Speech Perception in Noisy Environments

We study neural bases of speech perception (fMRI, EEG), when intelligibility relies on what we...

1) Hear: Acoustic cues (attending to pitch and space; suppressing echoes)
2) See: Audiovisual integration (often more benefit than aids)
3) Know: Top-down inference (prior knowledge of words, meanings, syntax)

Then the bartender says: “But I was talking to the pigeon!”

Overview

I. ATTENTION IN MULTI-TALKER ENVIRONMENTS (Kevin Hill)
What neural systems coordinate attention to one speaker among many, based on different acoustic attributes (space, pitch)?

II. HEARING vs UNDERSTANDING SPEECH IN NOISE (Chris Bishop)
What neural systems reflect understanding speech rather than merely hearing it (auditory and visual contributions)?

III. TOP-DOWN INFERENCE TO IMPROVE INTELLIGIBILITY (Tony Shahin)
What networks mediate illusory filling-in of degraded speech?

Attention in Multi-talker Environments

→ Experimental question:
  What neural systems coordinate attention to one speaker among many, based on different attributes (space, pitch)?

1) Attentional control networks (“sources” of attention) candidates: superior parietal & prefrontal
2) Attentional selection (“targets” of attention) candidates: space – dorsal stream; pitch – ventral stream

Stimuli and Task

• three simultaneous talkers differ in pitch and location
• continuous, meaningful sentences (Harvard/IEEE corpus)
• location in virtual acoustic space (head-related transfer function: HRTF)
• subjects press button at sentence boundaries for target talker
Trial Structure and Performance

- event-related fMRI
- two-part trials: cue (pitch, location, or rest) and stimuli
- performance using pitch and space equalized adaptively, pre-scan
- same stimuli, same performance: fMRI results cannot be attributed to different acoustics or effort

Attribute-general activity precedes stimuli (Control)

- [Attention Cues > Rest Cue]: pitch, space, or both
- attentional control towards either attribute activates same general network
- with biases to pitch (inferior frontal) or space (superior frontal-parietal)

Attribute-specific activity during stimuli (Selection)

- [Talker streamed by location & Talker streamed by pitch]
- attentional and speech areas generally all active for both conditions
- but streaming by different attributes leads to differential activation in:
  - posterior parietal for space (“where” stream)
  - superior temporal sulcus for pitch (“what” stream)

Summary: Multi-talker Attention

- IFG bias toward using pitch as cue
- Superior parietal “bottom up” areas modulated more when selecting by pitch
- Posterior parietal “bottom up” areas modulated more when selecting by space

Overview

I. ATTENTION IN MULTI-TALKER ENVIRONMENTS
   What neural systems coordinate attention to one speaker among many, based on different acoustic attributes (space, pitch)?

II. HEARING vs UNDERSTANDING SPEECH IN NOISE
   What neural systems reflect understanding speech rather than merely hearing it (auditory and visual contributions)?

III. TOP-DOWN INFERENCE TO IMPROVE INTELLIGIBILITY
   What networks mediate illusory filling-in of degraded speech?

I. ATTENTION IN MULTI-TALKER ENVIRONMENTS
   Fronto-parietal control network has biases for space or pitch, modulates earlier cortices in “where” or “what” streams during attentional selection.

II. HEARING vs UNDERSTANDING SPEECH IN NOISE
   What neural systems reflect understanding speech rather than merely hearing it (auditory and visual contributions)?

III. TOP-DOWN INFERENCE TO IMPROVE INTELLIGIBILITY
   What networks mediate illusory filling-in of degraded speech?
Experimental question:
What neural systems distinguish understanding from merely hearing speech?
(Naturalistic conditions: auditory & audiovisual)

1) We know little about the neural bases of this crucial perceptual distinction candidates: STS, posterior temporal
2) We do not know how visual speech contributes to improved understanding candidate: visual motion MT+

Design
- mixed block/event-related fMRI
- 4 vowel-consonant-vowels (VCVs) in 16-talker babble.
  Task: “Heard” vs “Understood”
- conditions: Auditory-only or Audiovisual (synch or offset)
- video averaged across VCVs – temporal cues preserved
- same stimuli, different percept: adaptively rove VCV signal-to-noise ratio near threshold

Understanding Speech in Noise
- Understanding > Hearing for all conditions
- reflects perceptual, not stimulus-driven distinction
- left STS primarily reflects auditory-only understanding
- LTO adjacent to functionally-identified visual motion area

Visual Contribution to Understanding
- these areas reflect improvement in intelligibility with vision:
  Understanding > Hearing and Audiovisual Synchronous > Offset
- LTO may use integration of physical stimulus attributes (motion)
- MTL may use integration of more abstract properties (e.g. identity)
- does not include anterior STS

Summary: Understanding Speech in Noise
- increased communication between LTO, left MTL, STS
- Understanding via auditory speech
- Understanding via more abstract multisensory integration, verbal recognition memory
Overview

I. ATTENTION IN MULTI-TALKER ENVIRONMENTS
   Fronto-parietal control network has biases for space or pitch, modulates earlier cortices in “where” or “what” streams during attentional selection.

II. HEARING vs UNDERSTANDING SPEECH IN NOISE
   What neural systems reflect understanding speech rather than merely hearing it (auditory and visual contributions)?

III. TOP-DOWN INFERENCE TO IMPROVE INTELLIGIBILITY
   What networks mediate illusory filling-in of degraded speech?

Phonemic Restoration Experiment

- Phonemic restoration \(\rightarrow\) illusion of continuity, even if speech artificially removed from signal
- This temporal “filling in” a general mechanism across stimuli, species

Psychophysics suggests two dissociable mechanisms

1) Unconscious repair: reconstructing low-level sensory representation
2) Continuity illusion: perceptual evaluation of stimulus coherence

→ Experimental questions:
   - Are Continuity and Repair neurally dissociable mechanisms?
   - Can we distinguish areas that apply high-level prior knowledge (word-level templates)?

Stimuli and Task

- 230 tri-syllabic English words with 1-3 fricatives/affricates
- Two stimulus conditions: fricative/affricate combined with or replaced by white noise burst
- Task 2AFC: “Is speech continuous through or interrupted by the noise?”
- One session with pseudoword controls (e.g. “bressor”)

Stimuli near Illusion Threshold

- Two stimulus conditions & two responses = 4 combinations
  - Same stimuli, different percepts and vice versa
  - Noise duration determines illusion: too long \(\rightarrow\) no illusion
  - Noise duration adaptively adjusted around subject’s threshold for illusion

Design and Interpretation

- Interpretations based on contrasting:
  - Continuous vs interrupted percepts (controlled for stimuli), and
  - Continuous vs interrupted stimuli (controlled for percept)
Continuity Illusion

• Illusion > Illusion-failure
• this contrast is between conditions with identical stimuli (interrupted), differing only in what subjects perceived

Unconscious repair of degraded speech

• Illusion > Natural
• this contrast is between conditions with identical percept (continuous), differing only in stimulus properties
• could give same areas as Continuity; rather, are dissociable

Mechanisms: Template Matching or Gestalt

• at least two mechanisms could account for phonemic restoration:
  1) gestalt completion (rule-based)
  2) template matching (object-based)

• comparing restoration of words vs nonwords identifies regions for word-level template matching
• may be evident in illusion areas and/or repair areas

Functional Connectivity

• Hypothesis: continuity areas should communicate with repair areas
• left Angular Gyrus communicates more with Broca’s, insula, and left preSMA when speech is repaired

Summary: Phonemic Restoration

• left Angular gyrus / STS template matching relies on communication with repair regions, e.g. Broca’s
• right Superior Temporal Sulcus mediates illusion using low-level (gestalt) completion

Mechanisms: Template Matching or Gestalt

• left Angular gyrus / STS mediates illusion more for words than pseudowords
  → template-matching

• right STS, Precuneus, SFS mediate illusion irrespective of word/pseudoword distinction
  → gestalt completion
  → template-matching

• Broca’s, Insula, and pre-SMA favor repairing degradations in words
  → template-matching
Overview

I. ATTENTION IN MULTI-TALKER ENVIRONMENTS
   Fronto-parietal control network has biases for space or pitch, modulates earlier cortices in “where” or “what” streams during attentional selection.

II. HEARING vs UNDERSTANDING SPEECH IN NOISE
   Both anterior and posterior temporal areas reflect understanding; LTO and MTL use audiovisual integration to improve it.

III. TOP-DOWN INFERENCE TO IMPROVE INTELLIGIBILITY
   What networks mediate illusory filling-in of degraded speech?

Acknowledgments

Kevin Hill
Chris Bishop
Antione (Tony) Shahin
Jess Kerlin
Kristina Backer

National Institutes of Health
National Institute on Deafness and Other Communication Disorders