Socially guided allocation of attention in the memory encoding of spoken language

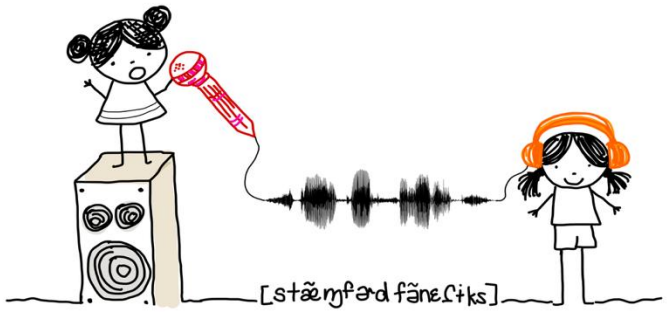
William Clapp

Advisor: Meghan Sumner

Committee: Rob Podesva, Dan Jurafsky, Hyo Gweon

May 13, 2025

Department of Linguistics
Stanford University



National Science Foundation

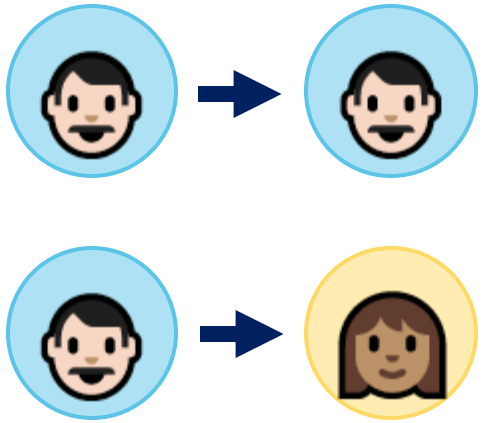
WOD FOUNDATION

Josephine de Karman Fellowship

Background

Talker-specific, ***acoustically-detailed*** memory for individual words.
(Goldinger, 1996; Palmeri et al., 1993)

Better memory for ***same*** talker than ***different*** talker.



Highly replicated over 30 years. (Bradlow et al., 1999; Nygaard & Queen, 2008; Pufahl & Samuel, 2014; Sheffert, 1998)

Memory is central to language understanding.
(Goldinger, 1998; Pierrehumbert, 2016; Wedel, 2012)

Talker-Specificity

Most specificity research: ***isolated words*** with ***full attention.***

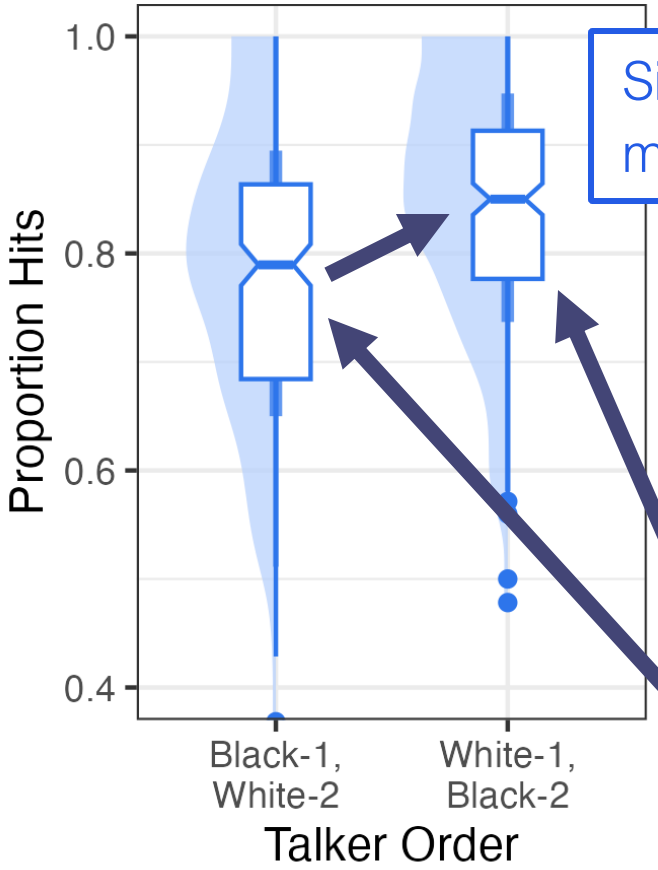
Most speech experiences are ***more complicated!***

- Longer utterances.
- Multi-tasking; planning responses.
- Talker information and messages interact in complex ways.

Fine-grained info is critical at the ***word level.***

How explanatory is this in ***longer utterances*** with ***competing cognitive demands?***

Puzzle: Memory Asymmetries



Clapp, Vaughn, & Sumner, 2023

Simply swapping the order of talkers, memory patterns change.

What cognitive behaviors drive asymmetries?

May result from asymmetric ***resource allocation.***
(Sumner, Kim, King, & McGowan, 2014; Sumner, 2015)

More resources? Relatively **strong encoding.**

Fewer resources? Relatively *weak encoding.*

Resource allocation must be dynamic and context-sensitive.

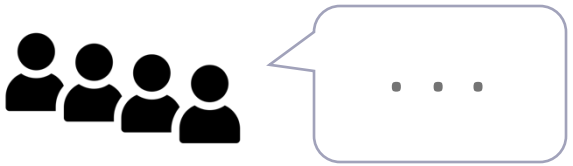
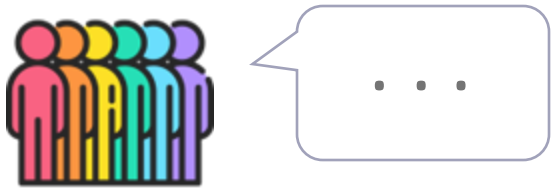
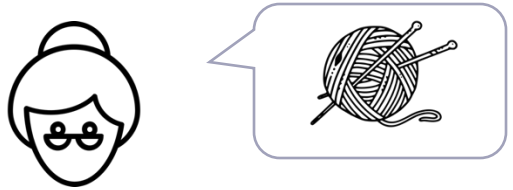
Central Hypothesis

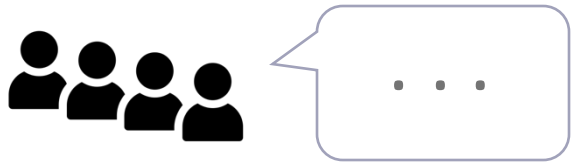
*Listeners **subconsciously** draw on fine-grained **phonetic** information and **social** associations to **dynamically** adapt **low-level** cognitive processes.*

All of this is crucial for language understanding.

Overview

Recognition memory with *Full* or *Divided* Attention;
Sentences repeated by Same or Different talker:

Study	*Question*	*Approach*
Study 1	*How does resource allocation affect talker-specific memory for sentences?*	
Study 2	***A:*** *Do memory patterns differ across individual talkers?* ***B:*** *How can we characterize talker-based memory asymmetries?*	
Study 3	*How do memory patterns differ based on relationship between talker and message?*	



How does resource allocation affect talker-specific memory for spoken sentences?

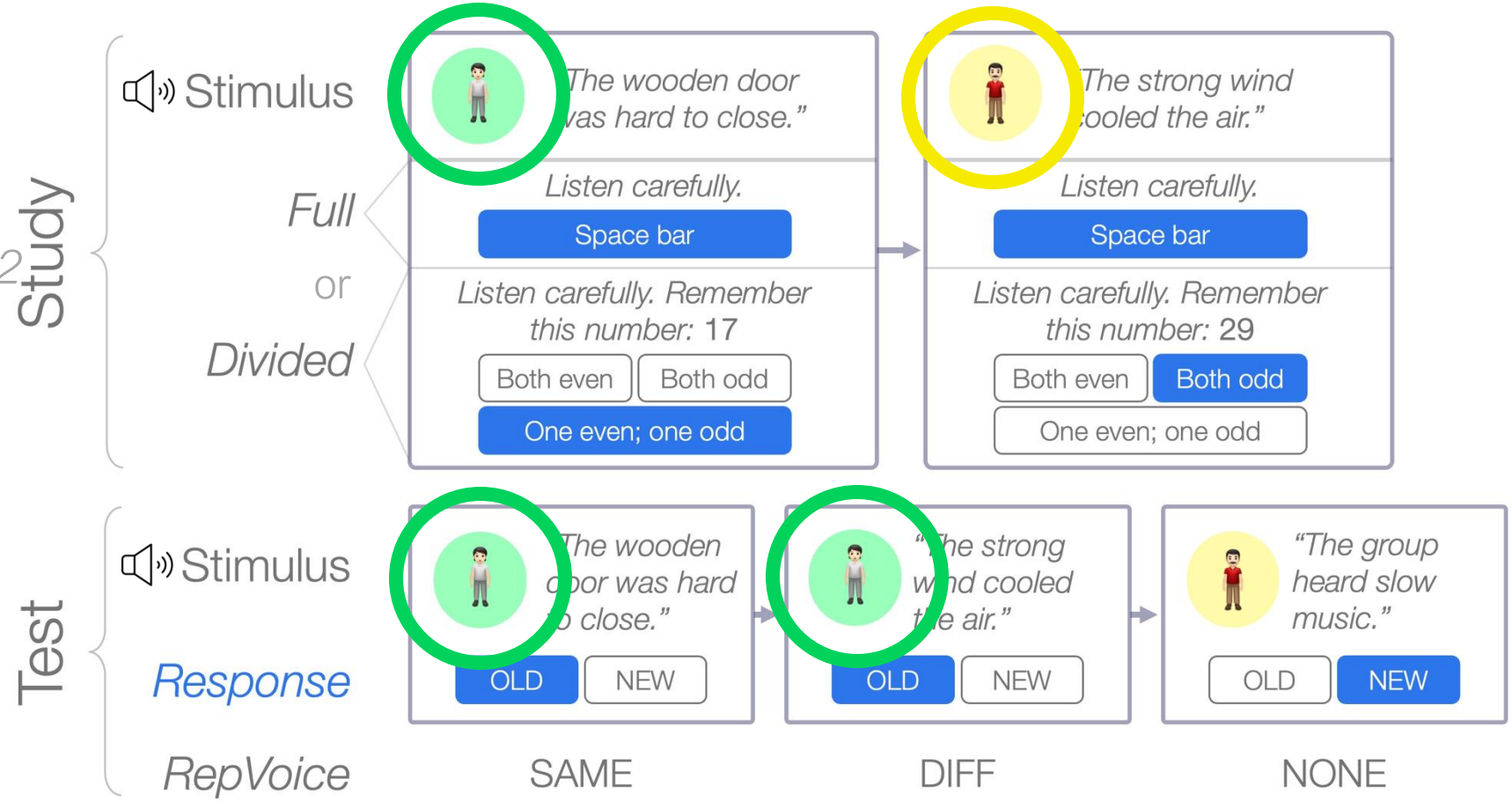
Methods

Participants: From Prolific; Full (*N* = 163), Divided (*N* = 159).

Talkers: *2 female; male GA speakers.*

Stimuli: *Basic English Lexicon sentence list (Rimikis, Smiljanic, & Calandruccio, 2013)*

RepVoice: *SAME vs. DIFF talker.*



Analysis

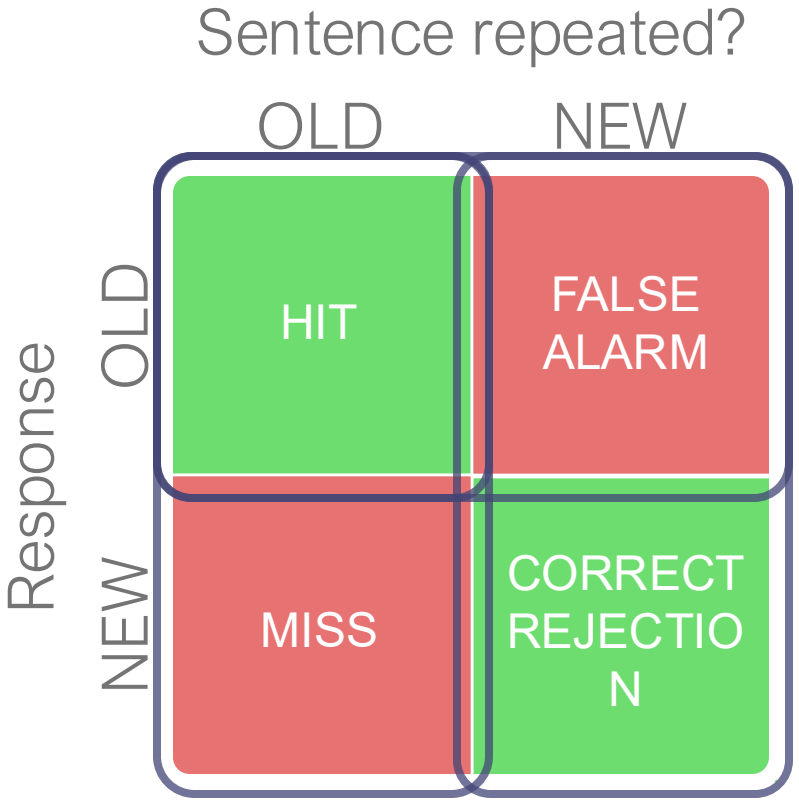
Specificity for Spoken Sentences

Hits: OLD responses on OLD sentences.

False alarms: OLD responses on NEW sentences.

D': z(hits) – z(false alarms)

logRT: Log response time on Hits, measured from stimulus offset.

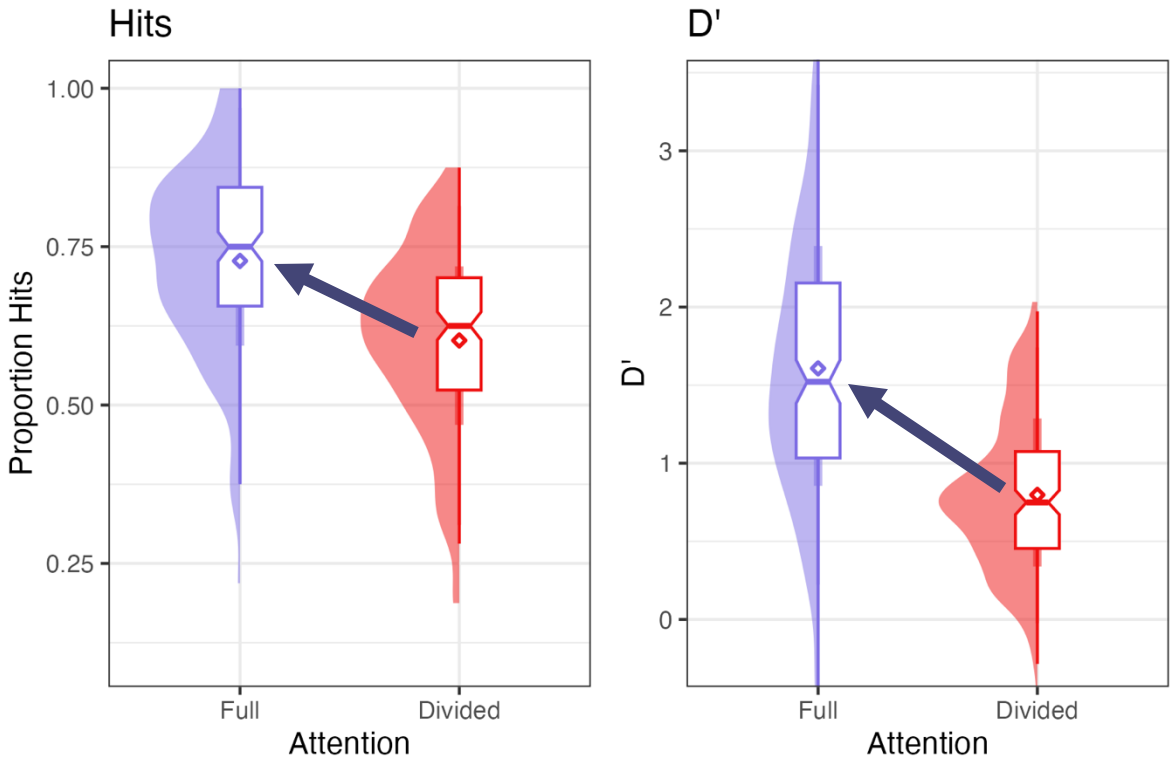


Attention

Specificity for Spoken Sentences

More OLD sentences recognized in Full than Divided.
p < 0.001

Overall, more accurate in Full than Divided.
p < 0.001



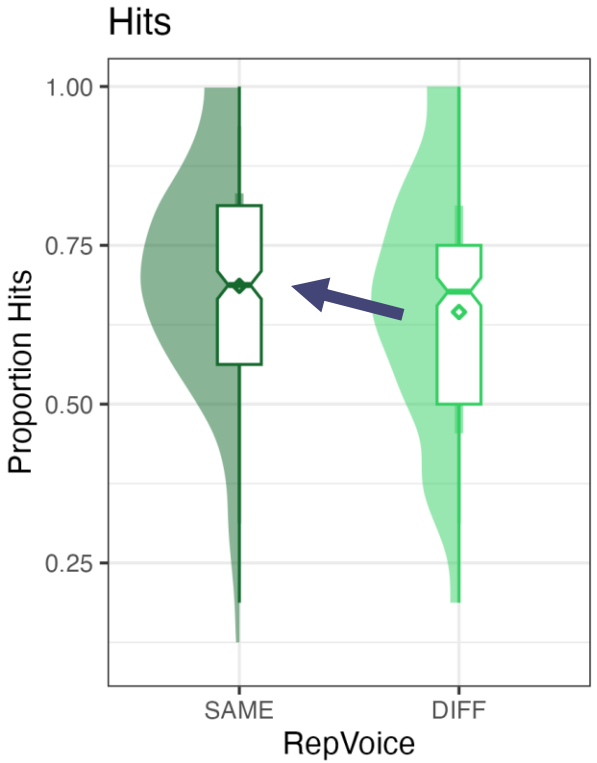
Proof of concept: Attention influences memory in the predicted way.

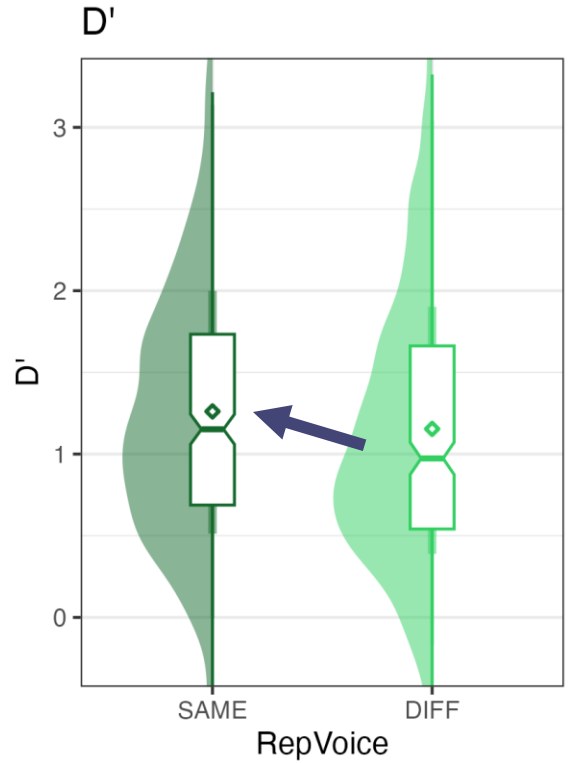
RepVoice

Specificity for Spoken Sentences

More OLD sentences recognized when repeated by than SAME than by a DIFF talker.
p < 0.001

Holds after correcting for False Alarms.
p < 0.001





Specificity replicated for spoken sentences!

Not just a lexical effect!

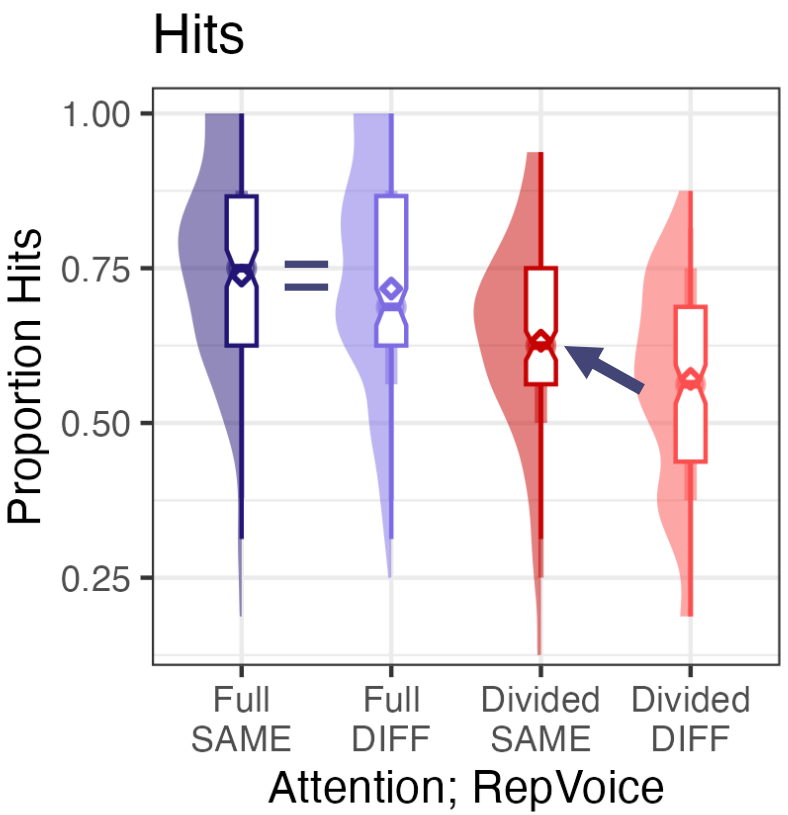
Attention & RepVoice

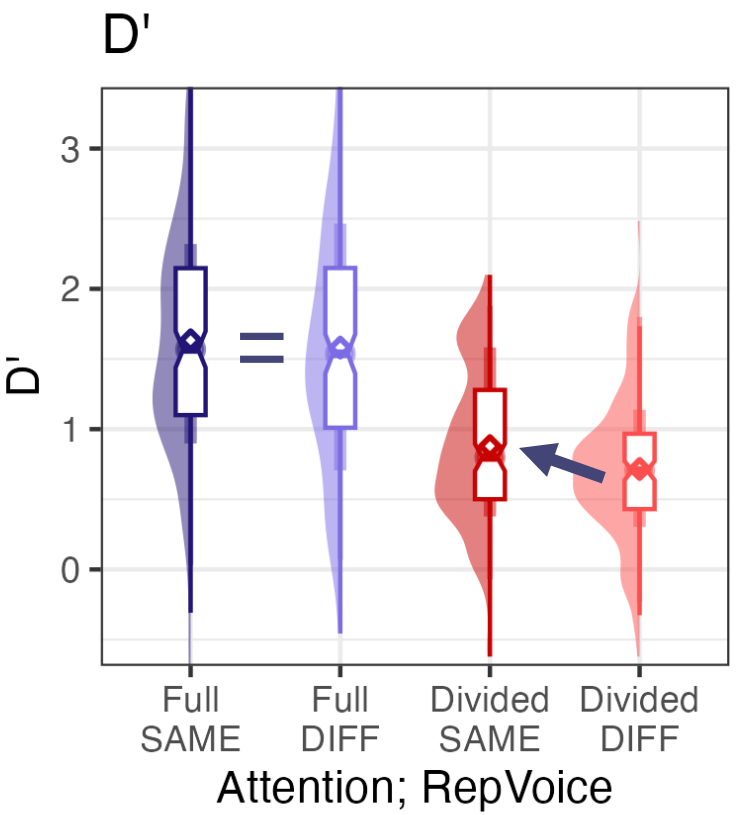
Specificity for Spoken Sentences

Talker-specificity effect driven by Divided Attention.
Divided: $p < 0.001$
Full: $p > 0.1$

This holds for D'.

Talker-specific detail is remembered automatically/implicitly.





Discussion

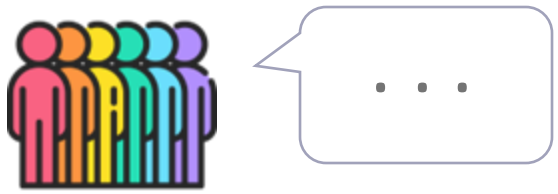
Talker-specificity effects for ***spoken sentences:*** not exclusively a lexical phenomenon.

Effect is ***stronger*** for Divided than Full attention.

Fine-grained acoustic memory is ***fundamental*** to the system!

This info is ***not sacrificed*** when cognitive resources are scarce.

Memory for spoken sentences is acoustically detailed and structured by attention. Are these patterns consistent across talkers?



How do memory patterns differ across individual talkers?

Memory for diverse talkers

Different talkers' speech ***varies*** widely.

Speech always carries ***social meaning.***

Previous work has treated memory encoding as ***indiscriminate*** at the individual-talker level.

If memory allocation is dynamic/social, we would predict asymmetric memory for ***individual talkers!***

Methods

Participants: From Prolific; Full (*N* = 380), Divided (*N* = 380).

Talkers: 12 diverse talkers recruited online, all identified as American.

Procedure, design, stimulus sentences all the same as previous study.

Talker	*Associates*				
T01	Woman	Hispanic	Store	Teaching	Cooking
T02	Man	White	Minnesota	Suburban	Library
T03	Woman	Grandma	White	Store	Knitting
T04	Man	Black	Older	Jazz, music	Store
T05	Woman	Southern	White	Farmer	Barbecue
T06	Man	Southern	Rural	Hardware	Middle-aged
T07	Man	Black	Basketball	Business	Urban
T08	Woman	Black	Southern	Cooking	Church
T09	Man	New York	Pizza, bagels	Italian	Sports
T10	Woman	Latina	Store	Immigrant	Angry
T11	Woman	Southern	White	School	Store
T12	Man	Young	College	Nerdy	Video games

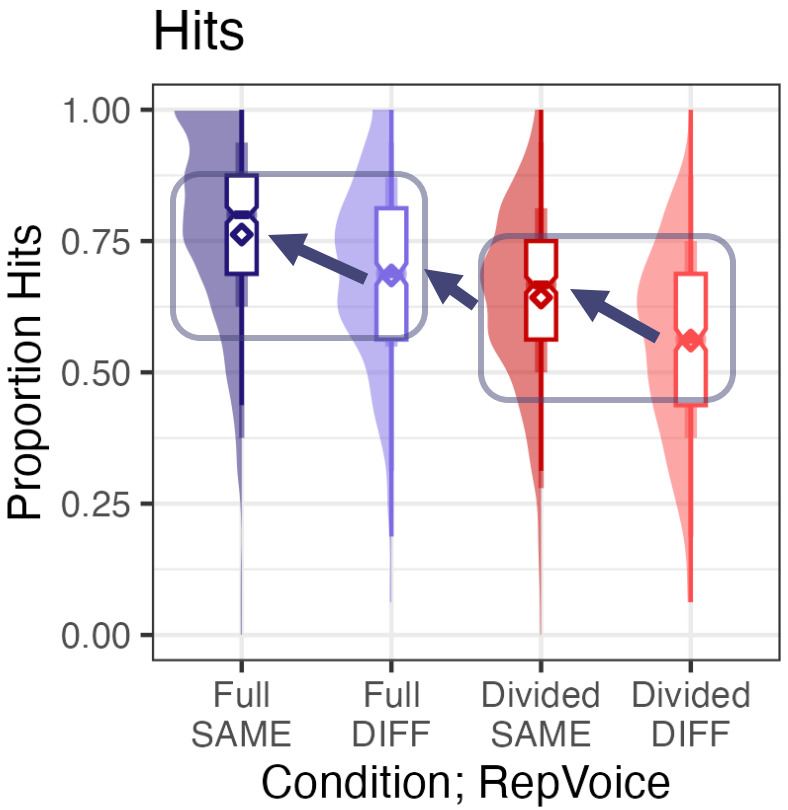
Attention & RepVoice

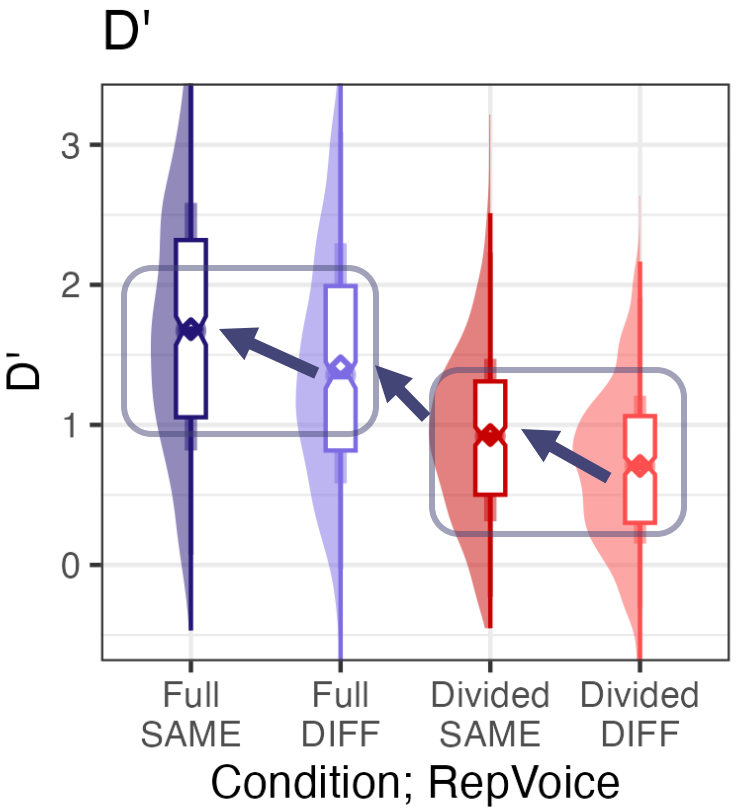
Memory for Diverse Talkers

More accurate in Full than Divided across all measures.
Both p < 0.001

More accurate for SAME than DIFF talker, regardless of attention.
Both p < 0.001

Talker-specificity replicates with diverse talkers.





Memorability

How can we tell whether memorability differs without repeated pairwise comparisons?

Memorability composite score, treated as independent variable:

$$Memorability = \sqrt{\frac{scale(Hits)^2 + scale(FAs)^2 + scale(RT)^2}{3}}$$

Sum-of-squares of each talker's memory performance (Hits, FAs, RTs re-scaled 0-1, worst to best).

Bootstrapped 1,000 times.

Memorability

How can we tell whether memorability differs without repeated pairwise comp

Memorability c t variable:

$$Memo \qquad \frac{}{(RT)^2}$$

Sum-of-squares of each talker's memory performance (Hits, FAs, RTs re-scaled 0-1, worst to best).

Bootstrapped 1,000 times.

N.B. DIFF trials involve two talkers!

DIFF trials contributed to the score of the talker heard in the *Study* block, not the *Test* block.

Memorability

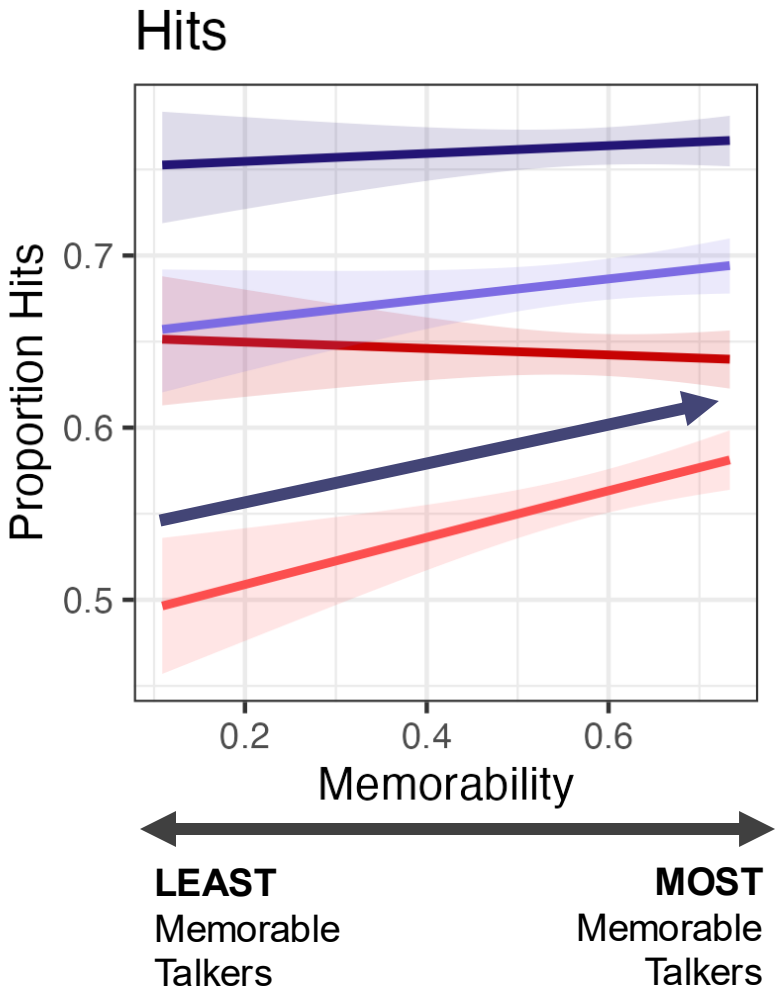
Memorability score was predictive of all dependent variables.

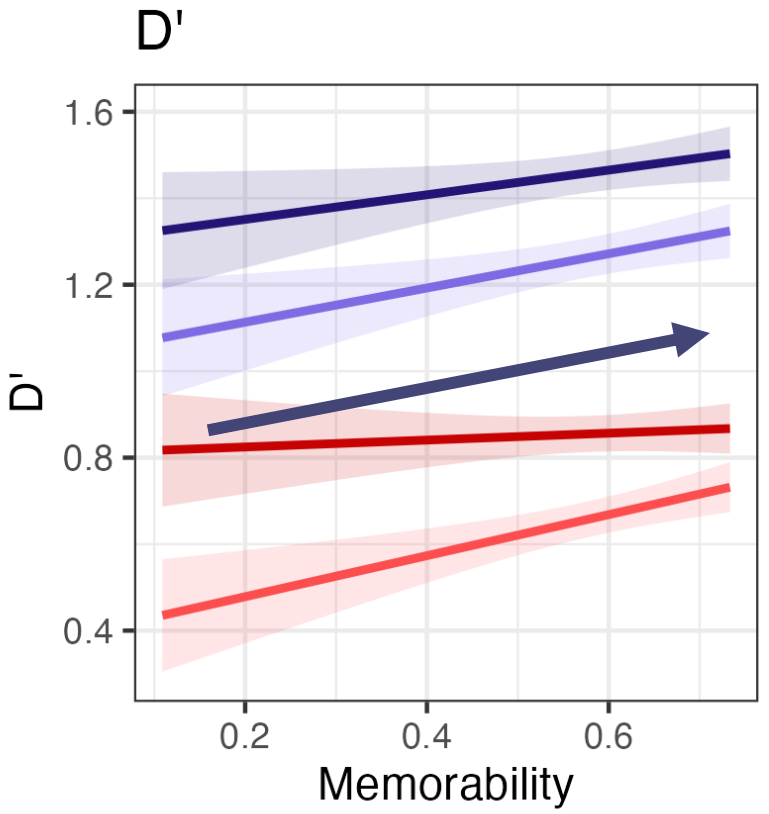
For more memorable talkers:

More Hits $p < 0.01$

Higher D' $p < 0.01$

Talkers were not remembered alike.





Full SAME | Div. SAME

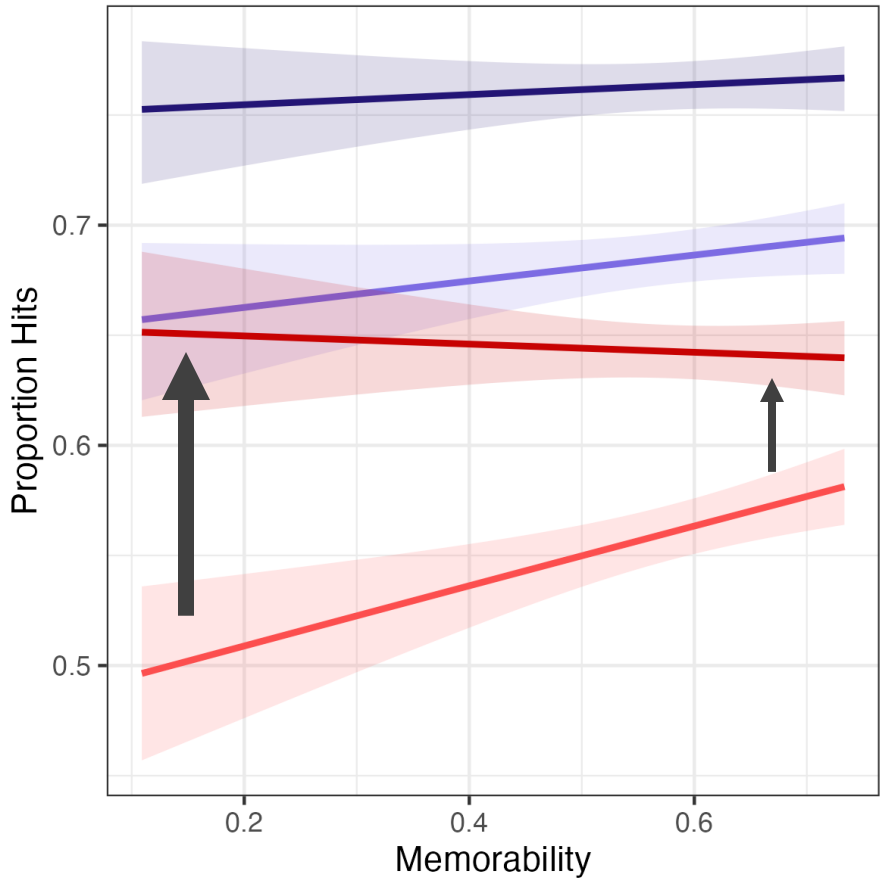
Full DIFF | Div. DIFF

Memorability: Hits

Talker-specificity effect size larger at lower than higher memorability.
p < 0.05

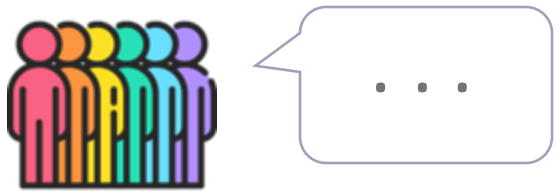
Performance stable across SAME repetitions.

Memory asymmetries driven by DIFF repetitions.



Full SAME | *Div. SAME*
Full DIFF | *Div. DIFF*

No interaction with attention. Effect stable across memorability.



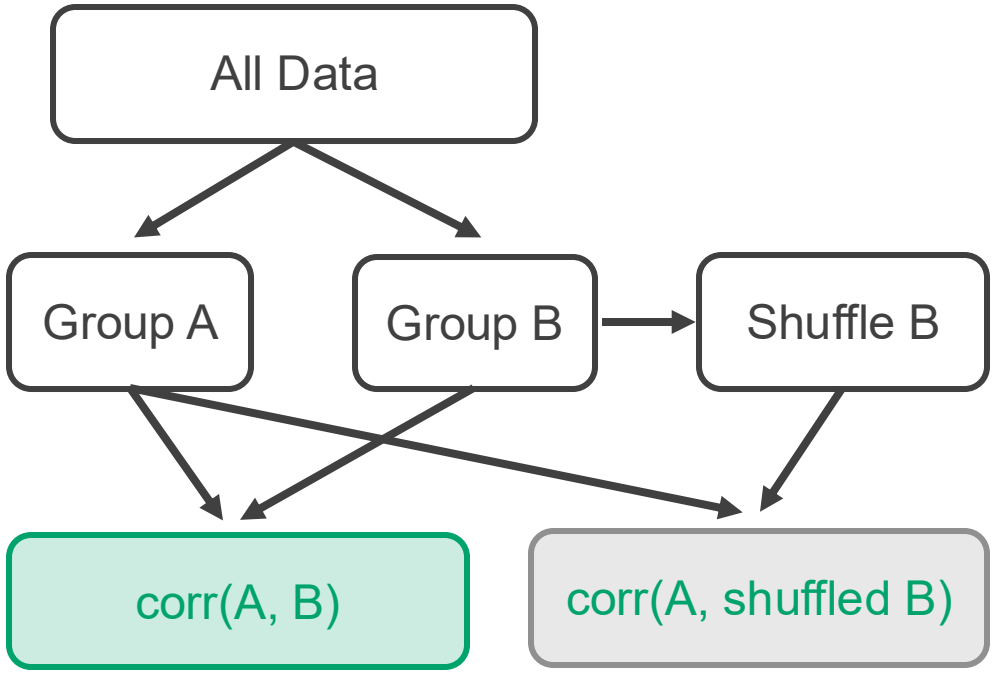
How consistent is the memorability of individual talkers for separate listeners?

Reliability of Memorability

How similar are Memorability scores across listeners?

Split-half consistency analysis:

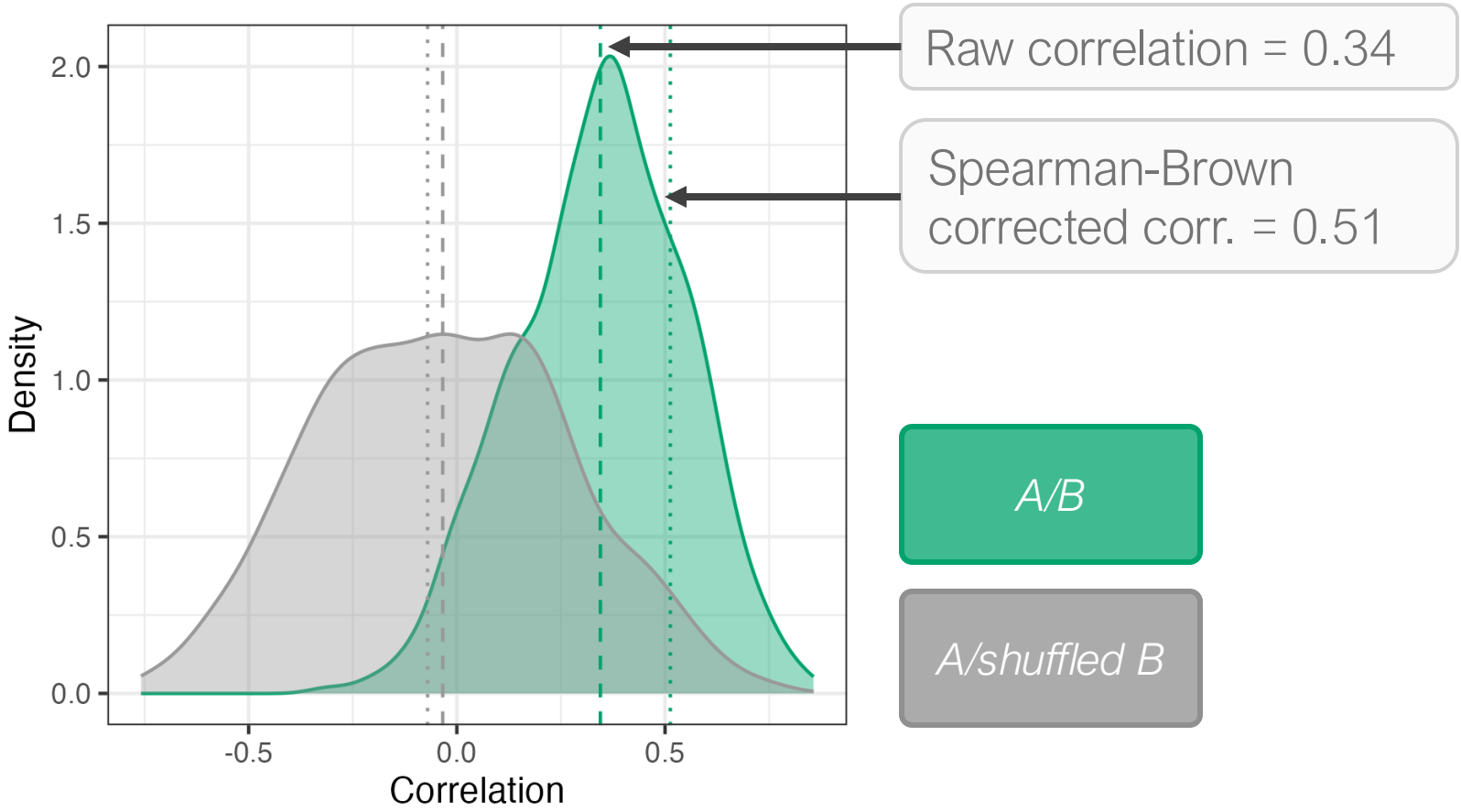
1. Divide participants in half.
2. Shuffle one group.
3. Compute talker memorability for all three.
4. Compute correlations.
5. Repeat 1,000 times.

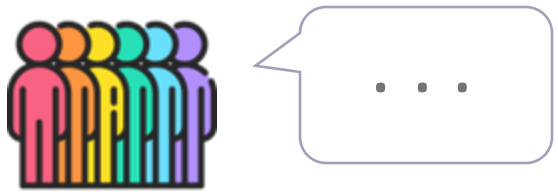


Reliability of Memorability

Stronger correlations in A/B than A/shuffled B.
p < 0.001

The relative memorability of individual talkers is consistent across listeners.





How do talkers' phonetic characteristics influence memorability?

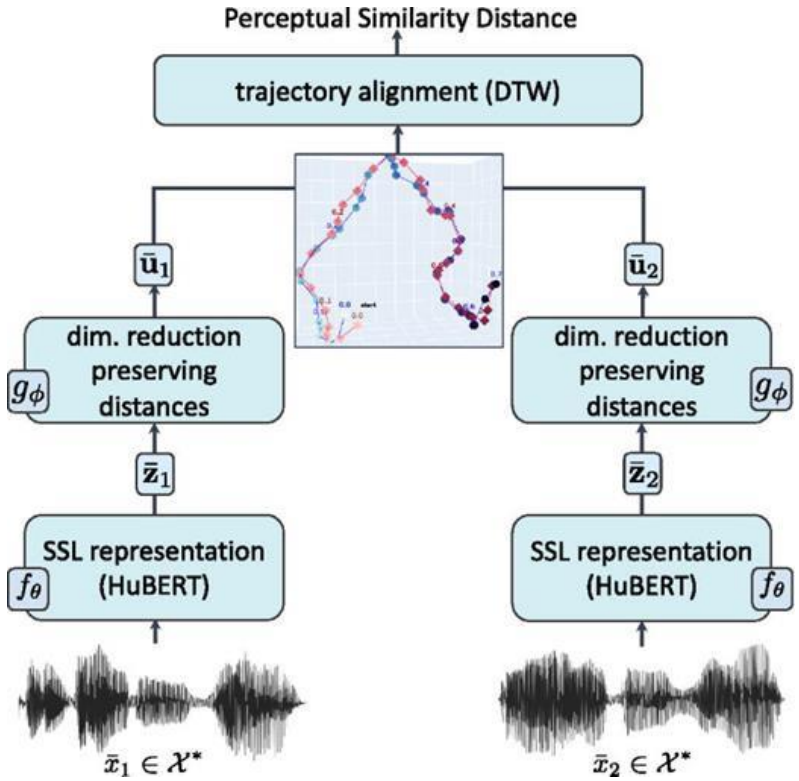
Phonetic similarity

DIFF accuracy rates are more variable than SAME.

More likely to to recognize phonetically similar than dissimilar repetitions?

Quantify similarity between utterances.
(Chernyak, Bradlow, Keshet, & Goldrick, 2024)

Analyze DIFF trials based on similarity between Study/Test tokens.

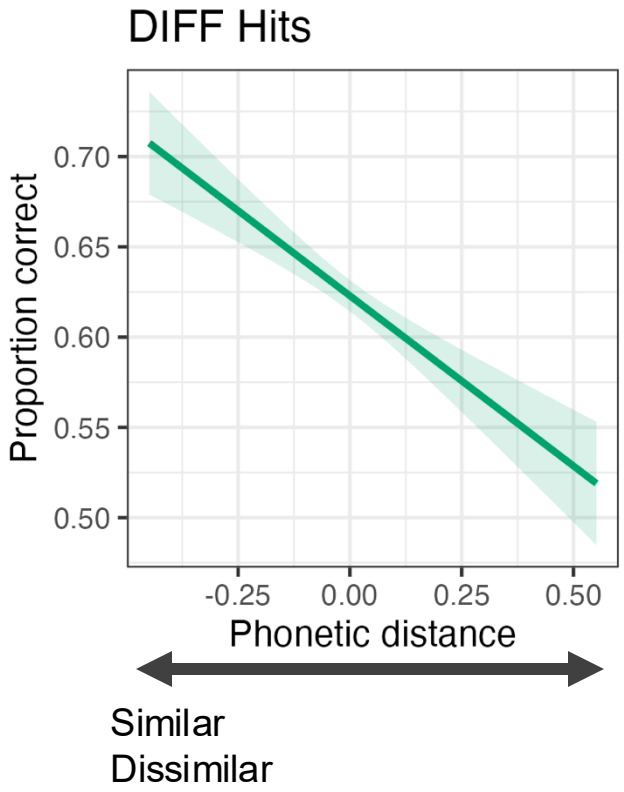


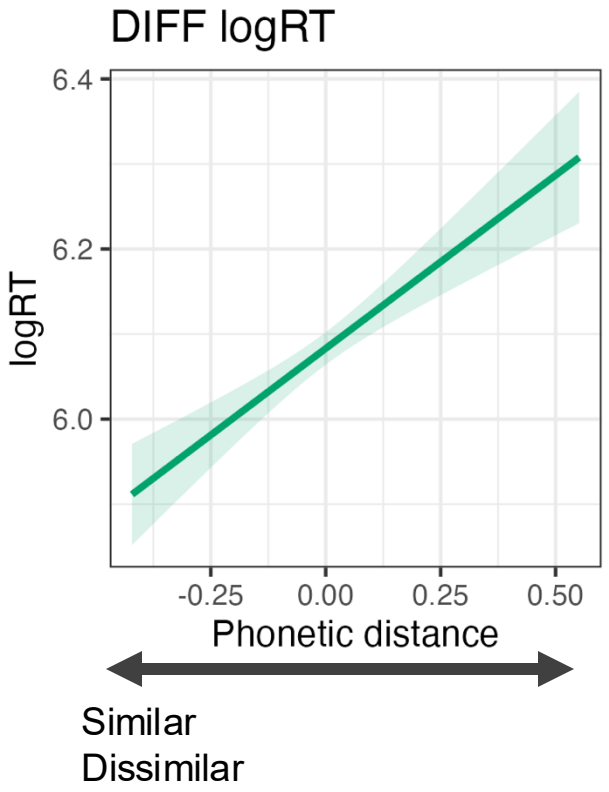
Chernyak, Bradlow, Keshet, & Goldrick, 2024

Phonetic similarity

Performance only on DIFF-talker trials.

More OLD sentences recognized for *similar* than *dissimilar* repetitions.
$p < 0.001$





logRT 6.4 = 602 ms
logRT 6.0 = 403 ms

Faster for *similar* than *dissimilar* repetitions.
$p < 0.001$

Specificity effects are gradient, not "all or none"!

Discussion

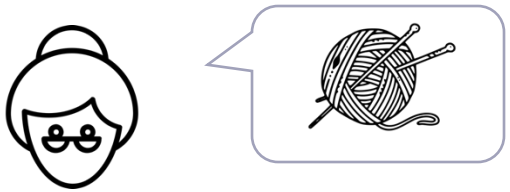
Memory ***differed*** across talkers – ***consistently*** across listeners.

Talker-specificity ***robust*** across talkers.

Less memorable talkers relied more on talker-specific detail for recognition than ***more memorable*** talkers.

Specificity effects are gradient!

Is talker memorability hard-coded or context-dependent?



How do memory patterns differ based on the relationship between talker and message?

Socially guided attention

Some views suggest that memorability is stimulus-intrinsic.
(Revsine, Goldberg, & Bainbridge, 2025)

Memory based on ***dynamic resource allocation*** must be flexible.

Hypothesis: *Congruence between speech and meaning leads to increased allocation of attention/memory resources.*

Expect to see ***memory boost*** when phonetic and semantic information are ***socially consistent***, particularly under divided attention.

Persona-Cuing Talkers

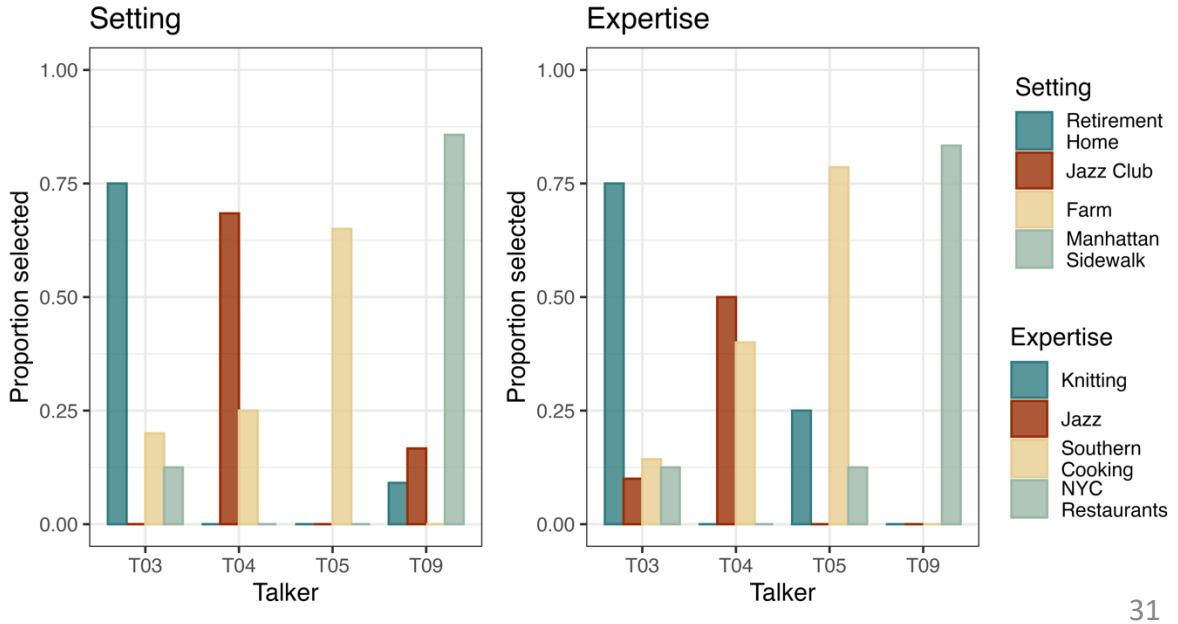
Personae: holistic, ideological social types that are recognizably linked with ways of being and speaking. (D'Onofrio, 2020)

Persona-cuing talkers: *Talkers with speech styles found to evoke consistent packages of social associations among naïve listeners.*

Four talkers selected from Study 2 via two-part norming.

Part 1: Free response.

Part 2: Multiple Choice.



Design

Participants: Full (*N* = 471), Divided (*N* = 476).

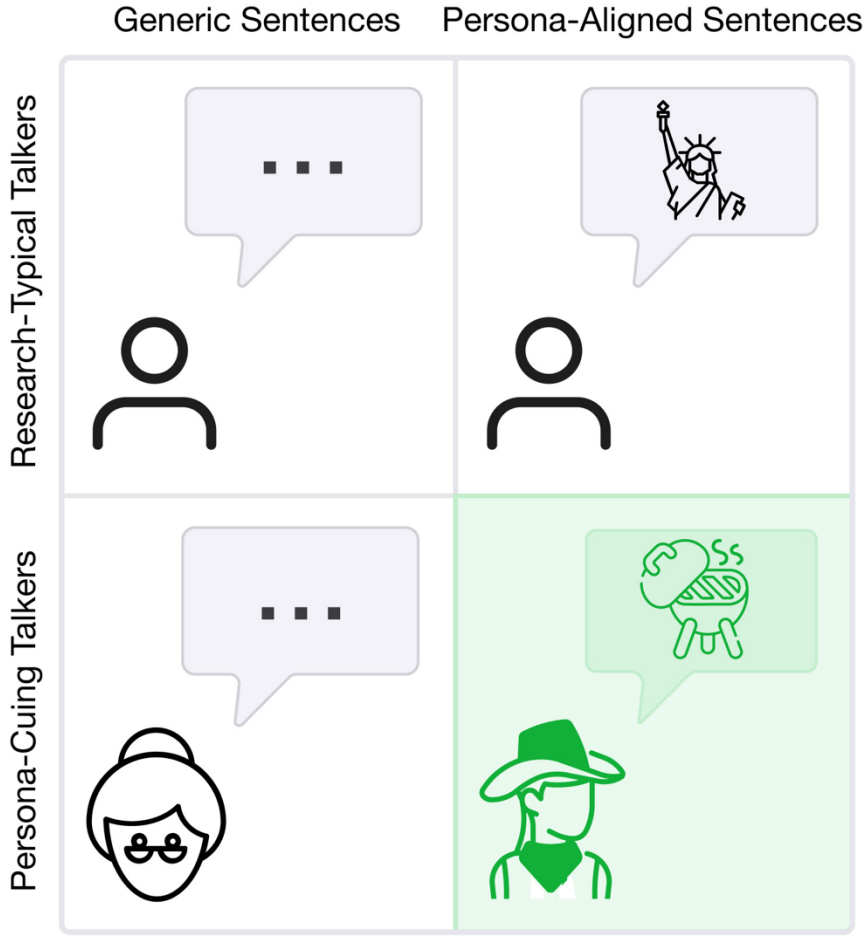
Procedure, attention/repVoice conditions same as previous studies.

Talkers:

- 4 Research-Typical: Same as Study 1.
- 4 Persona-Cuing: Selected via norming.

Stimulus sentences:

- Generic: Same as Studies 1 & 2.
- Persona-aligned: Constructed to emphasize social associations with Persona-Cuing talkers.



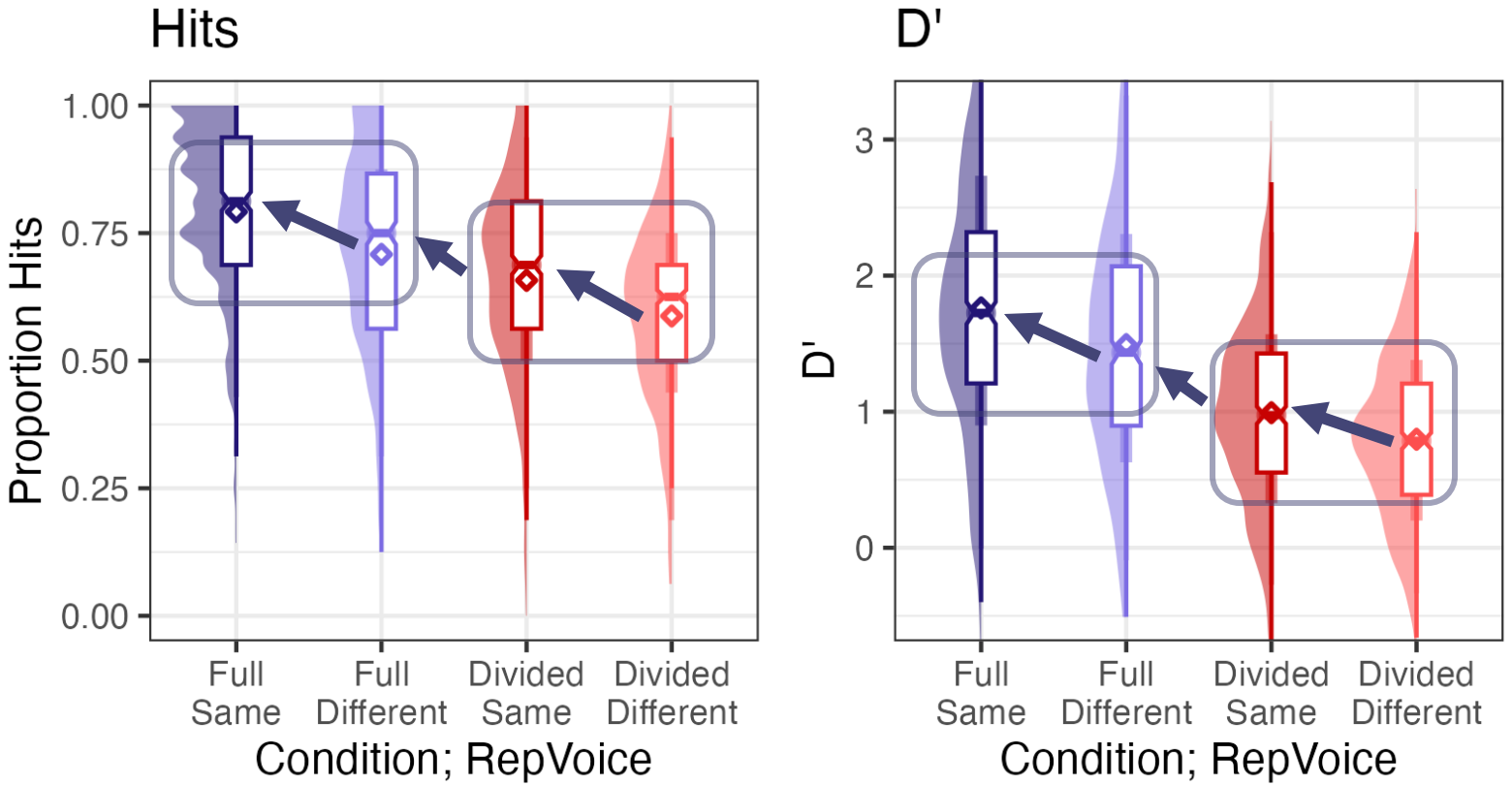
Congruent condition: *"The bikini was full of pokey."*

Attention & RepVoice

More accurate in Full than Divided across all measures.
Both p < 0.001

More accurate for SAME than DIFF talker, regardless of attention.
Both p < 0.001

Talker-specificity replicated again!



Full SAME *Div. SAME*

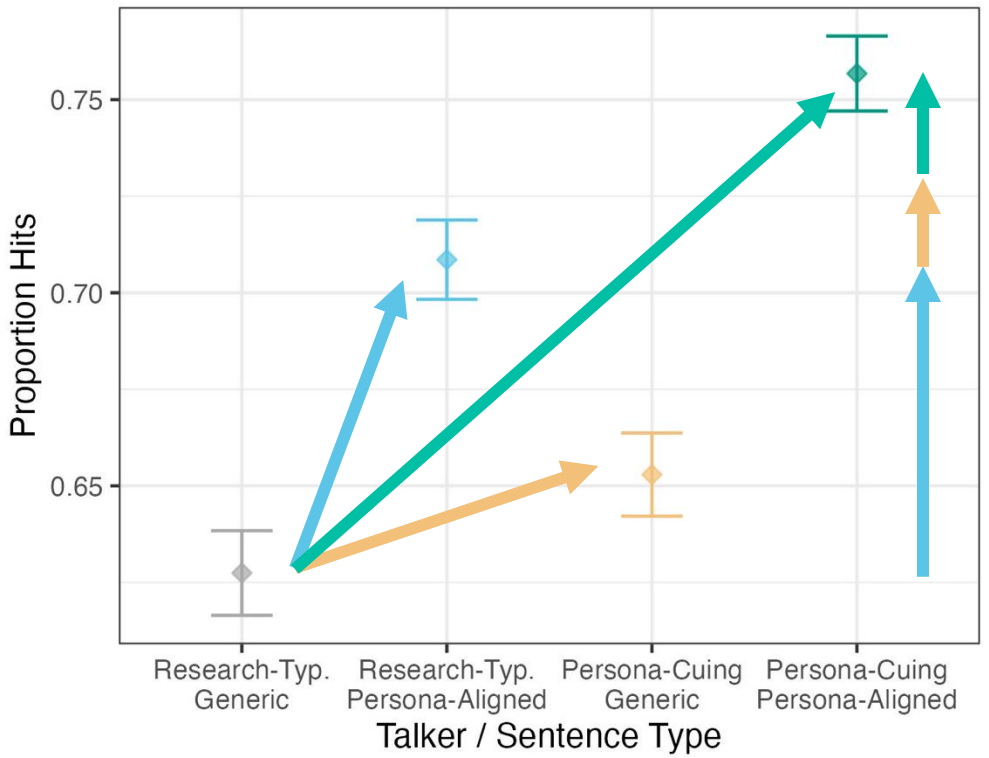
Full DIFF *Div. DIFF*

Talker & Sentence Type

Sentences: More accurate for *persona-aligned* than *generic*.
$p < 0.001$

Talkers: More accurate for *persona-cuing* than *research-typical*.
$p < 0.001$

Extra accuracy increase in *congruent condition.*
$p < 0.01$



Talker	Sentence
Research-Typical	Generic
Research-Typical	Persona-Aligned
Persona-Cuing	Generic
Persona-Cuing	Persona-Aligned

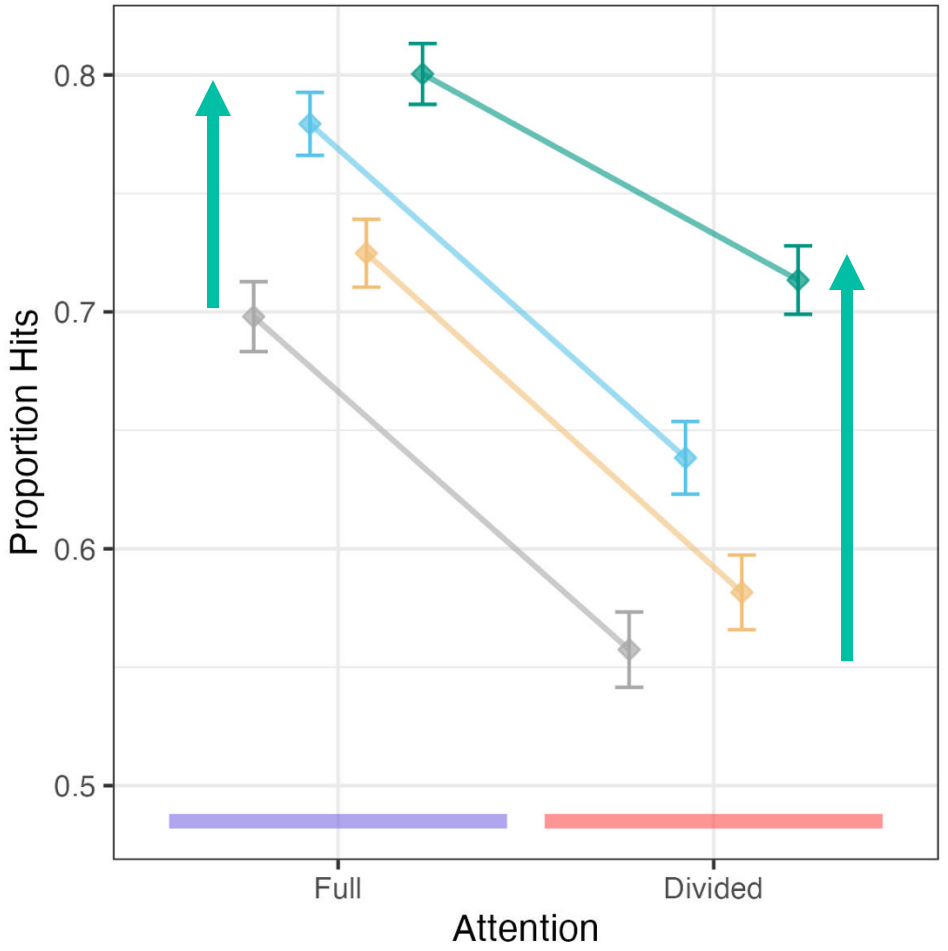
Listeners are sensitive to unique talker/message relationship!

Talker/Sentence & Attention

Congruence boost is even larger with Divided than Full attention.
$p < 0.01$

When the talker matched the message, participants ***reallocated*** attention to the stimulus.

Boost is strong enough that Congruent/Divided rivals some Full attention conditions.

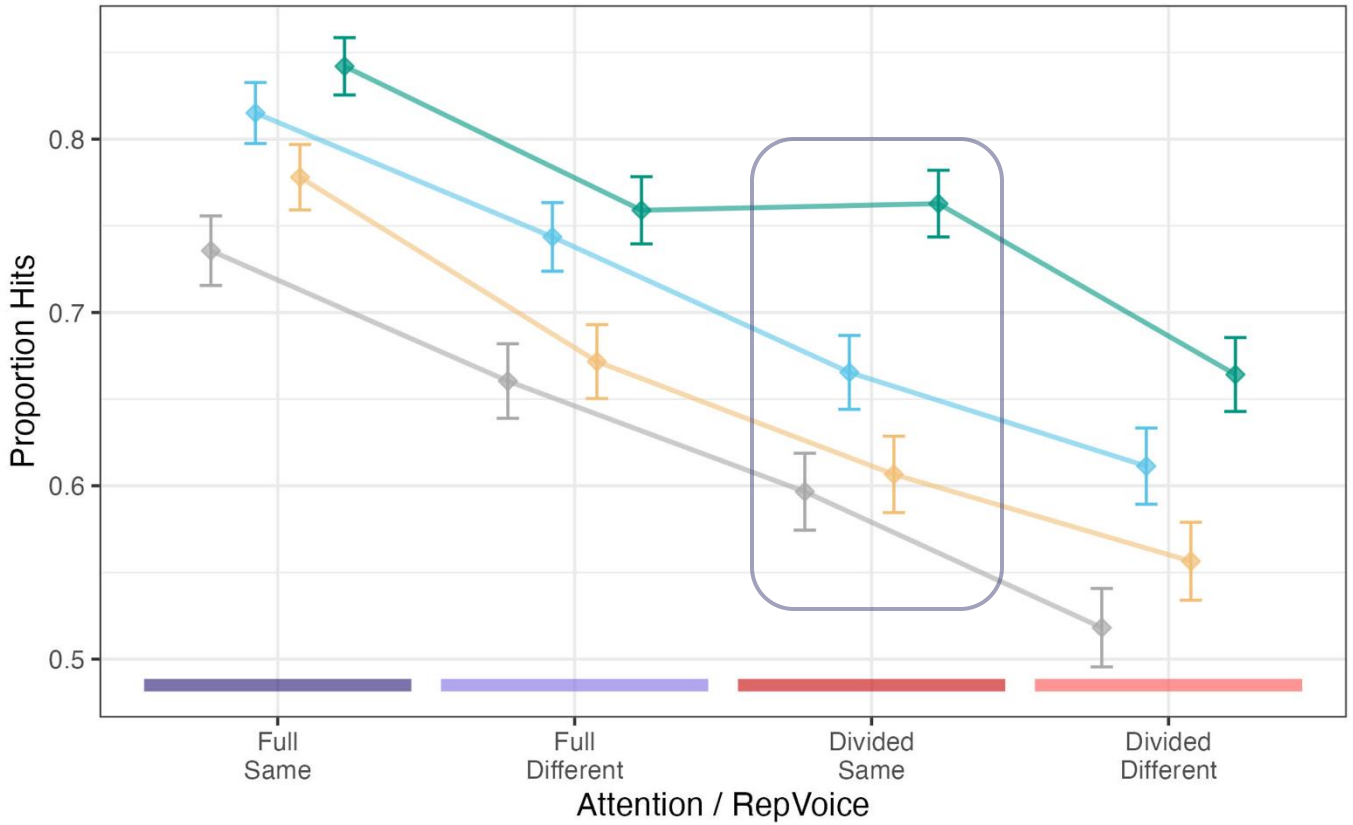


Talker	*Sentence*
Research-Typical	*Generic*
Research-Typical	*Persona-Aligned*
Persona-Cuing	*Generic*
Persona-Cuing	*Persona-Aligned*

All Variables

Effect was even larger for SAME than DIFF talker repetitions.
p < 0.05

Talker-specific info further magnified *Congruence* boost.



Talker	*Sentence*
Research-Typical	*Generic*
Research-Typical	*Persona-Aligned*
Persona-Cuing	*Generic*
Persona-Cuing	*Persona-Aligned*

Discussion

Central predictions successful:

- Memory boost in Congruent condition.
- Boost was even stronger in Divided condition than Full.

Social info central to allocation of memory/attention resources.

Talker memorability is not intrinsic, but ***context-dependent.***

Social associations learned in a particular cultural context fundamentally structure the way we perceive and remember language.

Recap

S1. SAME > DIFF, even for full sentences.
Effect larger in Divided than Full attention.

→ Fine-grained acoustic info encoded automatically; fundamental to the system.

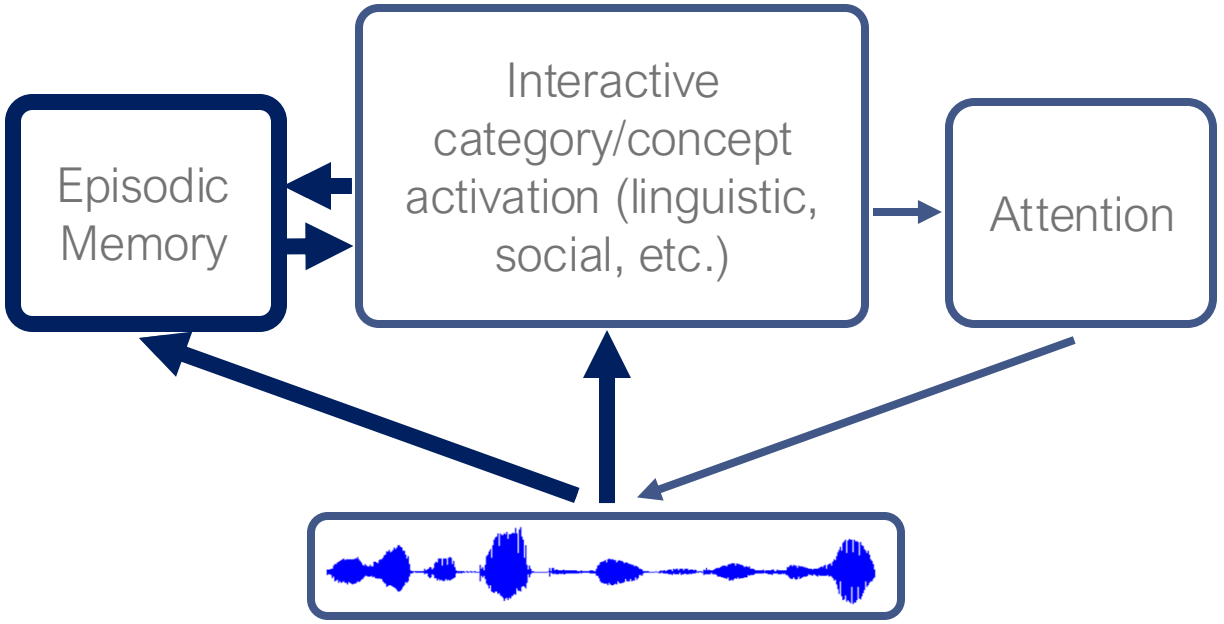
S2. Memory patterns ***different*** across talkers.
Memory patterns ***similar*** across listeners.

→ Listeners subconsciously allocate memory resources based on phonetic info.

S3. Talker/message Congruence led to better memory.
Memory boost most pronounced under Divided attention.

→ Talker memorability is context-dependent. Resource allocation is dynamic and socially guided.

Proposal – Socially Guided Attention



Behaviors depend on feedback between different types of info.

Resonance between ***linguistic*** and ***social*** categories enhanced attention.

Downstream consequences: *More robust representations of patterns we attentionally prioritize!*

Future Directions

How does memory for one talker depend on social context?

- Southern woman: With three Southern women? With three New Yorkers?

Mixing social associations.

- Southern woman talking about NYC?

Memory for acoustics vs. sentence meaning.

- Are we more likely to *internalize information* from some talkers than others?

Implications for language change. (Todd, Pierrehumbert, & Hay 2019)

Broader Implications

Contributor to speech-based biases?

- Are less-prioritized varieties at a memory disadvantage?
- Could be used to design interventions.

Introduces AI safety questions for ASR, TTS, & voice assistants.

Talker memorability could be considered for public outreach materials, e.g., PSAs or emergency communications.

Social and linguistic information are deeply integrated, and we process them dynamically.

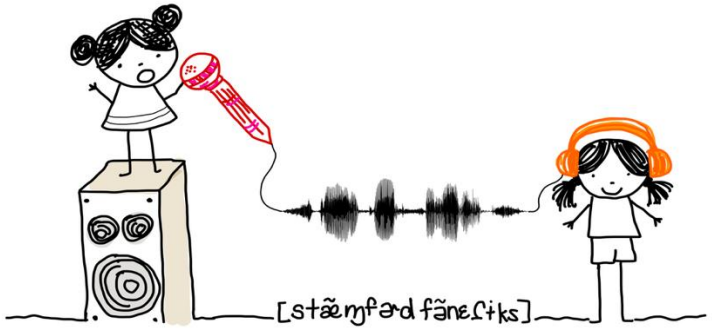
Variation isn’t an obstacle!
It’s a resource for language understanding.

Thank you!

Thanks first and foremost to Meghan Sumner, without whom none of this would have been possible! Thank you to my other committee members, Rob Podesva, Dan Jurafsky, and Hyo Gweon, as well as my University Chair, Takako Fujioka.

This project also benefited from conversations with Charlotte Vaughn, Ann Bradlow, Arty Samuel, and Steve Goldinger, and years of helpful comments from the Phonetics Lab.

Thanks also to my funding sources, including NSF DDRIG, William Orr Dingwall Foundations of Language Fellowship, and Josephine de Karman Fellowship Trust.



National Science Foundation

Josephine de Karman Fellowship

References

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. Journal of Experimental Psychology: Human Perception and Performance, 35 (2), 520–529.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics, 61*(2), 206–219.

Clapp, W., Vaughn, C., and Sumner, M. (2023). "The episodic encoding of talker voice attributes across diverse voices," *Journal of Memory & Language, 28*, 104376.

Chernyak, B. R., Bradlow, A. R., Keshet, J., & Goldrick, M. (2024). A perceptual similarity space for speech based on self-supervised speech representations. The Journal of the Acoustical Society of America, 155 (6), 3915–3929.

Gahl, S. (2008). Time and Thyme Are not Homophones: The Effect of Lemma Frequency on Word Durations in Spontaneous Speech. Language, 84 (3), 474–496.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251–279.

Hauk, O., & Pulverm¨uller, F. (2004). Effects of word length and frequency on the human event-related potential. Clinical Neurophysiology, 115 (5), 1090–1103.

Nygaard, L. C., & Queen, J. S. (2008). Communicating emotion: Linking affective prosody and word meaning. *J. Exp. Psychol. Hum. Percept., 34*, 1017–1030.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic Encoding of Voice Attributes and Recognition Memory for Spoken Words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(3), 309–328.

Pierrehumbert, J. B. (2016). Phonological Representation: Beyond Abstract Versus Episodic. Annual Review of Linguistics, 2 (1), 33–52.

Pufahl, A., & Samuel, A. G. (2014). How Lexical is the Lexicon? Evidence for Integrated Auditory Memory Representations. *Cognitive Psychology, 70*, 1–30.

Revsine, C., Goldberg, E., & Bainbridge, W. A. (2025). The memorability of voices s predictable and consistent across listeners. Nature Human Behaviour .

Rimikis, S., Smiljanic, R., & Calandruccio, L. (2013). Nonnative English Speaker Performance on the Basic English Lexicon (BEL) Sentences. Journal of Speech, Language, and Hearing Research, 56 (3), 792–804.

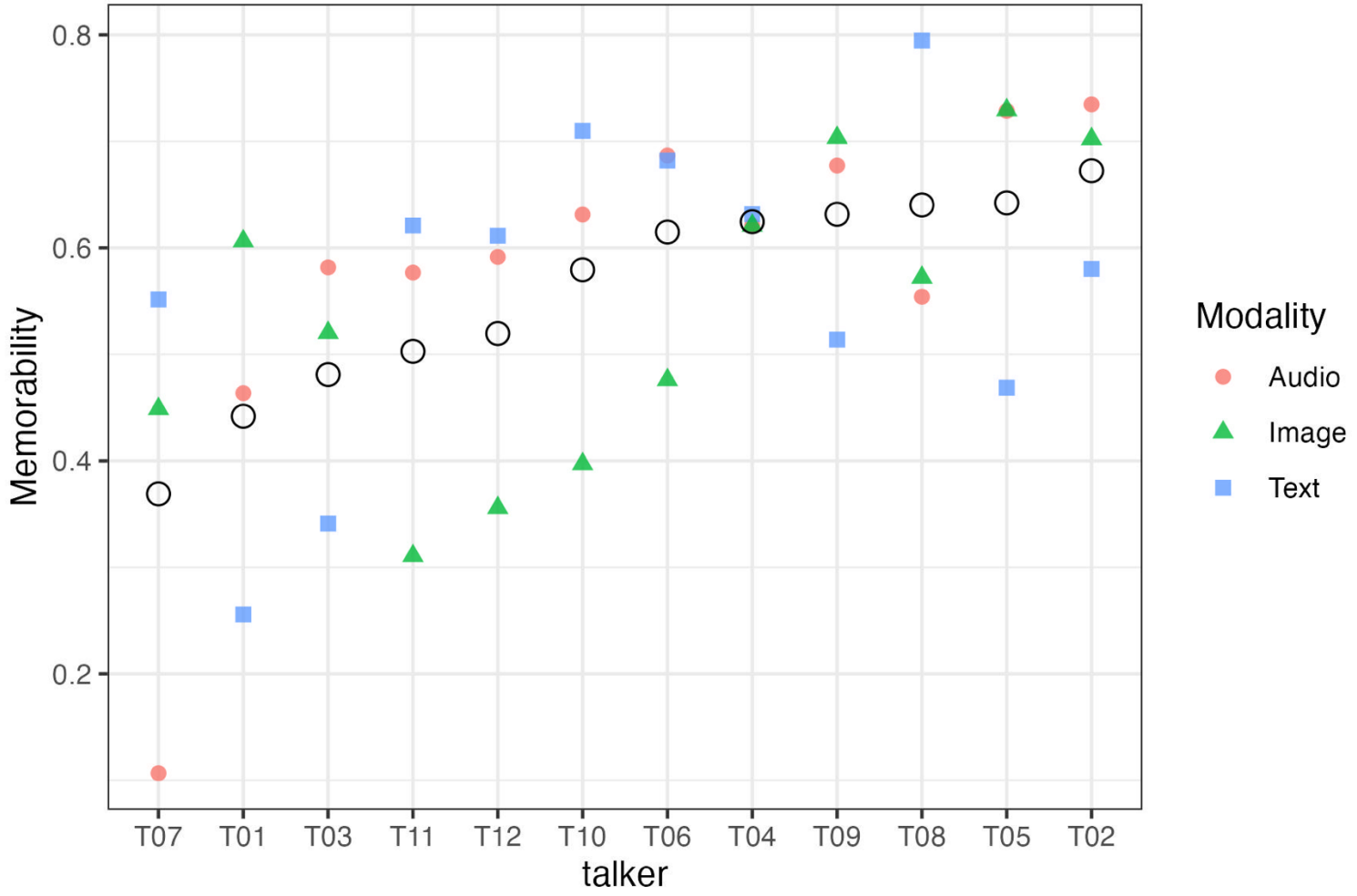
Sheffert, S. M. (1998). Contributions of surface and conceptual information to recognition memory. Perception & Psychophysics, 60 (7), 1141–1152.

Sumner, M., Kim, S. K., King, E., & McGowan, K. (2014). The socially-weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology, 4*, 1–13.

Todd, S., Pierrehumbert, J. B., & Hay, J. (2019). Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition, 185*, 1–20.

Wedel, A. (2012). Lexical contrast maintenance and the organization of sublexical contrast systems. Language and Cognition, 4 (4), 319–355.

Memorability & Modality



Three experiments with different target (test) modalities.

Memorability correlates across modalities.

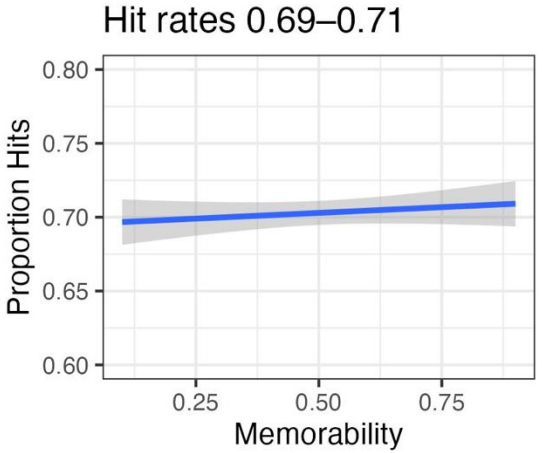
Correlation around 0.35 – similar to Study 2 split-half analysis.

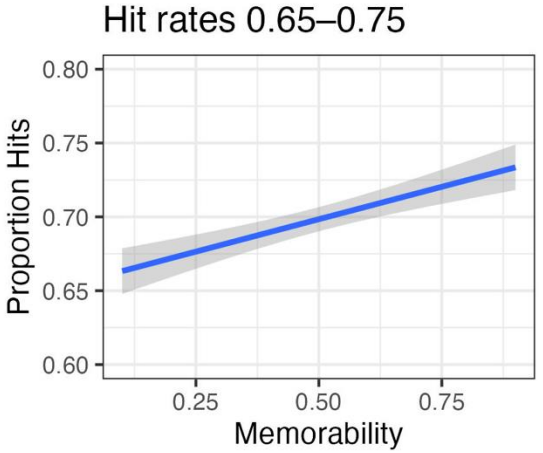
Memorability

Is it circular to use experimental outcomes as an independent variable?

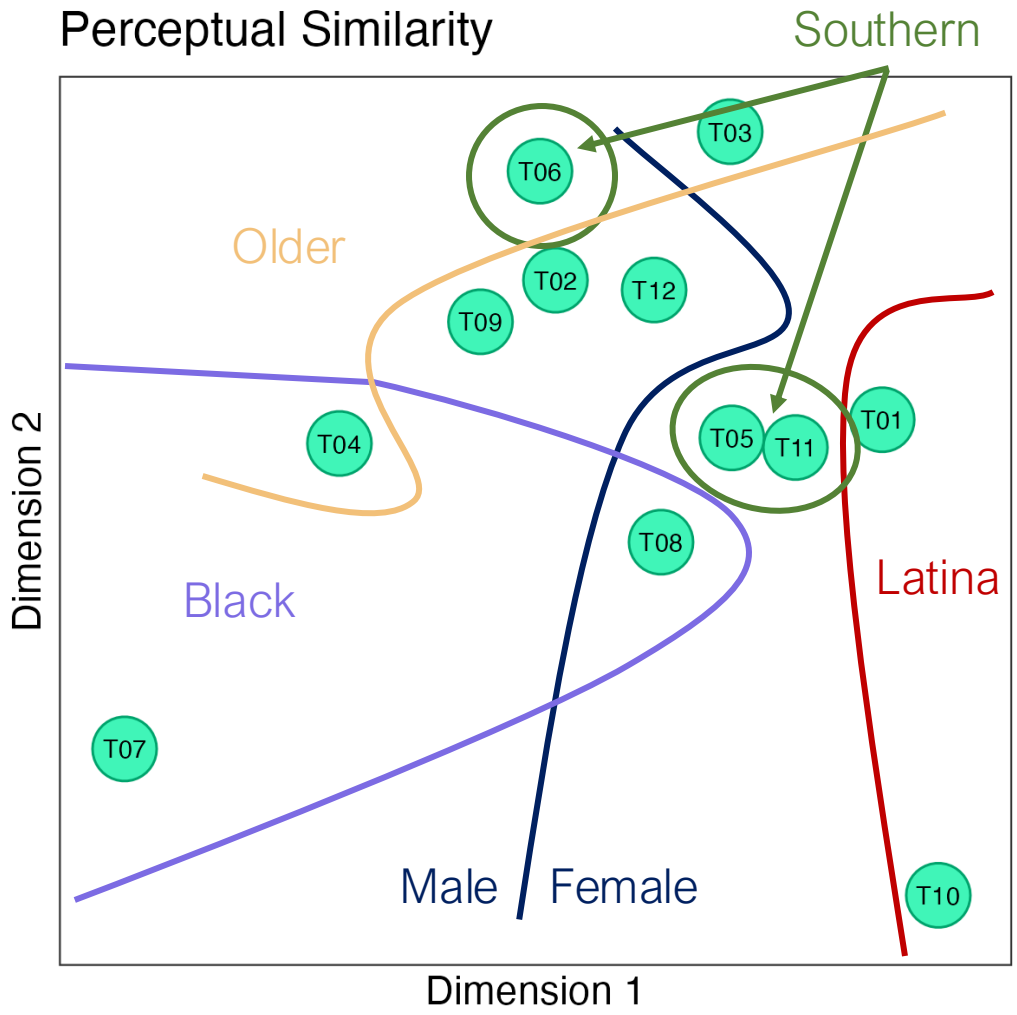
1. Scores calculated from Hits, False alarms, RTs together, but used to analyze them independently.
2. Low-level variability in memory performance wouldn't necessarily lead to significance.
3. Final values bootstrapped.

Synthetic data:





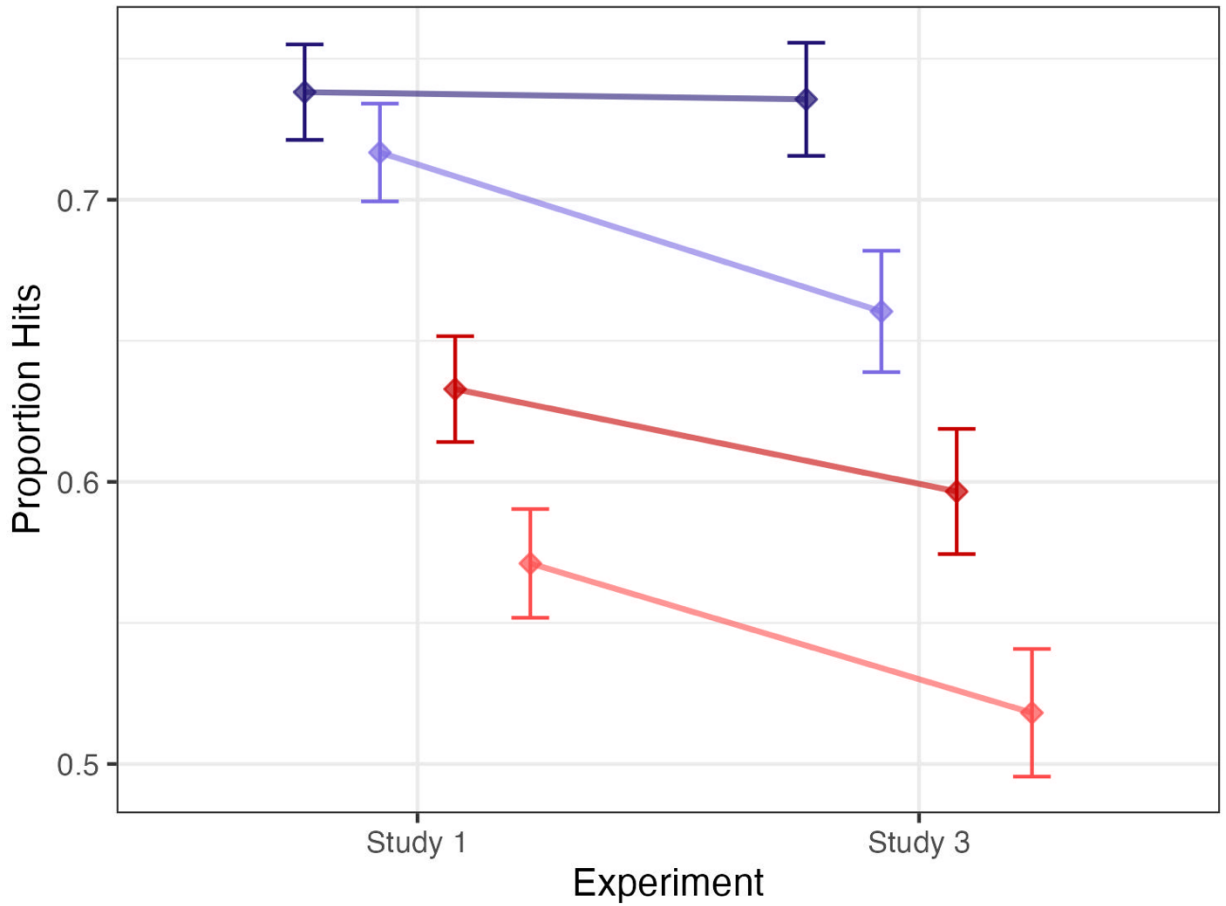
Perceptual Similarity



Talker	Mem
T01	0.47
T02	0.73
T03	0.57
T04	0.63
T05	0.73
T06	0.68
T07	0.11
T08	0.56
T09	0.68
T10	0.63
T11	0.58
T12	0.59

Memory cost?

Generic sentences; Research-typical talkers



Full SAME | *Div. SAME*

Full DIFF | *Div. DIFF*

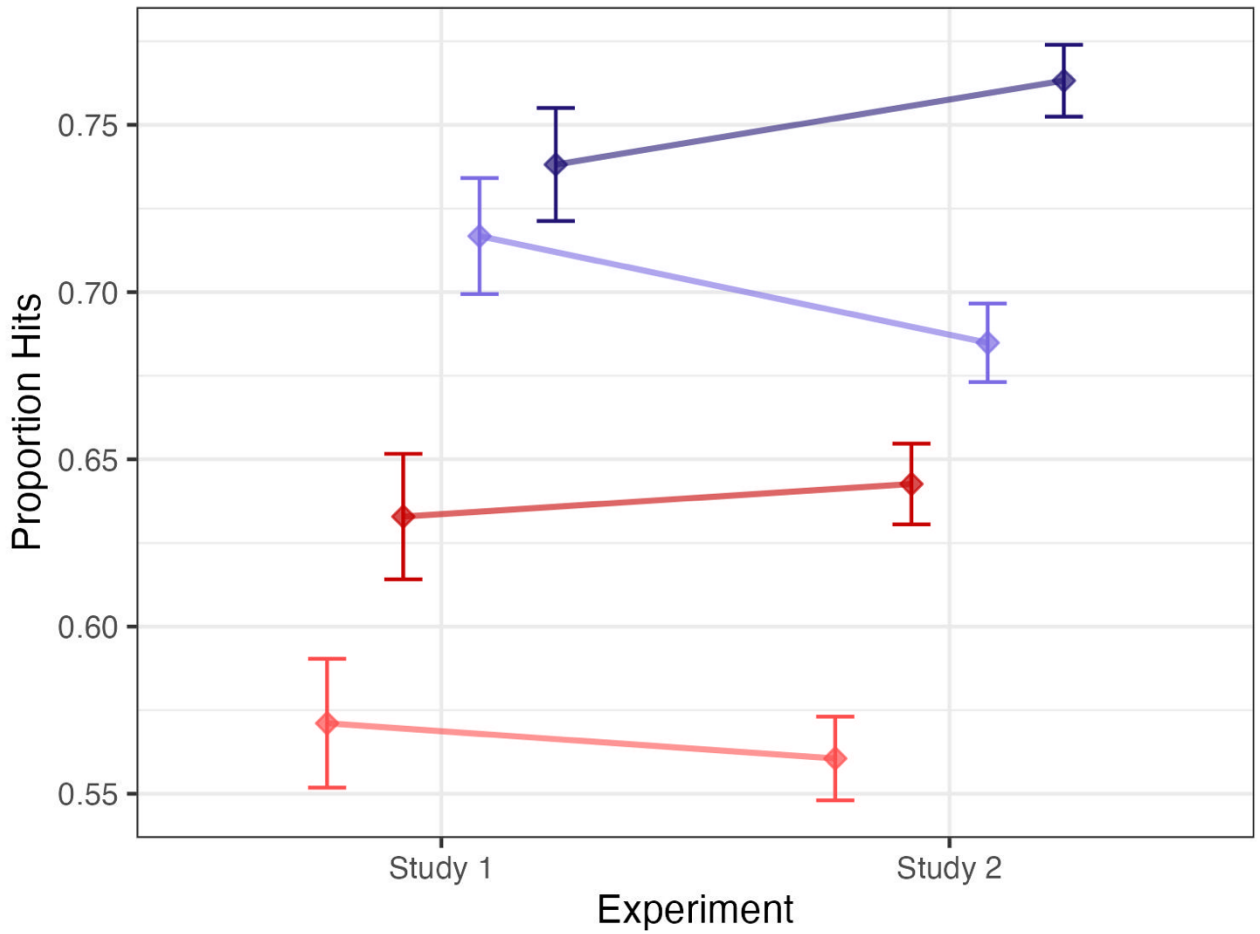
Overall higher accuracy in Study 1 than Study 3.
p < 0.001

Performance decrease more pronounced for DIFF than SAME.
p < 0.05

Boosts in other conditions may have diverted resources away from these talkers/sentences.

Study 1 vs. Study 2

Research-typical & diverse talkers



Full SAME | *Div. SAME*

Full DIFF | *Div. DIFF*

No difference in overall accuracy between studies.

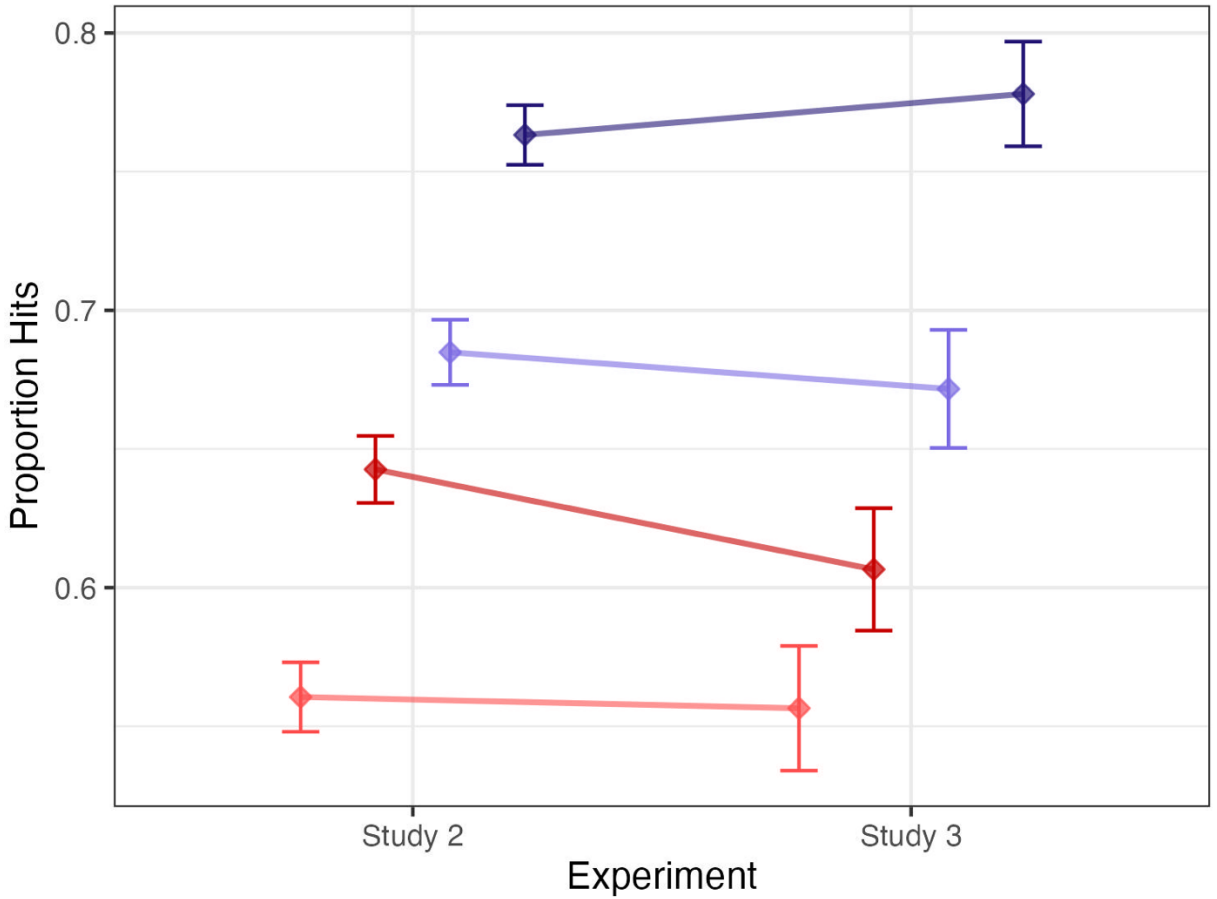
p > 0.1

Specificity effect size larger with diverse talkers than research-typical talkers.

p > 0.001

Study 2 vs. Study 3

Generic Sentences; Diverse Talkers



Full SAME | *Div. SAME*

Full DIFF | *Div. DIFF*

No difference between studies.

p > 0.1