly 6 dB improvement in S/N; roughly 30% improvement in intelligibility for NH subjects.

Possible Roles of Speechreading

- ***Provide segmental information that is redundant with acoustic information***

Possible Roles of Speechreading

- Provide segmental information that is redundant with acoustic information
- ***Provide segmental information that is complementary with acoustic information***

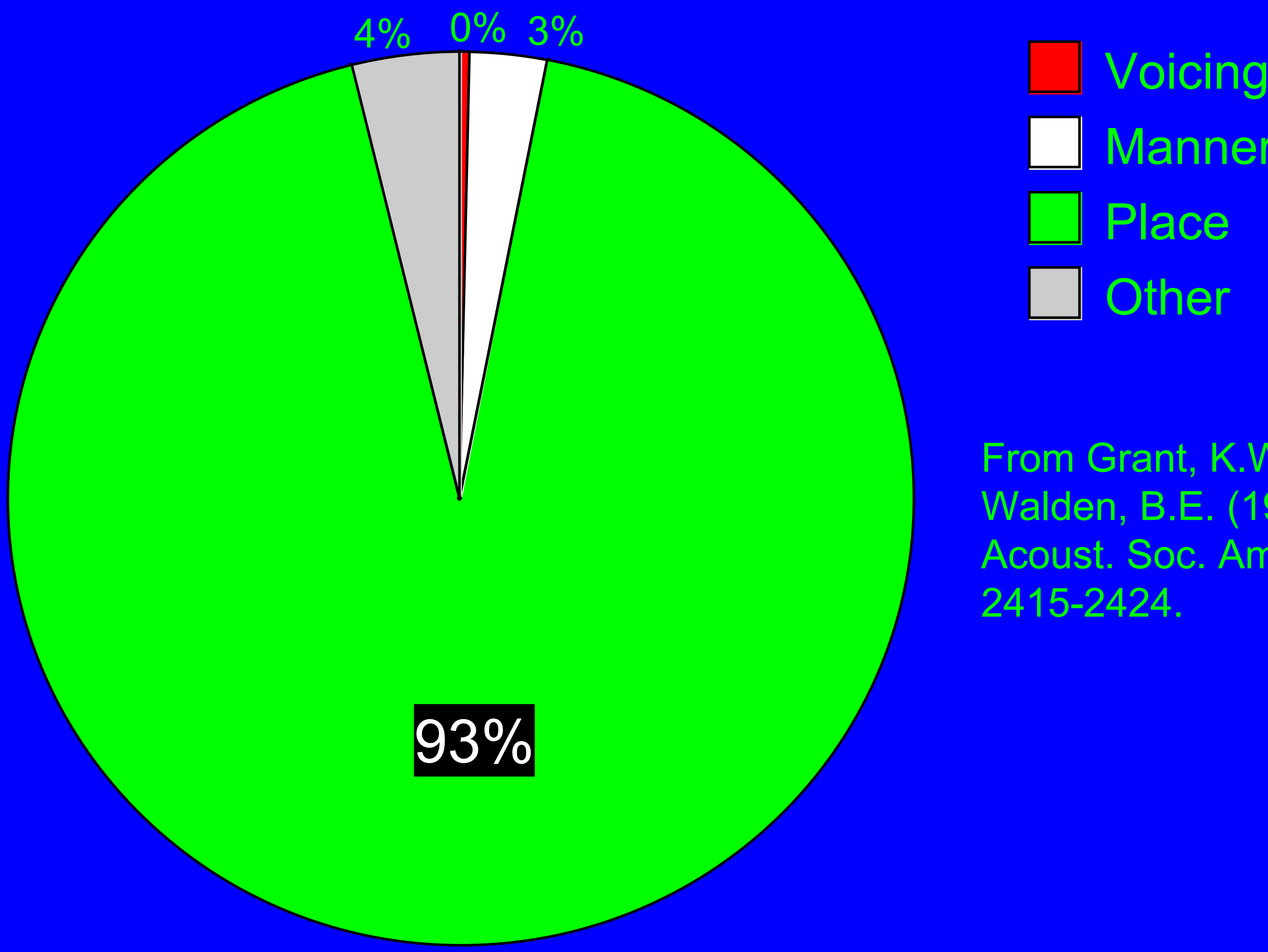
Possible Roles of Speechreading

- Provide segmental information that is redundant with acoustic information
- Provide segmental information that is complementary with acoustic information
- ***Direct auditory analyses to the target signal***
  - ***who, where, when, what (spectral)***

Auditory-Visual Speech Recognition: Consonants

- What information is available through speechreading?
- Which acoustic signals supplement speechreading?
- Are there significant audio-visual interactions in speech processing?

Visual Feature Recognition



From Grant, K.W., and Walden, B.E. (1996). J. Acoust. Soc. Am. 100, 2415-2424.

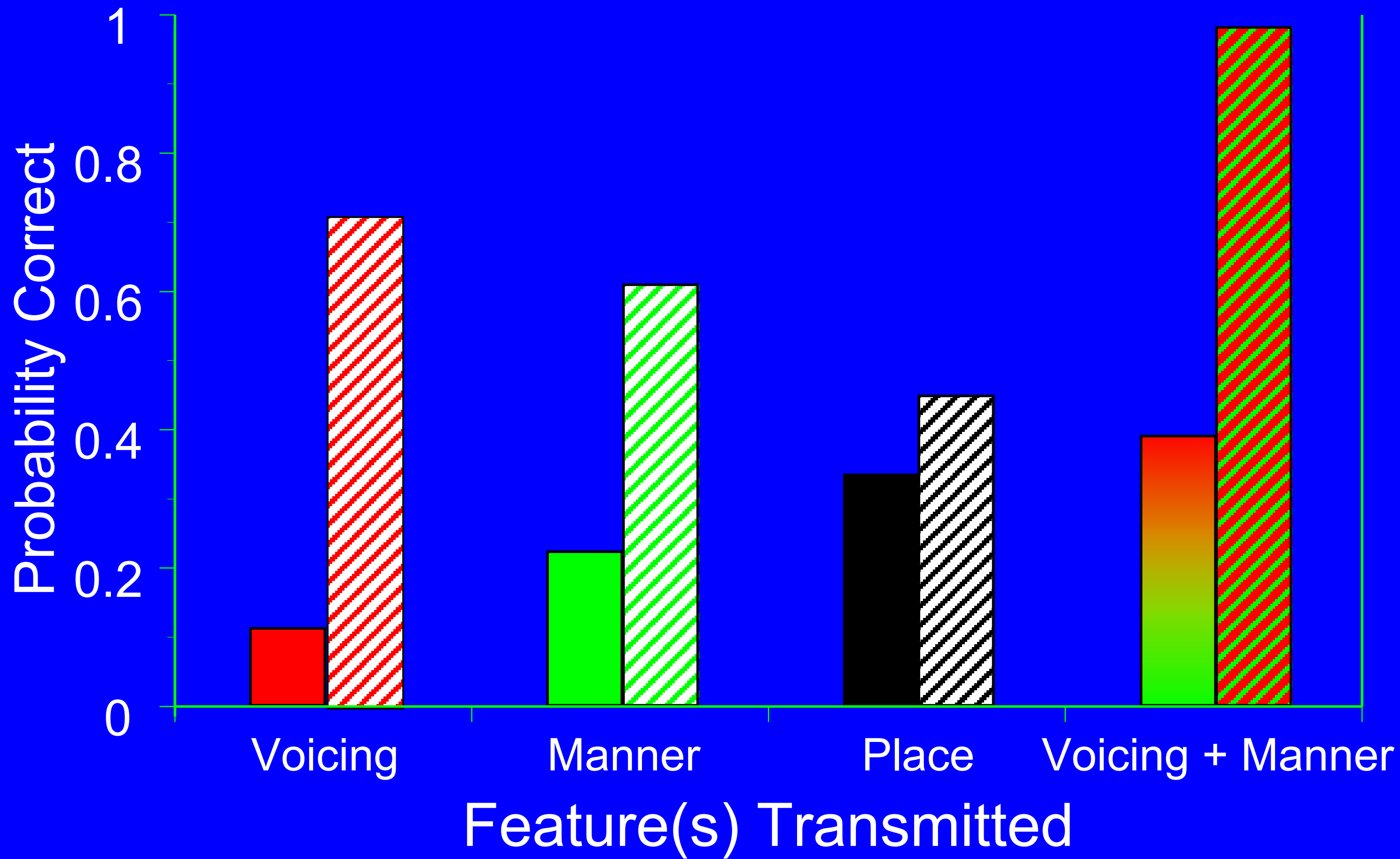
%Information Transmitted re: Total Information Received

Speech Recognition: Consonants

Linguistic feature contributions to visual speech recognition. The top row represents typical feature classifications for speechreading alone (visemes). Each subsequent row represents the effects of adding information about another linguistic feature via an additional input channel (in this case auditory). Note that as additional features are added, consonant confusions associated with speechreading are resolved to a greater and greater extent.

Speechreading	p,b,m t,d,n g,k f,v θ,ð s,z ʃ,tʃ,dʒ,ʒ l r w j
Voicing	p b,m t d,n g k f v θ ð s z ʃ,tʃ dʒ,ʒ l r w j
Nasality	p b m t d n g k f v θ ð s z ʃ,tʃ dʒ,ʒ l r w j
Affrication	p b m t d n g k f v θ ð s z ʃ tʃ dʒ ʒ l r w j

Hypothetical Consonant Recognition - Perfect Feature Transmission



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

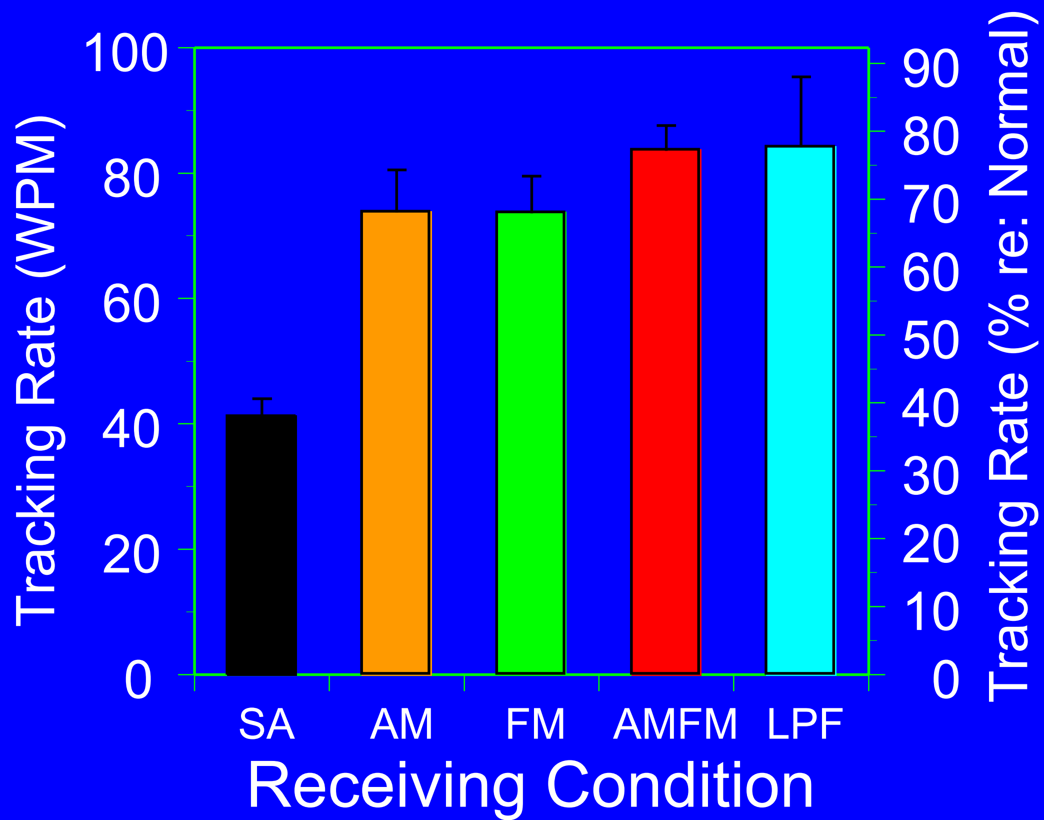
PRE — Predicted AV consonant recognition based on PRE model of integration (Braida, 1991).

Designer Acoustic Signals: Minimal Bandwidth, Maximum Benefit

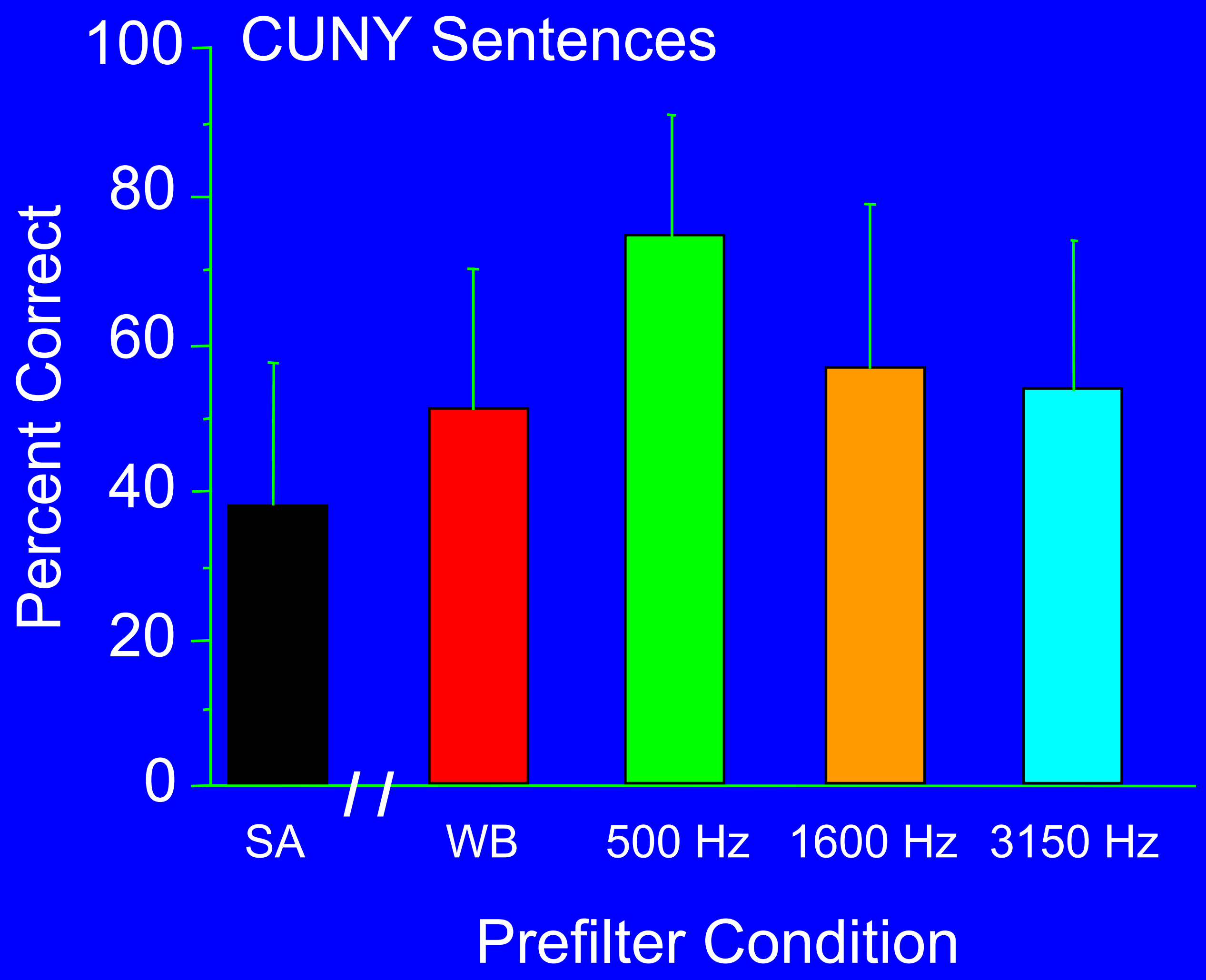
Frank and Joe Hardy scanned the wide valley that appeared before them as their yellow sports sedan rounded the crest of a hill. In the distance, a huge cylindrical tower rose from the valley floor. "Looks like a giant barnacle", Joe remarked to his older, dark-haired brother Frank. Biff Hooper, a tall, muscular high school friend of the two amateur detectives, leaned forward from the back seat. "You're looking at the cooling tower. The reactor itself is in the building next to it." Biff's uncle, Jerry Hooper, was a nuclear engineer at the Bayridge Nuclear Power Plant located outside Bayport. He had invited the three boys on a private afternoon tour of the facility. The summer had just begun, and the Hardy brothers were eager for new adventures.

WB LPF AM AMFM

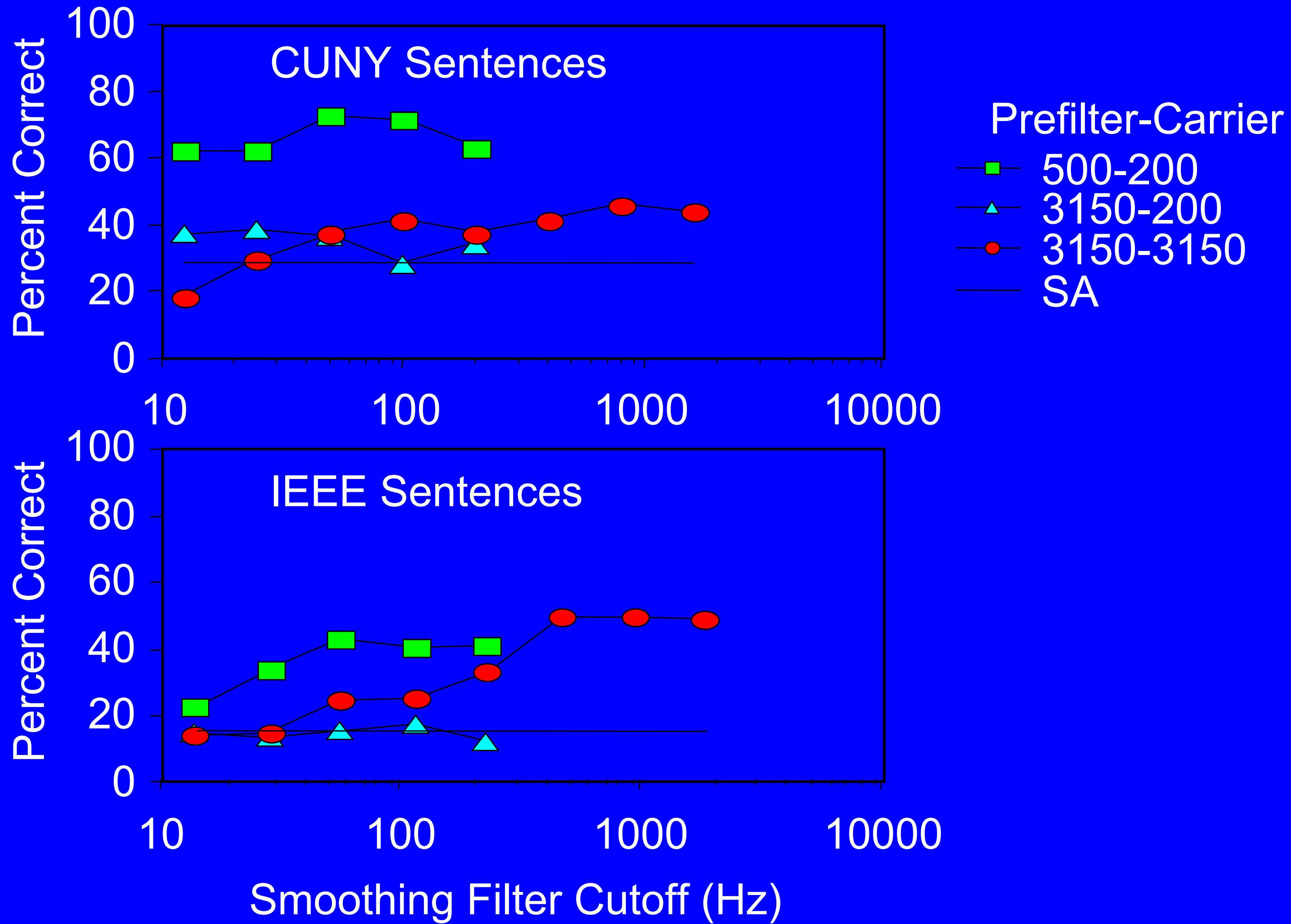
Designer Acoustic Signals: Minimal Bandwidth, Maximum Benefit



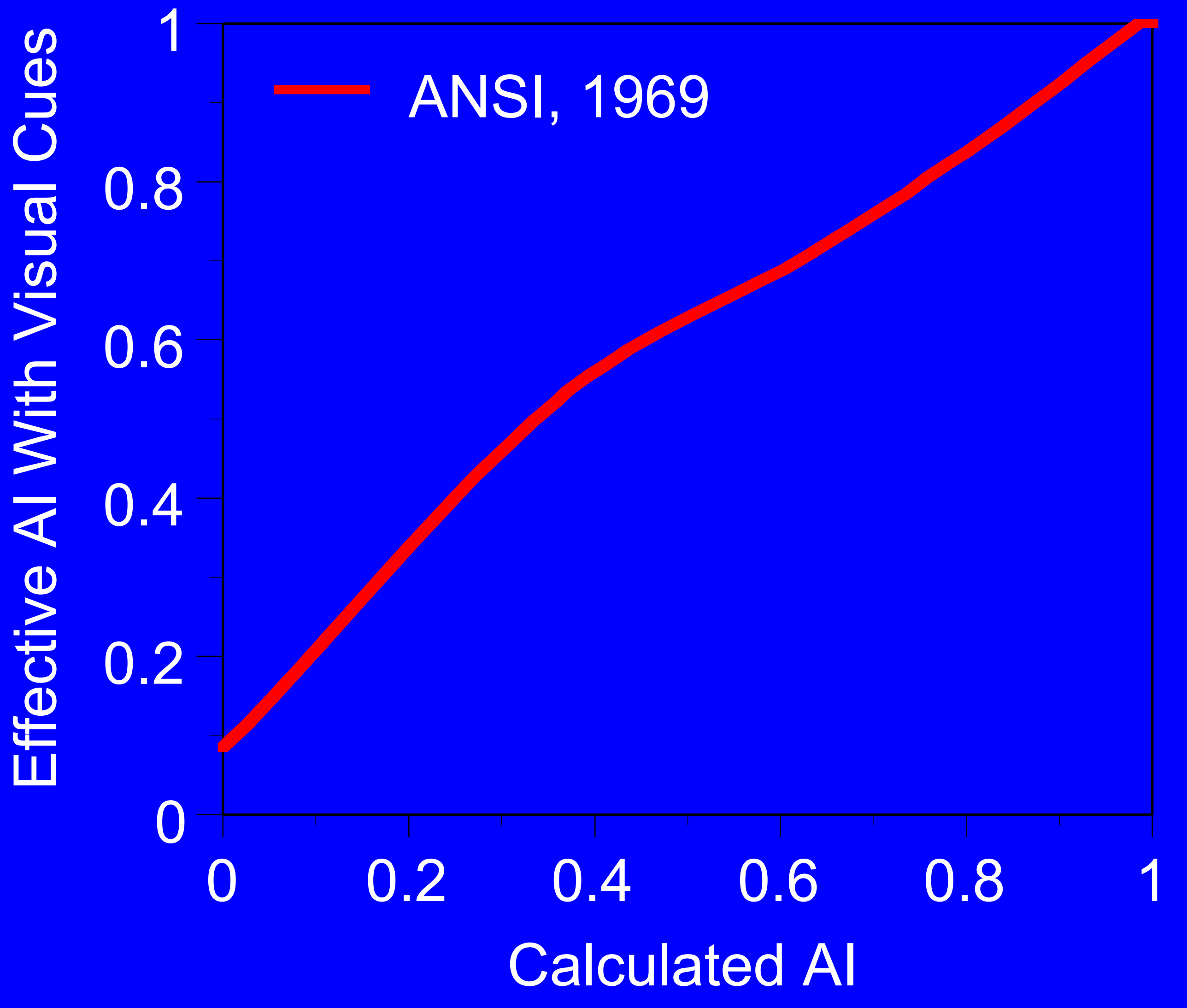
Designer Acoustic Signals: AM Bands



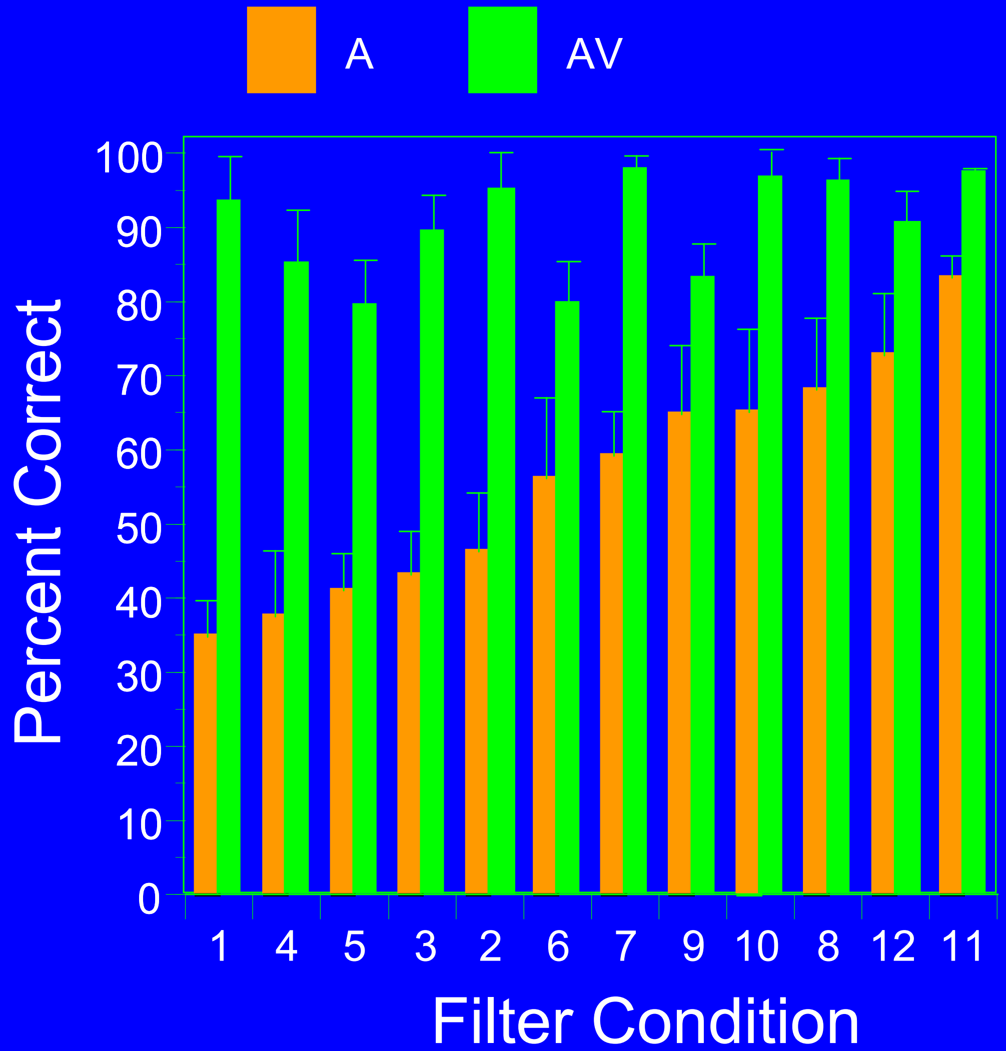
AM Bands - Smoothing-Filter Effects



Problems With Articulation Theory



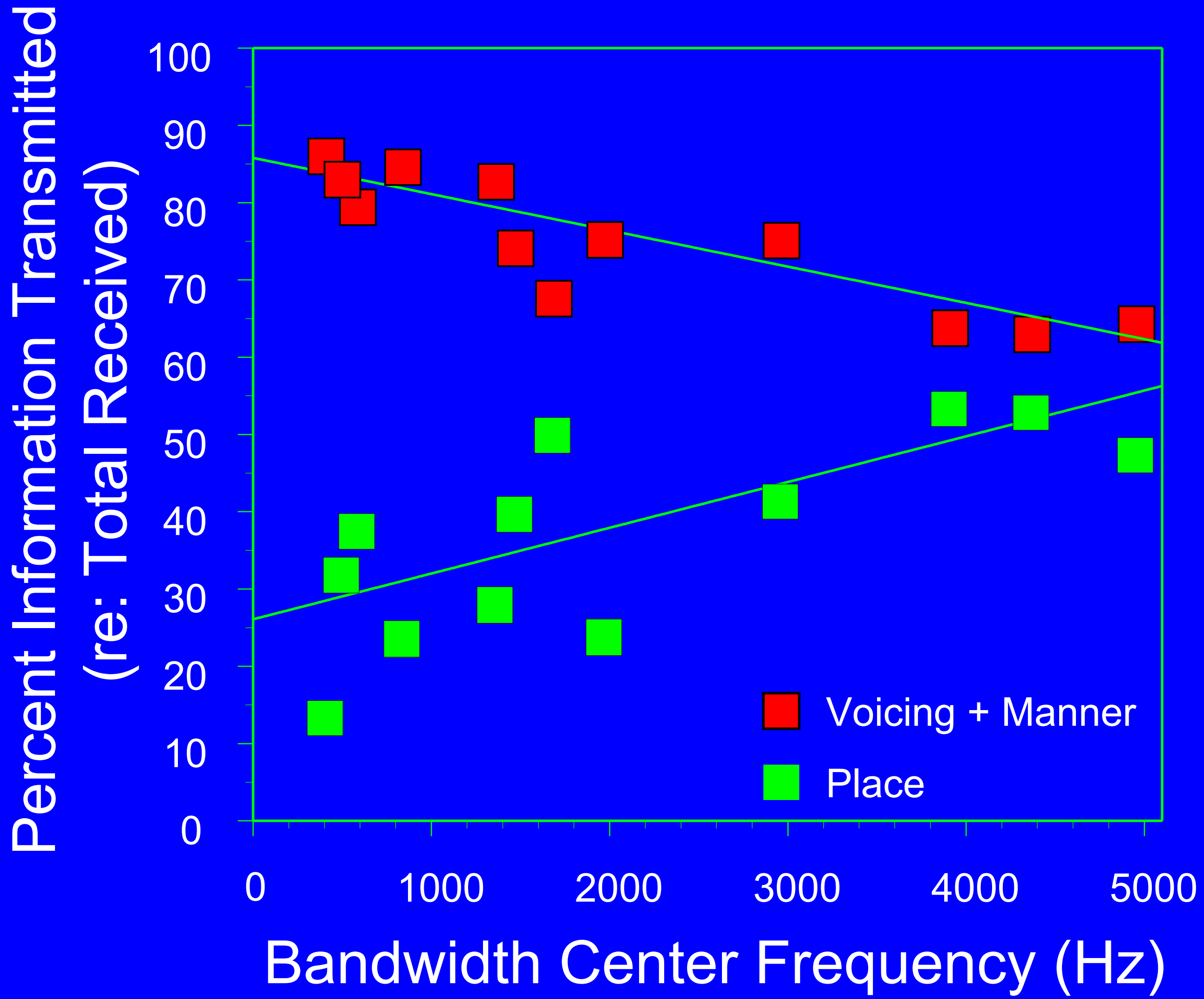
Auditory-Visual Spectral Interactions: Consonants



Condition (Hz)

1. 250-505
2. 645-955
3. 1130-1515
4. 1720-2140
5. 2600-3255
6. 4200-5720
7. 250-795
8. 1130-1930
9. 3255-5720
10. 250-1130
11. 1130-2355
12. 2600-5720

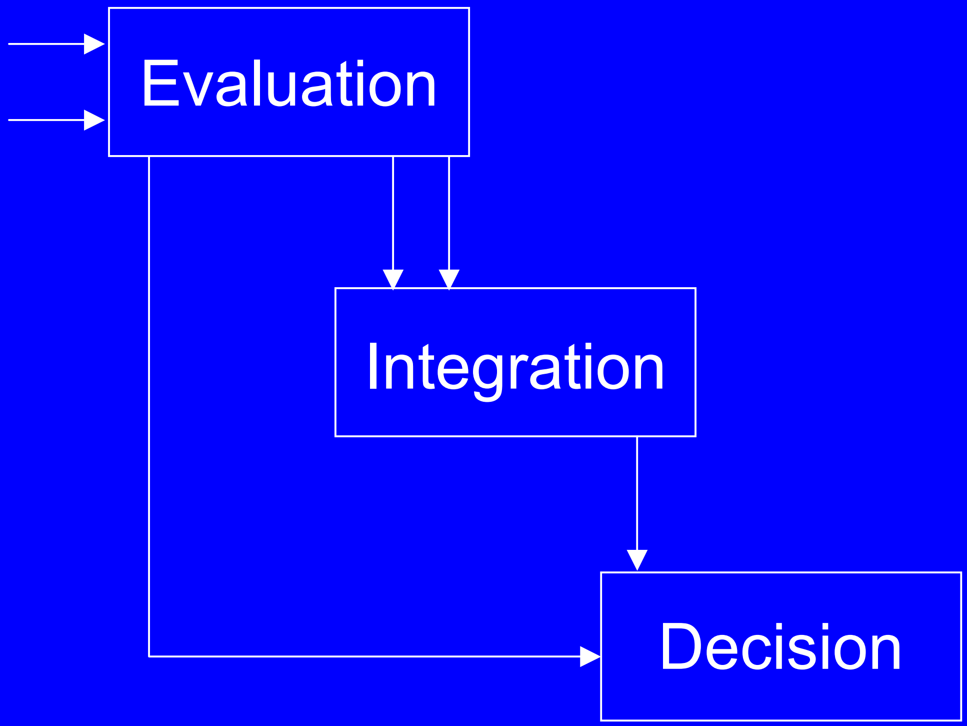
Feature Distribution re: Center Frequency



From Grant, K.W., and Walden, B.E. (1996). J. Acoust. Soc. Am. 100, 2415-2424.

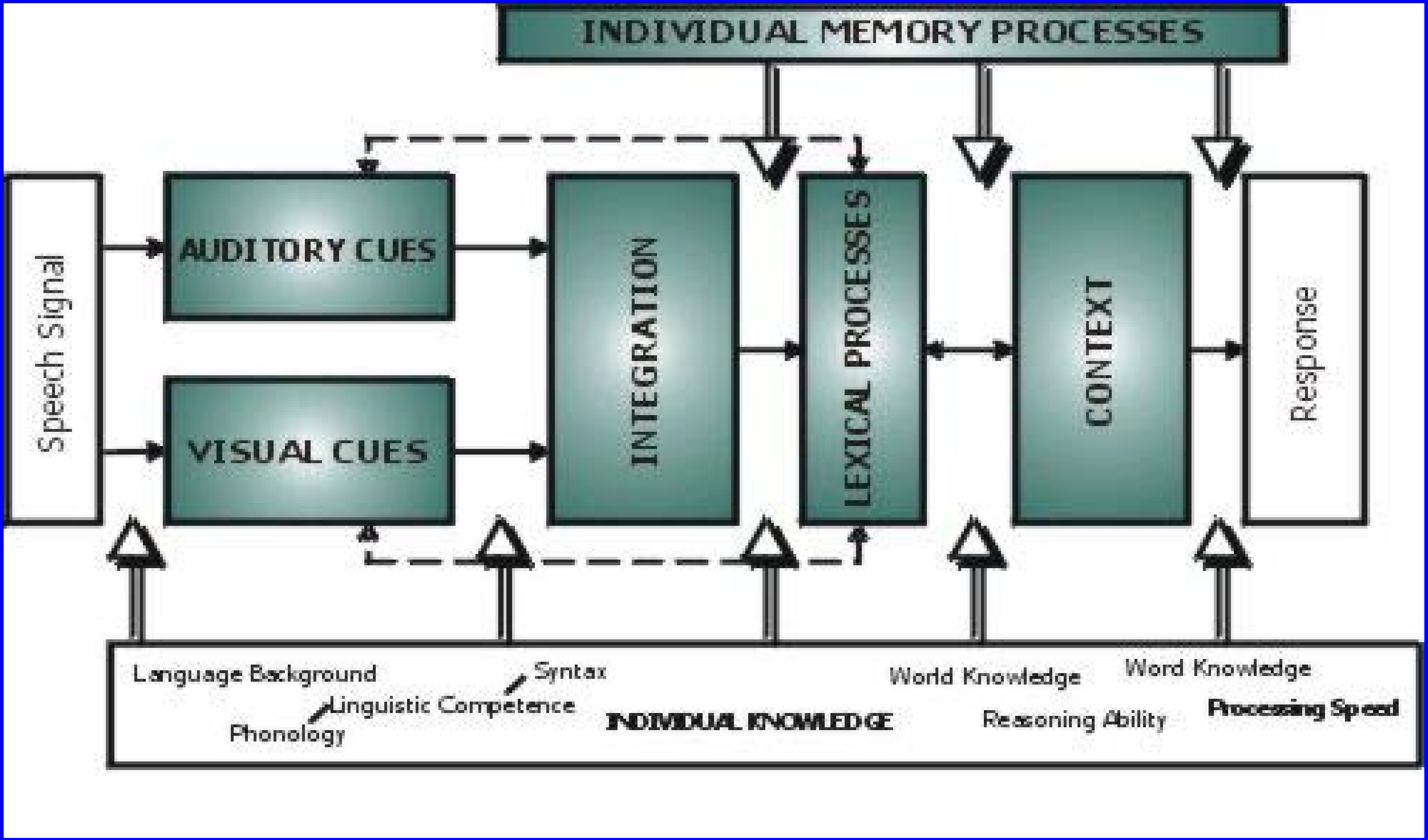
How It All Works - The Prevailing View

- Information extracted from both sources independently
- Integration of extracted information
- Decision statistic



From Massaro, 1998

A More Expanded View of the Process



Auditory Supplements to Speechreading

SUMMARY:

- Speechreading provides information mostly about place-of-articulation

Auditory Supplements to Speechreading

SUMMARY:

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

Auditory Supplements to Speechreading

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

Auditory Supplements to Speechreading

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

Auditory Supplements to Speechreading

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

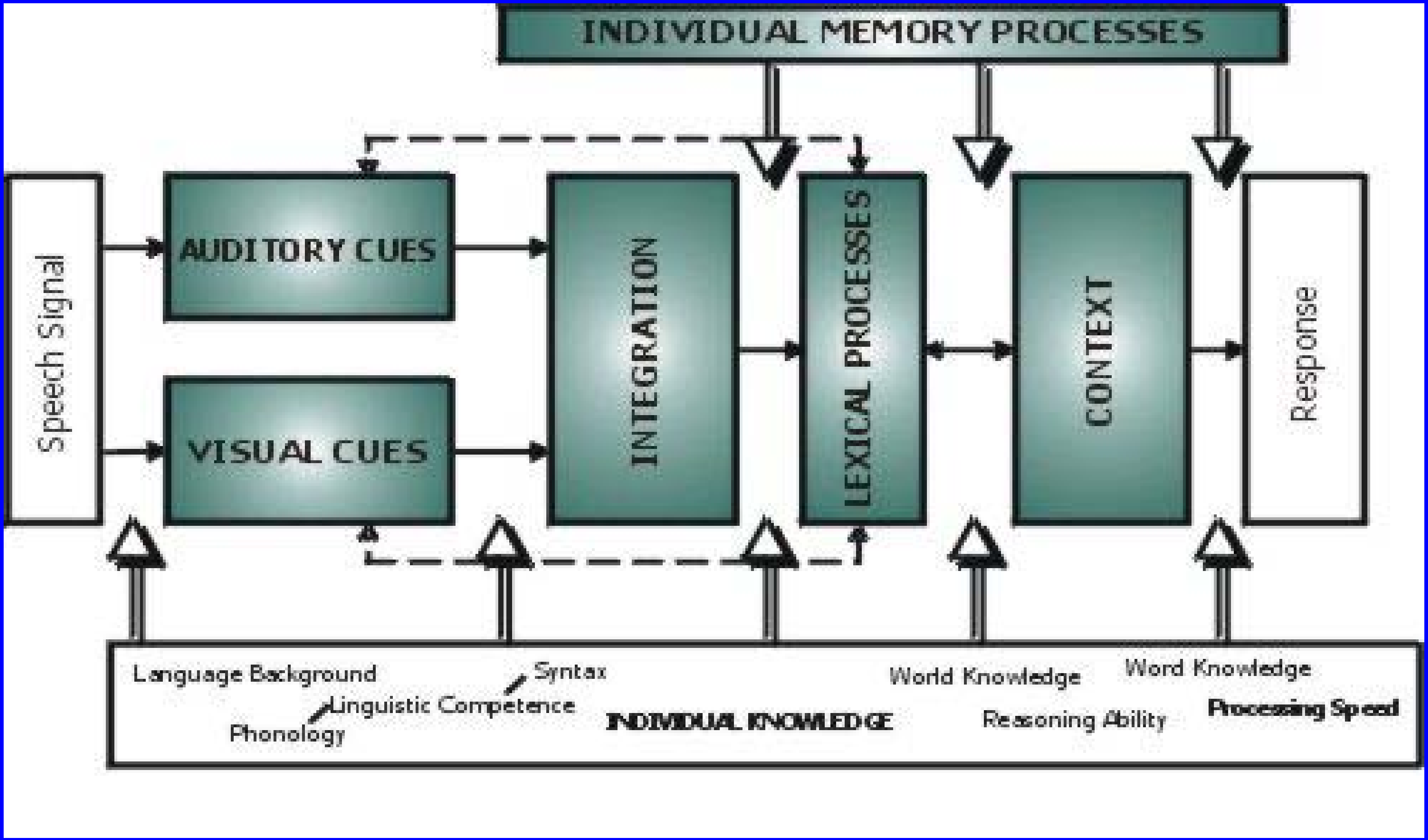
• These cues tend to be low frequency

Bimodal Coherence: Audio and Visual Comodulation

The Independence of Sensory Systems???

- ***Information is extracted independently from A and V modalities***
  - ***Early versus Late Integration***
  - ***Most models are "late integration" models***

Back to Our Conceptual Framework



The Independence of Sensory Systems ???

- Information is extracted independently from A and V modalities
  - Early versus Late Integration
  - Most models are "late integration" models

BUT

- ***Speechreading activates primary auditory cortex (cf. Sams et al., 1991)***

The Independence of Sensory Systems???

- Information is extracted independently from A and V modalities
  - Early versus Late Integration
  - Most models are "late integration" models

BUT

- Speechreading activates primary auditory cortex (cf. Sams et al., 1991)

- ***Population of neurons in cat Superior Colliculus respond only to bimodal input (cf. Stein and Meredith, 1993)***

Sensory Integration: Many Questions

- ***How is the auditory system modulated by visual speech activity?***

- ***What is the temporal window governing this interaction?***

Bimodal Coherence Masking Protection

- *BCMP (Grant and Seitz, 2000, Grant, 2001)*
  - Detection of speech in noise is improved by watching a talker (i.e., speechreading) as they produce the target speech signal, provided that the "*visible*" movement of the lips and acoustic amplitude envelope are highly correlated.

Basic Paradigm for BCMP: Exp. 1

- **Auditory-only speech detection**

  … speech

  noise noise

- **Auditory-visual speech detection**

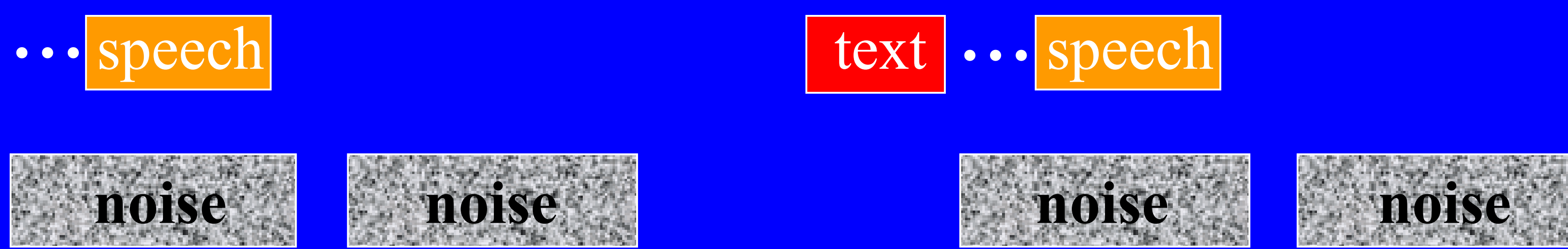
  … speech

  noise noise

Methodology for Orthographic BCMP: Exp. 2

- **Auditory-only speech detection**
- **Auditory + orthographic speech detection**



Methodology for Filtered BCMP: Exp. 3

F1 (100-800 Hz) F2 (800-2200 Hz)

- Auditory-only detection of filtered speech

• • • speech

noise

noise

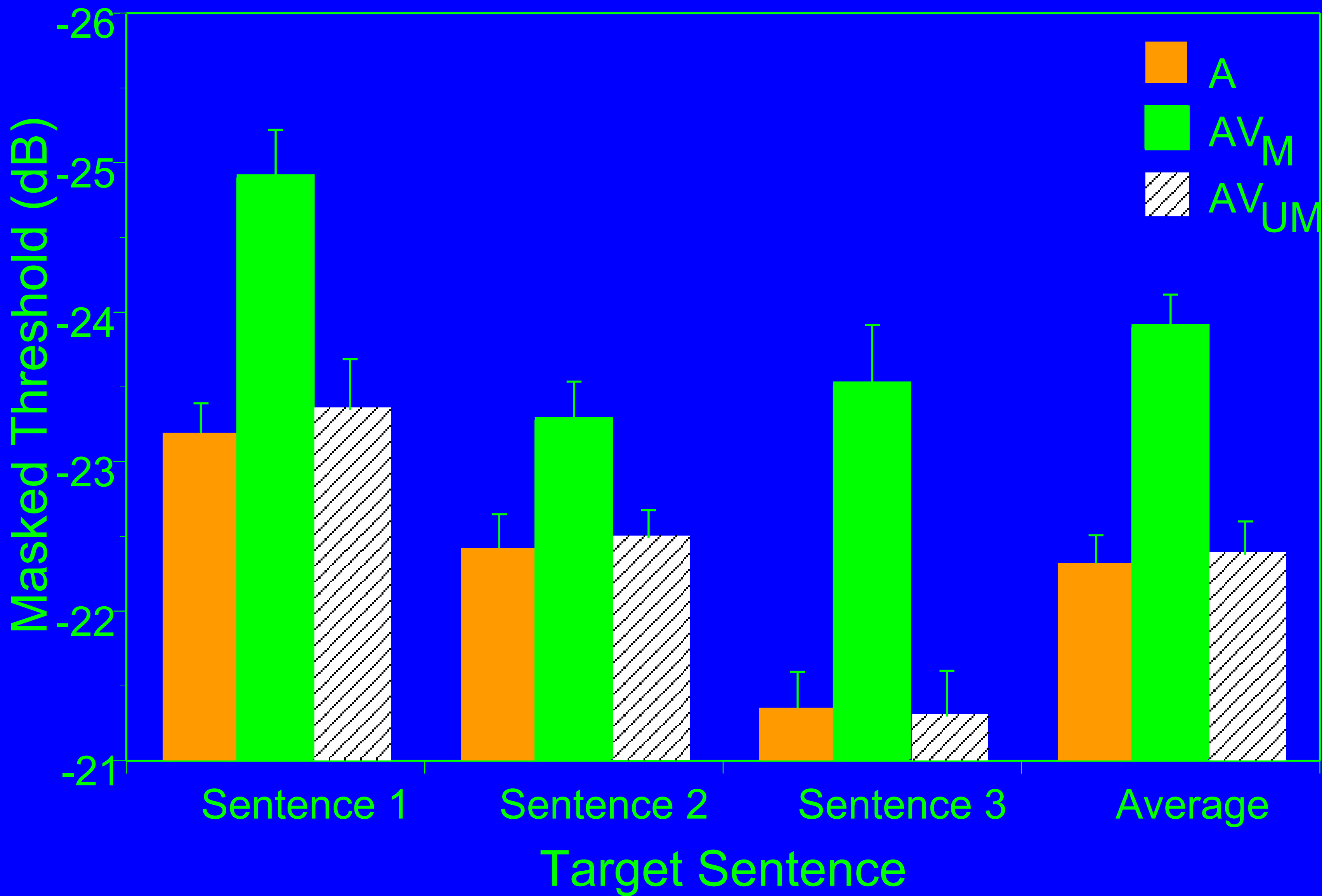
- Auditory-visual detection of filtered speech

• • • speech

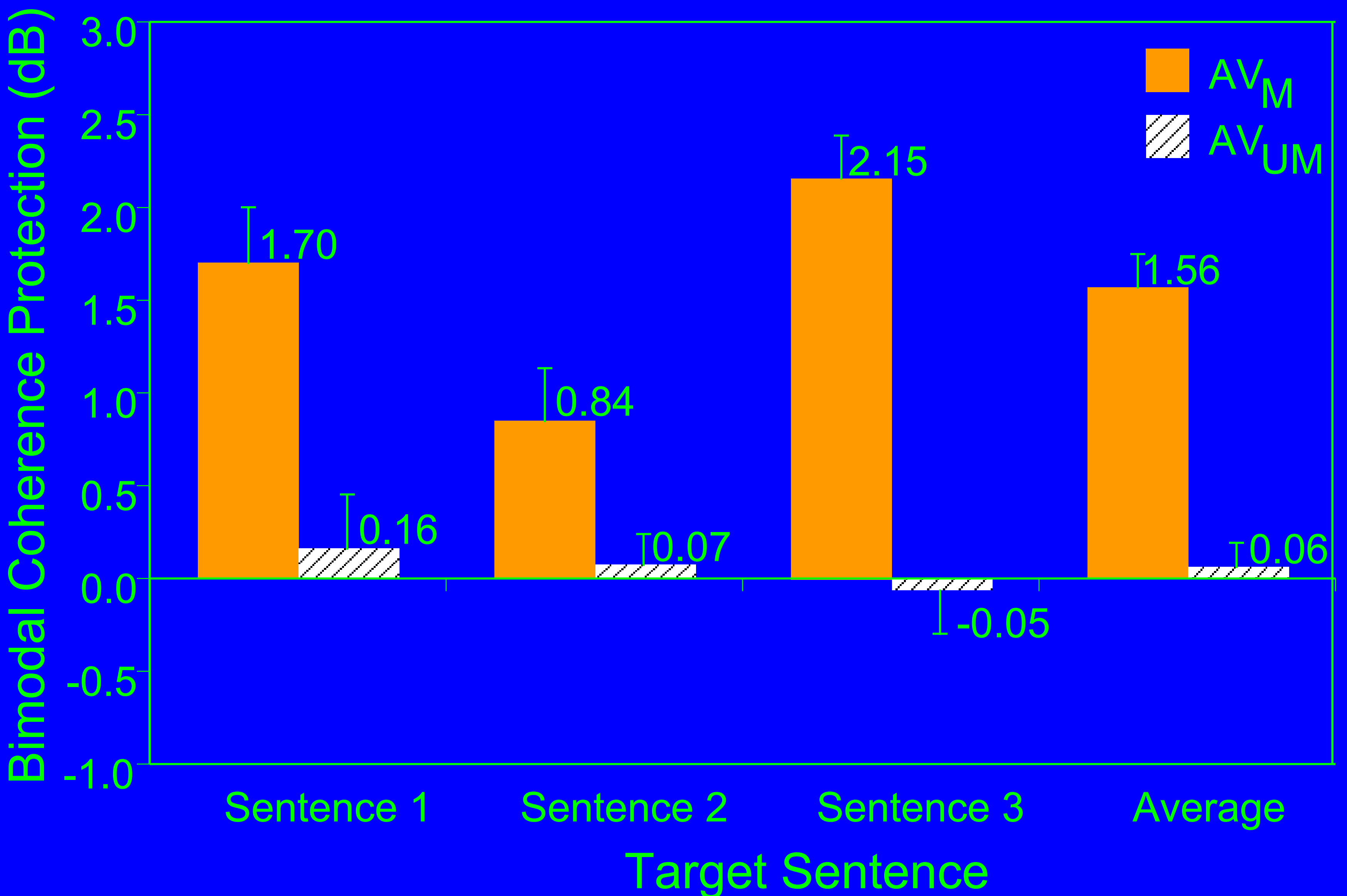
noise

noise

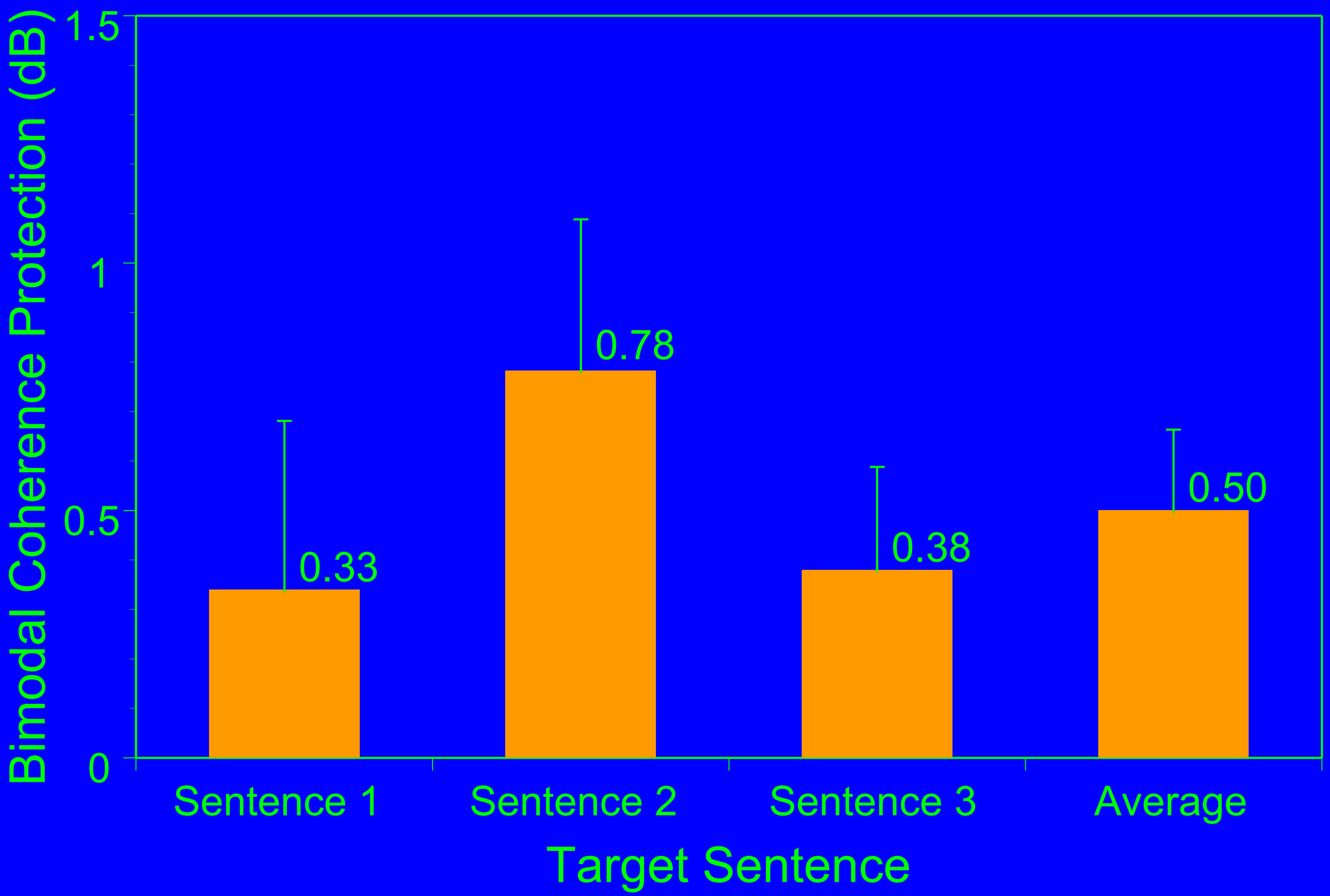
Congruent versus Incongruent Speech



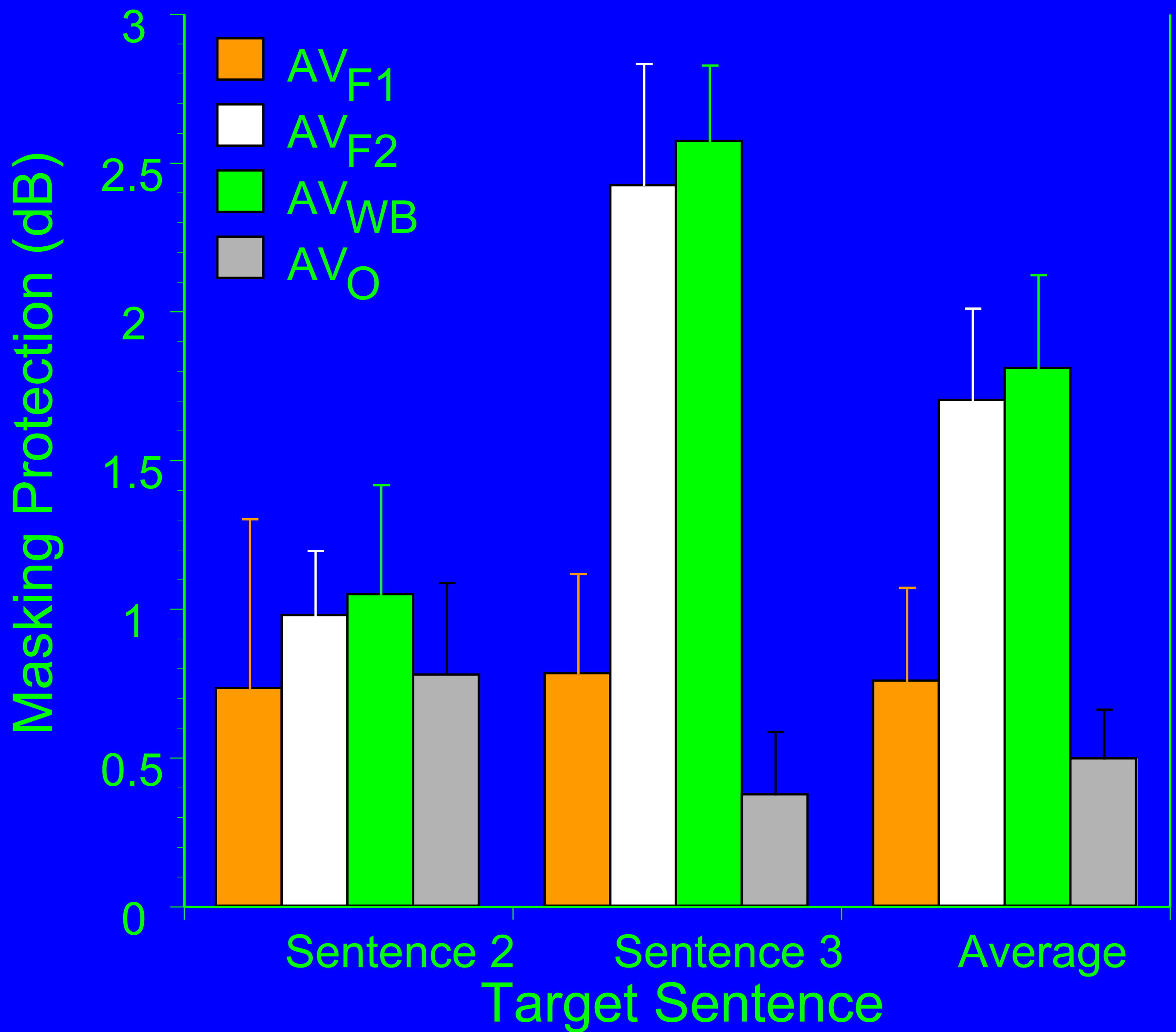
BCMP for Congruent and Incongruent Speech



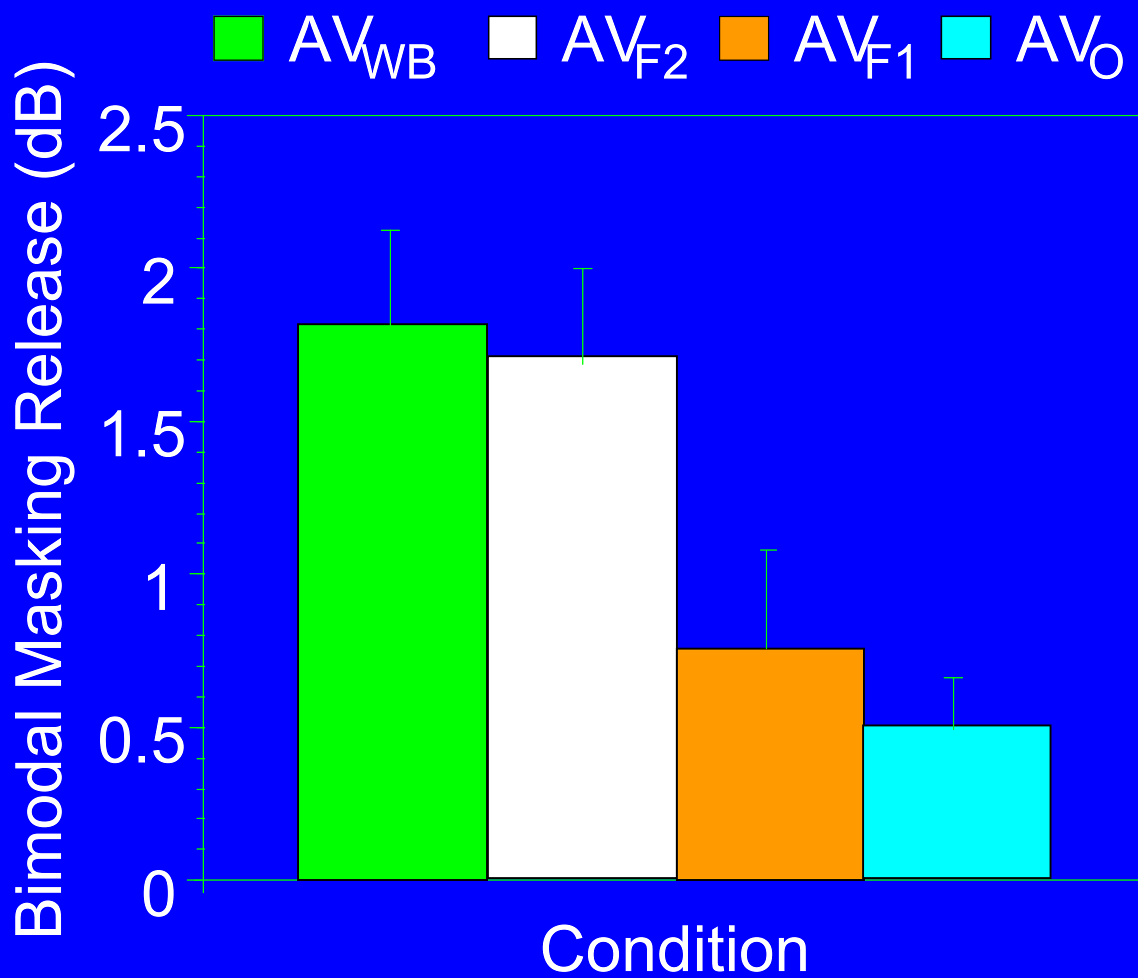
BCMP for Orthographic Speech



BCMP for Filtered Speech

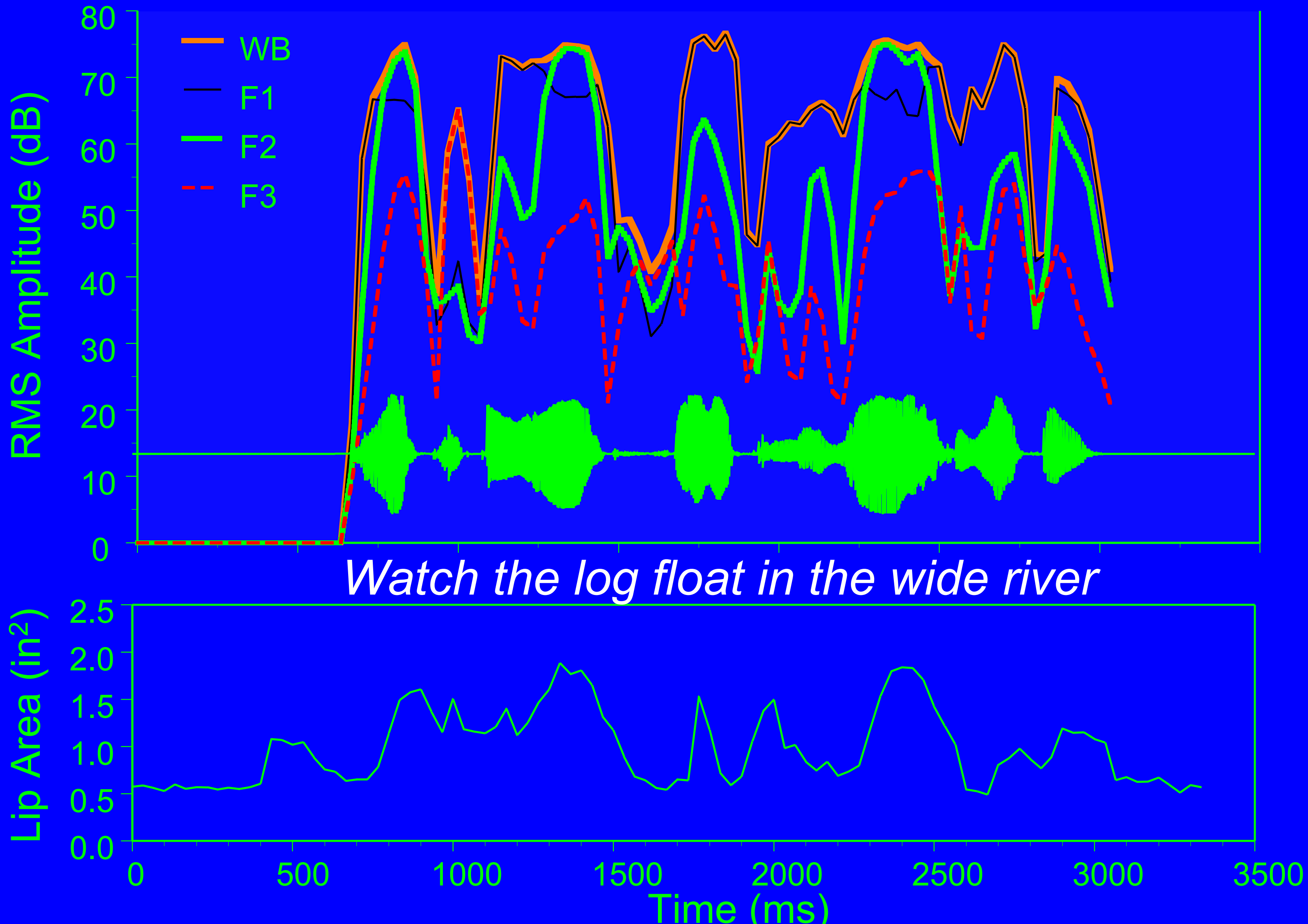


Average BCMP

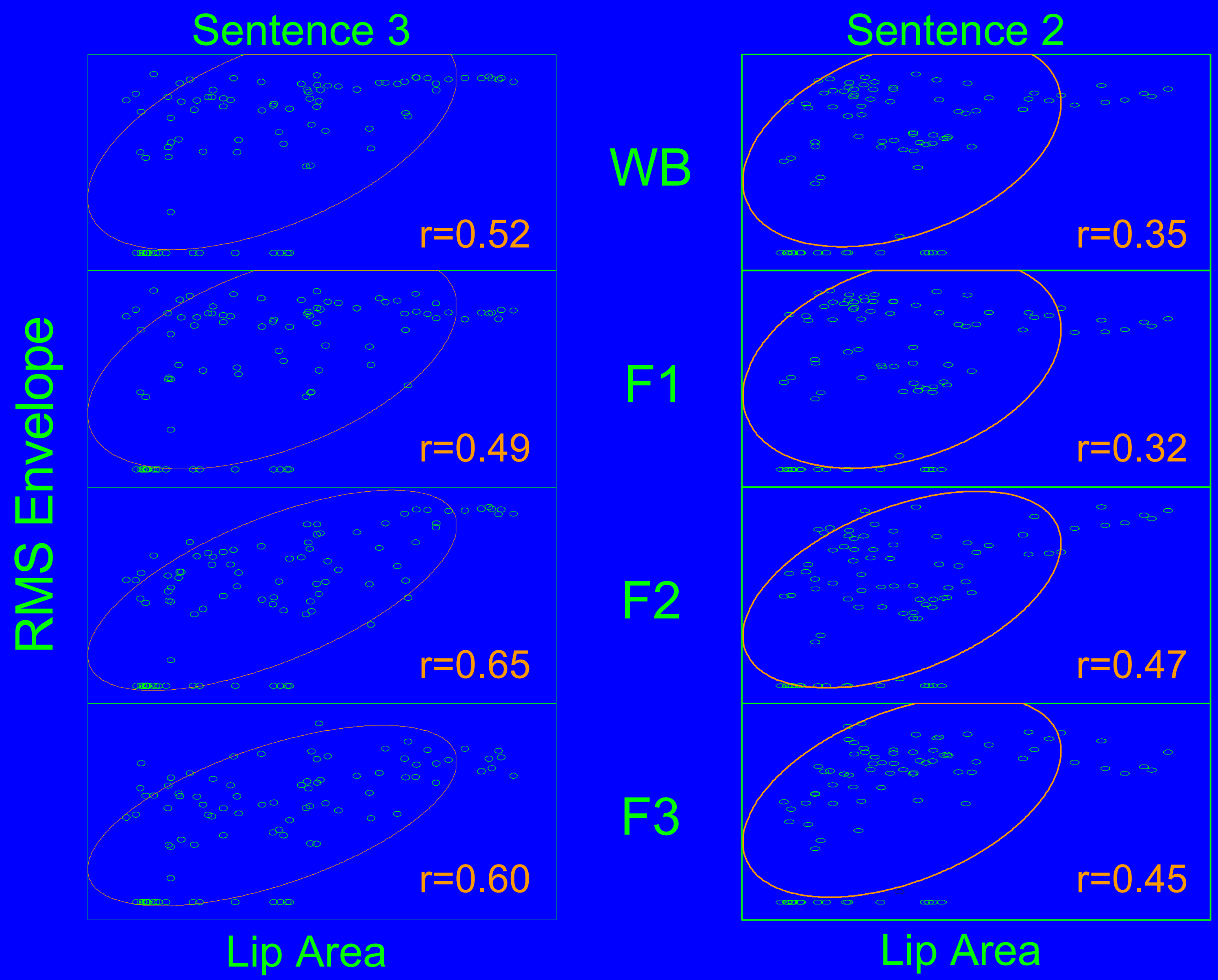


Bimodal comodulation masking protection for wideband speech (WB), filtered speech (F2 and F1), and for orthographically cued speech (O).

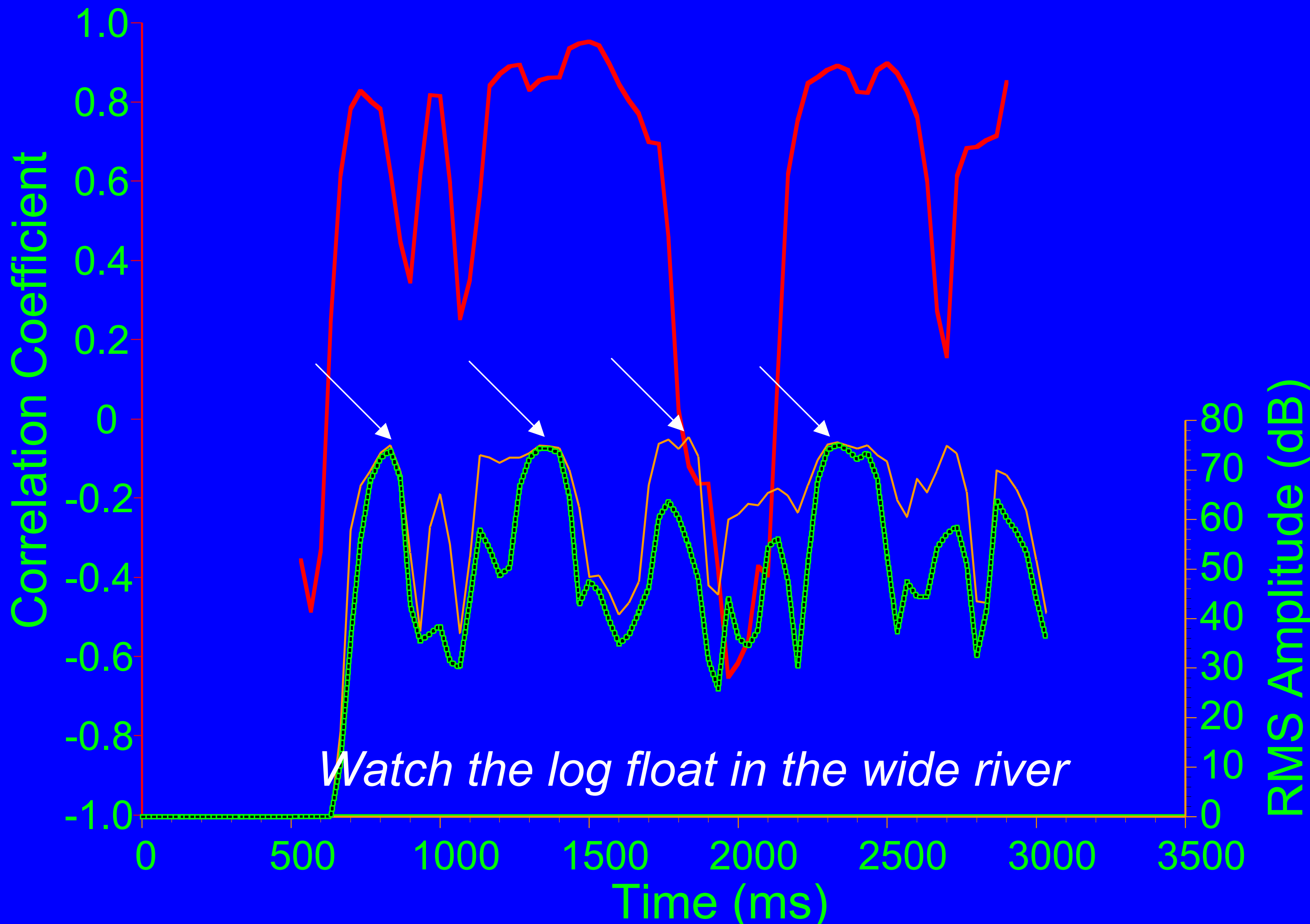
Acoustic Envelope and Lip Area Functions



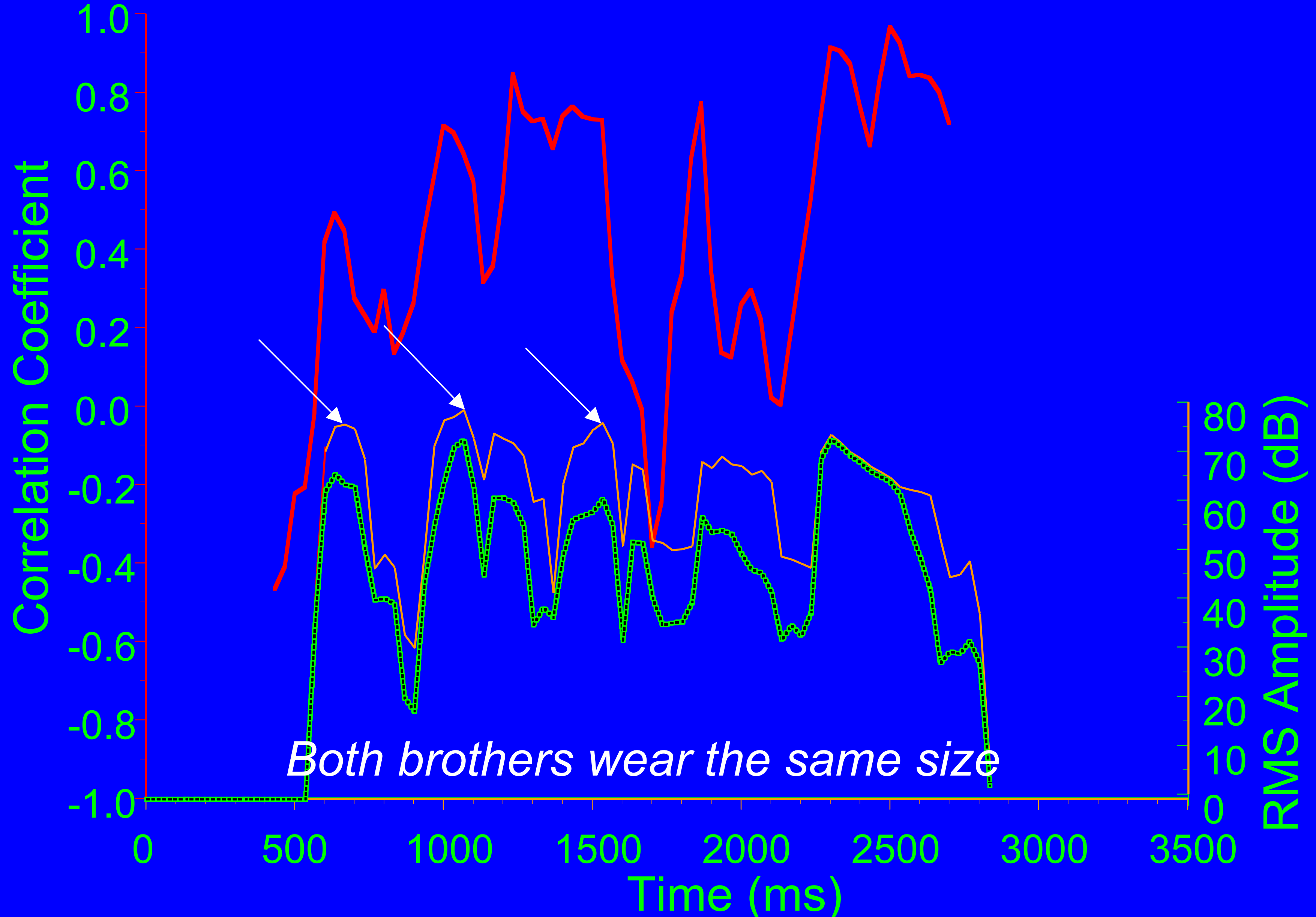
Cross Modality Correlation - Lip Area versus Amplitude Envelope



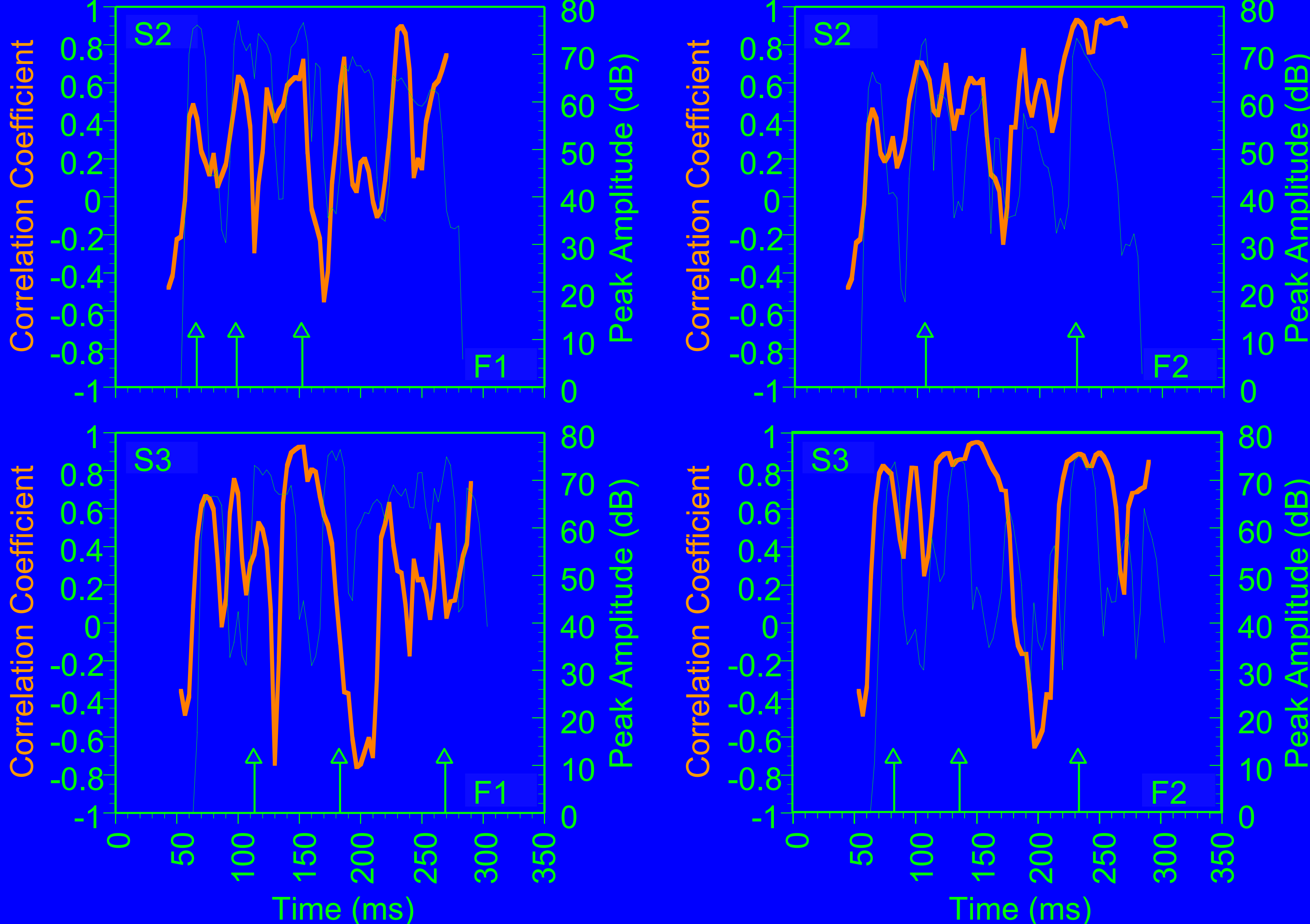
Local Correlations (Lip Area versus F2-Amplitude Envelope



Local Correlations (Lip Area versus F2-Amplitude Envelope



Cross Modality Correlations (lip area versus acoustic envelope)



BCMP – Summary

- ***Speechreading reduces auditory speech detection thresholds by about 1.5 dB (range: 1-3 dB depending on sentence)***

BCMP – Summary

- Speechreading reduces auditory speech detection thresholds by about 1.5 dB (range: 1-3 dB depending on sentence)
- ***Amount of BCMP depends on the degree of coherence between acoustic envelope and facial kinematics***

BCMP – Summary

- Speechreading reduces auditory speech detection thresholds by about 1.5 dB (range: 1-3 dB depending on sentence)
- Amount of BCMP depends on the degree of coherence between acoustic envelope and facial kinematics
- ***Providing listeners with explicit (orthographic) knowledge of the identity of the target sentence reduces speech detection thresholds by about 0.5 dB, independent of the specific target sentence***

BCMP – Summary

- Speechreading reduces auditory speech detection thresholds by about 1.5 dB (range: 1-3 dB depending on sentence)
- Amount of BCMP depends on the degree of coherence between acoustic envelope and facial kinematics
- Providing listeners with explicit (orthographic) knowledge of the identity of the target sentence reduces speech detection thresholds by about 0.5 dB, independent of the specific target sentence
- ***Manipulating the degree of coherence between area of mouth opening and acoustic envelope by filtering the target speech has a direct effect on BCMP***

Temporal Window for Auditory-Visual Integration

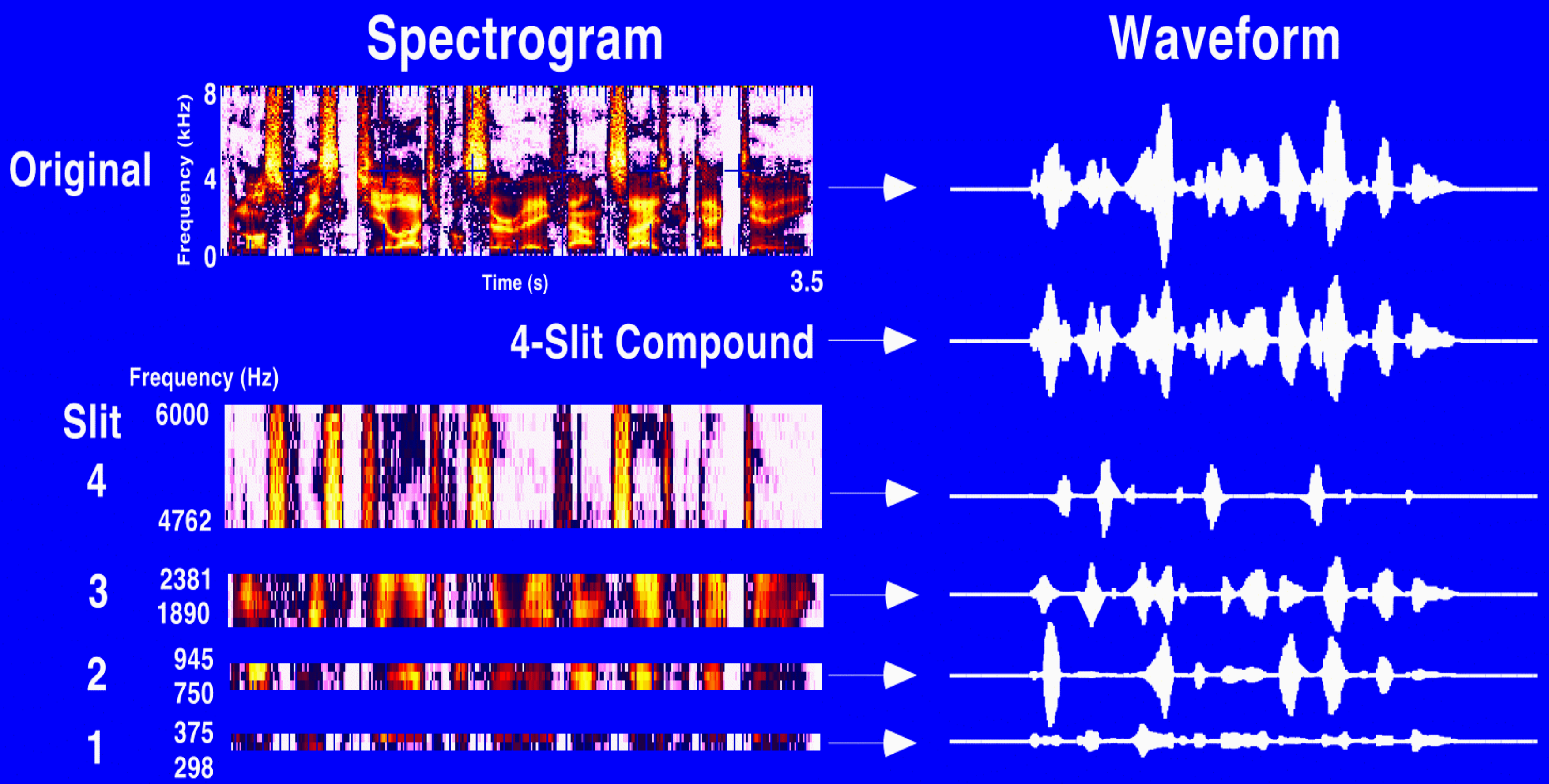
AUDIO-ALONE EXPERIMENTS

Audio (Alone) Spectral Slit Paradigm

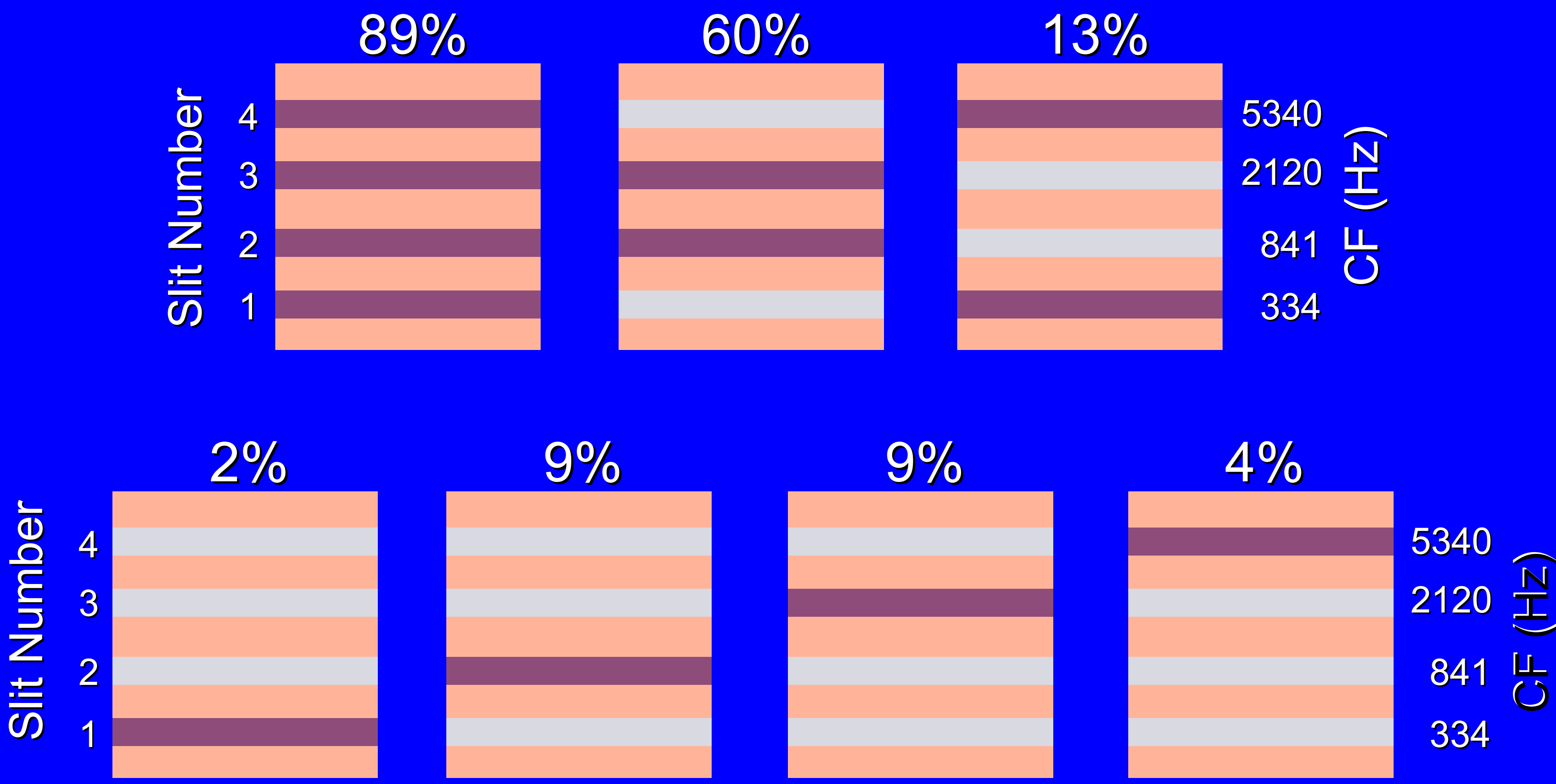
The edge of each slit was separated from its <u>nearest</u> neighbor by an octave

Can listeners decode spoken sentences using just four narrow (1/3 octave) channels ("slits") distributed across the spectrum? – YES (cf. next slide)

What is the intelligibility of each slit alone and in combination with others?



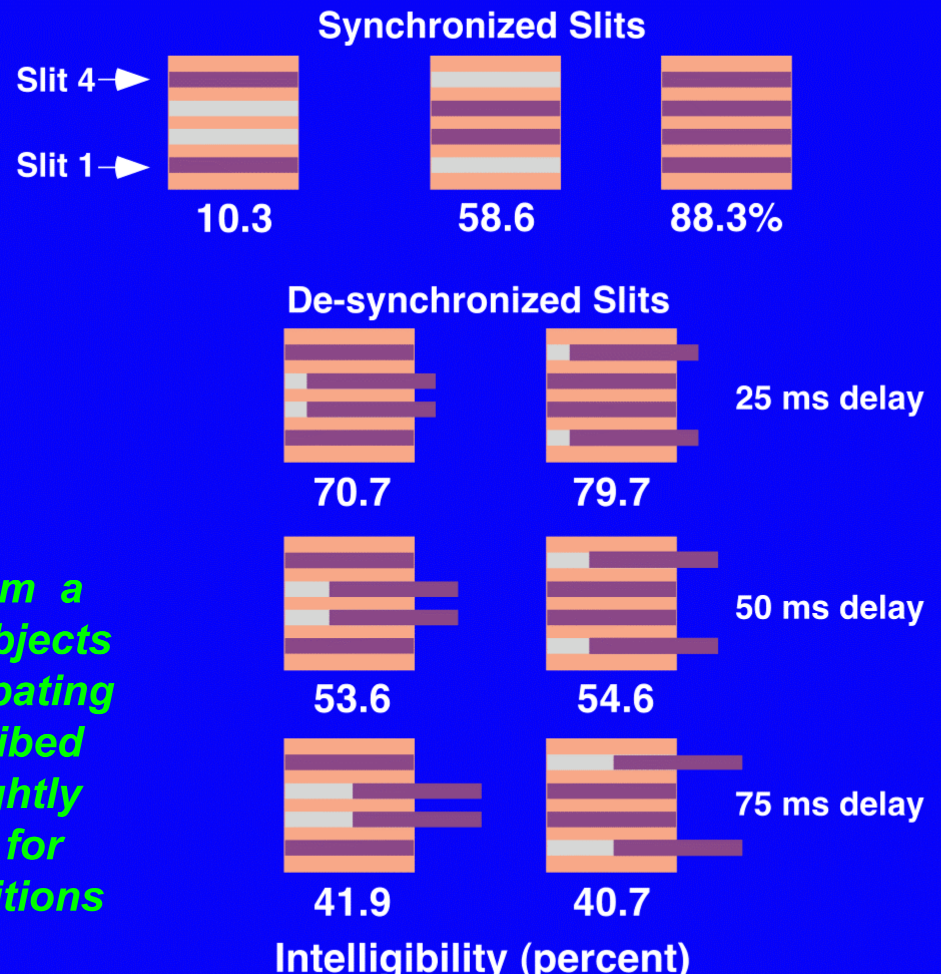
Word Intelligibility - Single and Multiple Slits



Slit Asynchrony Affects Intelligibility

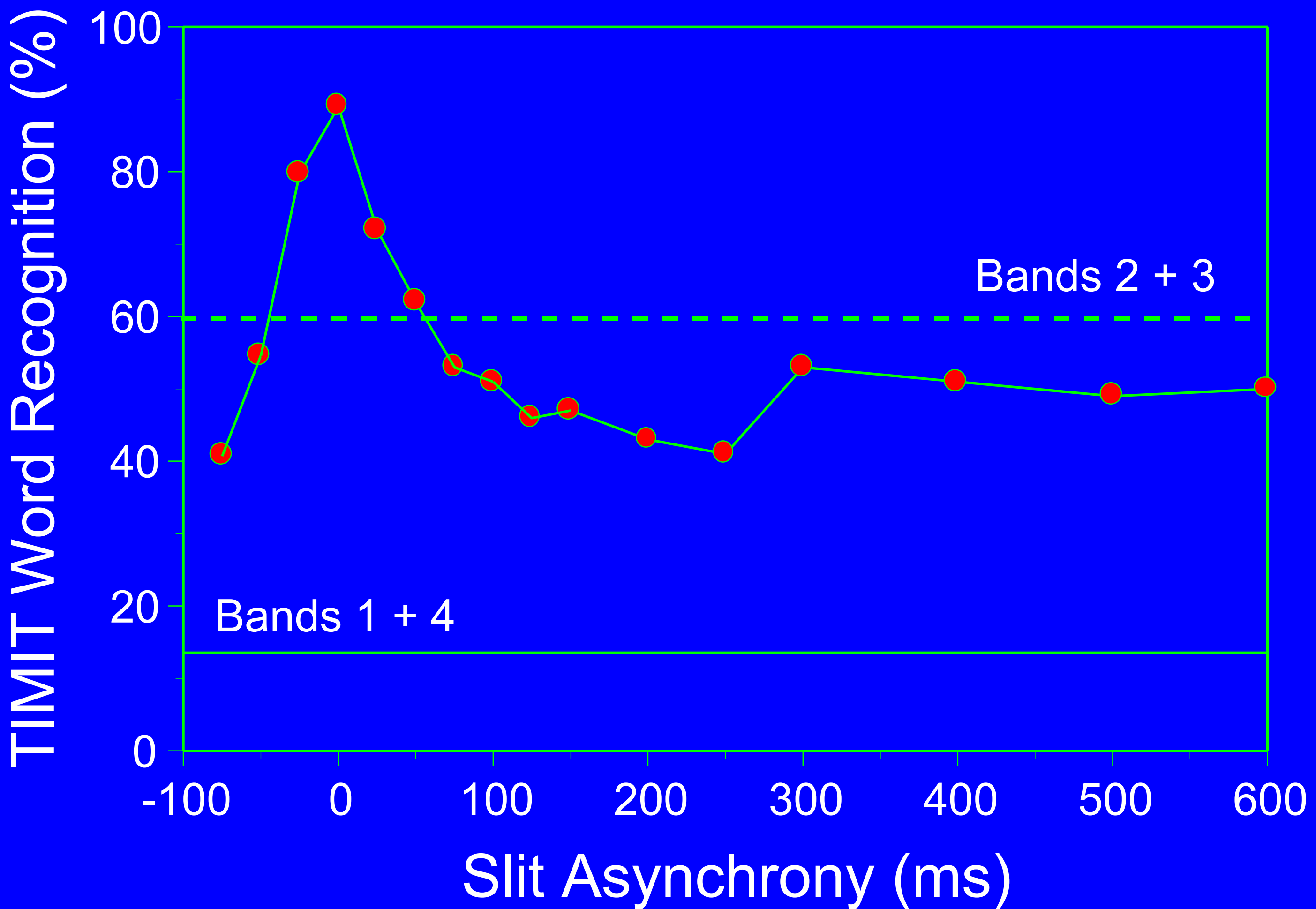
Desynchronizing the slits by more than 25 ms results in a significant decline in intelligibility

The effect of asynchrony on intelligibility is relatively symmetrical



These data are from a different set of subjects than those participating in the study described earlier - hence slightly different numbers for the baseline conditions

Cross-Spectral Temporal Asynchrony Effects



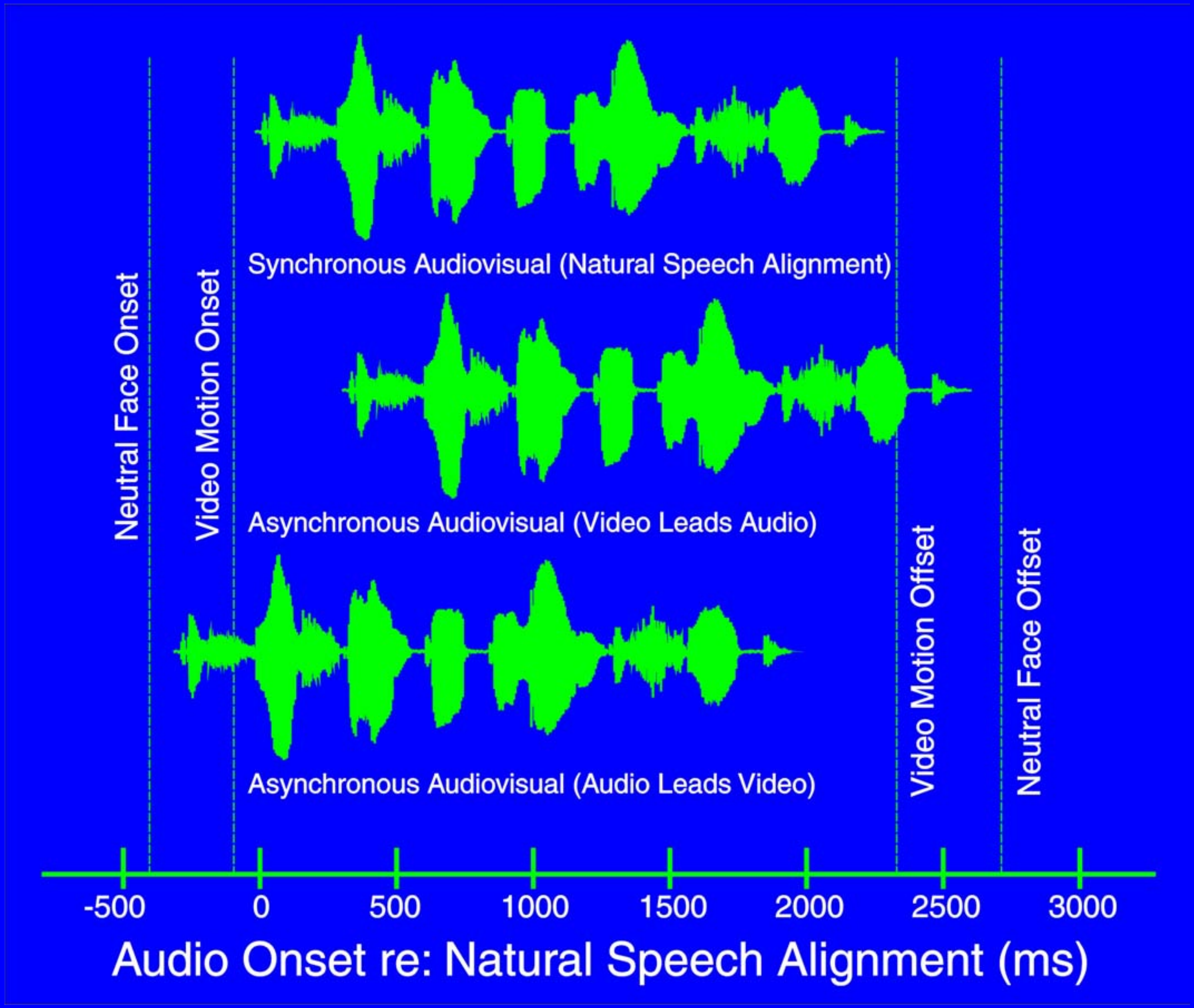
From Greenberg, Arai, and Silipo (1998). Proc. ICSLP, Sydney, Dec. 1-4.

AUDIO-VISUAL EXPERIMENTS

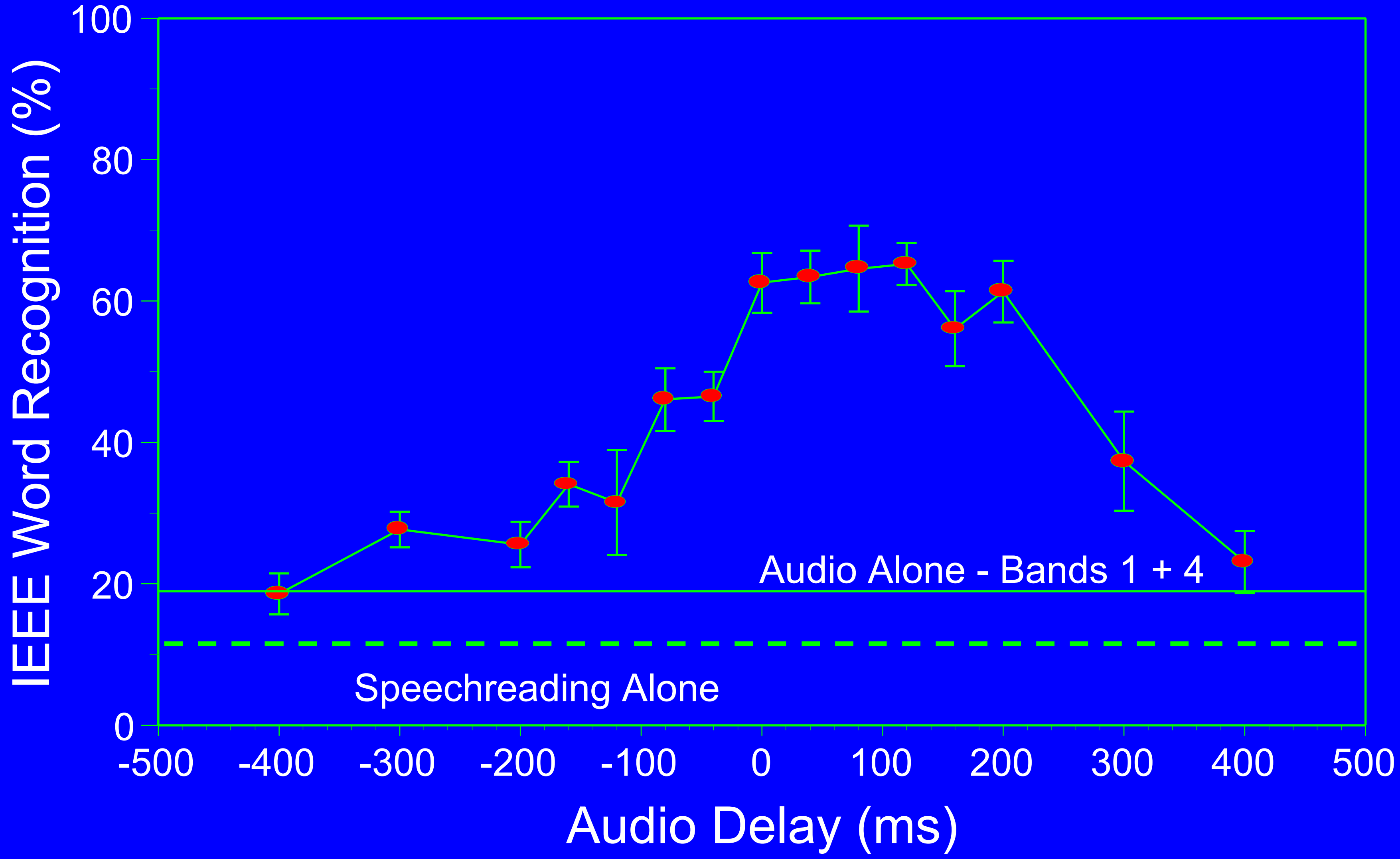
Auditory-Visual Tasks

- 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
    - 2IFC adaptive tracking

Auditory-Visual Asynchrony - Paradigm



Cross-Modality Temporal Asynchrony Effects: Sentences

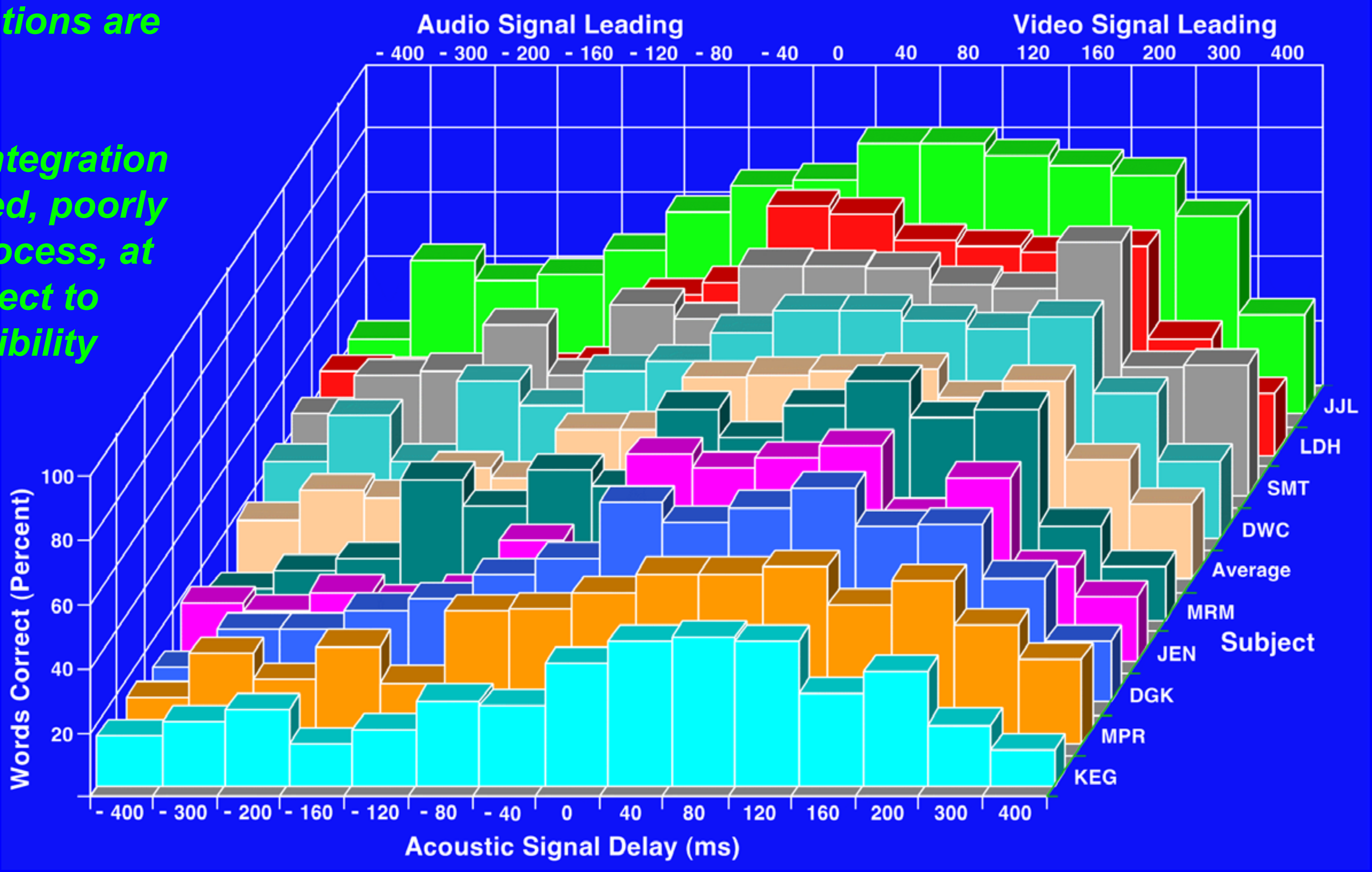


Auditory-Visual Integration - by Individual S's

These data are complex, but the implications are clear.

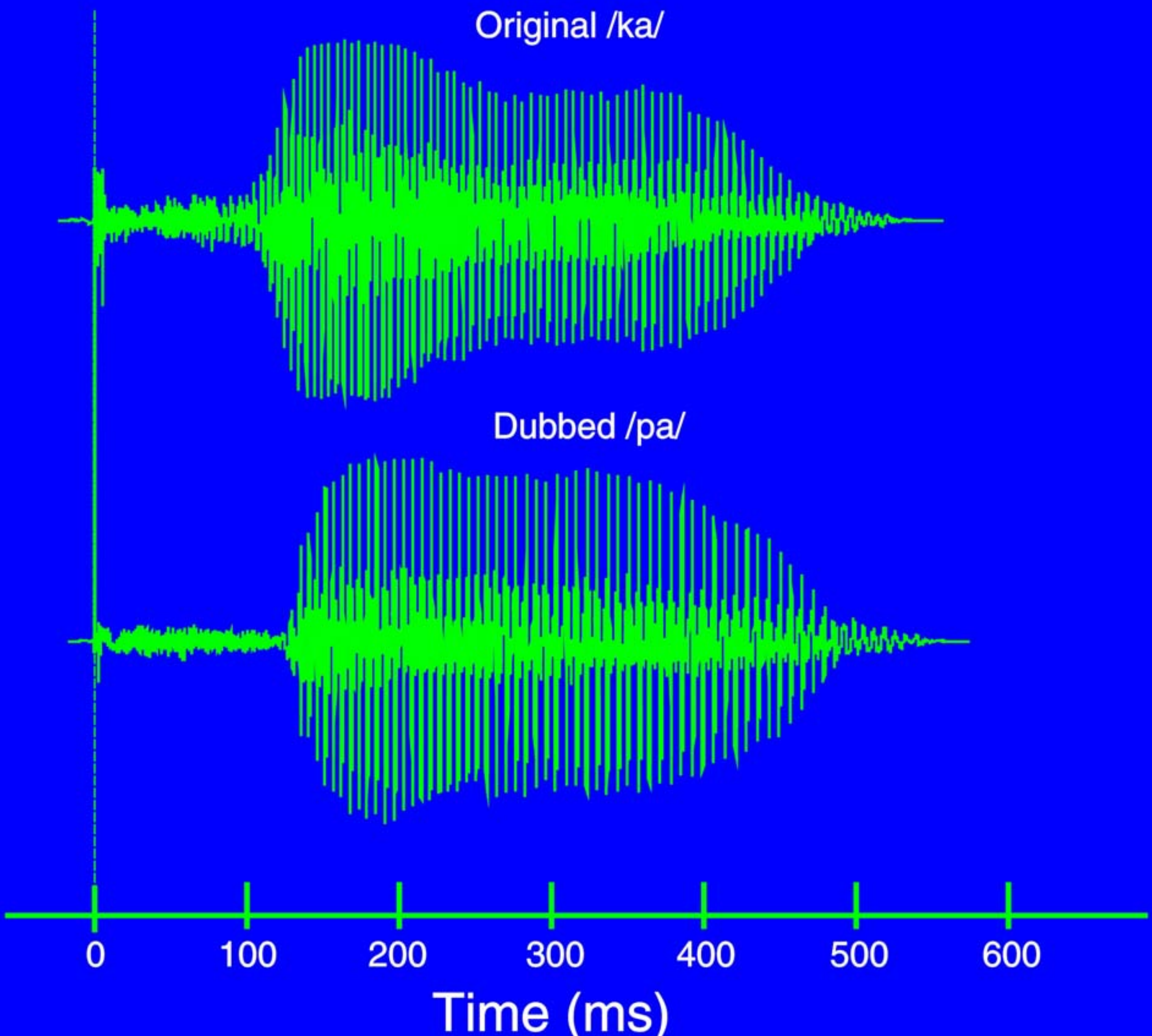
Audio-visual integration is a complicated, poorly understood process, at least with respect to speech intelligibility

Video signal leading is better than synchronous for 8 of 9 subjects

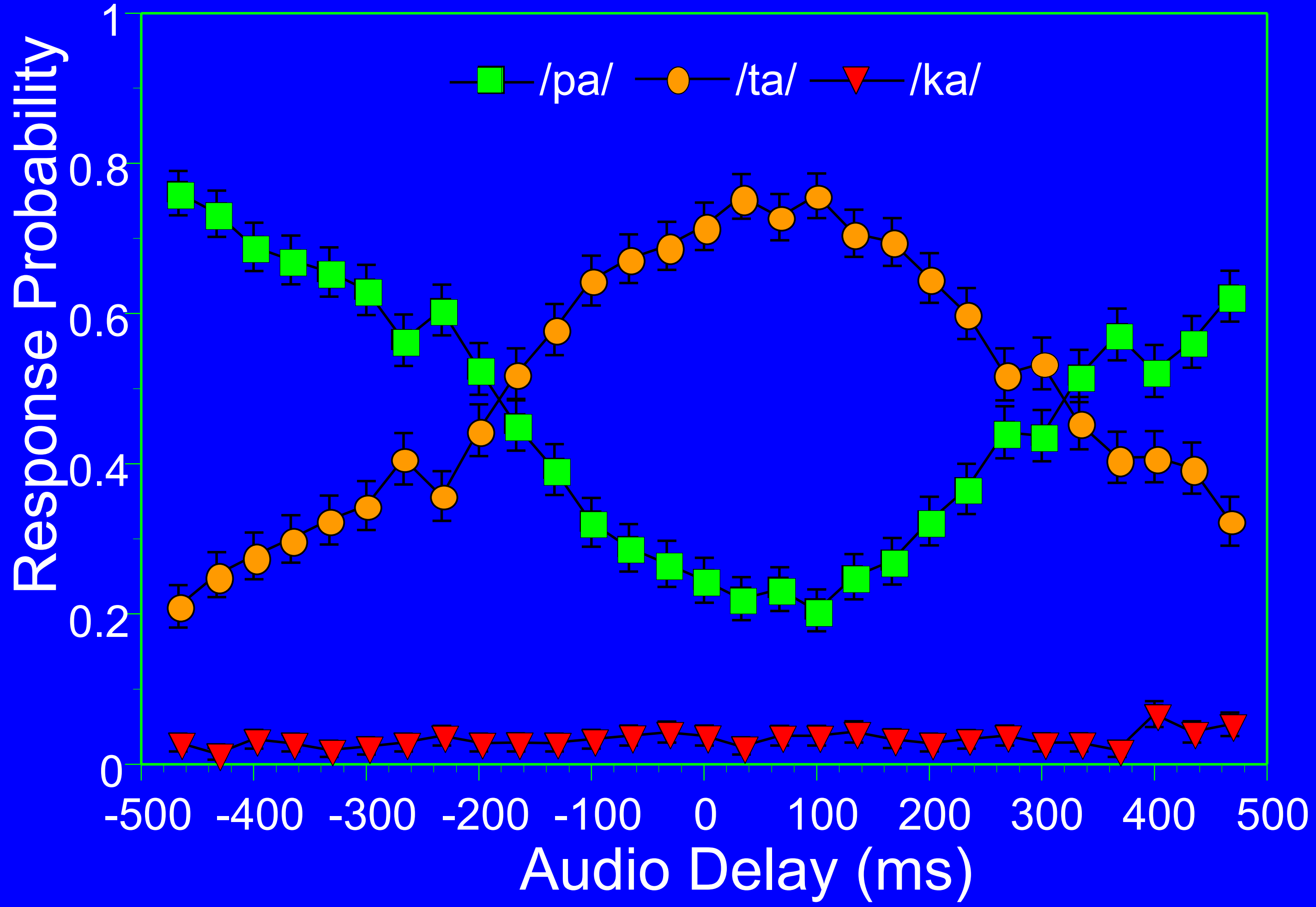


Variation across subjects

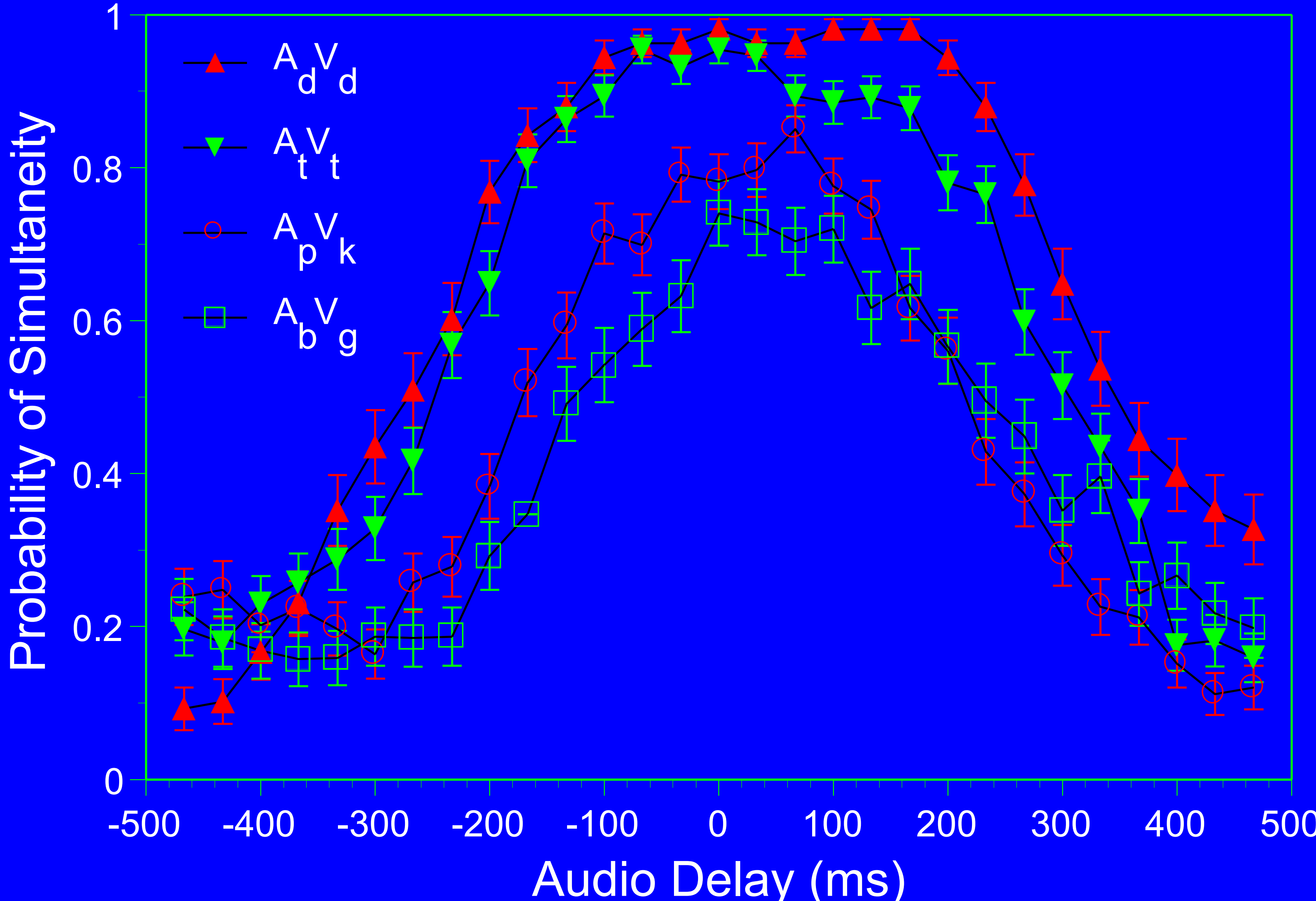
McGurk Synchrony Paradigm



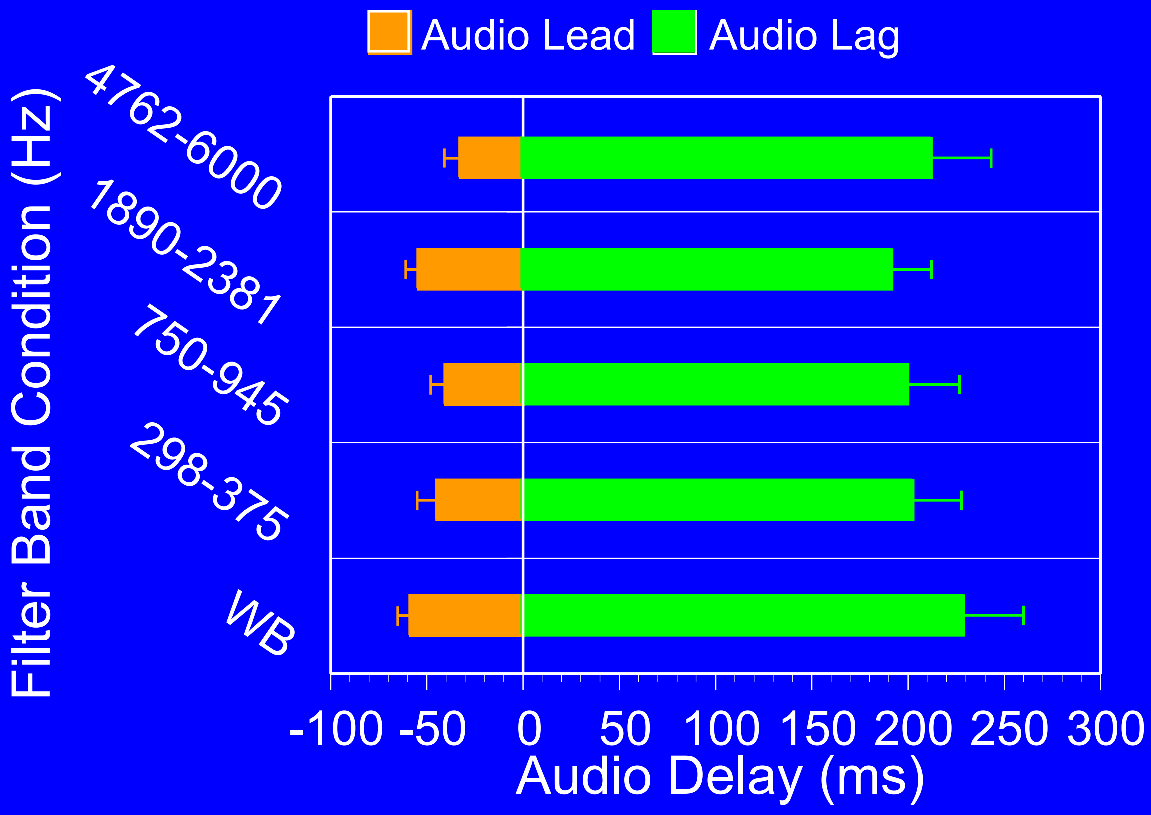
Temporal Integration in the McGurk Effect



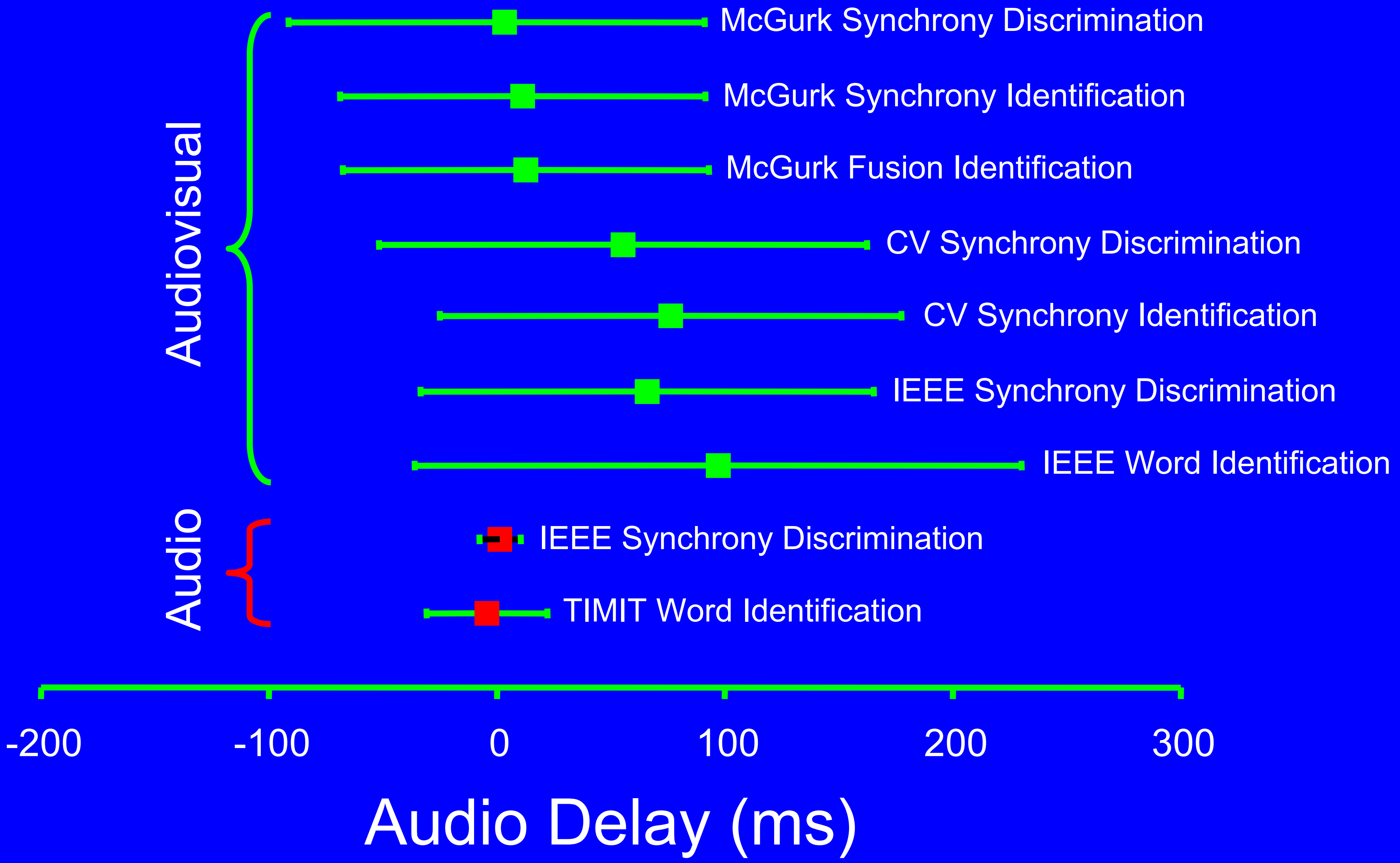
Simultaneity Judgements - Natural vs. McGurk AV Tokens



Spectro-Temporal Synchrony Discrimination



Temporal Window of Integration



Spectro-Temporal Integration: Summary

Within Modality (Cross- Spectral Auditory Integration)

Spectro-Temporal Integration: Summary

Within Modality (Cross- Spectral Auditory Integration)

- TWI is symmetrical

Spectro-Temporal Integration: Summary

Within Modality (Cross- Spectral Auditory Integration)

- TWI is symmetrical
- TWI roughly 50 ms or less (phoneme?)

Spectro-Temporal Integration: Summary

Within Modality (Cross- Spectral Auditory Integration)

- TWI is symmetrical
- TWI roughly 50 ms or less (phoneme?)

Across Modality (Cross-Modal AV Integration)

Spectro-Temporal Integration: Summary

Within Modality (Cross- Spectral Auditory Integration)

- TWI is symmetrical
- TWI roughly 50 ms or less (phoneme?)

Across Modality (Cross-Modal AV Integration)

- TWI is highly asymmetrical favoring visual leads

Spectro-Temporal Integration: Summary

Within Modality (Cross- Spectral Auditory Integration)

- TWI is symmetrical
- TWI roughly 50 ms or less (phoneme?)

Across Modality (Cross-Modal AV Integration)

- TWI is highly asymmetrical favoring visual leads
- TWI is roughly 160-220 ms (syllable?)

Spectro-Temporal Integration: Summary

Within Modality (Cross- Spectral Auditory Integration)

- TWI is symmetrical
- TWI roughly 50 ms or less (phoneme?)

Across Modality (Cross-Modal AV Integration)

- TWI is highly asymmetrical favoring visual leads
- TWI is roughly 160-220 ms (syllable?)
- TWI for Incongruent CV's (McGurk Stimuli) is not as wide as TWI for natural congruent CV's

Auditory-Visual Speech Perception Laboratory

Walter Reed Army Medical Center

Army Audiology and Speech Center

Washington, DC USA

http://www.wramc.amedd.army.mil/departments/aasc/avlab

grant@tidalwave.net