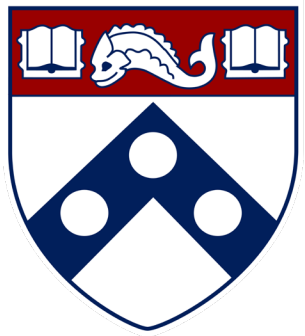
Now you hear me, later you don't:
The Markovian Nature of Phonological Categorization

Spencer Caplan
Alon Hafri (JHU)
John Trueswell

NECPhon 2019
November 16th, 2019



Penn
UNIVERSITY of PENNSYLVANIA

Speech Processing

Listeners convert speech from acoustic signal to an abstract linguistic (lexical/syntactic/semantic) representation

Speech Processing

Listeners convert speech from acoustic signal to an abstract linguistic (lexical/syntactic/semantic) representation

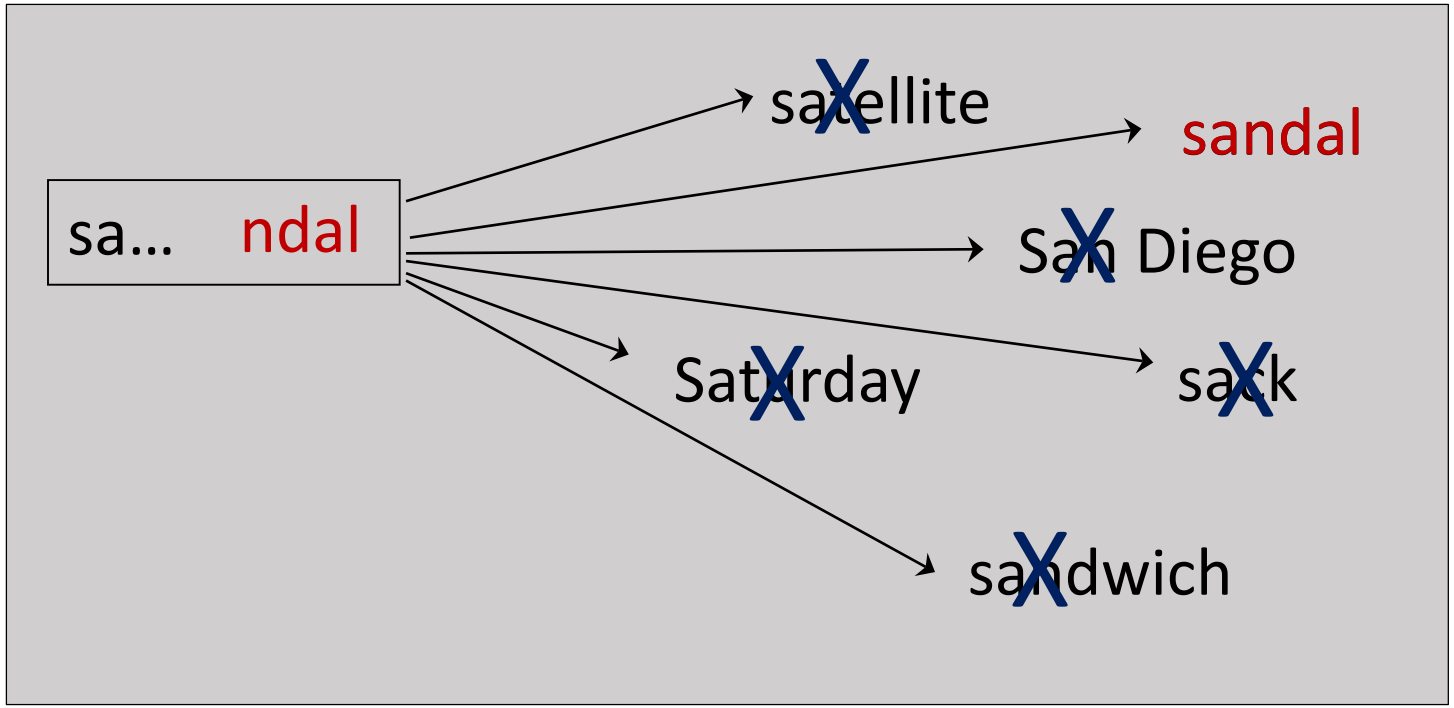
What mechanisms underlie this process?

What do *intermediate* representations contain?

Speech Processing in Time

Major constraint in speech processing: Time

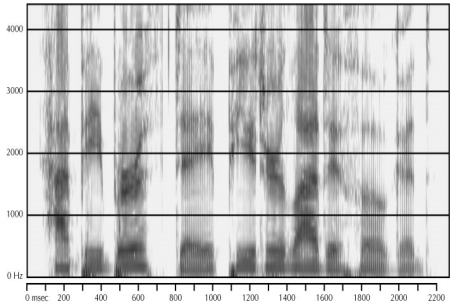
Information arrives sequentially and may be temporarily ambiguous



Integration of Multiple Sources

Bottom-Up

- **Gradient acoustics → discrete phones**
- **Phones → lexical items**
- **…..**



→

ʧ	ɜ	r	ʧ

Top-Down

- **Pragmatic/semantic/syntactic cues inform word choice**
- **Words entail phoneme selection**

→

Likelihood of discussing trees or camping or hiking increases

Ambiguity in Speech Processing

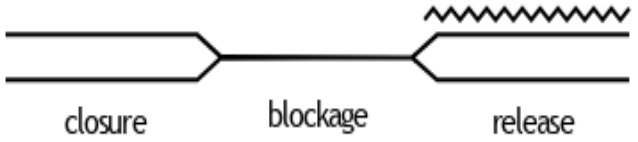
Speech processing is inextricably tied to local uncertainty

Given a time-slice of audio, it is not 100% deterministic what phoneme to map to

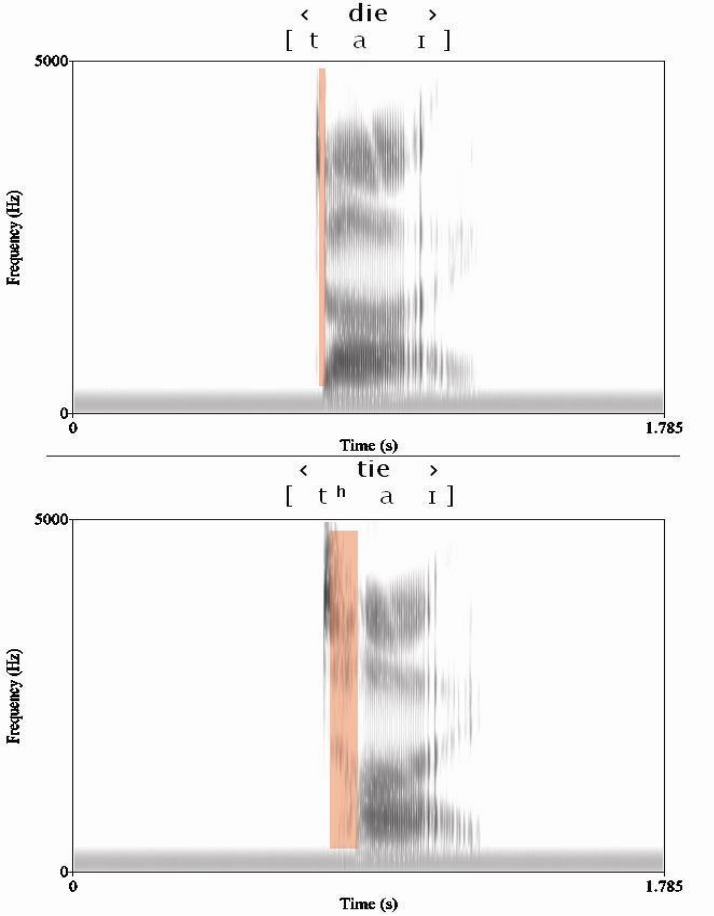
Tractable problem: (Primary) Acoustic cues for certain pairs of phonemes vary on particular, well-understood dimensions (e.g. VOT)

Voice Onset Time

Voiced /d/

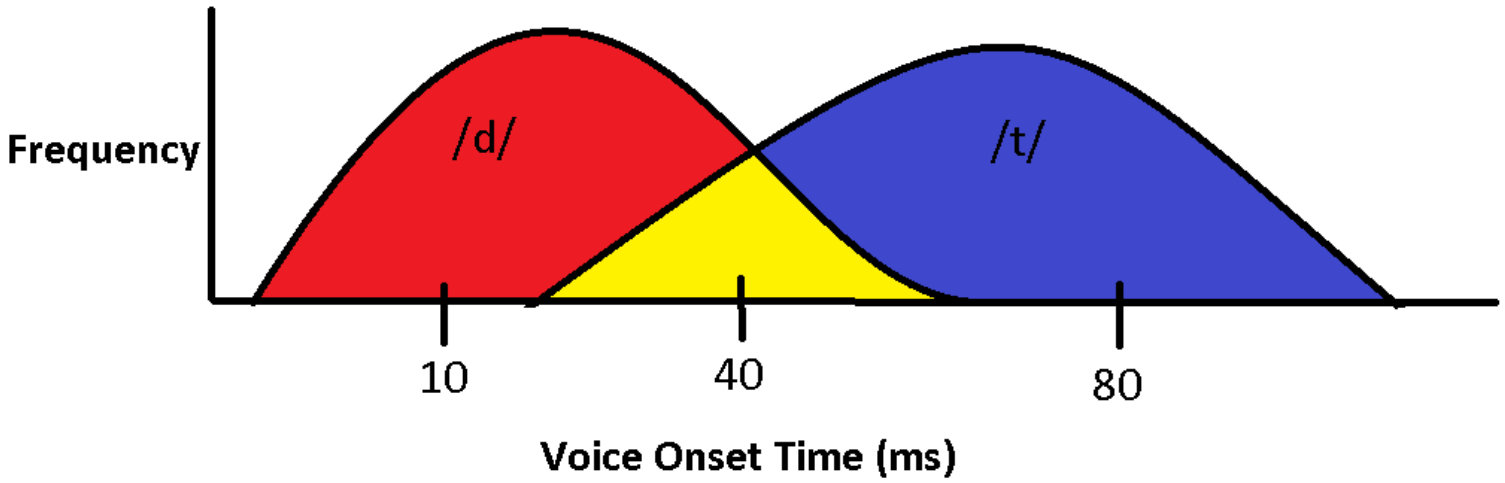


Voiceless /t/



Distributions over Phonetic Realization

Not every instance of production of a phoneme is acoustically identical



Maintenance during Processing

Listeners maintain an intermediate representation which includes some measure of 'uncertainty'

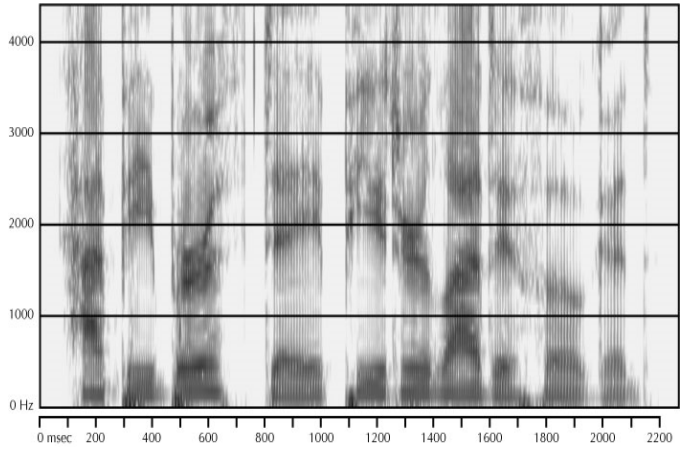
Otherwise information on the "*right*" couldn't integrate with information to the "*left*"

"In the **forest**, I saw a *t/dent*"

"I saw a *t/dent* in the **forest**"

Maintenance during Processing

What do *intermediate* representations contain?



Acoustic-Phonetic Signal

or

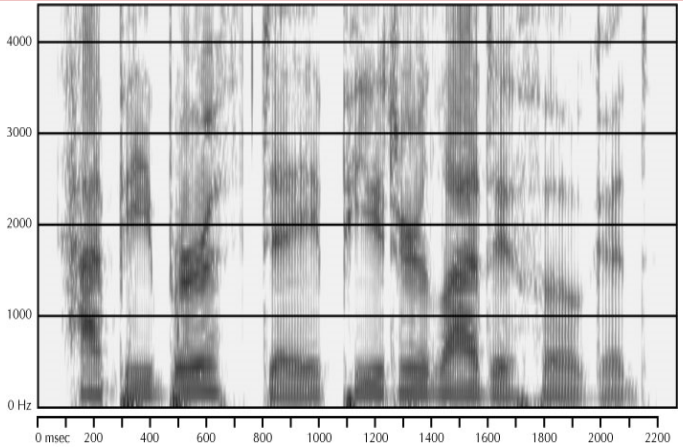
/t/	80%
/d/	20%

Activation over categories (phonemes, words, etc.)

Maintenance during Processing

The "activation over categories" (AOC) hypothesis is a *Markovian* process

Encodes a state of activation, but not the path that led to that belief



Acoustic-Phonetic Signal

or

/t/	80%
/d/	20%

Activation over categories (phonemes, words, etc.)

Outline

- **Background**
  - **Intermediate Representations in Speech Processing (signal retention vs. AOC)**
- Experiments 1 and 2
  - The Immediacy of Linguistic Computation
- Experiment 3
  - Mapping Between Categories
- Discussion

Bushong & Jaeger (2017)

"I saw a *t/dent* in the **forest**"

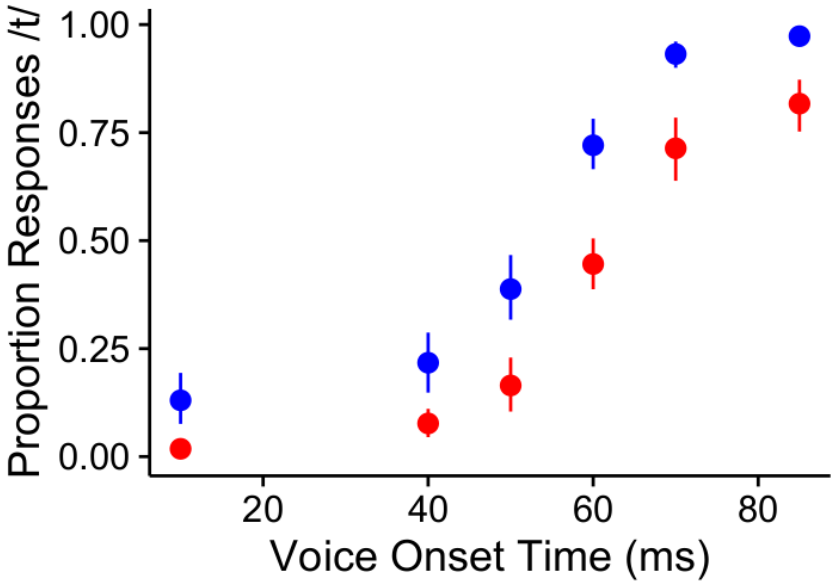
"I saw a *t/dent* in the **fender**"

- Subjects hear a sentence where VOT of the onset on some **target word** was modulated.
- The **disambiguating context** occurs to the right of the target
- Ask participants what (target) word they think they heard.

Bushong & Jaeger (2017)

Red: “dent”-contexts

Blue: “tent”-contexts

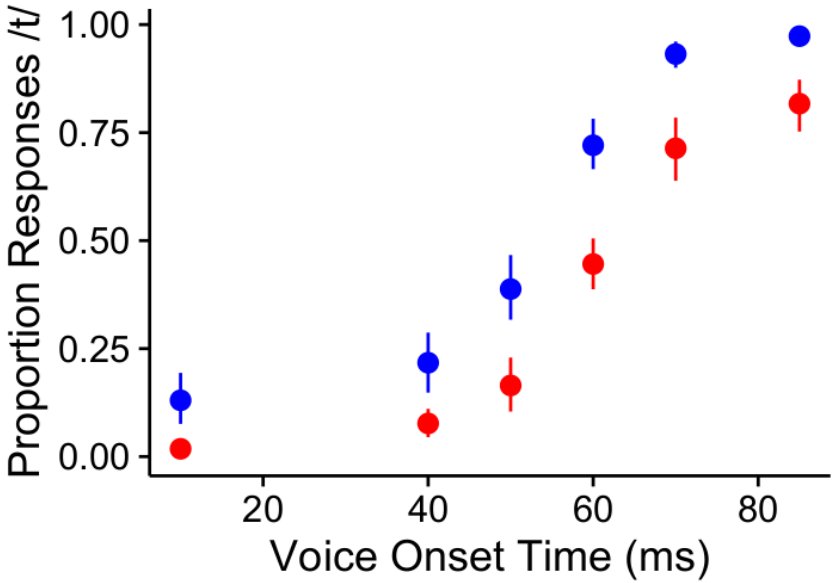


See also: Connine (1991)

Bushong & Jaeger (2017)

Red: “dent”-contexts

Blue: “tent”-contexts



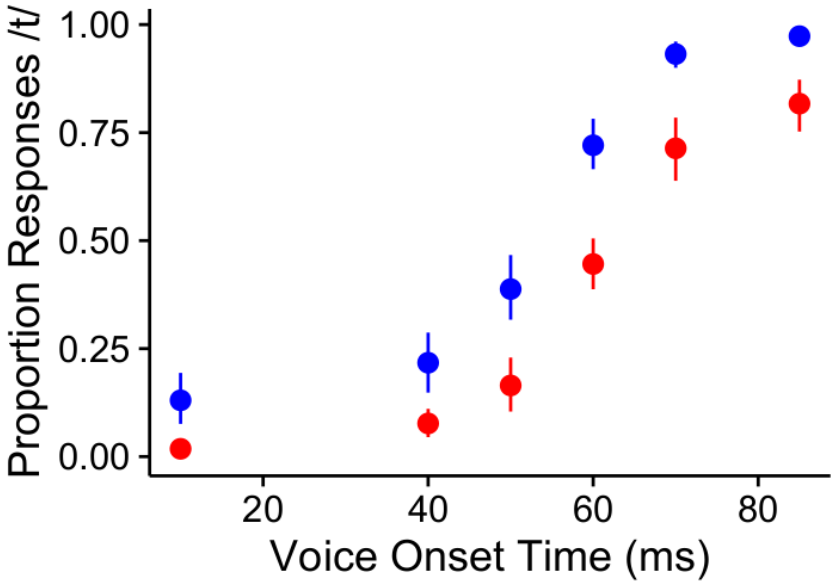
- Maintenance of some kind of intermediate representation
- Integration between temporally disjoint cues

See also: Connine (1991)

Bushong & Jaeger (2017)

Red: “dent”-contexts

Blue: “tent”-contexts



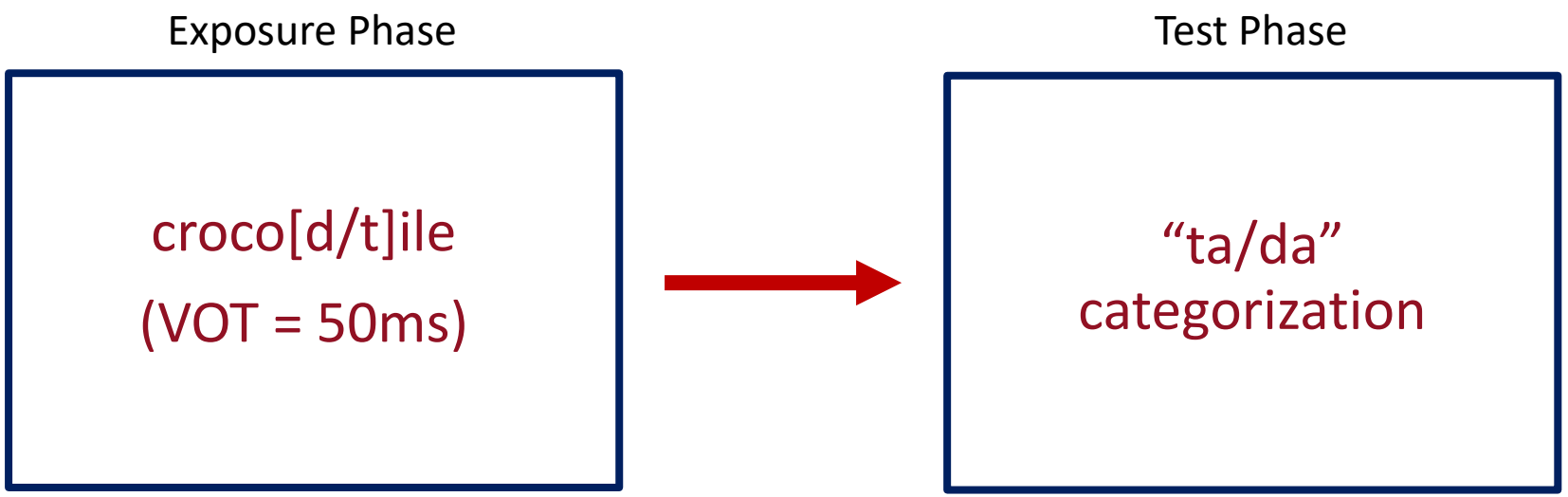
Does not differentiate between maintenance of:

acoustic-phonetic signal vs. **probabilistic activation over discrete categories (AOC)**

See also: Connine (1991)

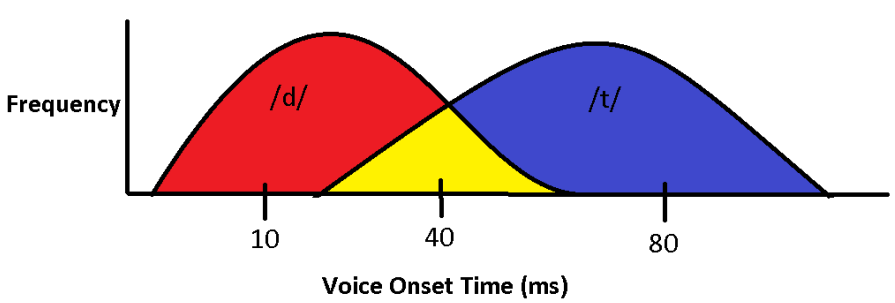
Accent Adaptation

Listeners are able to rapidly and effectively update speaker-specific models of speech processing (Bradlow & Bent, 2008; Burchill et al., 2018, Kraljic & Samuel, 2006, among others)

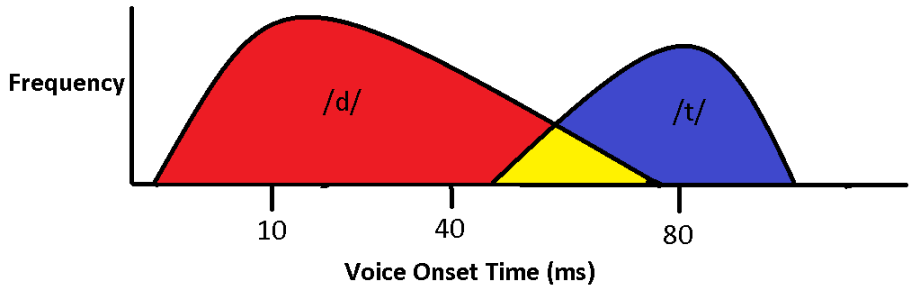


Accent Adaptation

Pre-Exposure Phase



Post-Exposure Phase



However, the cognitive mechanisms responsible for such accent adaptation effects are understudied

Outline

- Background
  - Intermediate Representations in Speech Processing (signal retention vs. AOC)
- **Experiments 1 and 2**
  - **The Immediacy of Linguistic Computation**
- Experiment 3
  - Mapping Between Categories
- Discussion

Experiment

New Accent Adaptation paradigm

Same exposure / test-phase design

Use minimal pairs (“tent/dent”) and orthographic labels to control the temporal availability of disambiguating cues for integration

Do intermediate representations contain acoustic-phonetic information or only activation over categories (AOC)?

Experiment

Exposure Phase

Audio-Text Order

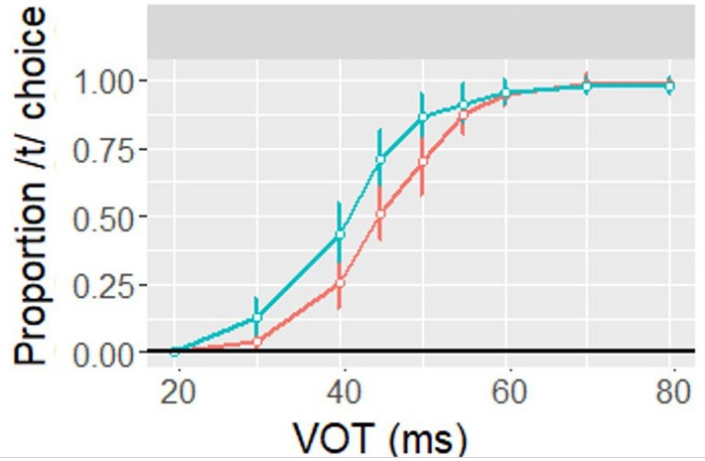
Expanded-Phoneme

Expand-/d/ Text-Before	Expand-/d/ Text-After
Expand-/t/ Text-Before	Expand-/t/ Text-After

Test Phase

(Same for all participants)

Categorize audio as “ta” vs. “da”



Experiment

Test Phase

- Identical for all participants

- 162 trials of phoneme categorization
  - 2 exemplar “ta/da” tokens
  - 9 VOT levels (between 20ms and 80ms)
  - 9 repetitions for each exemplar and VOT

Experiment

Exposure Phase

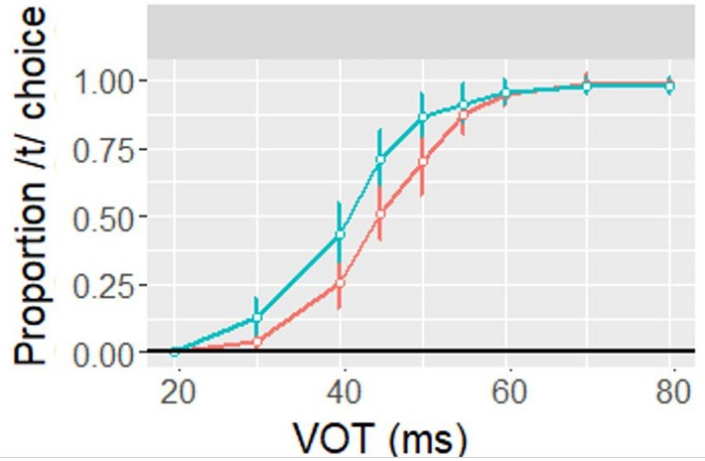
Audio-Text Order

Expanded-Phoneme		
	Expand-/d/ Text-Before	Expand-/d/ Text-After
	Expand-/t/ Text-Before	Expand-/t/ Text-After

Test Phase

(Same for all participants)

Categorize audio as “ta” vs. “da”



Experiment

Exposure Phase

- All participants hear a sequence of 142 text/audio pairings
  - 44 target items
  - 98 filler items
- Instructed to press a button to confirm whether or not the text/audio "match"
  - All targets are matches
  - 20 of 98 filler items include explicit mismatch (e.g. audio is "coffee" but text is "green")

Consistent for all groups

- Relative order of text vs. audio (**Text-Before vs. Text-After**)
- Pairing of target audio to text (**expand-/d/ vs. expand-/t/**)

Varies by group

Exposure Words

Target Words

- **Tab / Dab**
- **Tally / Dally**
- **Teem / Deem**
- **Tense / Dense**
- **Time / Dime**
- **Tusk / Dusk**
- **….**

- Minimal pairs differentiated only by onset position /t/ vs. /d/
- Same within-pair part-of-speech
- Frequency matched
- Manually selected 22 pairs (44 words)

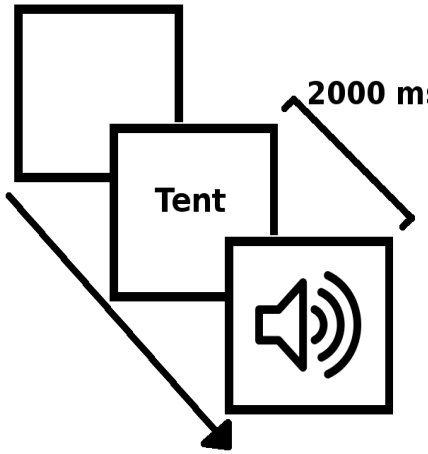
Filler Words

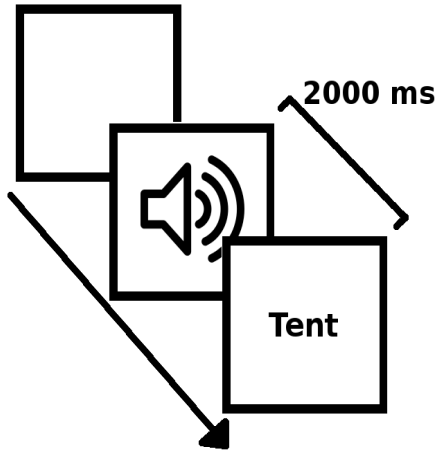
- **Acre**
- **Embrace**
- **Frame**
- **Jealous**
- **None**
- **Slip**
- **….**

- No phones /t/ or /d/
- No orthographic letters “t” or “d”
- No proper nouns, capital letters, etc.
- At least four letters, no longer than four syllables
- CELEX frequency > 150
- Randomly sampled 98 words

Timing

Disambiguating subtitle appears either before or after audio





Text-Before condition

Both Acoustic Maintenance and AOC predict adaptation

(in line with Kraljic & Samuel 2006 etc.)

Text-After condition

Acoustic Maintenance predicts adaptation

AOC predicts no adaptation

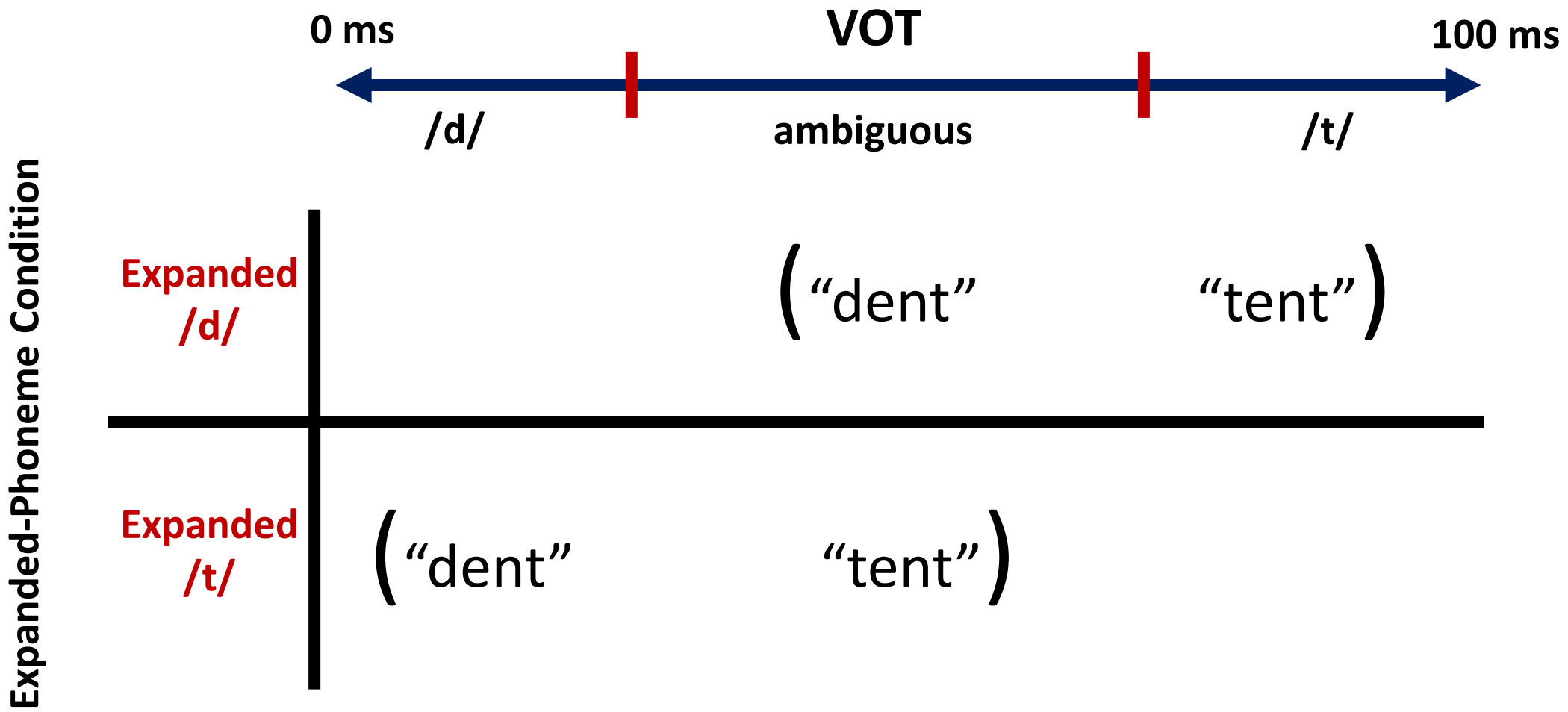
Audio Manipulation

Audio editing for Targets and Test Items

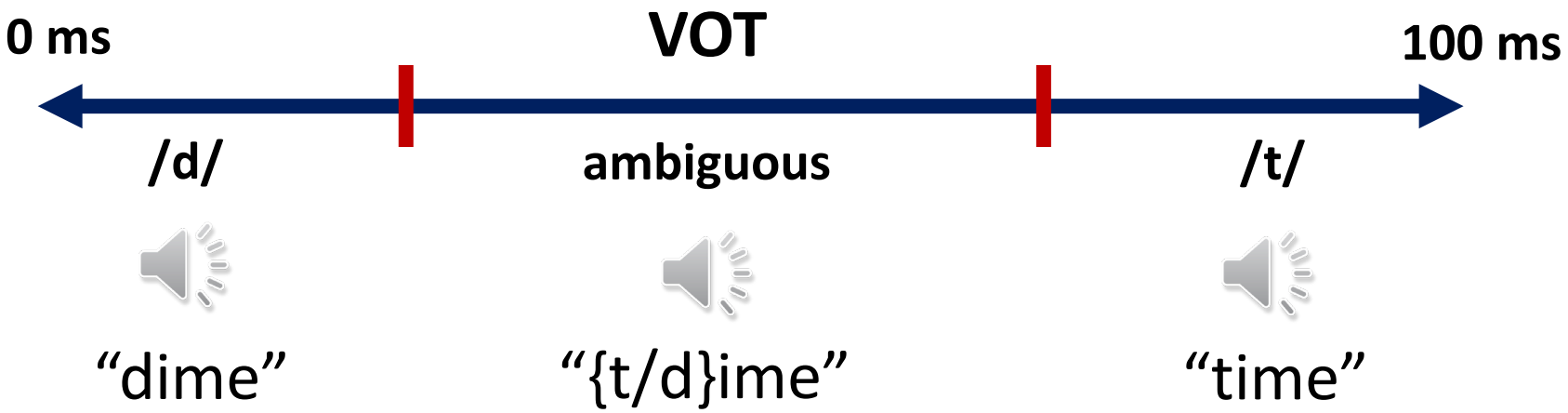
Manipulate VOT by splitting onset of /t/-word with rime of /d/-word at nearest zero-crossing

- Unambiguous /d/ used 10ms VOT
- Unambiguous /t/ used 100ms VOT
- Ambiguous targets used 60ms for Experiment 1 and 45ms for Experiments 2 and 3

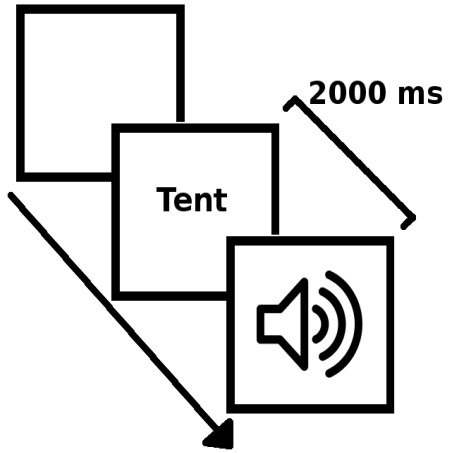
Target Text-Audio Pairing



Example Audio



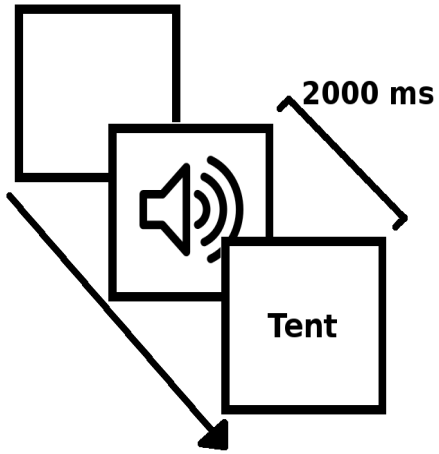
Predictions



Text-Before condition

Both Acoustic Maintenance and AOC predict adaptation

(in line with Kraljic & Samuel, 2006 etc.)

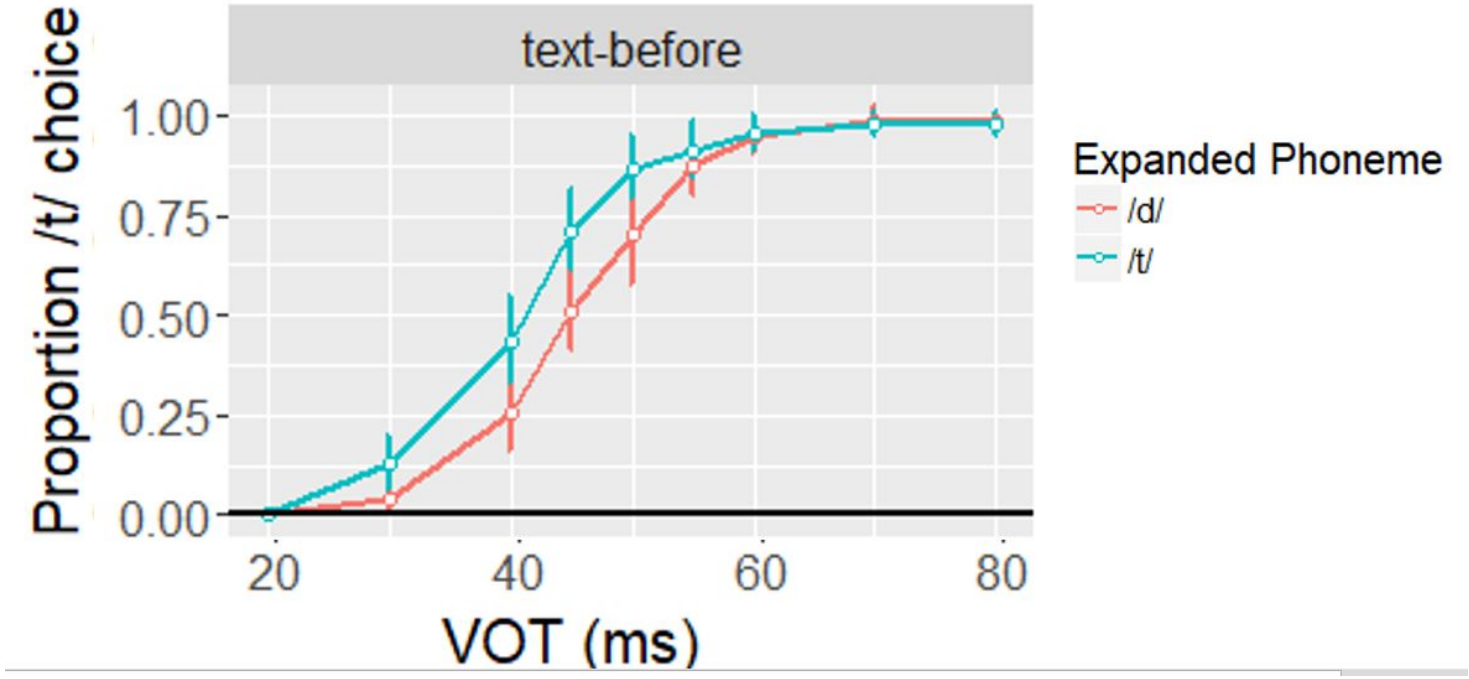


Text-After condition

Acoustic Maintenance predicts adaptation

AOC predicts no adaptation

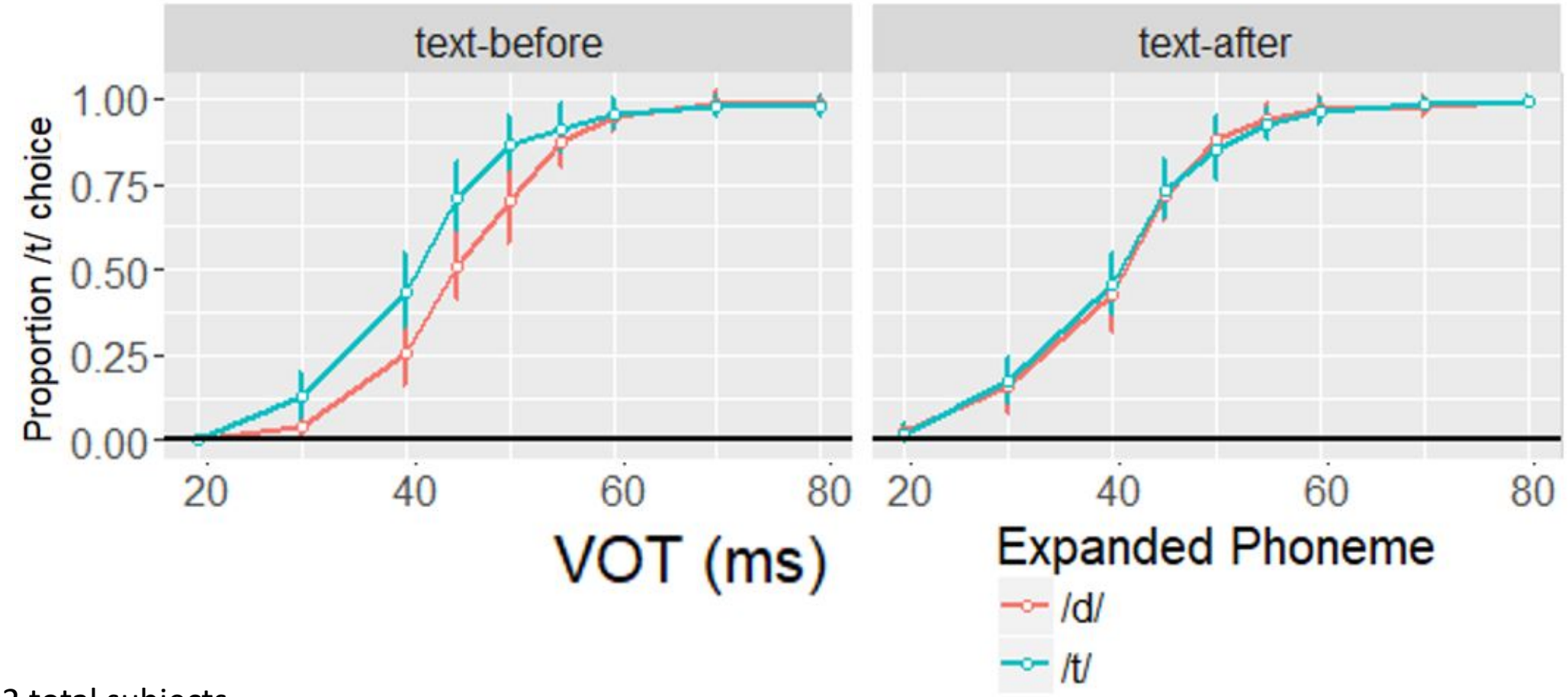
Text-Before condition



- Evaluated via mixed effects logistic regression to predict individual test-trials
- Additionally fit psychometric functions to each participant's data (maximum likelihood estimation) to predict the 50% threshold for phone categorization
- Results confirmed via repeated measures ANOVA across subjects

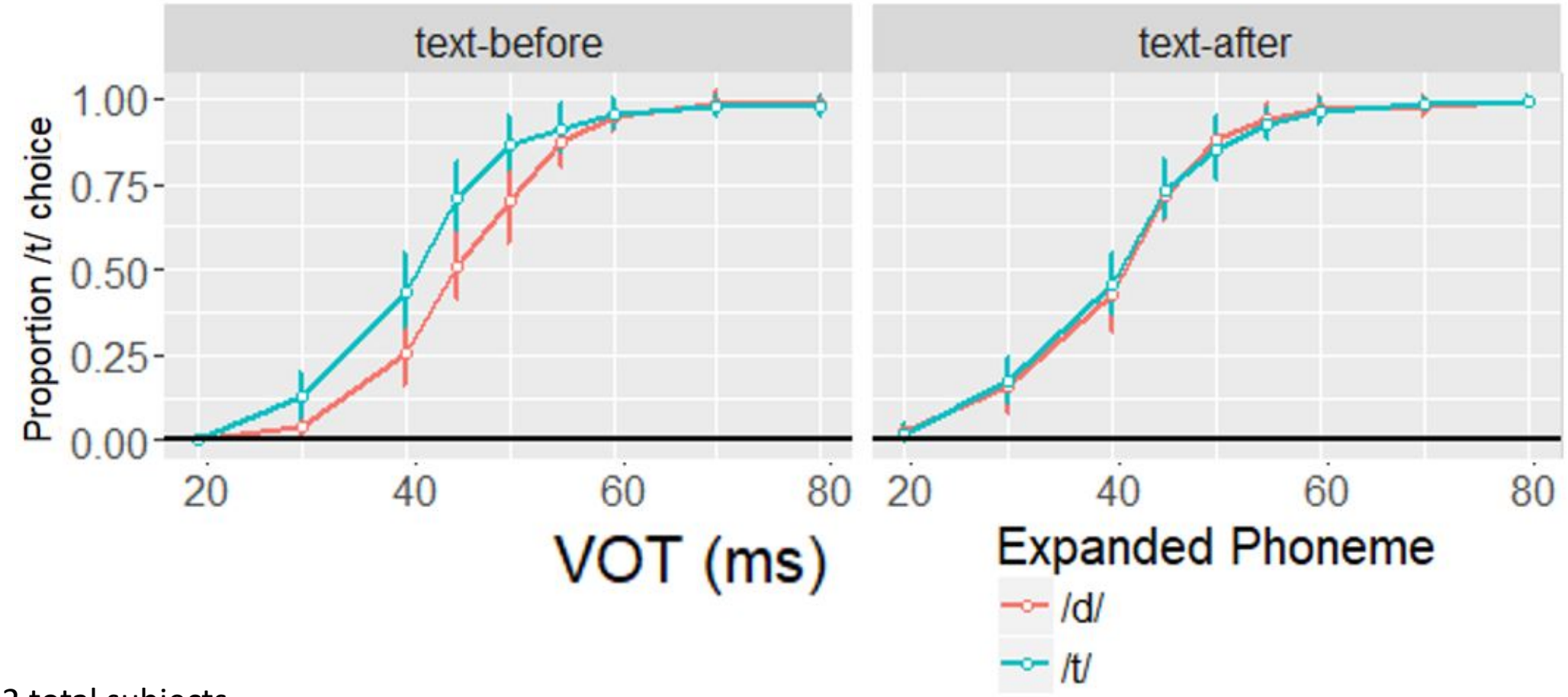
92 total subjects

No adaptation in Text-After condition



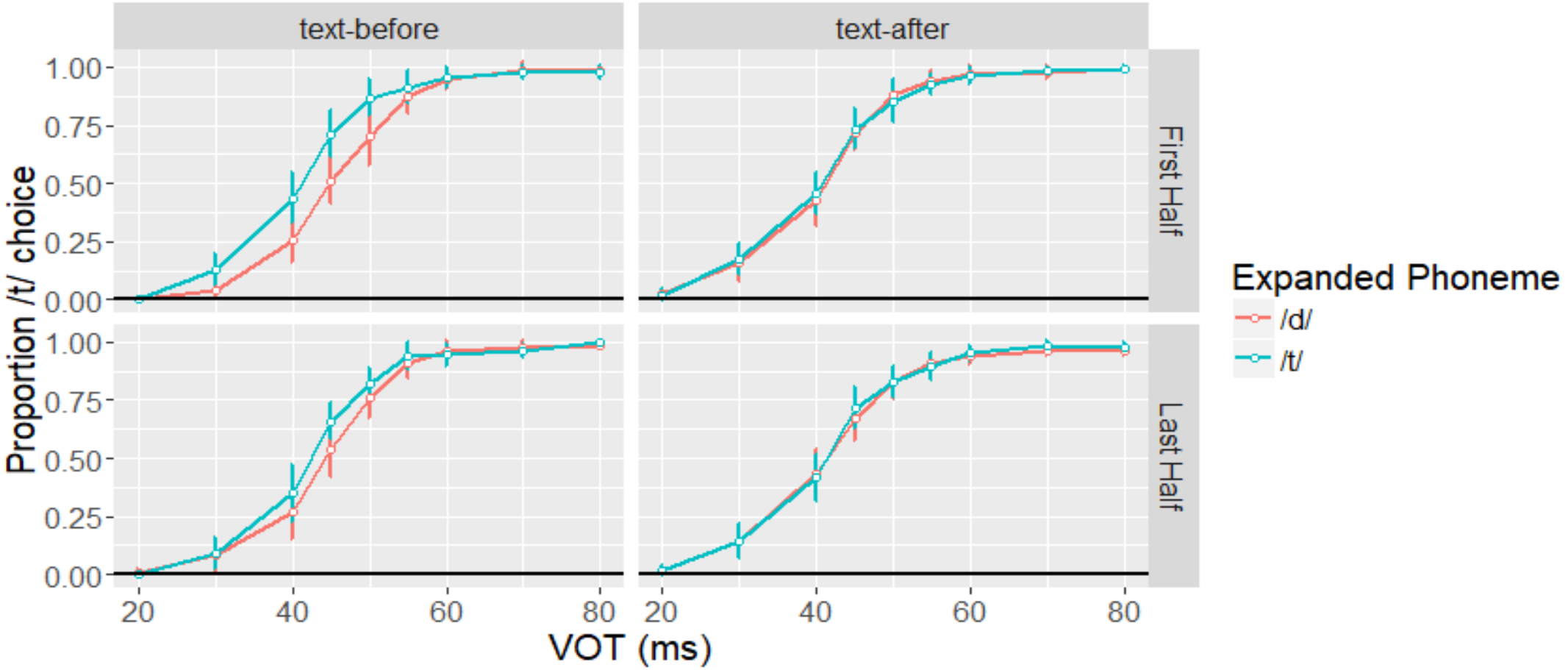
92 total subjects

Supports AOC rather than acoustic-maintenance



92 total subjects

Adaptation Fades Over Time



92 total subjects

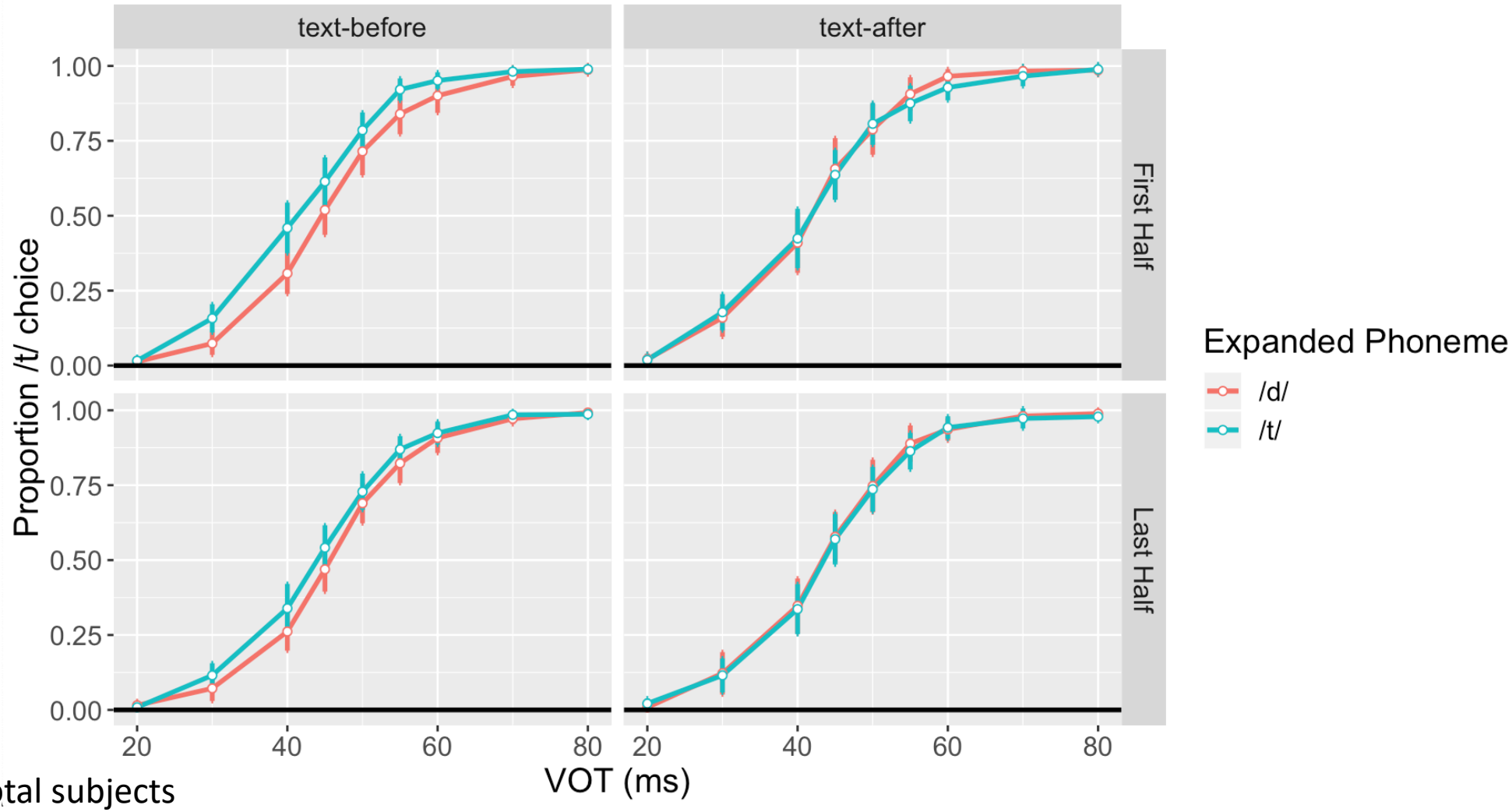
Experiment 2

- This pattern fully replicated in online (Mturk) follow-up
  - Additionally manipulated pitch-contour (f0) to remove a secondary cue to voicing

- 154 participants (after exclusion)

- Happy to talk offline if interested!

Experiment 2



154 total subjects

Interim Summary

- Adaptation present when phonological category is active *before* the audio
- No adaptation when phonological category not determined until after the audio

Intermediate representation during processing is "activation over categories" (AOC)

***Markovian* process: Encodes state of activation, but not the path that led to that belief**

Immediacy of Linguistic Computation: acoustic/phonetic buffer is flushed by categorization process

Outline

- Background
  - Intermediate Representations in Speech Processing (signal retention vs. AOC)
- Experiments 1 and 2
  - The Immediacy of Linguistic Computation
- **Experiment 3**
  - **Mapping Between Categories**
- Discussion

Mapping Between Categories

Markovian nature of speech processing under AOC

Acoustic-phonetic signal is transient → **Cannot learn signal-to-category mapping in text-after condition**

Category representations / activation are stable → **Category-to-category mappings could be learned even when cues are temporally disjoint**

Mapping Between Categories

Supra-phonemic categories

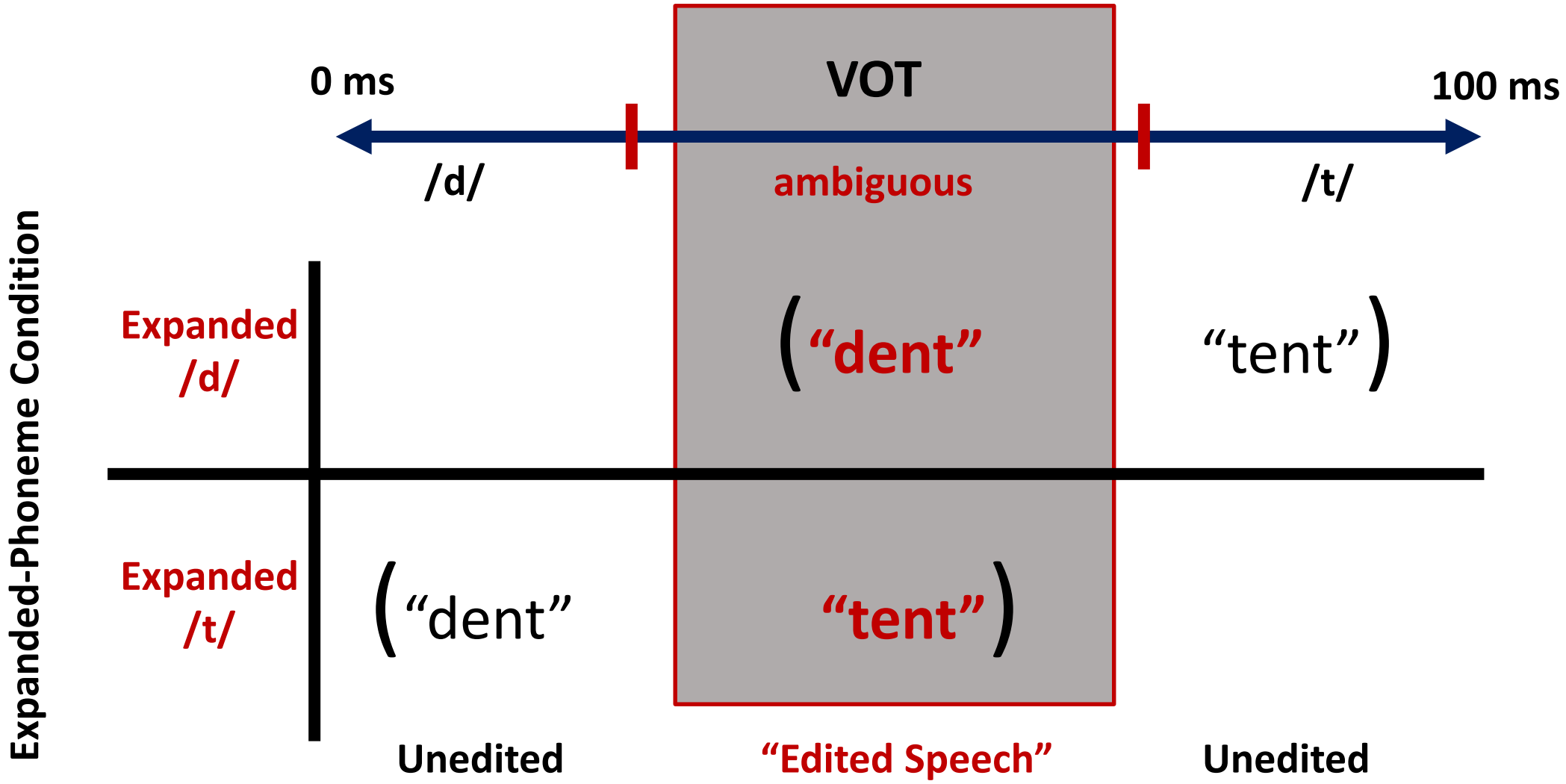
Not all speech categories are phonemic

- e.g. intonational contours, gender, speaker, etc.

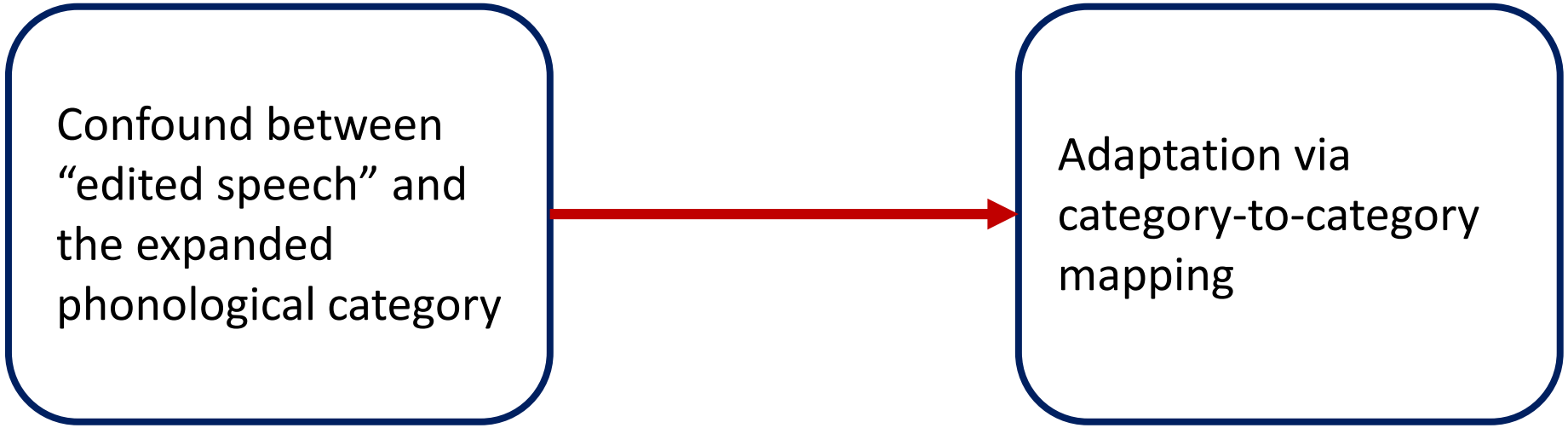
Supra-phonemic categories are dynamic / flexible

- Listeners can parse into a novel category such as “edited speech”
- In Experiments 1 and 2: “edited speech” was evenly balanced during training between the ambiguous and unambiguous targets

Target Text-Audio Pairing



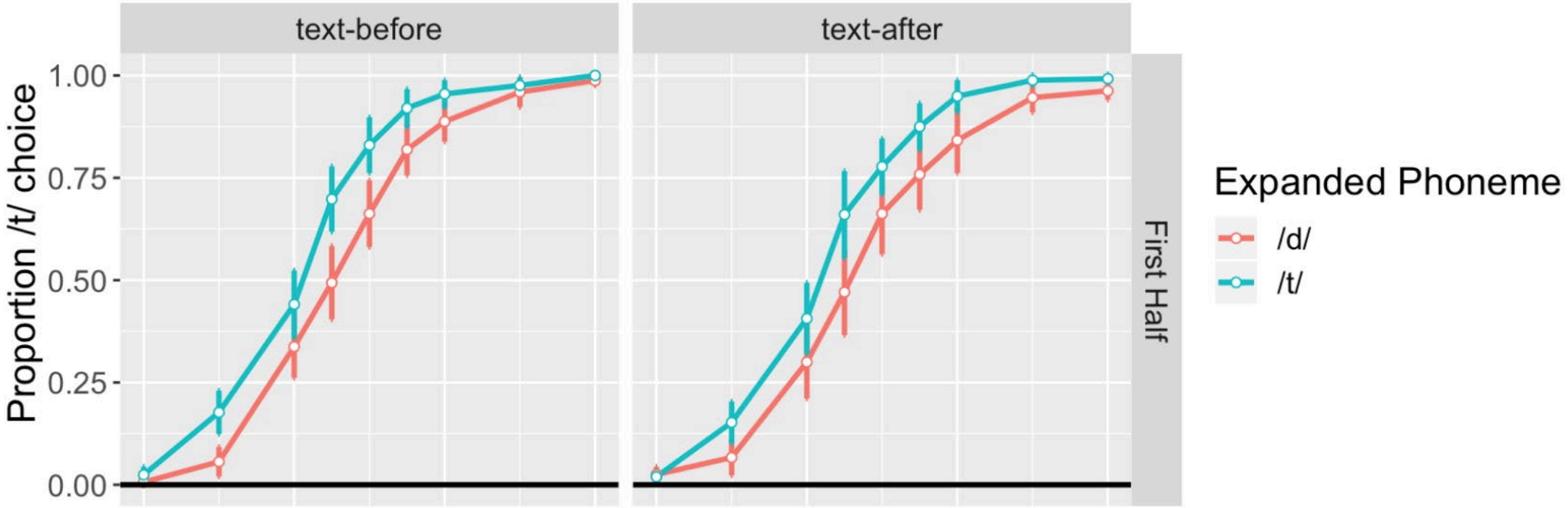
Mapping Between Categories



Experiment 3: non-ambiguous target items paired with unedited audio.

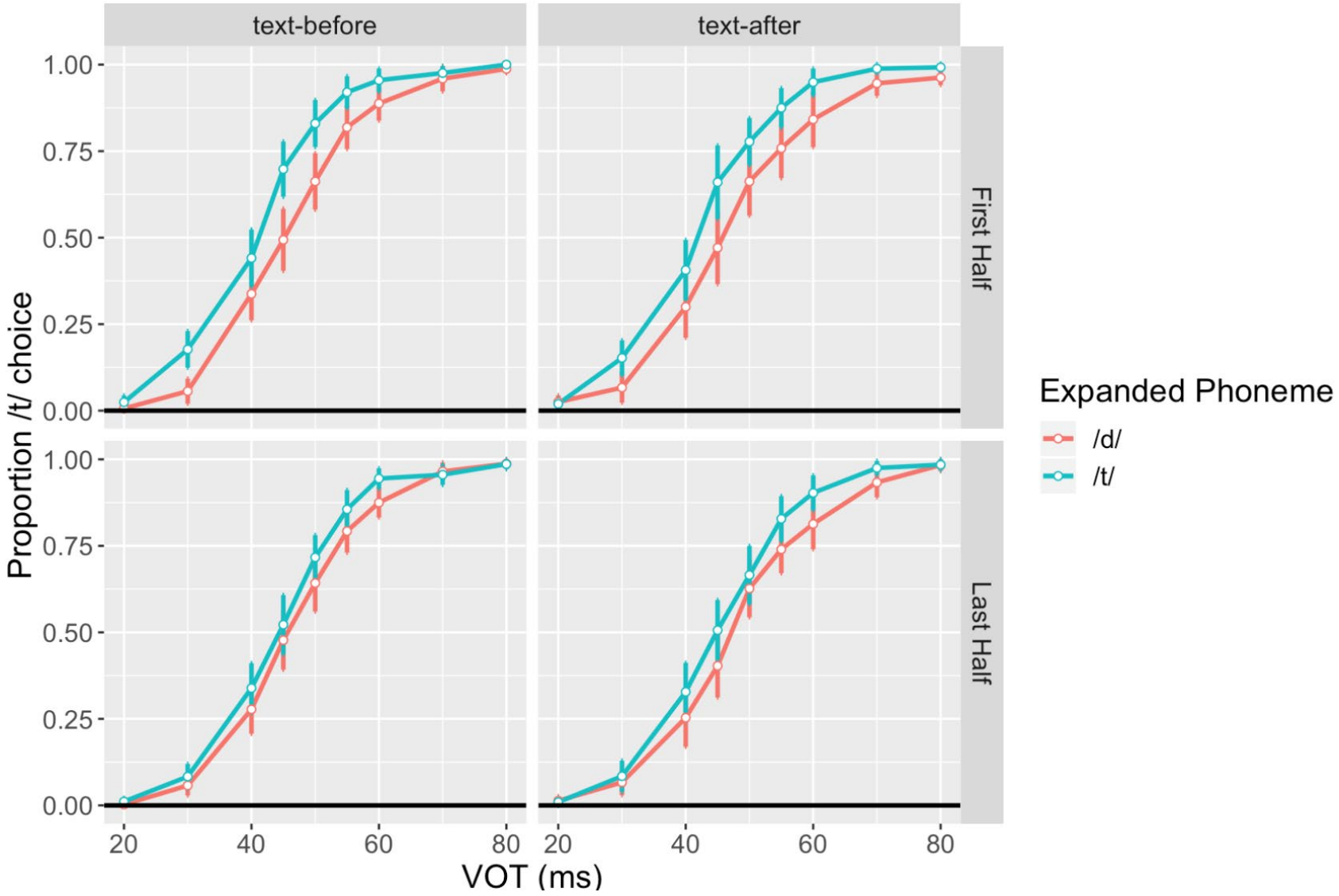
- AOC predicts adaptation in both text-before and text-after conditions

Experiment 3: Results



138 total subjects

Experiment 3: Results



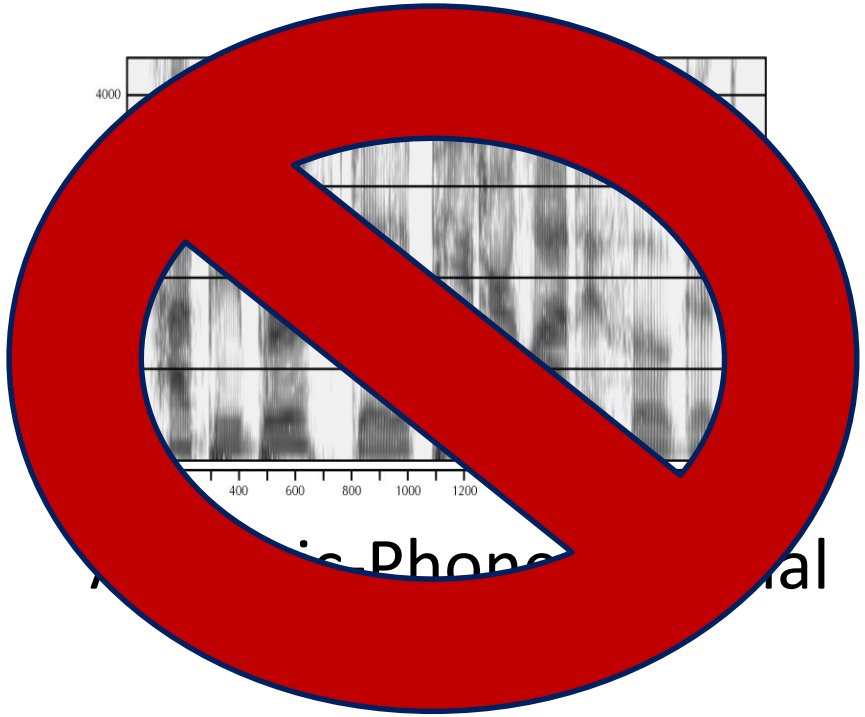
138 total subjects

Outline

- Background
  - Intermediate Representations in Speech Processing (signal retention vs. AOC)
- Experiments 1 and 2
  - The Immediacy of Linguistic Computation
- Experiment 3
  - Mapping Between Categories
- **Discussion**

Now you hear me, later you don't

What do *intermediate* representations contain?



or

/t/	80%
/d/	20%

Activation over categories
(phonemes, words, etc.)

Now you hear me, later you don't

Speech processing under AOC is a *Markovian* process

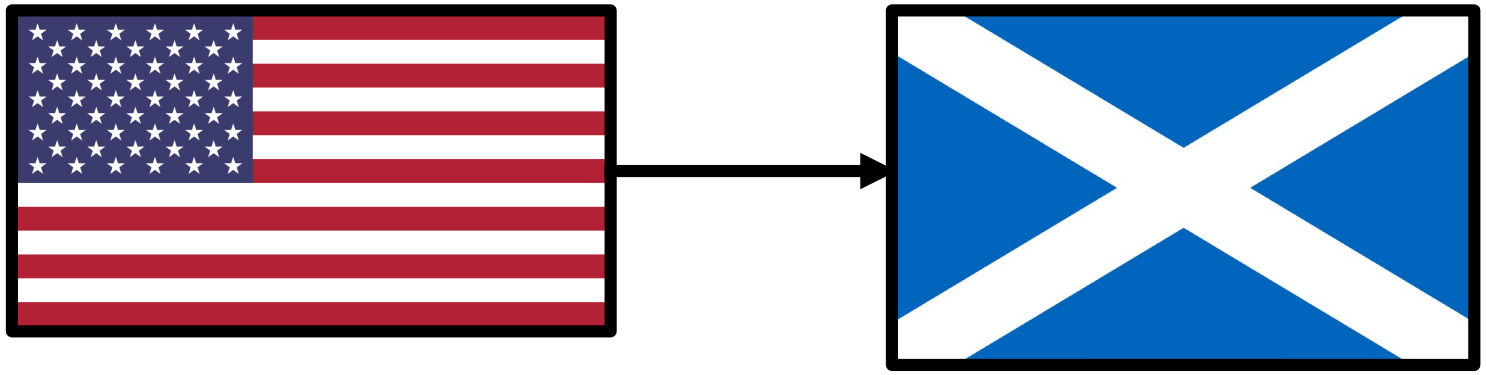
Encodes a state of activation, but not the path that led to that belief

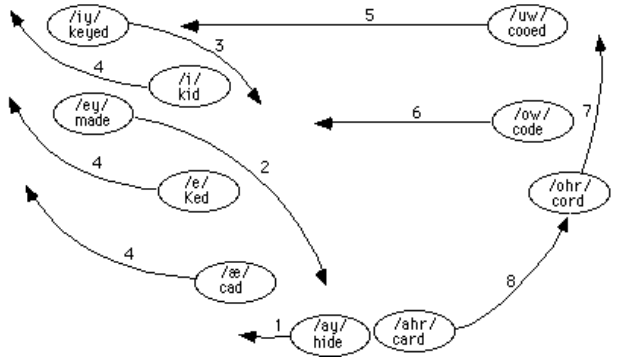
Immediacy of Linguistic Computation

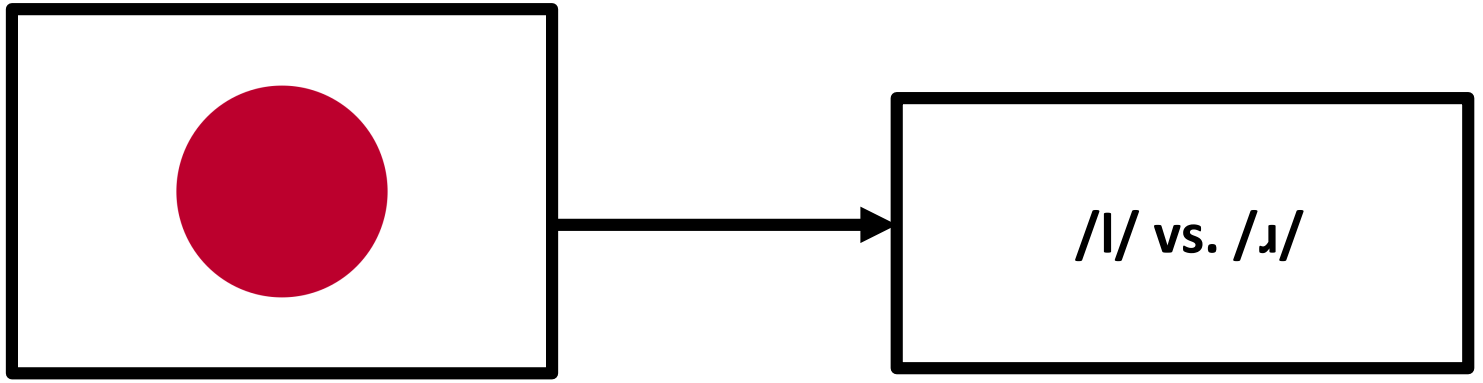
Parsing into discrete categories clears the buffer for the underlying signal

Now you hear me, later you don't

Category-to-category mapping matches real-world conditions







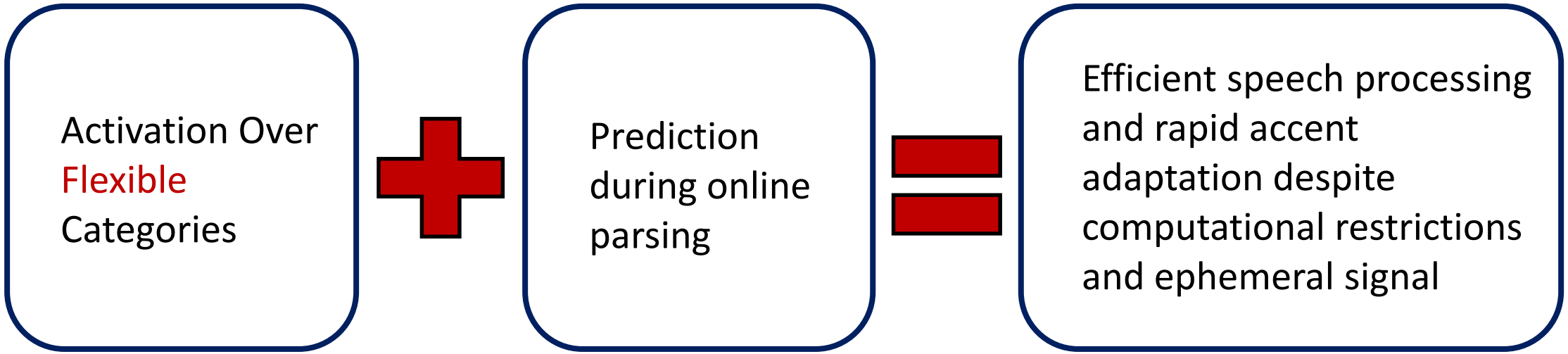
Now you hear me, later you don't

AOC is consistent with the outcomes of previous related studies

See: Connine (1991), Kraljic and Samuel (2006), Bushong & Jaeger (2017), Burchill et al. (2018)

Happy to talk more offline!

Now you hear me, later you don't



Acknowledgments

My collaborators Alon Hafri (JHU) and John Trueswell

Also thank you to Charles Yang, Mitch Marcus, Ryan Budnick, and the Penn Language Development & Language Processing Lab for helpful discussions and advice

Funding Acknowledgments

UPenn Integrated Language Science and Technology (ILST)

mindCORE ILST

Department of Computer and Information Science

Penn Engineering

Contact

Contact Spencer Caplan with questions or comments

spcaplan@sas.upenn.edu