OPTIMIZING THE PERFORMANCE OF MULTITALKER SPEECH DISPLAYS

Douglas S. Brungart

Brian D. Simpson

Air Force Research Laboratory

Introduction

Information Transfer in Audio Displays

Many audio features can convey information:

- **Pitch and Timbre**
- **Rhythm and Temporal Characteristics**
- **Melody**
- **Apparent Location… and many others**

Introduction
Information Transfer in Audio Displays

Many audio features can convey information:

- **Pitch and Timbre**
- **Rhythm and Temporal Characteristics**
- **Melody**
- **Apparent Location… and many others**

However, speech is often maximally efficient

- **Can convey *almost any* information quickly**
- **Requires little or no *additional* training**

Introduction
Information Transfer in Audio Displays

Many audio features can convey information:

- Pitch and Timbre
- Rhythm and Temporal Characteristics
- Melody
- Apparent Location… and many others

However, speech is often maximally efficient

- Can convey *almost any* information quickly
- Requires little or no *additional* training
- Allows *person-to-person* transfer of information
  - Optimized for human *production* and *perception*

Introduction

Info Transfer in Communication Systems

For communications systems:

- Speech is currently the only viable audio signal
  - Other alternatives are theoretically possible, but
  - Advanced *AI* and *Natural Language Processing* required
    -- Speech input → situation analysis → warning tone

Introduction

Info Transfer in Communication Systems

For communications systems:

- Speech is currently the only viable audio signal
  - Other alternatives are theoretically possible, but
  - Advanced *AI* and *Natural Language Processing* required
    -- Speech input → situation analysis → warning tone

- How should audio displays present speech?
  - What factors influence Intelligibility in audio displays?
  - How can multitalker listening ability be enhanced?
    - Important in command and control, ATC, etc.

Methods

The Coordinate Response Measure (CRM)

Data collected with Coordinate Response Measure

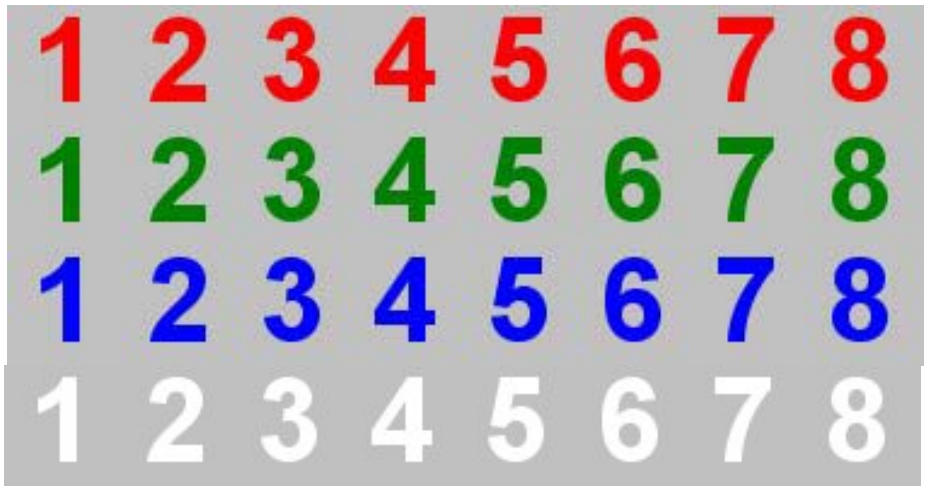
-CRM Originally developed by Moore (1981)

- Format: Ready (Call Sign) go to (Color) (Number) now.
- Target is indicated by call sign Baron
- Maskers indicated by other call signs
- Complete CRM corpus is available (Bolia et. al, 2001)
- 8 Talkers in corpus (4 M, 4 F), 2048 Phrases
    - 8 Talkers x 4 Colors x 8 Numbers x 8 Call Signs
- Embedded call-sign ideal for multitalker studies
    - Similar to many multichannel monitoring tasks

Methods

Response

Listeners responded by selecting the appropriate colored digit with the computer mouse



Methods

Pros and Cons of CRM

Advantages of CRM:

Rapid data collection: training and scoring

Sentences are reusable

Embedded call sign to designate target

- does not require *a priori* designation

Methods

Pros and Cons of CRM

Advantages of CRM:

Rapid data collection: training and scoring

Sentences are reusable

Embedded call sign to designate target

- does not require *a priori* designation

Disadvantages of CRM:

Limited vocabulary

- partially offset by lack of context
- not phonetically balaced

Not “conversationally” realistic

Methods

Pros and Cons of CRM

Advantages of CRM:

- Rapid data collection: training and scoring
- Sentences are reusable
- Embedded call sign to designate target
  - does not require *a priori* designation

Disadvantages of CRM:

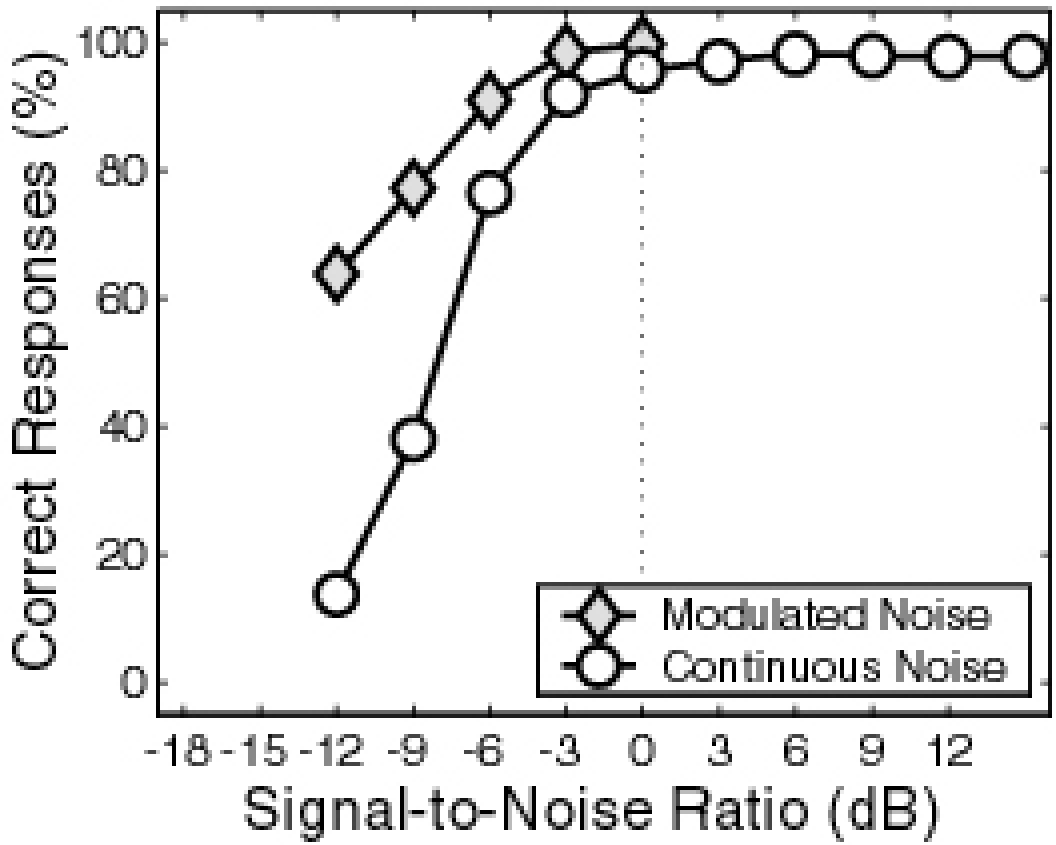
- Limited vocabulary
  - partially offset by lack of context
  - not phonetically balaced
- Not “conversationally” realistic

CRM emphasizes “speech on speech” masking

Factors Influencing Intelligibility

Signal-to-Noise Ratio in Noise

Drop-off in performance from noise is very rapid

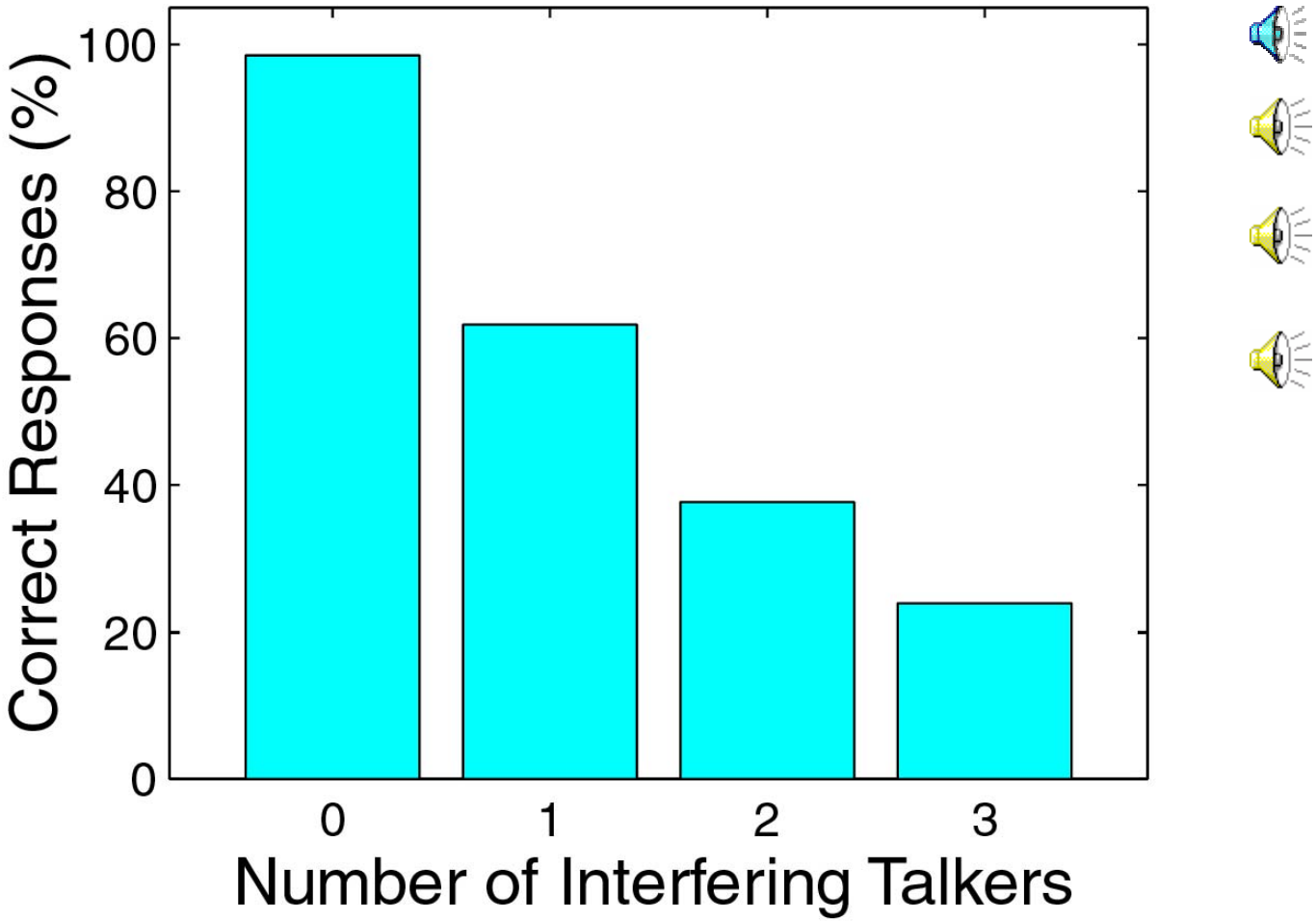


Factors Influencing Intelligibility

Number of Competing Talkers

Same-sex talkers at same level as target talker

Each additional talker decreases performance 40%

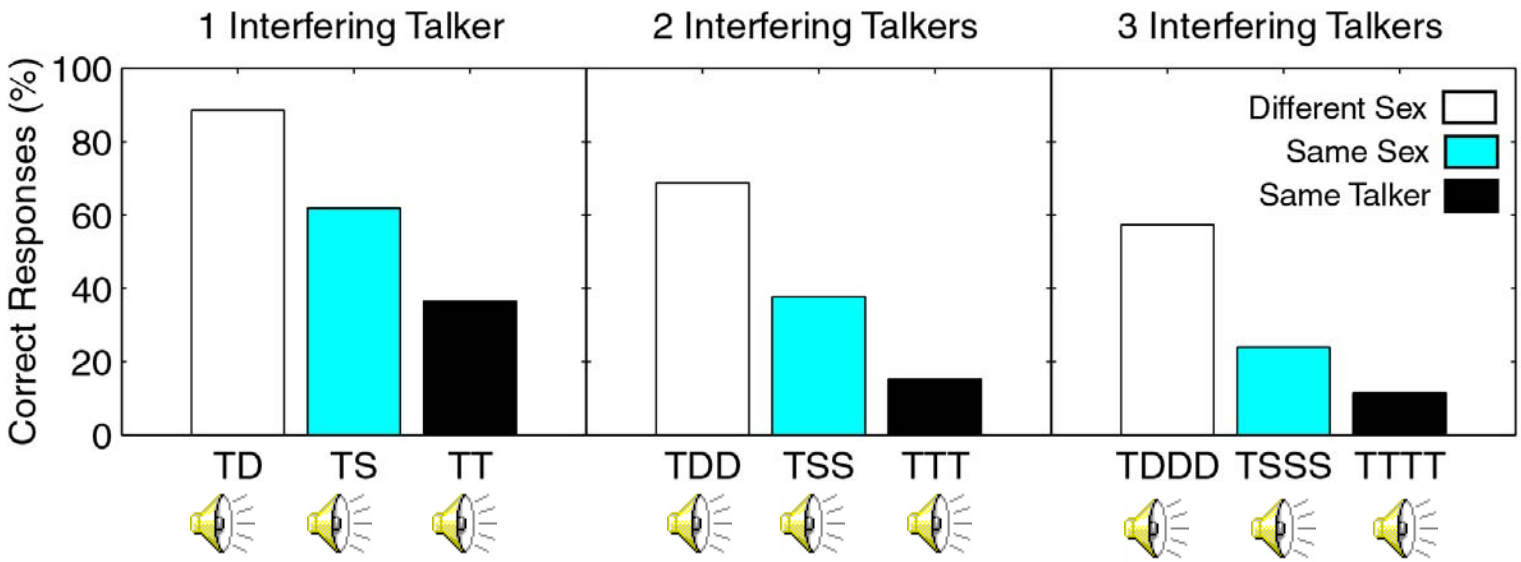


Factors Influencing Intelligibility

Voice Characteristics

Different-sex interfering talkers are better than same-sex

Same-sex interfering talkers are better than same talker

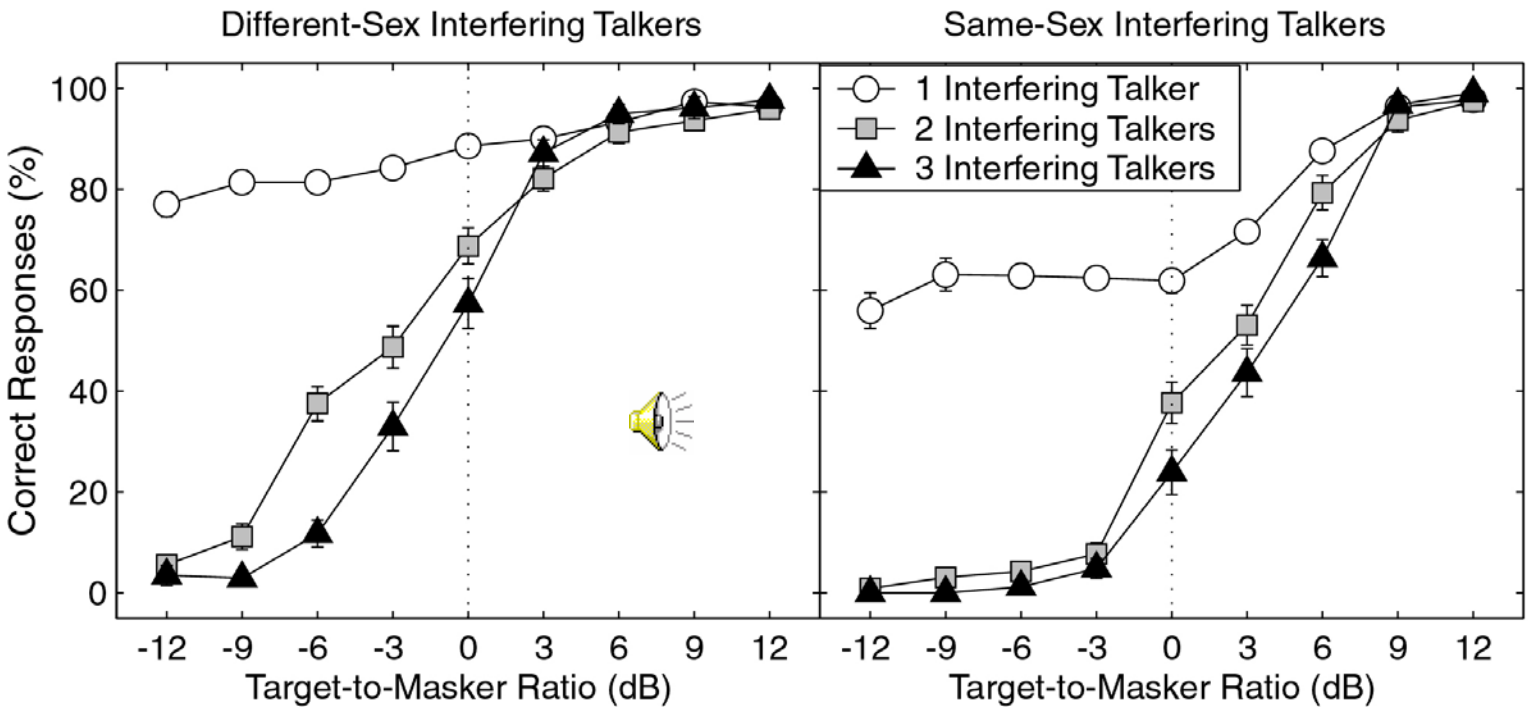


Factors Influencing Intelligibility

Target-to-Masker Ratio

Target-to-Masker Ratio:

Ratio of Target Speech Level (RMS) to Each Masker



Performance good at negative TMRs with one interfering talker

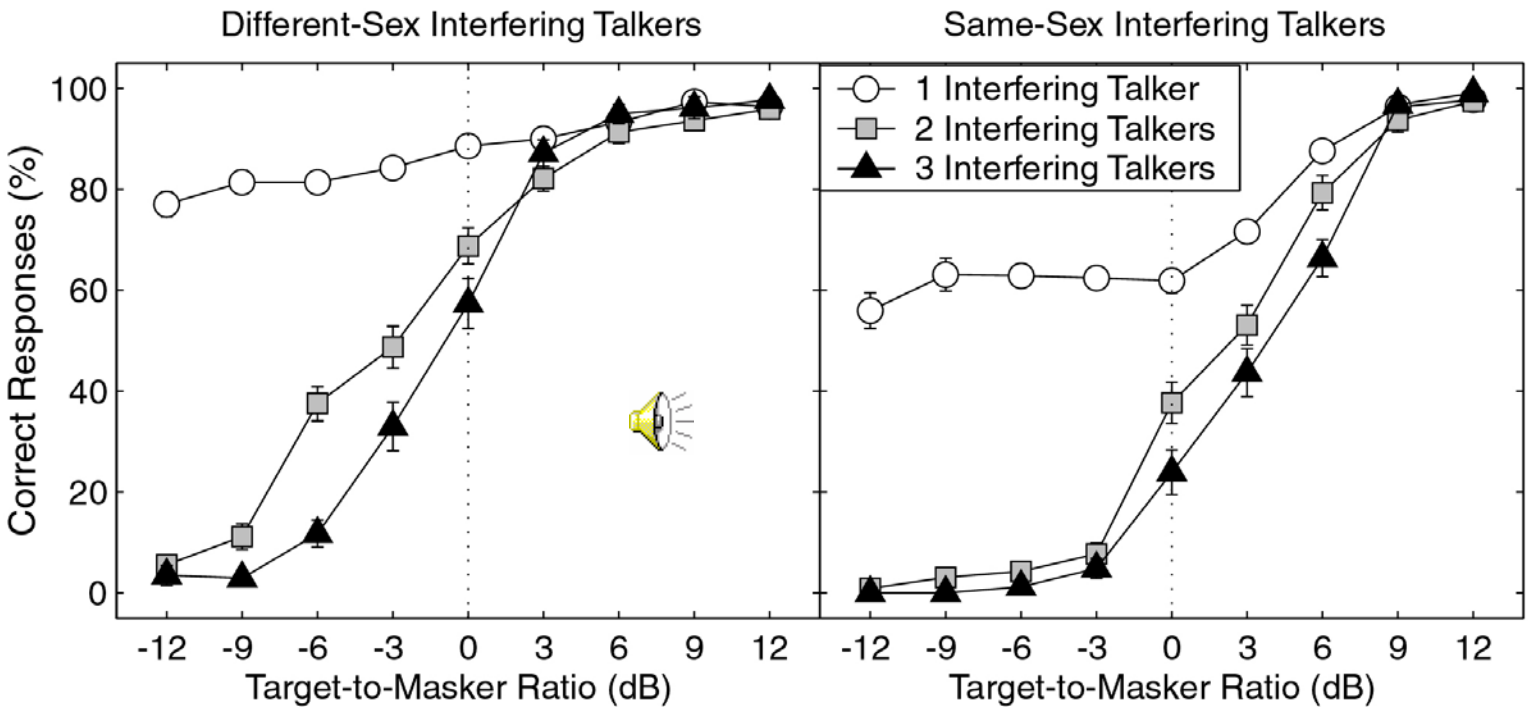
… suggests the use of volume control to segregate talkers

Factors Influencing Intelligibility

Target-to-Masker Ratio

Target-to-Masker Ratio:

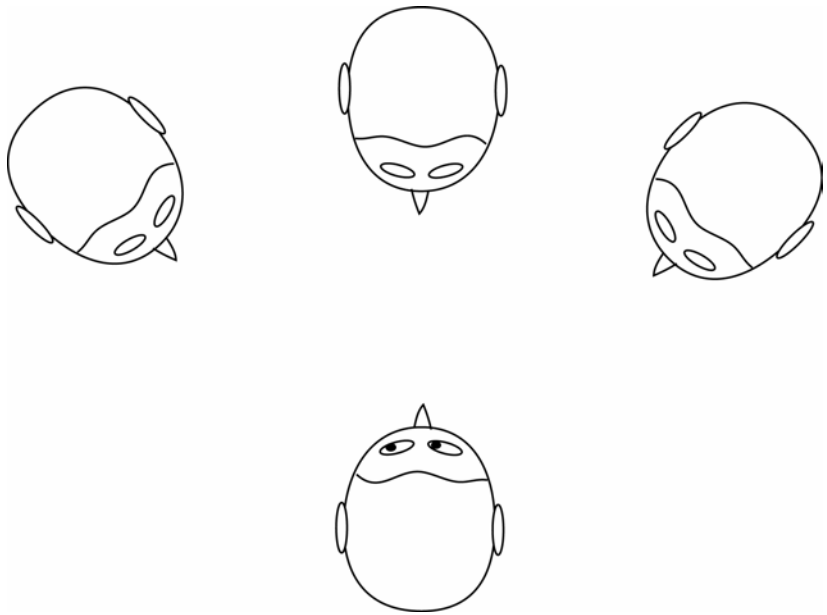
Ratio of Target Speech Level (RMS) to Each Masker



Level cues do not work with more than two simultaneous talkers

Factors Influencing Intelligibility
Spatialization

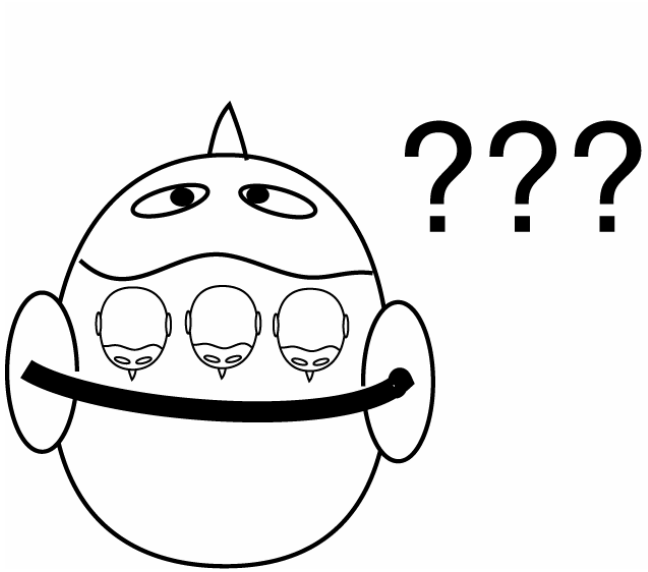
In the real world, competing talkers are spatially separated:



This makes it easy to selectively listen to one talker, and keep track of who said what

Factors Influencing Intelligibility
Spatialization

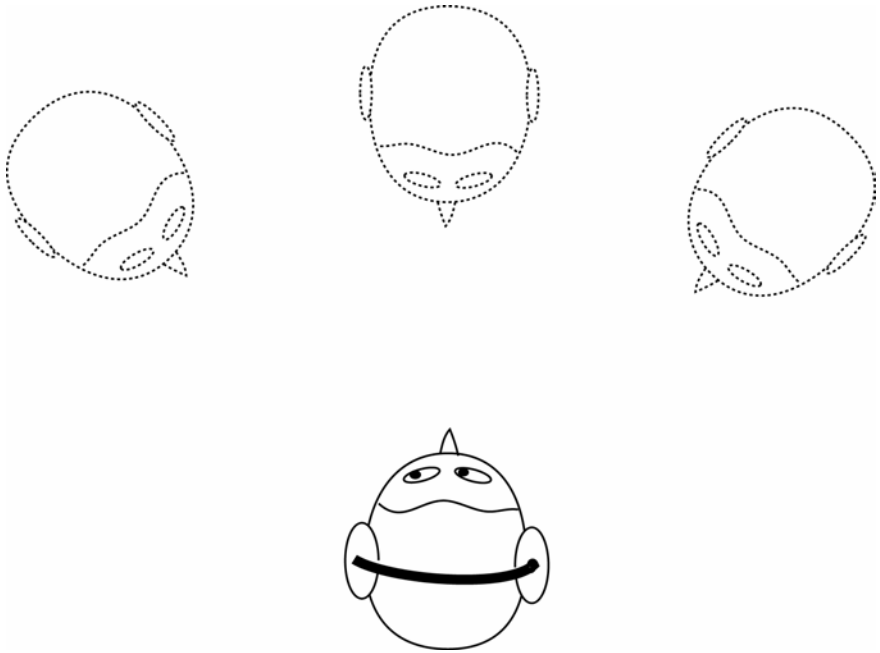
Most current intercom systems are monaural…



This causes all talkers to be heard "inside the head," making it difficult to tell what was said and who said it

Factors Influencing Intelligibility
Spatialization

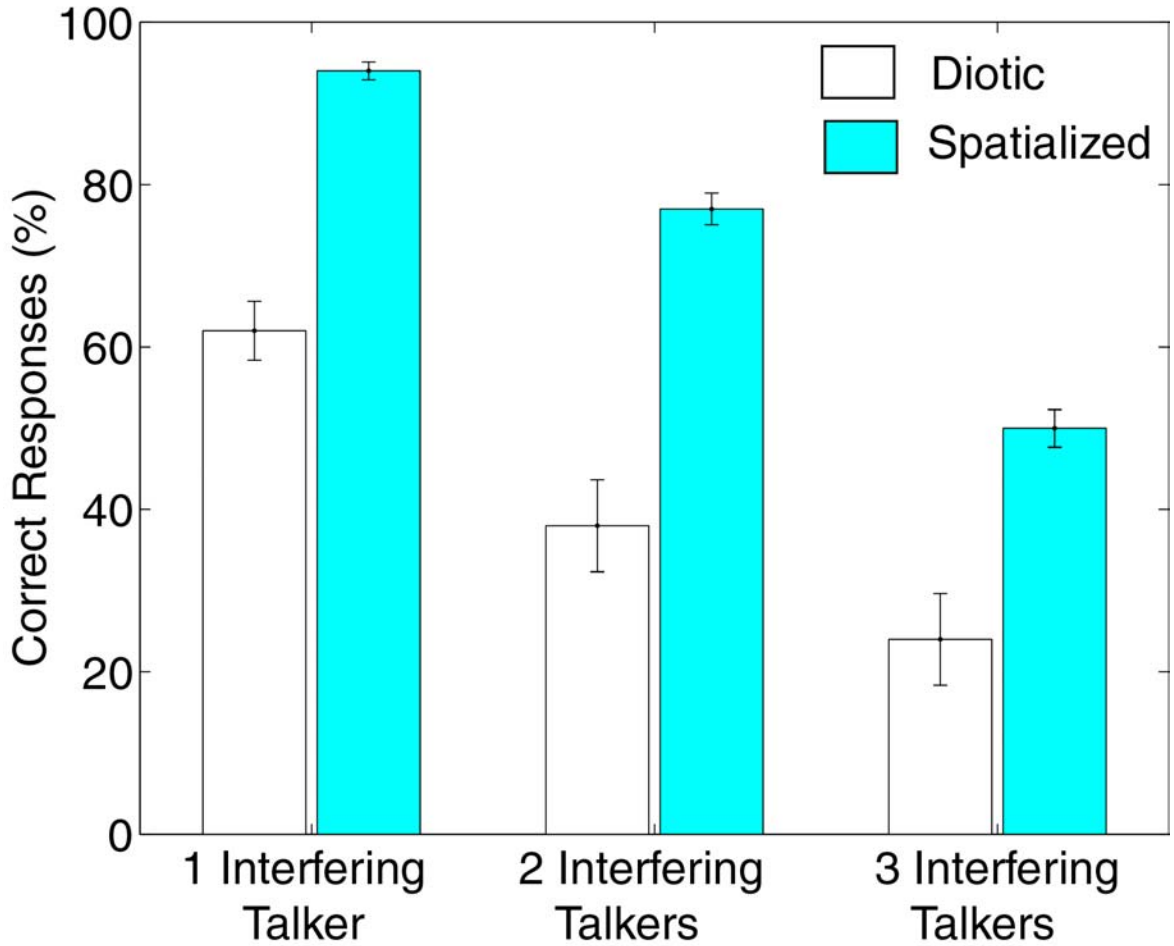
3D Audio uses stereo headphones to simulate spatially separated talkers



Speech is heard in different locations, and it is again easier to selectively attend to one talker and keep track of who is talking

Factors Influencing Intelligibility
Spatialization

Spatial separation improves performance

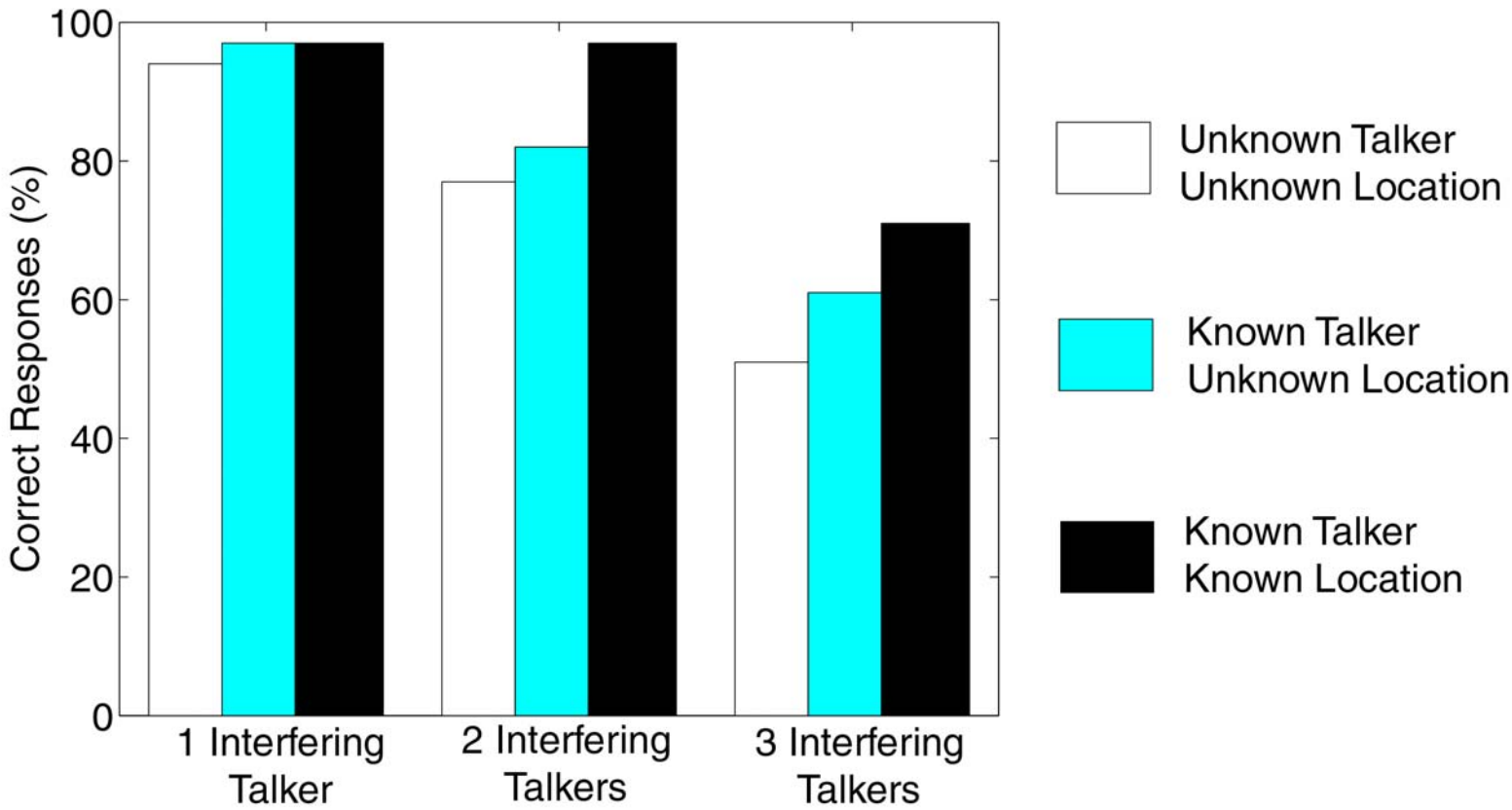


Diotic vs. 45˚ Separation, same-sex talkers

Factors Influencing Intelligibility

A Priori Information

Performance improves when the listener knows the voice or location of the target

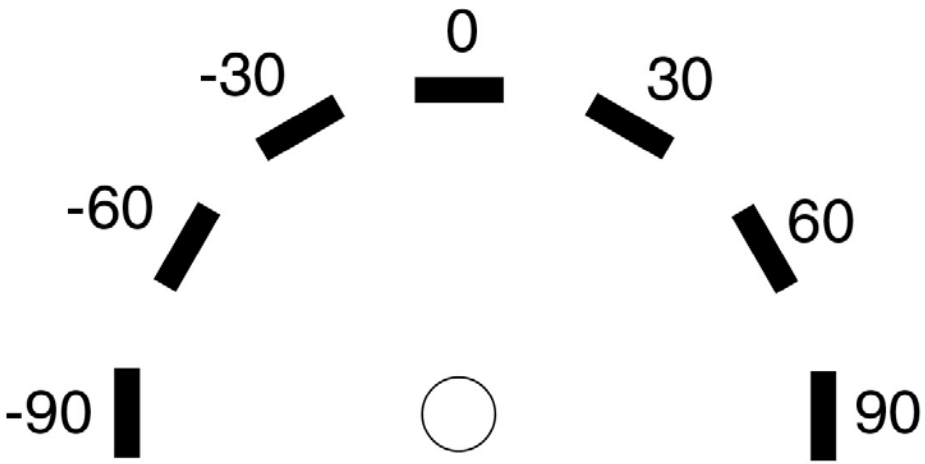


Optimizing 7-Talker Configuration

- Spatial Separation is known to improve intelligibility
- Little is known about “optimal” spatial configuration
- Experiment to find optimal 7-talker placement

“Standard” Configuration

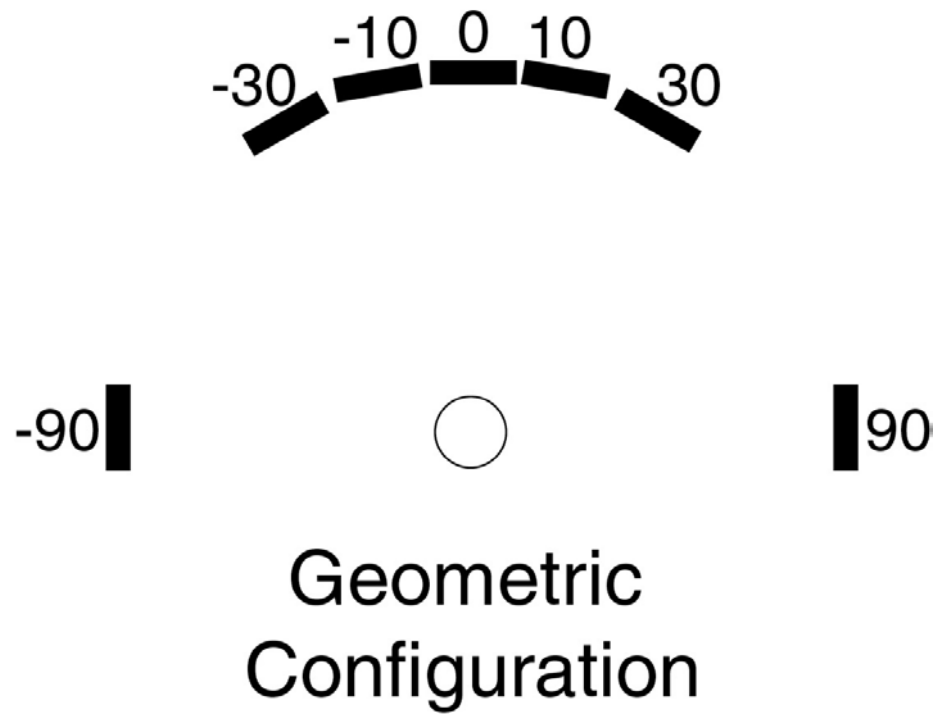
- Talkers equally spaced at 30 degree intervals
- Used in almost all previous multitalker studies
- Ignores enhanced angular sensitivity in front



Standard
Configuration

“Geometric” Configuration

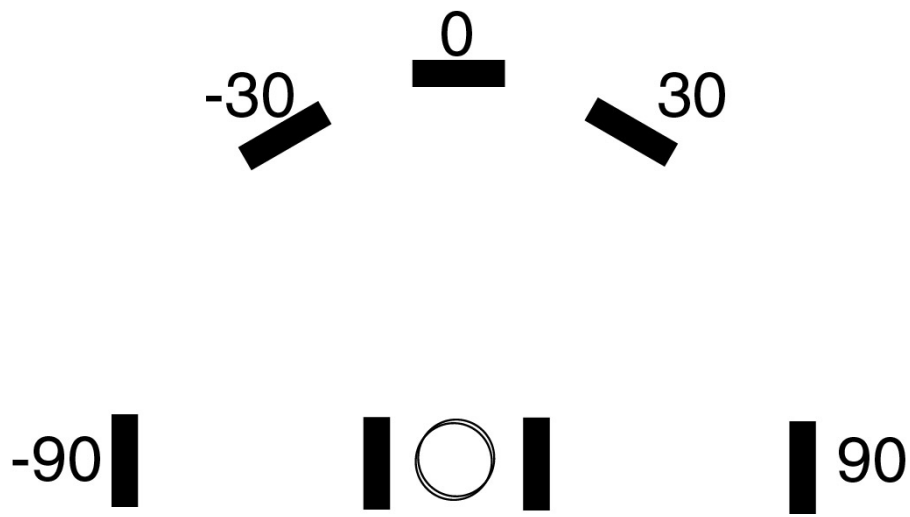
- **Increasing spatial separation between talkers**
- **Takes advantage of enhanced resolution in front**



Geometric Configuration

“Near-Far” Configuration

- **Uses geometric far-field configuration**
- **Combines with two “near-field” or dichotic talkers**



Near-Far
Configuration

Level Normalization

- In “real-world” environments, levels of talkers are determined by production level and relative distances from talkers to listener
  - *Center-of-Head*
    - All talkers equally intense in the free field at a position at the center of the head (with the head removed)
      - No effect when talkers were all equidistant from listener
      - Removed the ~18 dB increase in overall level for sources at 12cm re: far-field level

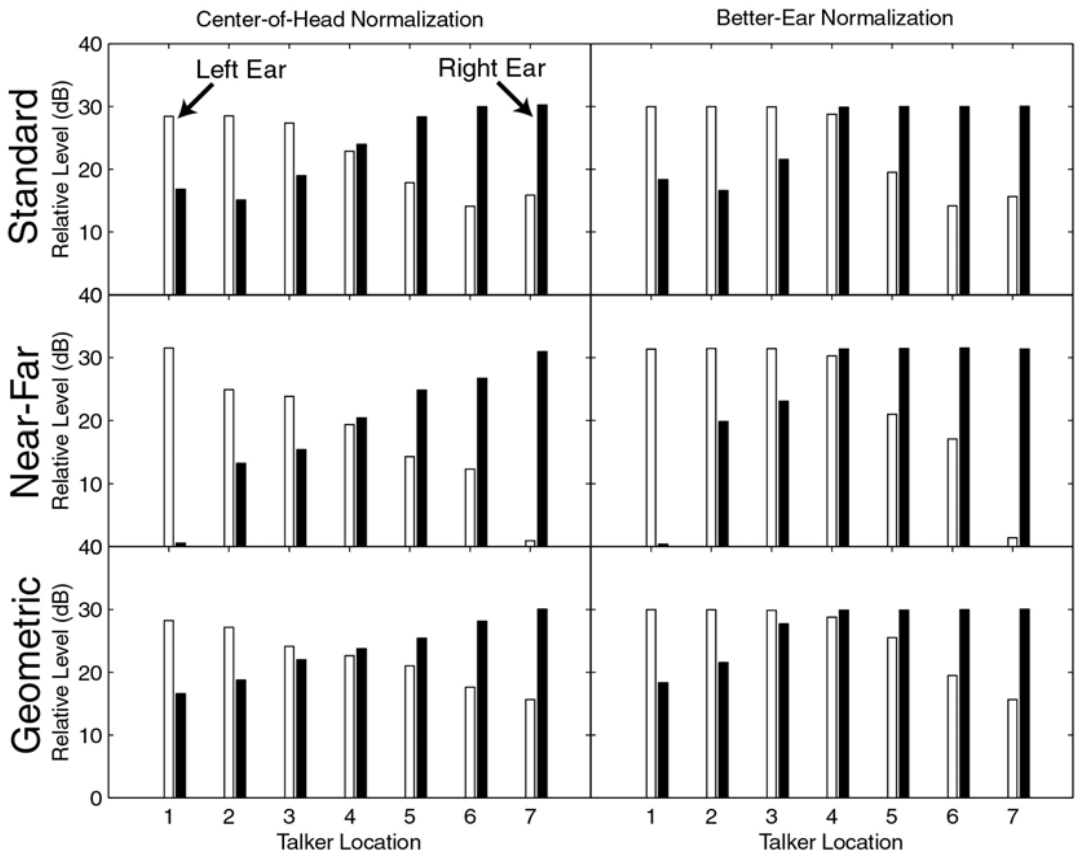
Level Normalization

– *Better-Ear*

- Levels of talkers adjusted such that they are all the same level at the more intense ear
  - –i.e., talkers in right hemifield are all the same level in the right ear; talkers in left hemifield are all the same level in the left ear
  - –Accomplished by adjusting the levels of the speech signals after they are convolved with the appropriate HRTFs
  - –Ensures that all talker locations have approximately the same effective SNR at the better ear – no location is favored

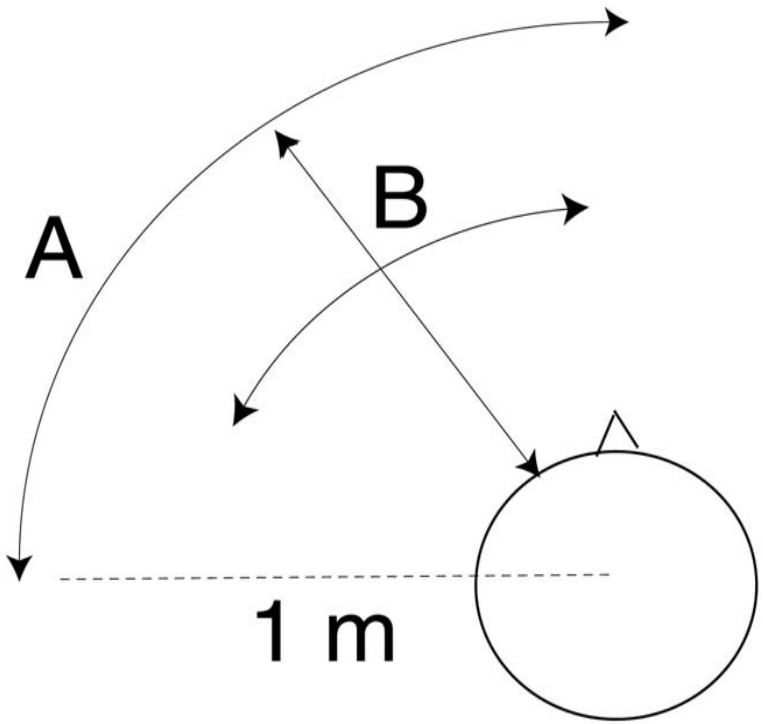
Relative Levels of Talkers

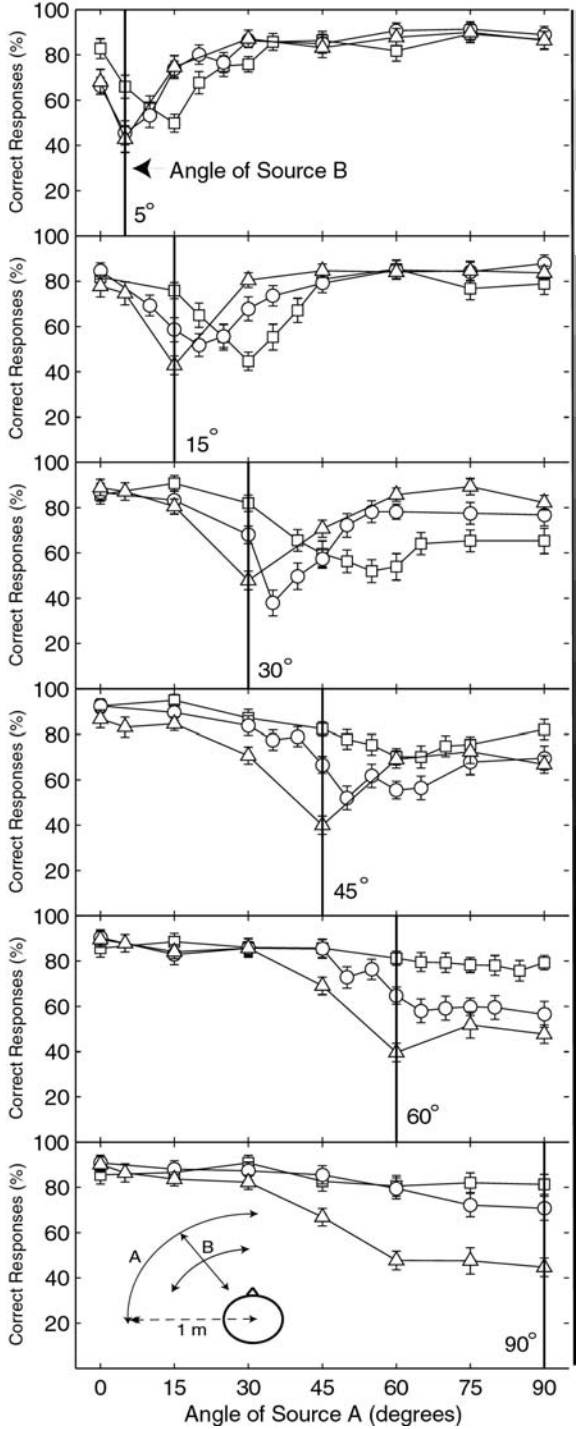
- **Levels measured from the RMS power of speech-shaped noise that was passed through the HRTFs for each talker location**
  - **Resulting increase in relative levels of talker at 0° in "better-ear" normalization scheme**



Experiment 1 - Methods

- Listeners
  - 7 normal hearing listeners
- Speech Stimuli
  - Two-talker stimulus – same talker
  - Talker A varied in angle from 5° to 90°
    - Distance of 1m
  - Talker B fixed at one of 6 angles (5,15,30,45,60,90°)
    - Distance of 12cm, 25cm, 1m
  - “Better-ear” normalization





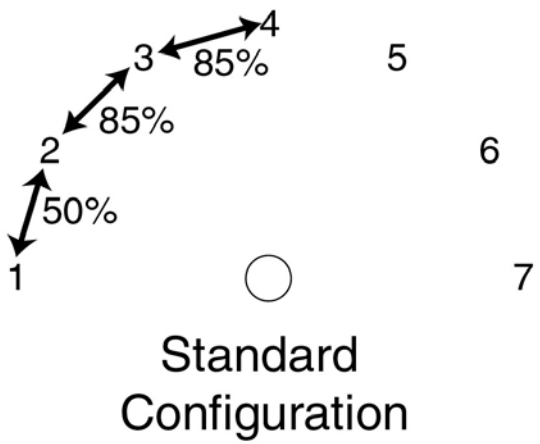
1 m - 1 m
25 cm - 1 m
12 cm - 1 m

Results

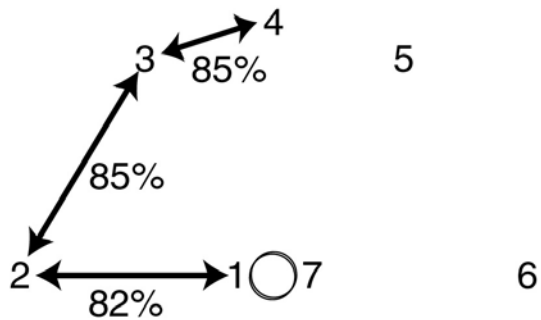
Experiment with 7 simultaneous CRM talkers

- **With 2 randomly located talkers**

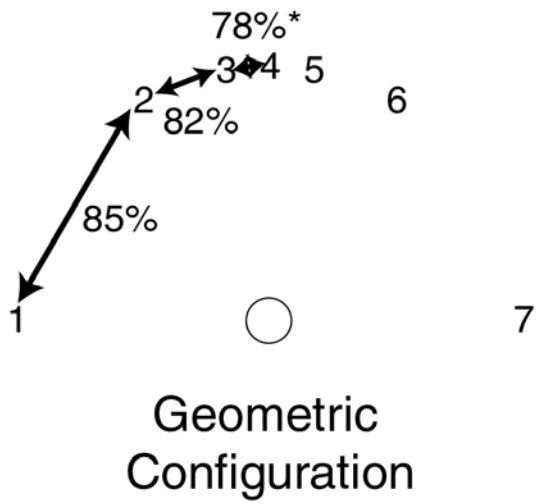
  - Standard configuration causes interference at lateral locations



Standard Configuration



Normalized Near-Far Configuration

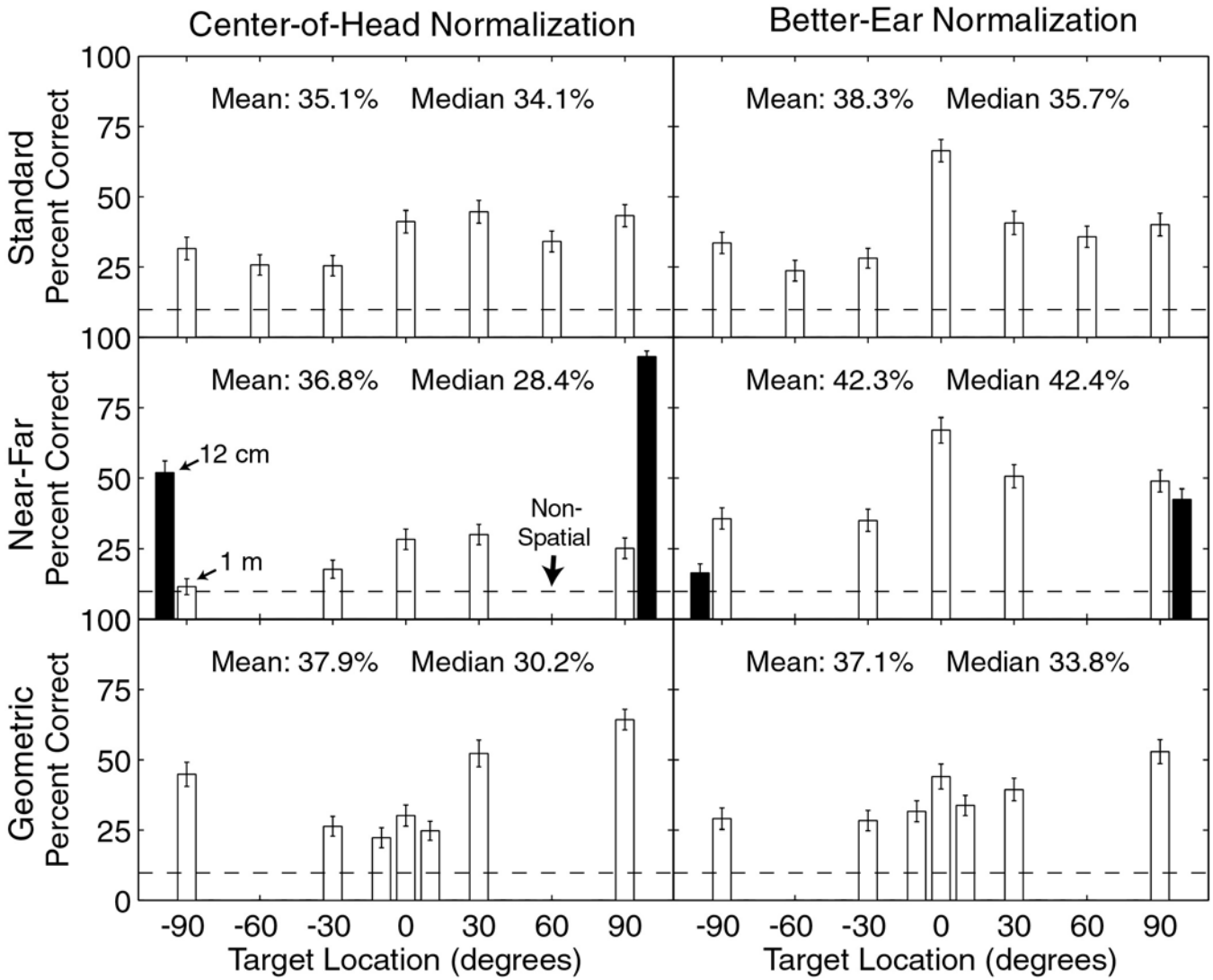


Geometric Configuration

Experiment 2 - Methods

- Seven-talker speech display
- Listeners
  - 10 normal hearing listeners
- Speech Stimuli
  - 7 “male” talkers
    - 4 actual male speech signals
    - 3 female speech signals processed using PSOLA synthesis to scale the F0 by factor of .59 and the vocal tract size by 1.16
  - 3 spatial configurations x 2 normalization schemes
    - 1 non-spatialized condition (all from 0°)
  - Onset of target speech led competing speech onsets by 100ms

Experiment 2 - Results



Results

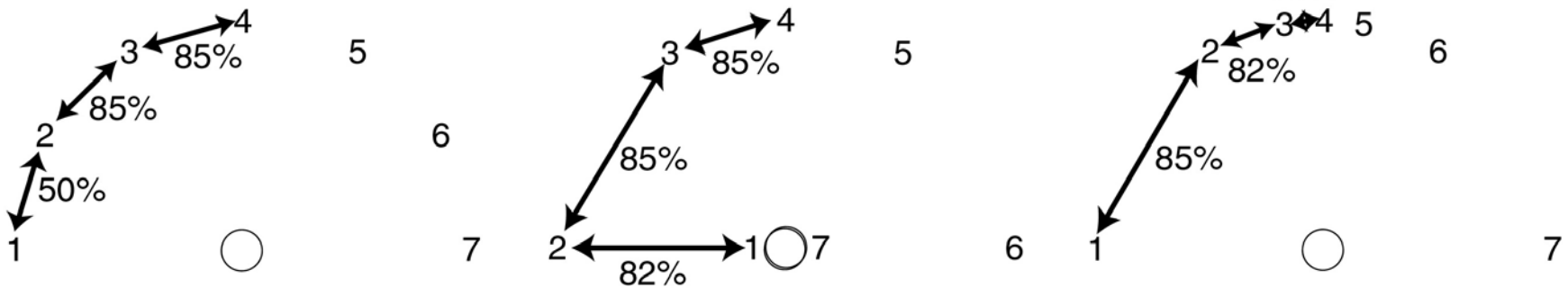
Experiment with 7 simultaneous CRM talkers

- **With 2 randomly located talkers**

  - Standard configuration causes interference at lateral locations

- **With 7 randomly located talkers (100 ms lead on target)**

  - Near-far configuration is 20% better than standard



- Standard Configuration

35%

Normalized Near-Far Configuration

42%

Geometric Configuration

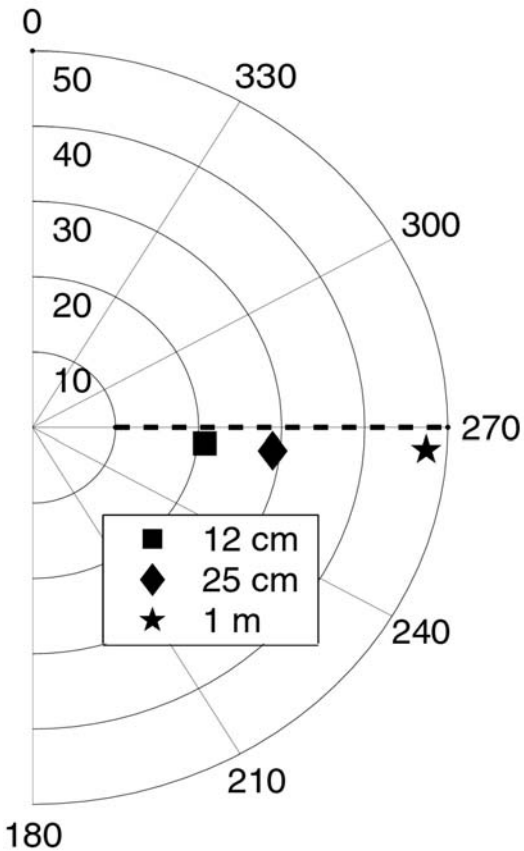
38%

(Non-Spatialized=8 % Correct)

Spatial Location

Do listeners hear near-field sources at different distances?

Yes!



Conclusions

- All 6 spatialized conditions in Experiment 2 led to much better performance than the non-spatialized condition
- The Near-Far configuration with the Better-Ear normalization scheme was the best
- However, range of performance across spatialized conditions was small (35-42%)
  - Perhaps it is not critical
  - But….cost of implementing these changes to a spatialized auditory display is small, and even small improvement might be beneficial

Further Tuning

7-Talker Near-Far Configuration appears optimal…

but is it really "worth it" for real world applications?

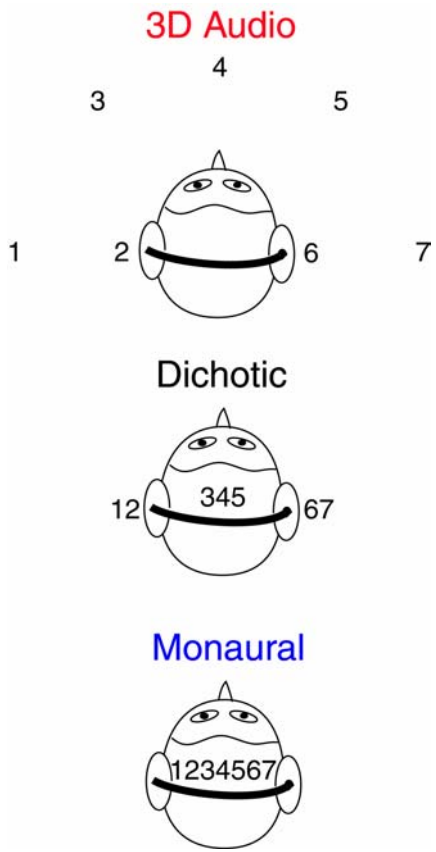
Seven simultaneous talkers rarely occur in real world…

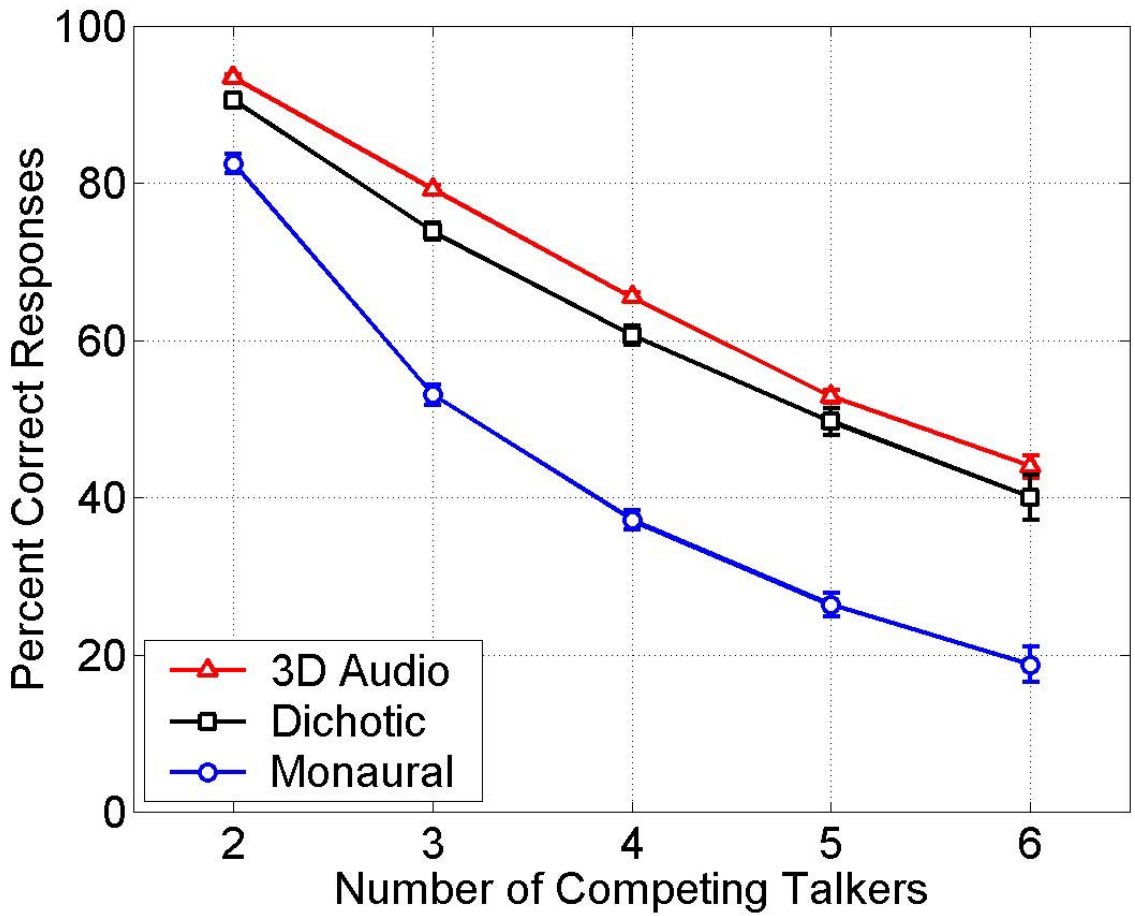
does 3D Audio still provide a benefit with fewer talkers

Is 3D Audio better than simpler Dichotic presentation?

Advantages of 3D Audio

3D Audio is Consistently Better than Mono or Dichotic





Room for Improvement?

- **Free-field errors are exaggerated by virtual displays**
  - **Increased front/back confusions**
  - **Reduced accuracy in elevation**

- **What helps?**
  - **Broadband signals**
  - **Individualized HRTFs**
  - **Headtracking**

- **Are accurate *localization* cues critical for *intelligibility?***

Spatial Audio and Speech Displays

- Bandwidth?
  - Probably Not - Speech is low frequency (But see recent results by Carlile)
- Individualized HRTFs?
  - No - HRTFs are most similar across listeners in the lower frequencies, where most speech information occurs (Drullman and Bronkhorst, 2001)
- Headtracking?
  - Does headtracking improve performance in spatialized speech displays?
    - Particularly when target location is random…

Methods

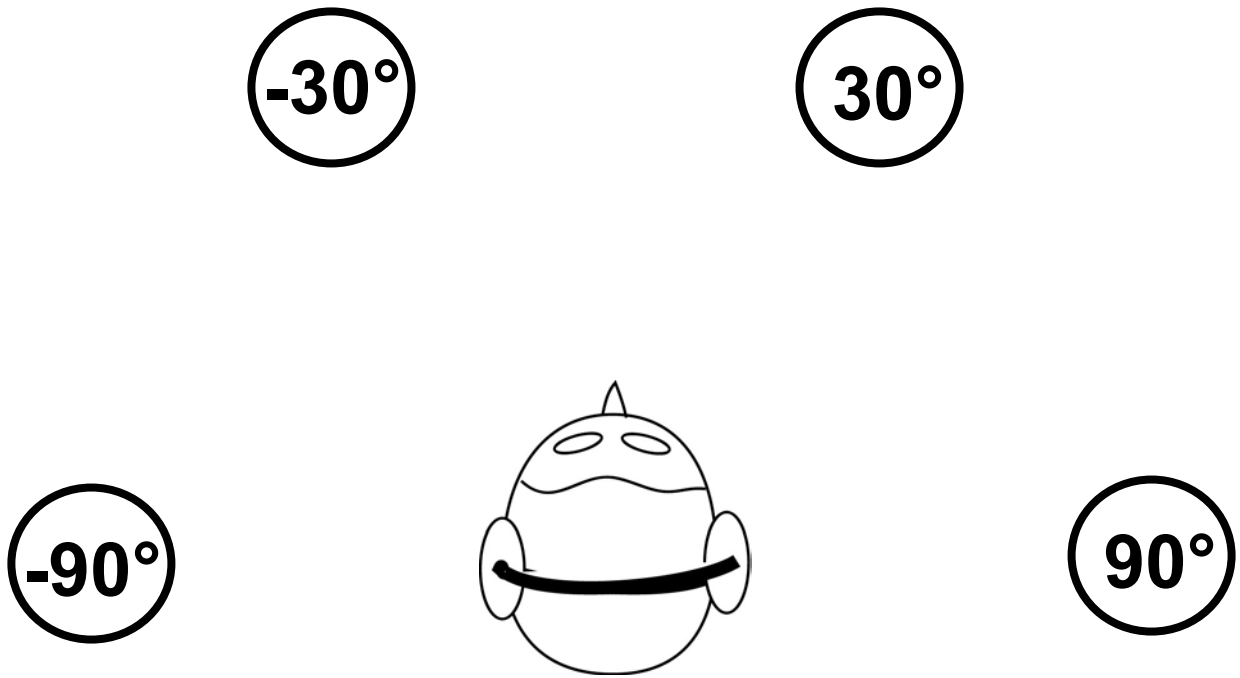
- Spatialization
  - Veridian 3-D VALS, gyroscopic headtracker
    - Kemar HRTFs, 1° spacing on horizontal plane
- Talker configuration
  - 4 talkers randomly assigned to location at beginning of block
    - Remained at location throughout 60-trial block
    - Each talker occurred in each of 4 starting locations in all conditions
- Instructions
  - "might benefit from head movement"

    or
  - "head motion will have no effect"

Methods – Spatial Configurations

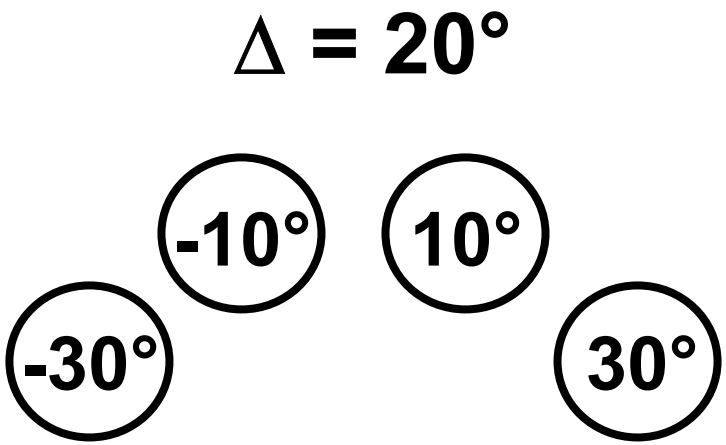
- 4 simultaneous male talkers (wide separation)

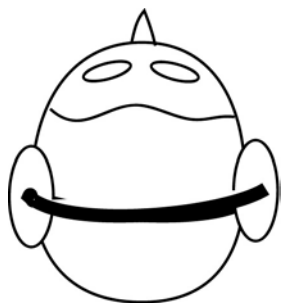
$\Delta = 60°$



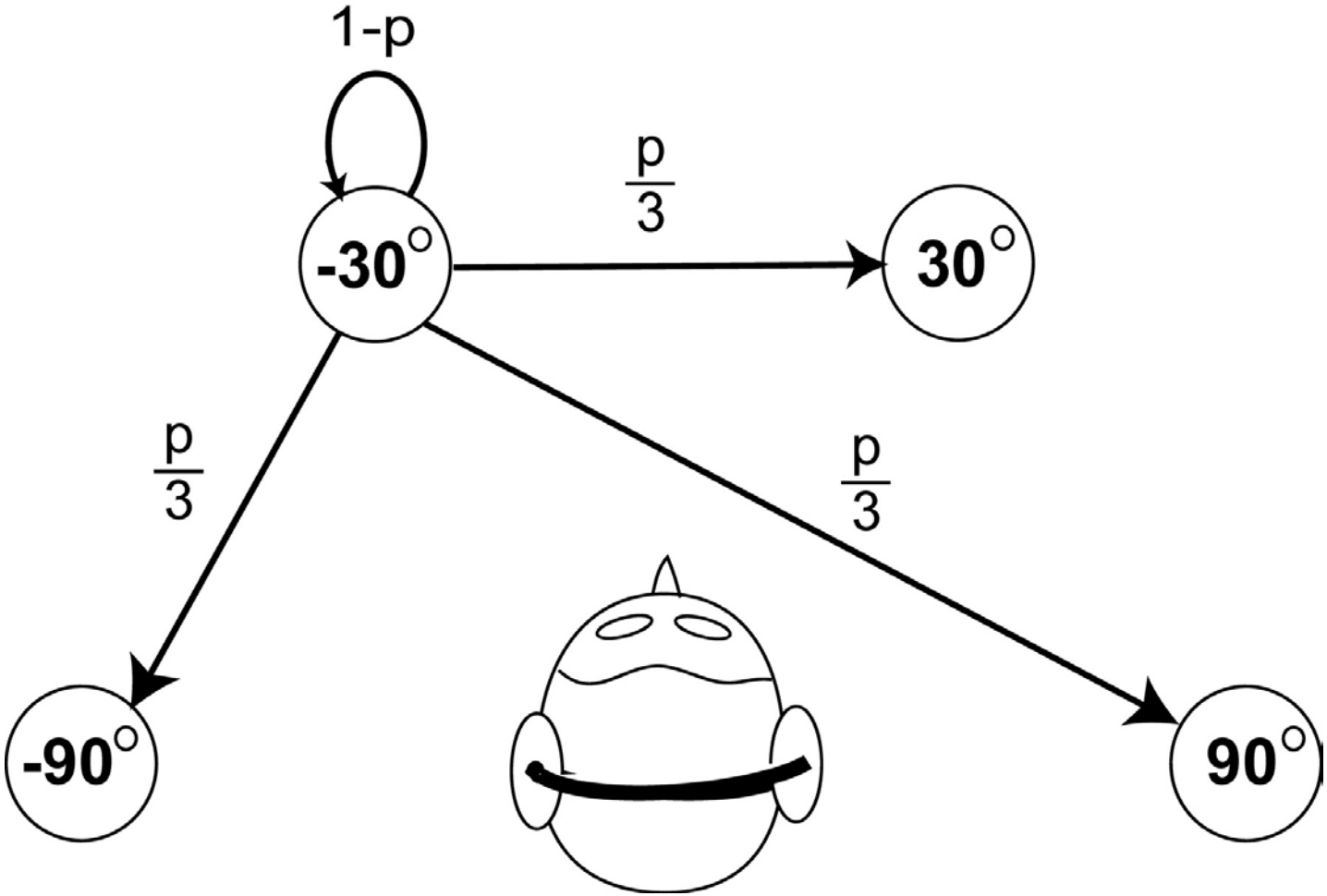
Methods – Spatial Configurations

- 4 simultaneous male talkers (close separation)





Methods – Transition Probability

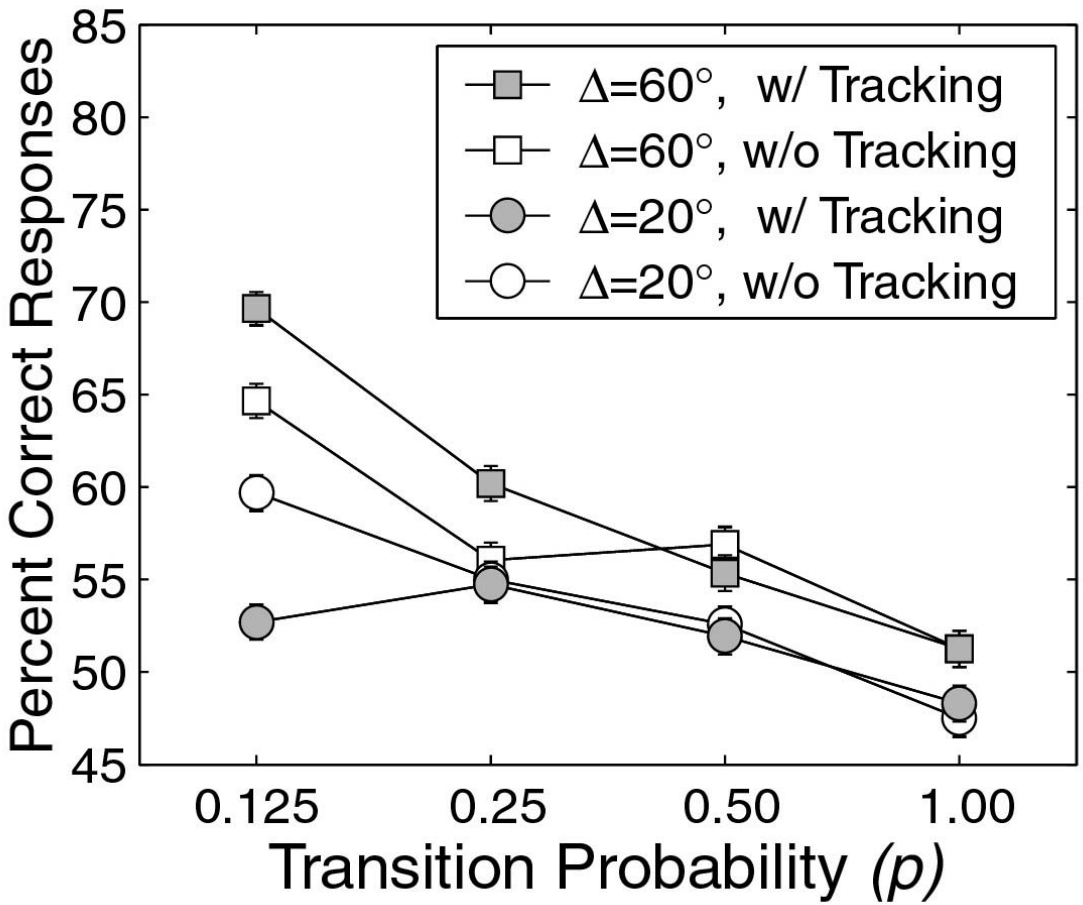


***p* = .125, .25, .50, 1.0**

Experimental Design

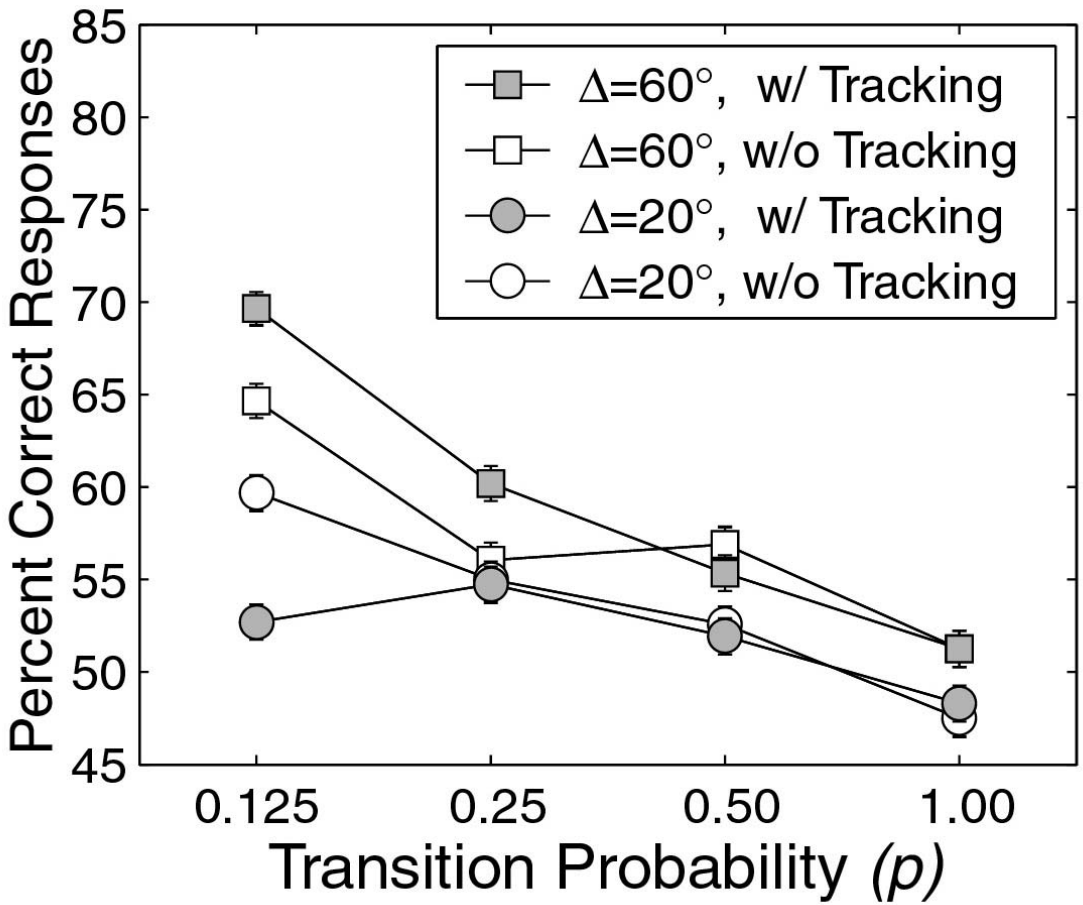
- 2 (spatial configurations) x 2 (headtracking) x 4 (transition probability) design
- Each listener ran 4 blocks of 60 trials in each of 16 conditions
- Total of 42,240 trials overall
  - (3820 per listener)

Results



When transition probability was high (.5 or 1.0), headtracking had almost no effect on performance

Results

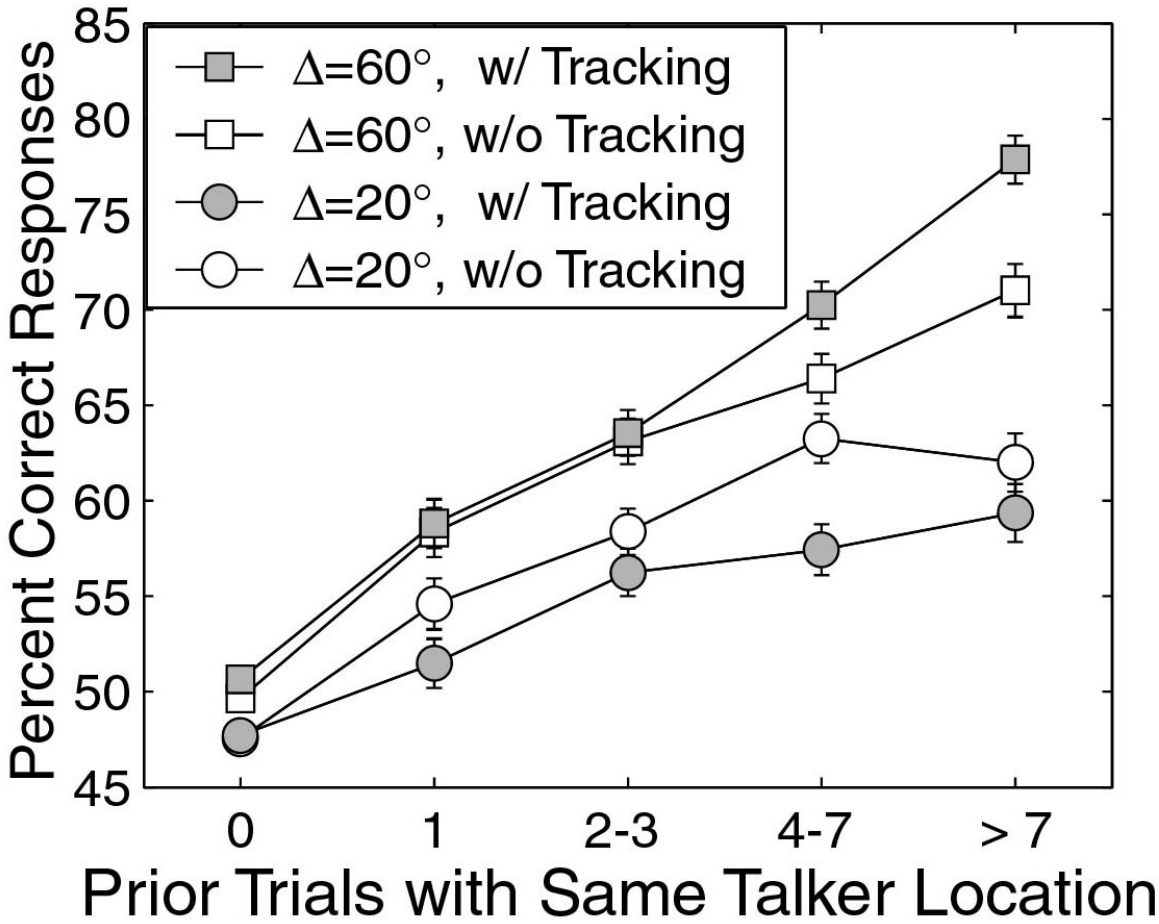


When transition probability was low (.125),

headtracking improved performance with wide separation…

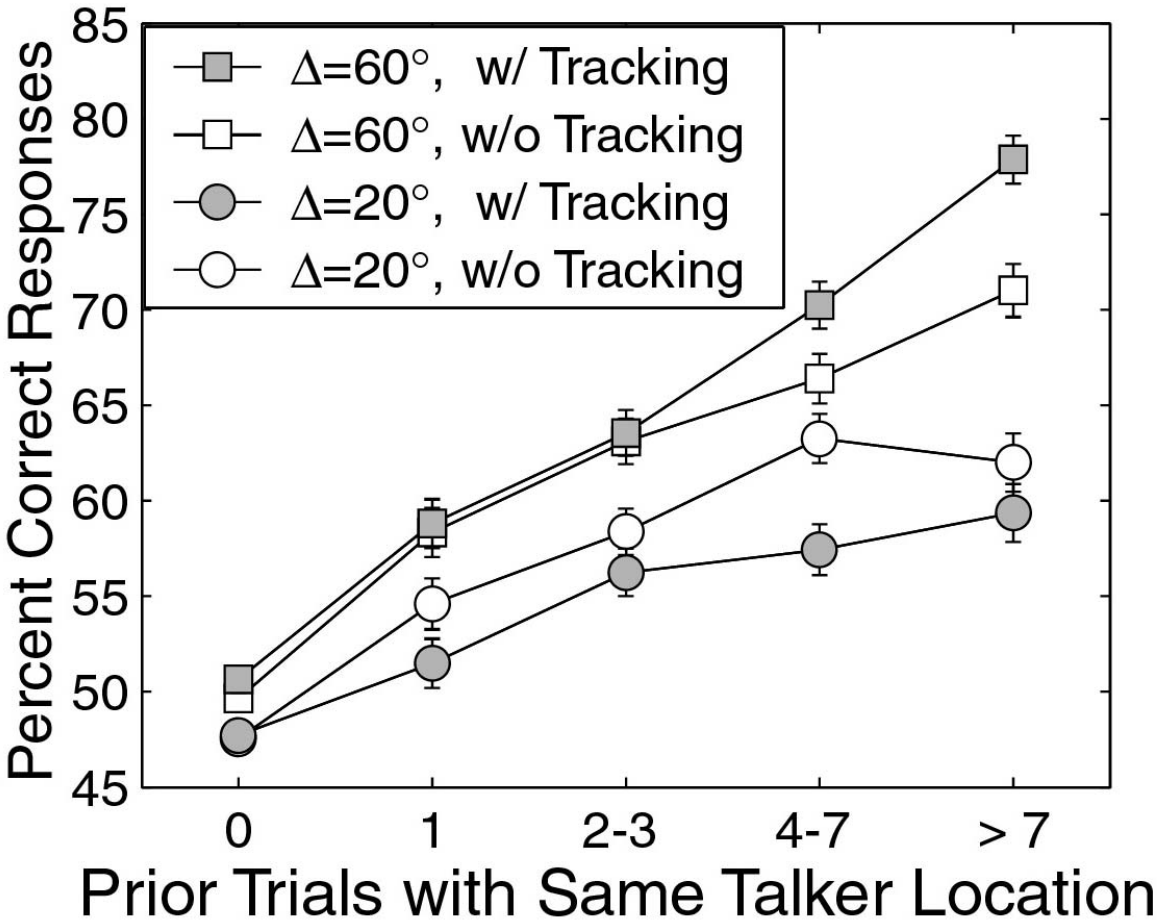
But decreased it with narrow separation

Results



Performance improves when talkers stays in same location

Results



Headtracking only helps when

1) Separation is wide and 2) Talker position is fixed for > 3 trials

Conclusions

- **Results are somewhat surprising**
  - **Headtracking was expected to lead to performance that was *at least as good* as performance without headtracking in all conditions**
  - **But, headtracking actually led to a *reduction* in performance in the $\Delta = 20°$ condition**
- **WHY?**
  - **Localization acuity greatest near midline (Mills, 1958)**
    - **Two talkers at ±10° (starting location for $\Delta = 20°$ condition) might be easier to segregate than talkers that are 20° apart but asymmetric with respect to the midline (e.g., at 10° and 30°)**
    - **Head motion can lead to a situation in which talkers are located off of midline**

Conclusions

- In the $\Delta = 60°$ condition
  - Headtracking did lead to improved performance, but……
    - Only in cases where *p* was low
    - Even here, performance only increased slightly (5-8 %)
  - Improvement with spatialized speech displays over diotic speech displays is much greater

Conclusions

- Is headtracking in multitalker speech displays a good thing?
  - Typically most costly capability
- But….may be integrated in a system requiring headtracked audio for conveying sound localization information or integrated with HMD that already has headtracking capability
  - auditory-cued visual target acquisition
  - navigation/waypoint finding
  - maintaining awareness of, e.g., wingman location
- Should be implemented with a spatial display in which multiple channels are sufficiently separated