Cognition of Synthetic Speech: An Eye Tracking Account

Rajakrishnan Rajkumar

Department of Humanities and Social Sciences,
Indian Institute of Science Education and Research (IISER), Bhopal

Reference

Michael White, **Rajakrishnan Rajkumar**, Kiwako Ito, and Shari R. Speer. *Eye tracking for the online evaluation of prosody in speech synthesis*.

In Amanda Stent and Srinivas Bangalore, editors, Natural Language Generation in Interactive Systems, Chapter 12, pages 281-301, 2013.

Research Question

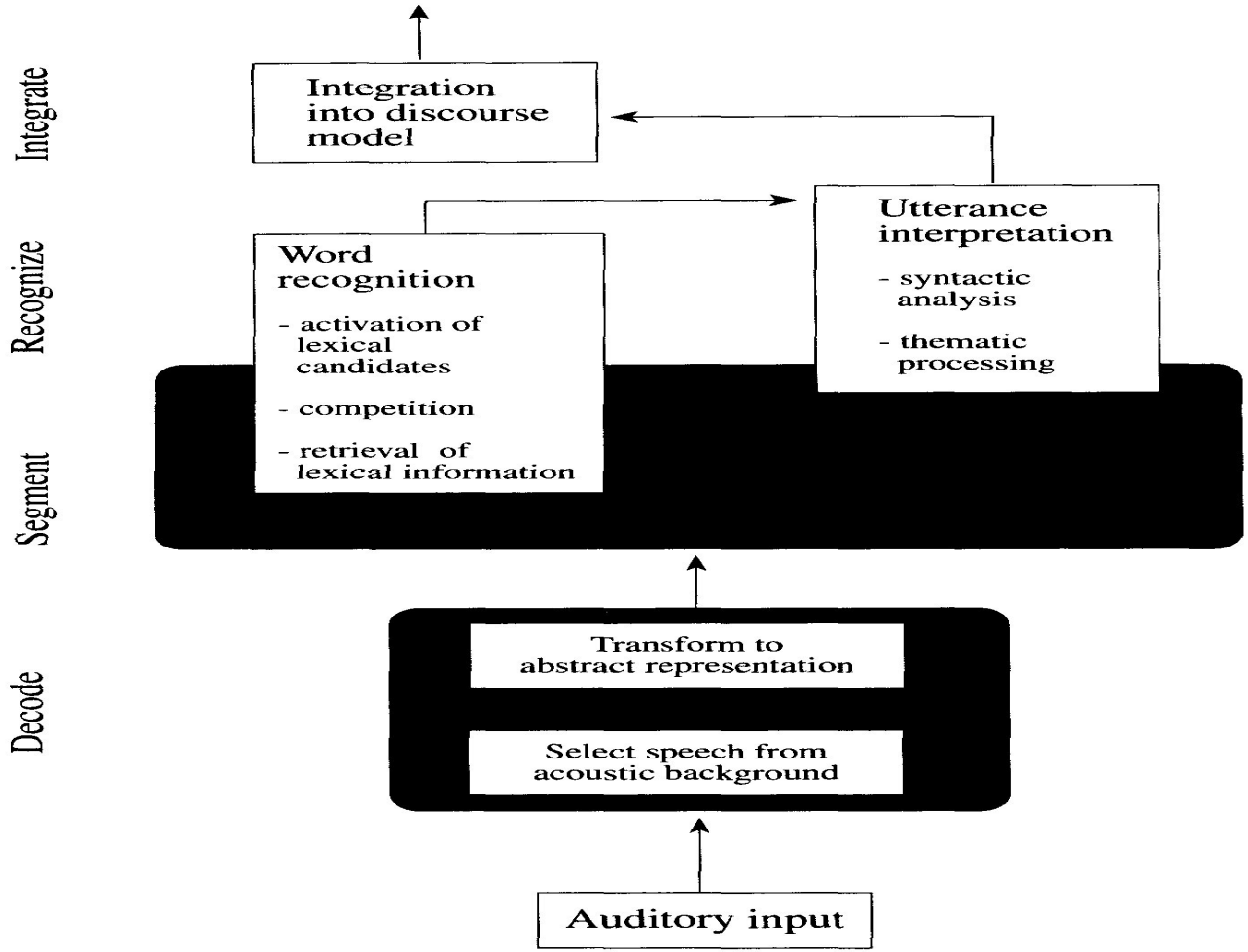
1. Is synthetic speech processed differently from natural speech?

Research Question

1. Is synthetic speech processed differently from natural speech?

   Yes, the results of our eye-tracking experiments suggest this

Spoken Language Comprehension (Cutler and Clifton, 1999)



Eye Movements and Spoken Language Comprehension

1. **Altmann and Kamide (1999)**: Evidence of incremental interpretation
2. **Ito and Speer (2008, 2009)**: Prosody in incremental interpretation

Eye Movements and Spoken Language Comprehension

Participants facing a picture display heard instructions to perform a task

1. **Altmann and Kamide 1999**: *the boy will* **eat** *the cake/candle*

Eye Movements and Spoken Language Comprehension

Participants facing a picture display heard instructions to perform a task

1. **Altmann and Kamide 1999**: *the boy will* **eat** *the cake/candle*
   Anticipatory eye movements to picture of *cake* (NOT *candle*) at onset of *eat*

Eye Movements and Spoken Language Comprehension

Participants facing a picture display heard instructions to perform a task

1. **Altmann and Kamide 1999**: *the boy will* **eat** *the cake/candle*
   Anticipatory eye movements to picture of *cake* (NOT *candle*) at onset of *eat*
2. **Dahan et al. 2002**: Prominent accent vs. lack of accent

   "Click on the candle. Now, click on the CAN/can ..."
   CAN ==> looks to candy
   can ==> looks to candle

Eye Movements and Spoken Language Comprehension

Participants facing a picture display heard instructions to perform a task

1. **Altmann and Kamide 1999**: *the boy will* **eat** *the cake/candle*
   Anticipatory eye movements to picture of *cake* (NOT *candle*) at onset of *eat*
2. **Dahan et al. 2002**: Prominent accent vs. lack of accent

   "Click on the candle. Now, click on the CAN/can ..."
   CAN ==> looks to candy
   can ==> looks to candle
3. **Ito and Speer (2008, 2009)**: To be discussed shortly

Altmann and Kamide (1999)

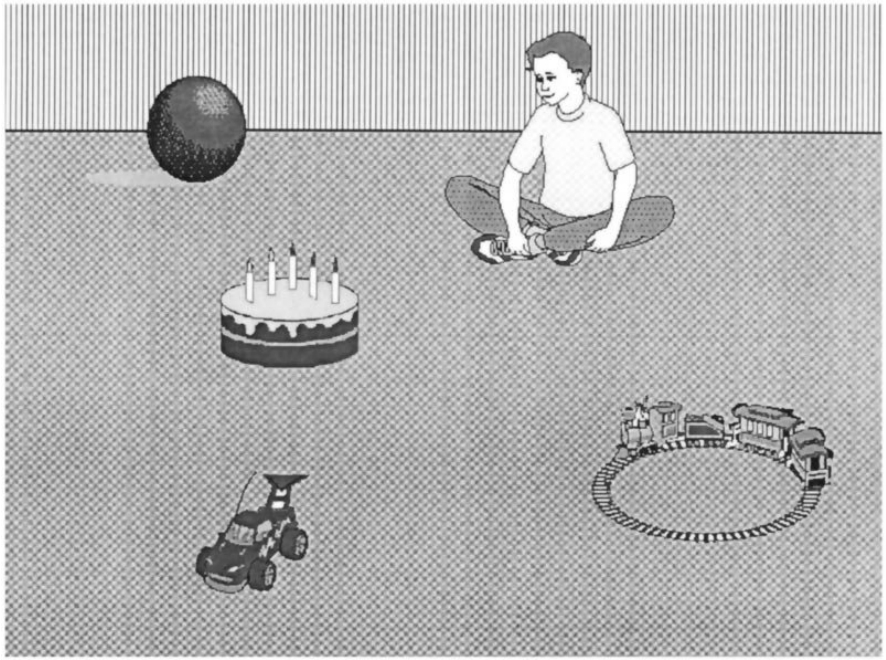


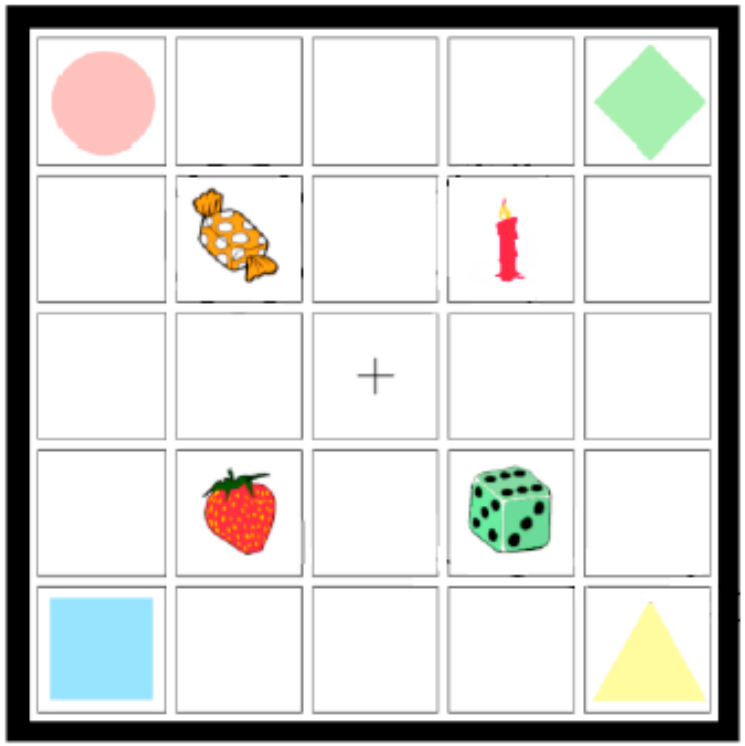
Figure: *the boy will* **eat** *the cake/candle*

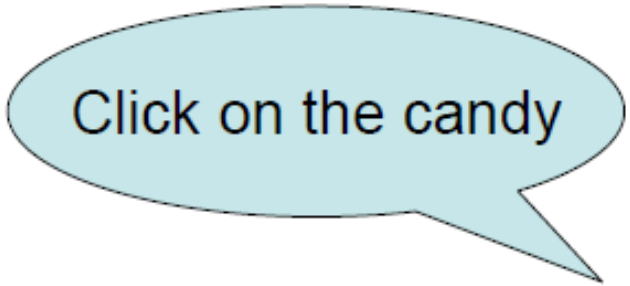
Eye Movements and Spoken Language Comprehension

Participants facing a picture display heard instructions to perform a task (Tanenhaus et al. 95)

Eye Movements and Spoken Language Comprehension

Participants facing a picture display heard instructions to perform a task (Tanenhaus et al. 95)

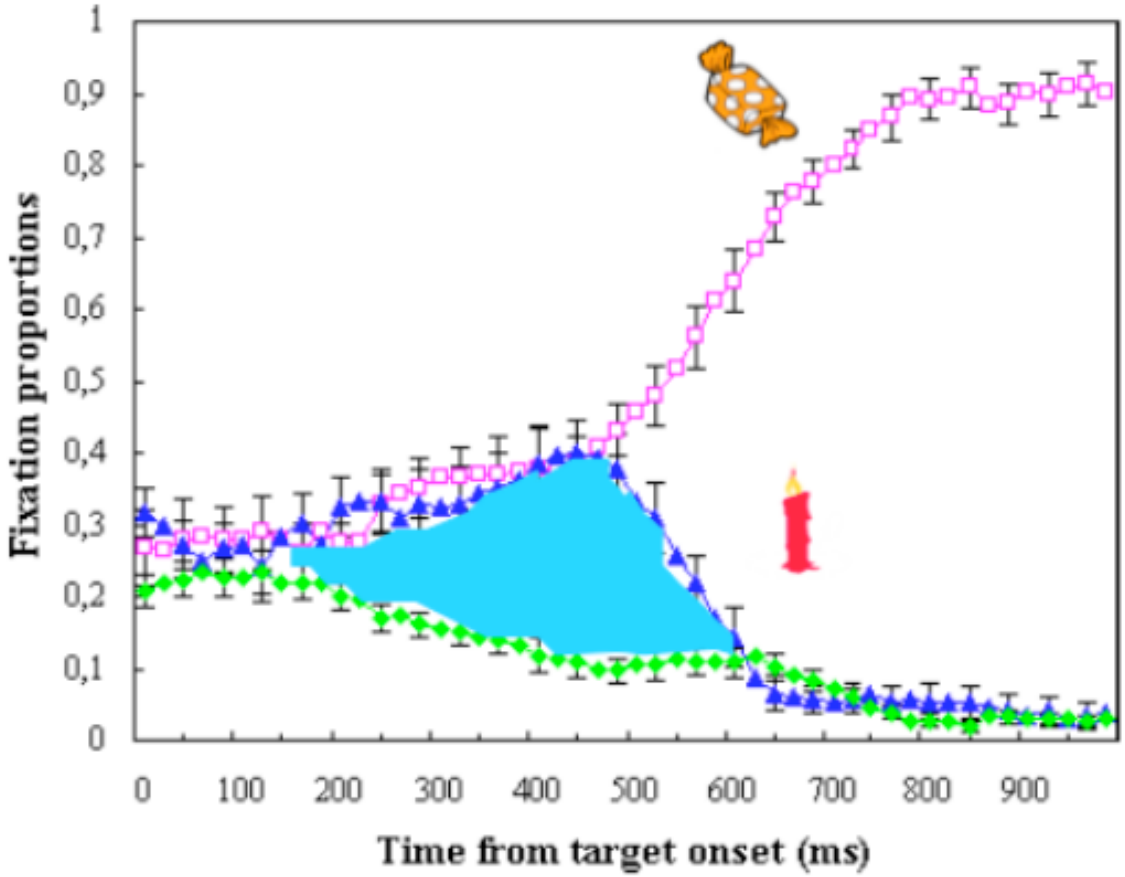




Target object: candy

Cohort (competitor): candle

Distractors: strawberry, dice


Fixation proportions
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
100
200
300
400
500
600
700
800
900
Time from target onset (ms)

Types of Eye Movements (Feng, 2010)

1. *Saccades*: Rapid, ballistic movements of our gaze (primary means to acquire new visual information)

Types of Eye Movements (Feng, 2010)

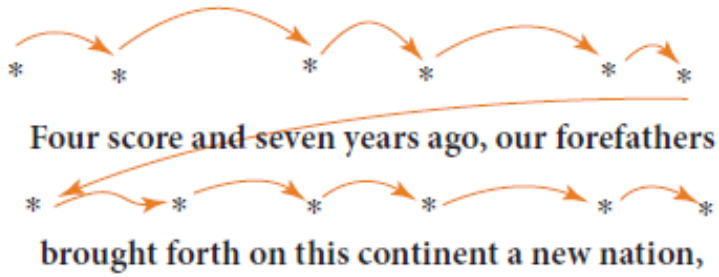
1. *Saccades*: Rapid, ballistic movements of our gaze (primary means to acquire new visual information)
2. *Fixations*: Between saccades, eyes stay relatively still to allow for visual perception

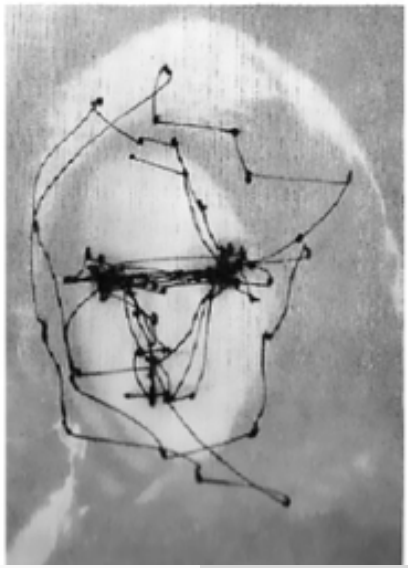
Types of Eye Movements (Feng, 2010)

1. *Saccades*: Rapid, ballistic movements of our gaze (primary means to acquire new visual information)
2. *Fixations*: Between saccades, eyes stay relatively still to allow for visual perception

(Insights go back to 1879: *Dr. Louis Emile Javal*)

Saccades and Fixations





Our Hypothesis

Even high quality synthetic speech results in processing delays

Our Hypothesis

Even high quality synthetic speech results in processing delays

Confirmed using:

Our Hypothesis

Even high quality synthetic speech results in processing delays

Confirmed using:

1. Eye tracking experiment (White, Rajkumar, Ito and Speer 2009)

Our Hypothesis

Even high quality synthetic speech results in processing delays

Confirmed using:

1. Eye tracking experiment (White, Rajkumar, Ito and Speer 2009)
2. An acoustic analysis of the eye tracking stimuli (Rajkumar, White, Speer and Ito 2010)

Our Hypothesis

Even high quality synthetic speech results in processing delays

Confirmed using:

1. Eye tracking experiment (White, Rajkumar, Ito and Speer 2009)
2. An acoustic analysis of the eye tracking stimuli (Rajkumar, White, Speer and Ito 2010)
3. Offline speech rating task (Rajkumar, White, Speer and Ito 2010)

Related Work: Speech Synthesis Evaluation and Eye-tracking

Swift et al. (2002)

- Real-world object manipulation paradigm
- Found that segmental information in synthetic speech processed incrementally at both lexical and discourse levels
- Processing delayed in comparison to human speech

Related Work: Speech Synthesis Evaluation and Eye-tracking

Swift et al. (2002)

- Real-world object manipulation paradigm
- Found that segmental information in synthetic speech processed incrementally at both lexical and discourse levels
- Processing delayed in comparison to human speech

van Hooijdonk et al. (2006)

- Additionally looked at supersegmental information in two different discourse contexts, comparing a diphone voice and a unit selection voice to human speech

Related Work: Speech Synthesis Evaluation and Eye-tracking

Swift et al. (2002)

- Real-world object manipulation paradigm
- Found that segmental information in synthetic speech processed incrementally at both lexical and discourse levels
- Processing delayed in comparison to human speech

van Hooijdonk et al. (2006)

- Additionally looked at supersegmental information in two different discourse contexts, comparing a diphone voice and a unit selection voice to human speech
- Found more anticipatory looks to the competitor referent with the diphone voice
- Also found processing delays with synthetic speech

Departure from Previous Work

- Swift et al. (2002) do not examine suprasegmental phenomena
- van Hooijdonk et al. (2006) did not investigate the effect of different accent patterns

Departure from Previous Work

- Swift et al. (2002) do not examine suprasegmental phenomena
- van Hooijdonk et al. (2006) did not investigate the effect of different accent patterns
- They do not provide any acoustic analysis

Departure from Previous Work

- Swift et al. (2002) do not examine suprasegmental phenomena
- van Hooijdonk et al. (2006) did not investigate the effect of different accent patterns
- They do not provide any acoustic analysis
- First attempt to replicate psycholinguistic results with different accent patterns

Departure from Previous Work

- Swift et al. (2002) do not examine suprasegmental phenomena
- van Hooijdonk et al. (2006) did not investigate the effect of different accent patterns
- They do not provide any acoustic analysis
- First attempt to replicate psycholinguistic results with different accent patterns
- We provide a detailed acoustic analysis connecting various properties of speech to online processing effects

Departure from Previous Work

- Swift et al. (2002) do not examine suprasegmental phenomena
- van Hooijdonk et al. (2006) did not investigate the effect of different accent patterns
- They do not provide any acoustic analysis
- First attempt to replicate psycholinguistic results with different accent patterns
- We provide a detailed acoustic analysis connecting various properties of speech to online processing effects

Contrastive Accent

- Pierrehumbert & Hirschberg (1990) on function of L+H*: *the accented item — and not some alternative related item- should be mutually believed*

Contrastive Accent

- Pierrehumbert & Hirschberg (1990) on function of L+H*: *the accented item — and not some alternative related item- should be mutually believed*

  "I made a reservation for FIFTEEN, not fifty!"

Contrastive Accent

- Pierrehumbert & Hirschberg (1990) on function of L+H*: *the accented item — and not some alternative related item- should be mutually believed*

  "I made a reservation for FIFTEEN, not fifty!"

- Eye-tracking and L+H*: Ito & Speer (2008) report prosodic facilitation and garden-path effects associated with the L+H* tone

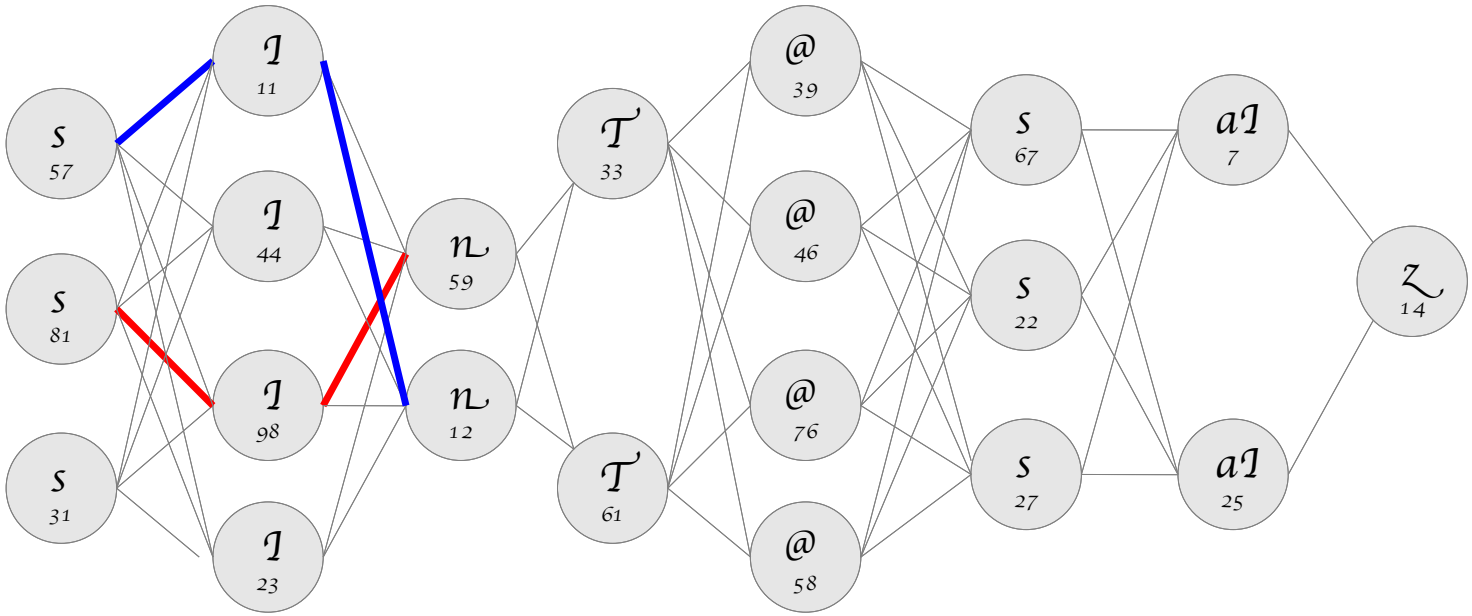
Unit Selection Synthesis

Basic Idea

- record utterances with natural prosody
- record many samples of each sound (unit)
- at runtime, select sequence of units that minimize the target and join costs
  - target cost: linguistic context match
  - join cost: acoustic fit
- concatenate units with little or no signal processing

Viterbi Search

"synthesize" = [s I n T @ s aI z]

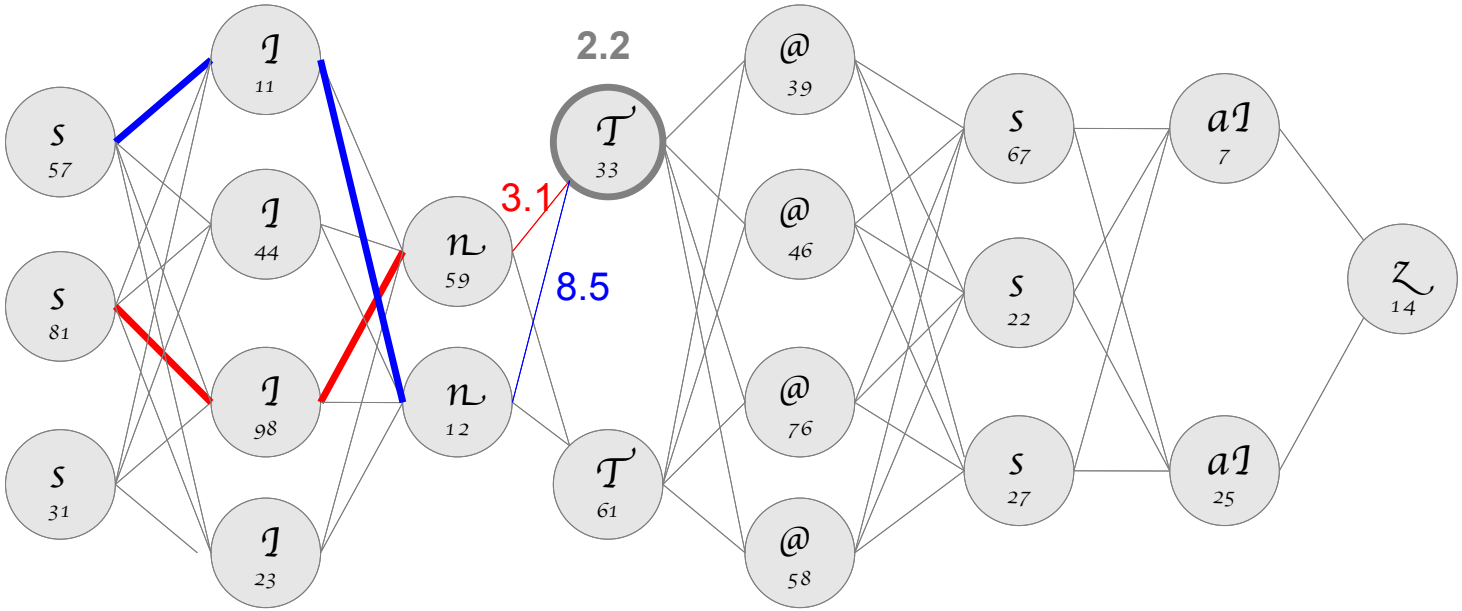


Partial path costs:
53.1
17.3

(Schröder, 2008)

Viterbi Search (2)

"synthesize" = [s I n T @ s aI z]



Partial path costs:

~~58.1 + 3.1 + 2.2 = 58.1~~

17.3 + 8.5 + 2.2 = 28.0

(Schröder, 2008)

Experiment I: Eye-tracking Experiment

We investigate whether different accent patterns in synthetic speech yield significant differences in anticipatory eye movements

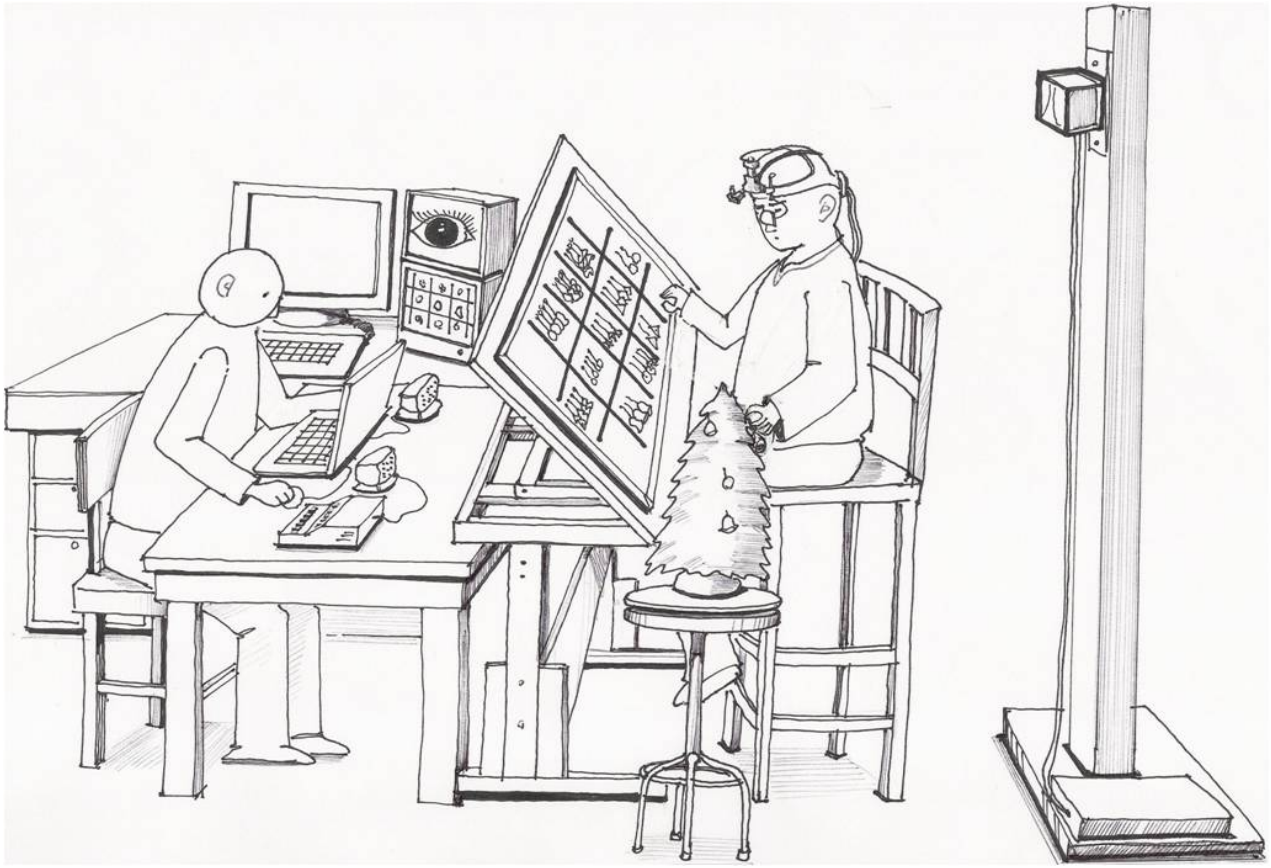
Experiment I: Eye-tracking Experiment

We investigate whether different accent patterns in synthetic speech yield significant differences in anticipatory eye movements

- Replicate Ito & Speer's (2008, 2009) experiment using synthetic speech instead of human speech
- Task: decorate holiday trees with ornaments laid out on a grid

Setup

- ASL Eye-Trac 6000
- Sampling rate: 60Hz



Contrasts

Local Instruction Sequences

- **contrastive** *Hang a red star. Next, hang a yellow star.*
- **non-contrastive** *Hang a yellow tree. Next, hang a green ball.*

Contrasts

Local Instruction Sequences

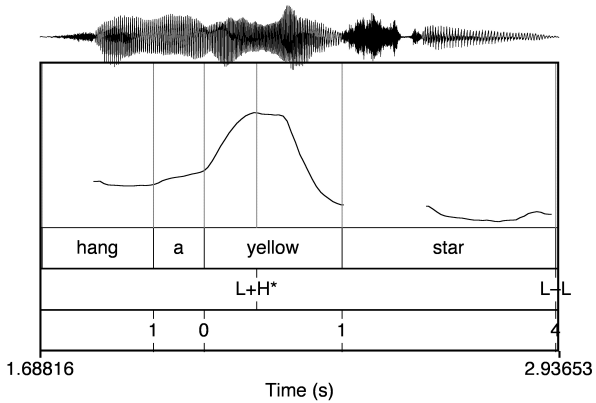
- **contrastive** *Hang a red star. Next, hang a yellow star.*
- **non-contrastive** *Hang a yellow tree. Next, hang a green ball.*

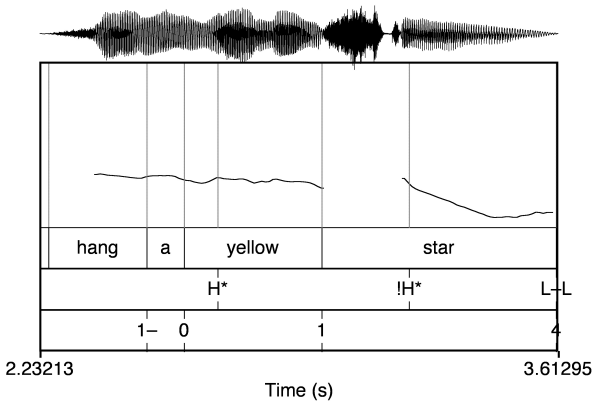
Accent Patterns

- **contrastive** *Hang a YELLOW$_{L+H*}$ star$_{\emptyset}$*
- **non-contrastive** *Hang a yellow$_{H*}$ star$_{!H*}$*

Stimuli

F0 traces and ToBI annotations





Natural	Play	Play
Synthetic	Play	Play

Ito & Speer (2008, 2009) Findings

- Facilitative effect of felicitous use of contrastive accent pattern: more and faster looks to the target

Ito & Speer (2008, 2009) Findings

- Facilitative effect of felicitous use of contrastive accent pattern: more and faster looks to the target

  Hang a red **star**. *Next, hang a YELLOW*$_{L+H*}$ *star*$_{\emptyset}$

Ito & Speer (2008, 2009) Findings

- Facilitative effect of felicitous use of contrastive accent pattern: more and faster looks to the target

  Hang a red **star**. *Next, hang a YELLOW*$_{L+H*}$ *star*$_{\emptyset}$

- 'Garden path' effect of infelicitous use of contrastive pattern: more looks to the competitor, delayed looks to the target

Ito & Speer (2008, 2009) Findings

- Facilitative effect of felicitous use of contrastive accent pattern: more and faster looks to the target

  Hang a red **star**. *Next, hang a YELLOW*$_{L+H*}$ *star*$_{\emptyset}$

- 'Garden path' effect of infelicitous use of contrastive pattern: more looks to the competitor, delayed looks to the target

  Hang a red drum. Next, hang a YELLOW$_{L+H*}$ **star**$_{\emptyset}$

Ito & Speer (2008, 2009) Findings

- Facilitative effect of felicitous use of contrastive accent pattern: more and faster looks to the target

  Hang a red **star**. *Next, hang a YELLOW*$_{L+H*}$ *star*$_{\emptyset}$

- 'Garden path' effect of infelicitous use of contrastive pattern: more looks to the competitor, delayed looks to the target

  Hang a red drum. Next, hang a YELLOW$_{L+H*}$ **star**$_{\emptyset}$

Processing Account

- Suggests immediate, parallel processing of segmental and suprasegmental information (Snedecker and Trueswell, 2003, Dahan et al., 2002)

Processing Account

- Suggests immediate, parallel processing of segmental and suprasegmental information (Snedecker and Trueswell, 2003, Dahan et al., 2002)
- Pitch accent invokes a set of possible referents from the discourse context

Processing Account

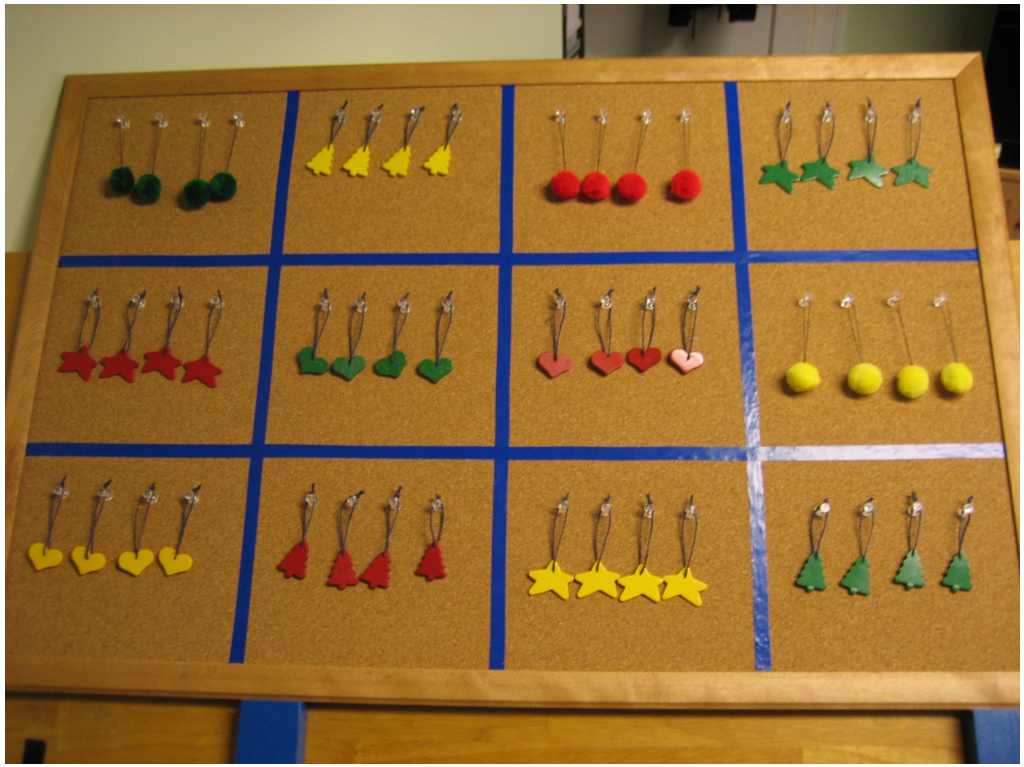
- Suggests immediate, parallel processing of segmental and suprasegmental information (Snedecker and Trueswell, 2003, Dahan et al., 2002)
- Pitch accent invokes a set of possible referents from the discourse context
- Finally eyes fixate on one possible referent

Example Interaction

[Video]

Ornament Board

- Three trees, three grids
- Four types of ornaments (3 targets, 1 filler)
- Three colors



Participants and Procedure

- Data from 29 native speakers of American English was analyzed
- Participants wore lightweight headgear fitted with eye tracking equipment
- Experimenter monitored participants' eye locations and body orientations, and pressed a key to cue each instruction

Results: Traditional ANOVA Analysis

- Each participant had 9 trials in each of the 4 critical conditions
- Dependent variables: mean proportion of fixations to the target and competitor
- Repeated measures ANOVA for subjects and items in 100ms windows

Facilitation

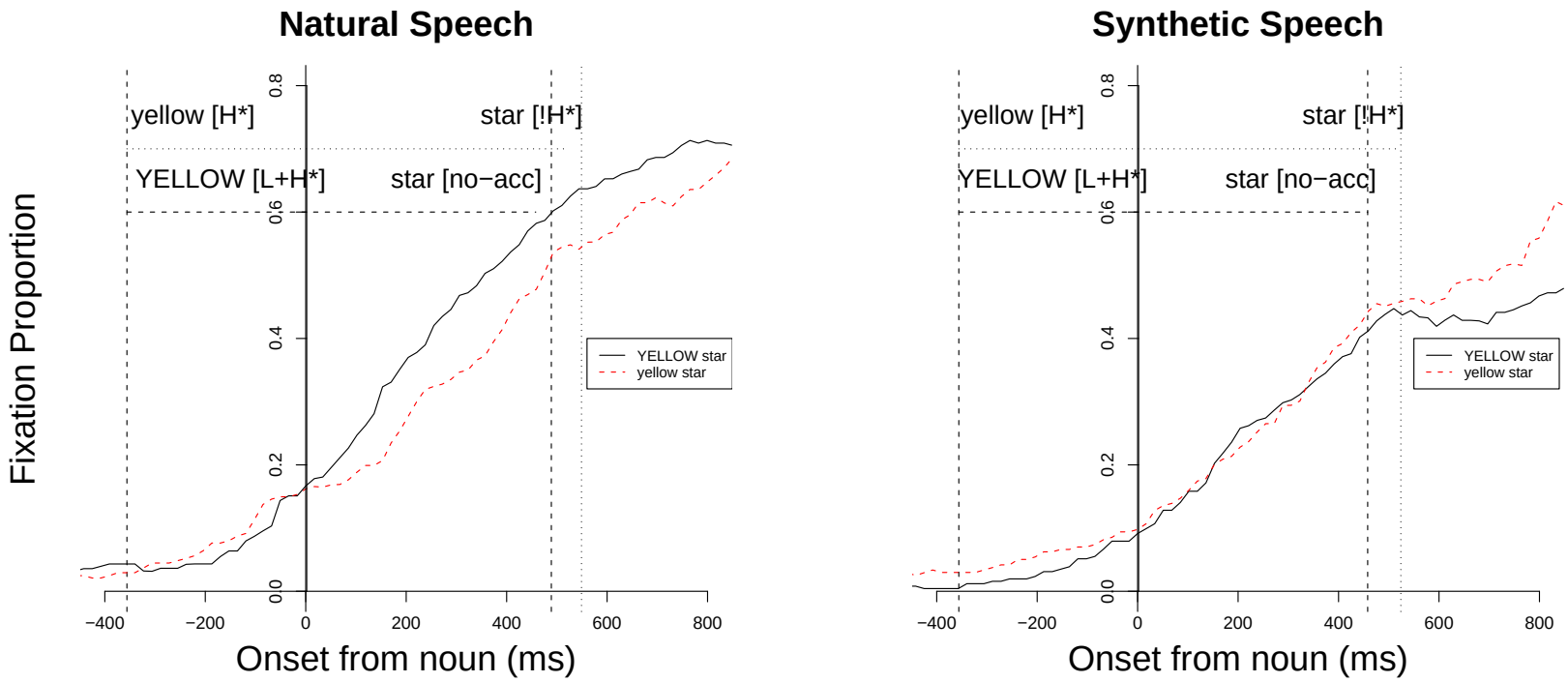


Figure: Fixation proportions to the target in two contrastive sequences, e.g. *red star* → *YELLOW/yellow star*

Facilitation

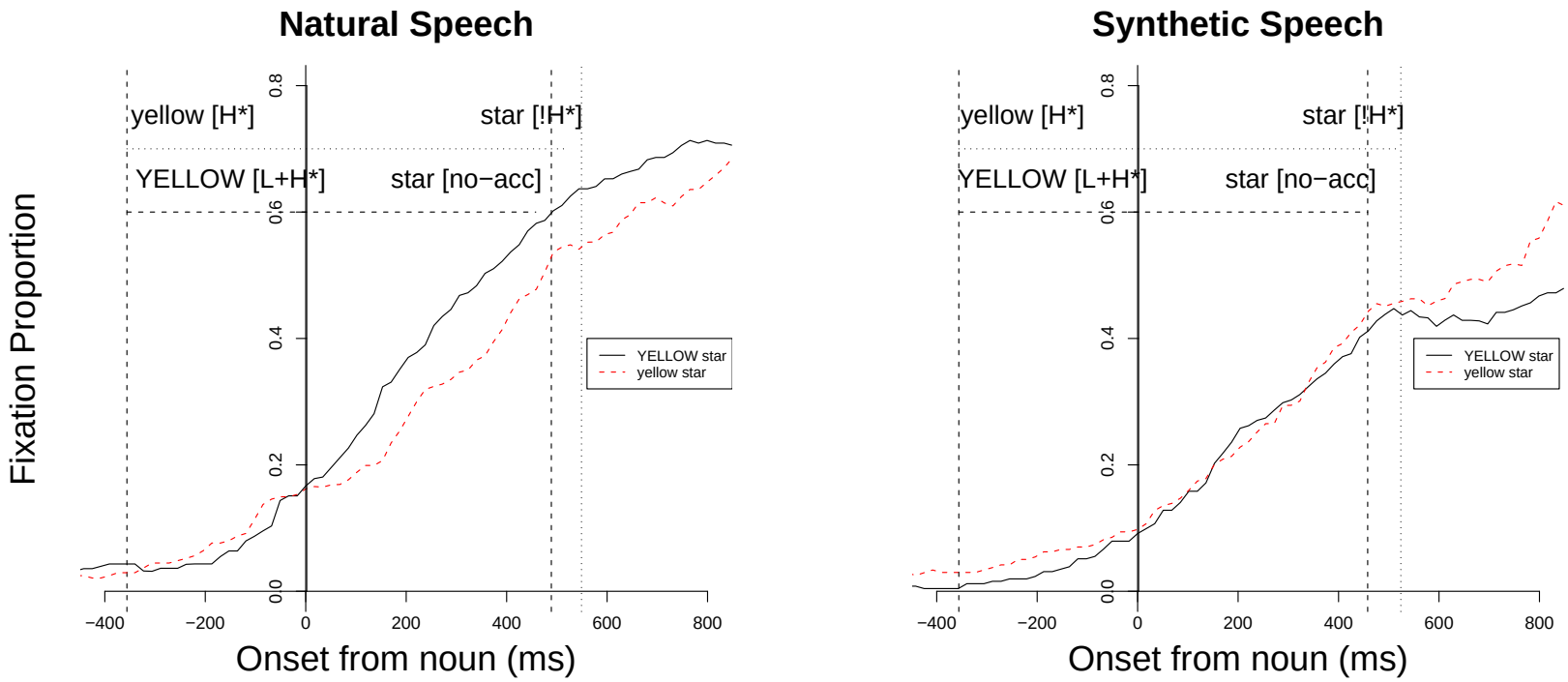


Figure: Fixation proportions to the target in two contrastive sequences, e.g. *red star* → *YELLOW/yellow star*

Processing Delays

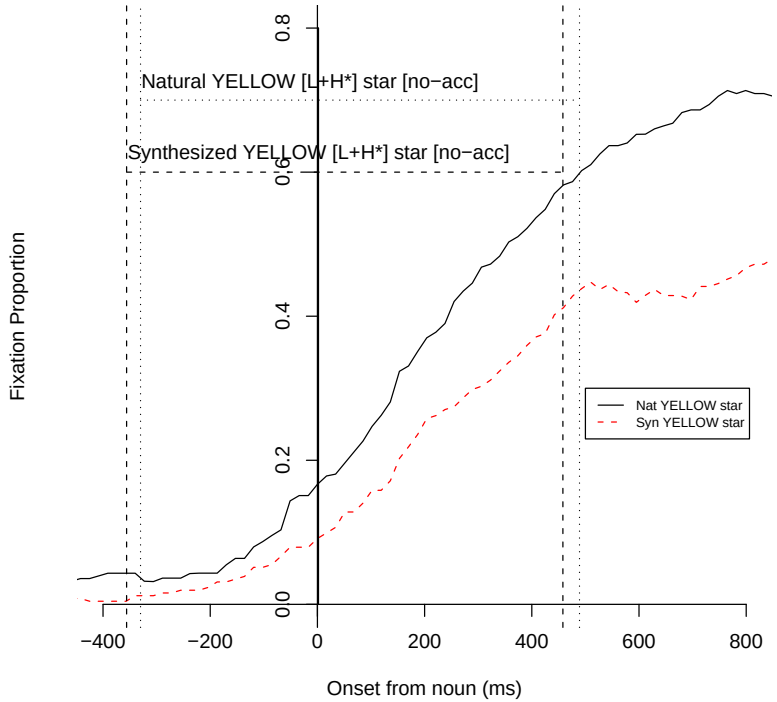


Figure: Fixation proportions to the target due to contrastive accent in contrastive sequences with natural and synthetic speech

Garden Pathing

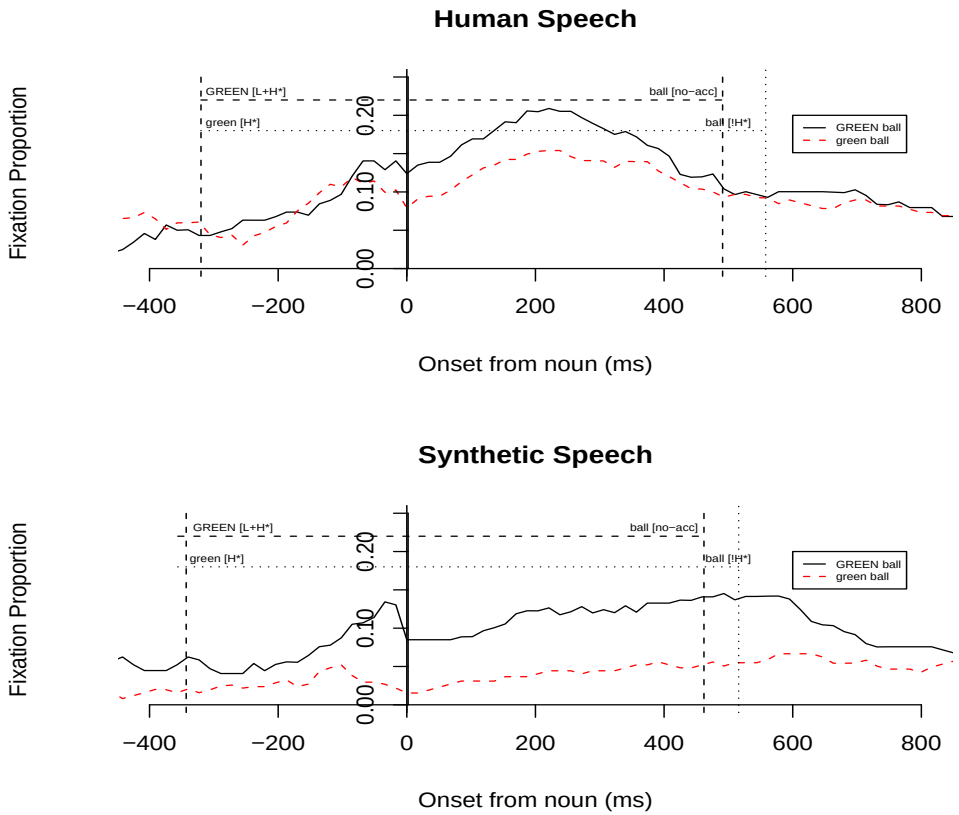


Figure: Fixation proportions to the contrastive competitor in non-contrastive sequences with natural and synthetic speech

Discussion

- Eye tracking able to clearly distinguish human and synthetic speech, despite high quality synthetic stimuli

Discussion

- Eye tracking able to clearly distinguish human and synthetic speech, despite high quality synthetic stimuli
- An offline rating task suggests that the stimuli were all of excellent quality, with only quite subtle artefacts

Discussion

- Eye tracking able to clearly distinguish human and synthetic speech, despite high quality synthetic stimuli
- An offline rating task suggests that the stimuli were all of excellent quality, with only quite subtle artefacts
- Trained prosody annotator had no trouble identifying the intended tune

Possible Explanation

- No facilitation, stronger garden pathing — better to not risk getting the tune wrong with expressive prosody?

Possible Explanation

- No facilitation, stronger garden pathing — better to not risk getting the tune wrong with expressive prosody?
- No, processing delays could explain lack of facilitation

Possible Explanation

- No facilitation, stronger garden pathing — better to not risk getting the tune wrong with expressive prosody?
- No, processing delays could explain lack of facilitation
- a delay in interpreting the segmental information in the adjective means that the disambiguating information in the noun in a sense arrives too soon

Possible Explanation

- No facilitation, stronger garden pathing — better to not risk getting the tune wrong with expressive prosody?
- No, processing delays could explain lack of facilitation
- a delay in interpreting the segmental information in the adjective means that the disambiguating information in the noun in a sense arrives too soon
- Do longer adjectives help?

Possible Explanation

- No facilitation, stronger garden pathing — better to not risk getting the tune wrong with expressive prosody?
- No, processing delays could explain lack of facilitation
- a delay in interpreting the segmental information in the adjective means that the disambiguating information in the noun in a sense arrives too soon
- Do longer adjectives help?

Stronger Garden Path Effect

- With processing delays, listeners might be updating their referential domains when the conflicting information from the noun arrives, causing additional delays

Stronger Garden Path Effect

- With processing delays, listeners might be updating their referential domains when the conflicting information from the noun arrives, causing additional delays
- Or, with somewhat less intelligible segmental information, listeners may be relying more heavily on prosody

Findings

1. Synthetic speech tunes do not facilitate looks to the target (**unlike** natural speech)

Findings

1. Synthetic speech tunes do not facilitate looks to the target (**unlike** natural speech)
2. Synthetic speech tunes contribute to robust garden-pathing effects (**akin** to natural speech)

Findings

1. Synthetic speech tunes do not facilitate looks to the target (**unlike** natural speech)
2. Synthetic speech tunes contribute to robust garden-pathing effects (**akin** to natural speech)
3. Processing delays contribute to above effects

Findings

1. Synthetic speech tunes do not facilitate looks to the target (**unlike** natural speech)
2. Synthetic speech tunes contribute to robust garden-pathing effects (**akin** to natural speech)
3. Processing delays contribute to above effects

Conclusion

1. Eye-tracking can help us gain insights into the processing of synthetic speech

Conclusion

1. Eye-tracking can help us gain insights into the processing of synthetic speech
2. Can potentially lead to better speech synthesis evaluation

Future Work

- Future experiments involving longer or multiple adjectives, or a more complicated visual search task

Future Work

- Future experiments involving longer or multiple adjectives, or a more complicated visual search task
- Design an eye-tracking experiment where items are controlled for the acoustic factors deemed significant in this experiment

Acknowledgements

- Arts & Humanities Innovation Grant
- Dominic Espinosa
- Cynthia Clopper
- Julie McGory, Laurie Maynell & Ross Methusalem
- Ping Bai

Appendix I: Comparable Duration and F0 of Target NPs

Contr? / Tune	Adj Dur (ms)	Adj F0 (Hz)	N dur (ms)	N F0 (Hz)
Y / L+H* ∅	356 *330*	332 *299*	458 *489*	148 *148*
Y / H* !H*	366 *332*	223 *207*	524 *549*	192 *164*
N / L+H* ∅	343 *320*	332 *300*	462 *491*	152 *150*
N / H* !H*	368 *316*	223 *208*	516 *558*	197 *163*

(Natural speech in italics)

Appendix II: Tone (To) and Break (B) Indices (I)

Accent	Function
H*	New info
L*	Old info
L*+H	
L+H*	Contrast
H+!H*	

Table: Pitch accents in American English

Appendix II: Tone (To) and Break (B) Indices (I)

Accent	Function
H*	New info
L*	Old info
L*+H	
L+H*	Contrast
H+!H*	

Table: Pitch accents in American English

- *: Phonetic alignment between tone and stressed syllable
- !: Contextually triggered lowering of tone