Evaluating the Effect of Pauses on Number Recollection in Synthesized Speech

Mikey Elmers, Raphael Werner, Beeke Muhlack, Bernd Möbius & Jürgen Trouvain

IDeaL SFB 1102

UNIVERSITÄT DES SAARLANDES

Why is this important?

- Speech synthesis systems are popular in banking and telephone industries
- Often, situations involve strings of numbers
- Exchanges are complicated by a necessity for high accuracy and show no redundancy

Introduction

- Telephone numbers are usually grouped prosodically, which helps in recollection (Baumann & Trouvain, 2001)
- Prosodic grouping is usually realized by rhythmic features within a minor prosodic phrase
- These boundaries are sometimes marked by a short pause

Research Question

- What effect does the presence of a pause have on recollection accuracy for synthesized digits?

Methods

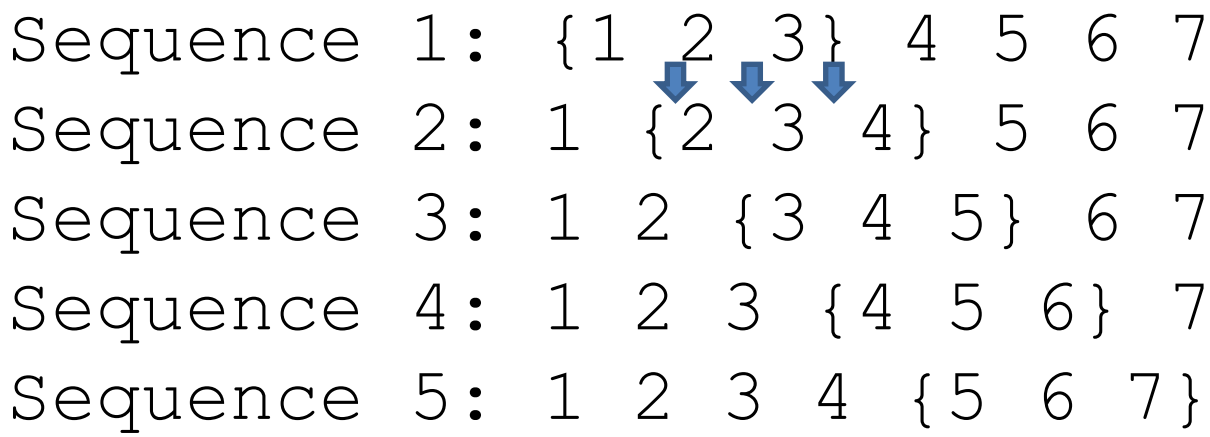
- Participants listened to synthesized audio for randomly generated 7-digit numbers (e.g. 3852791)
- Participants were asked to type a 3-digit sequence
- The 3-digit sequence was chosen to mask the critical digit

Pause Duration Information

- A pause was inserted before one of the digits (critical digit)
- 200 ms and 500 ms durations represented short and normal length pauses (Campione & Veronis, 2002)
- A non-inserted (0 ms) pause duration was also included

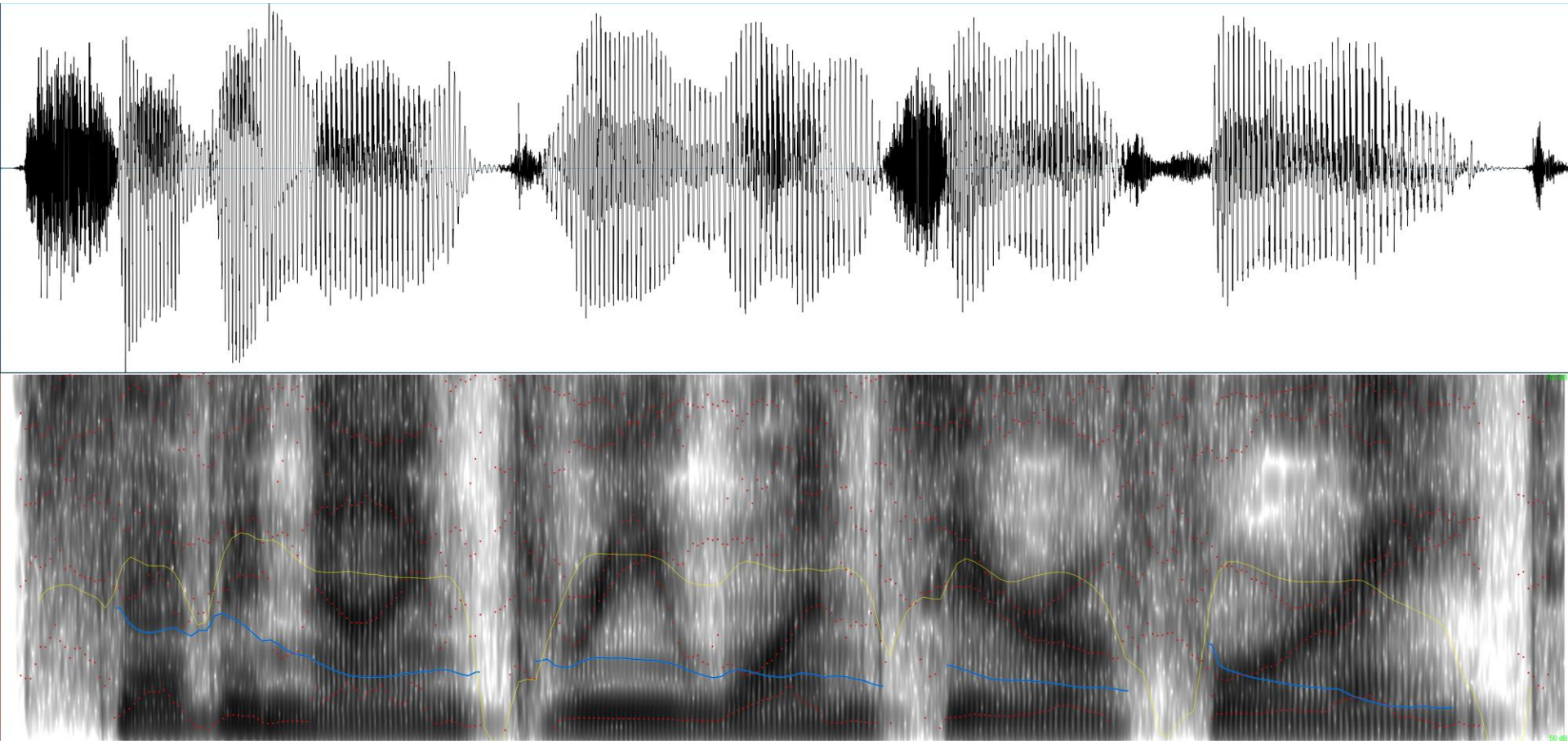
Stimulus Sequences

- Five possible locations for the 3-digit target sequences within the 7-digit sequences

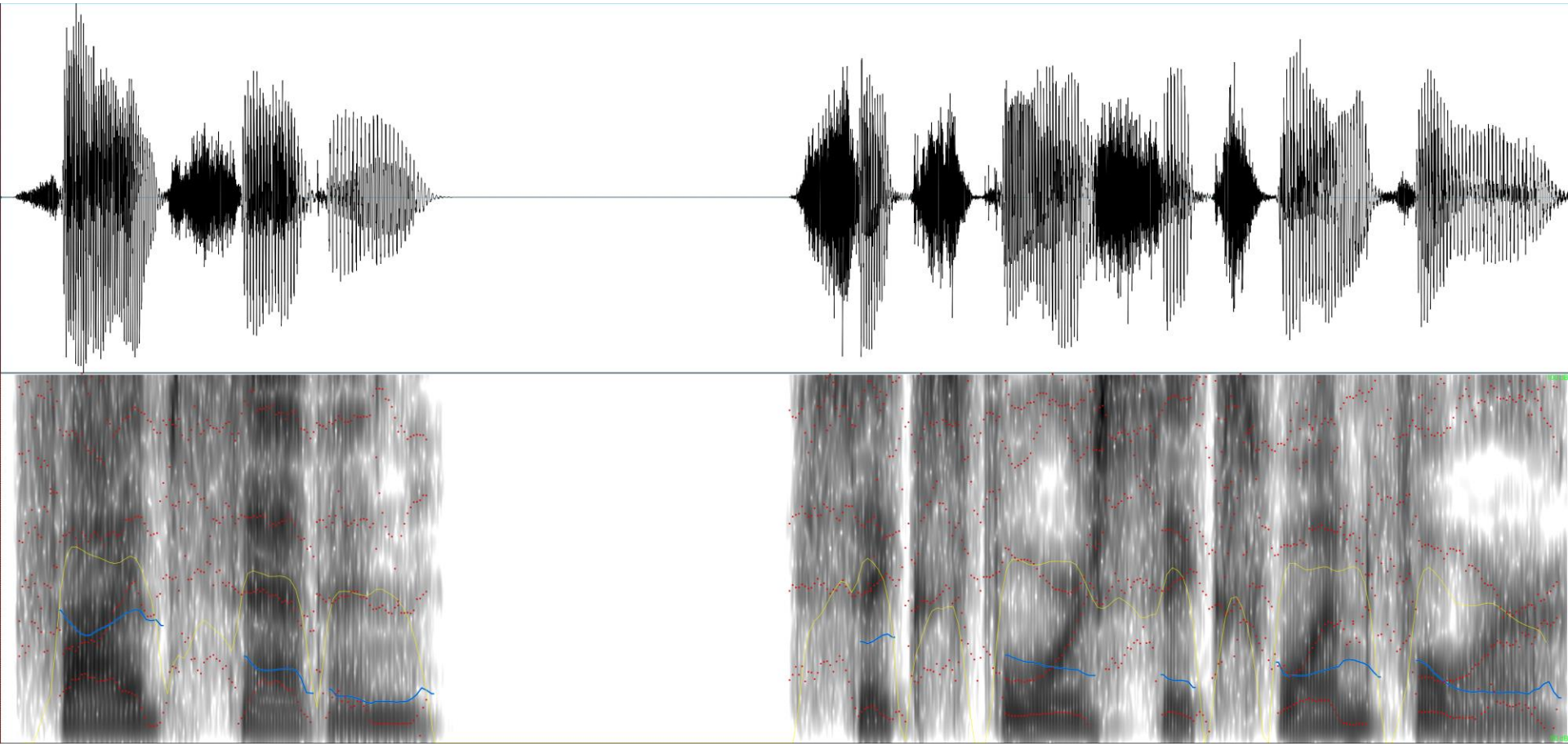


- First and last digit included to confirm primacy and recency effects (McLeod, 2008)

Amazon Polly Ex – 0 ms



Amazon Polly Ex – 500 ms



TTS Systems

- We considered multiple TTS systems
  - MaryTTS (Schröder & Trouvain, 2003)
  - Festival (Taylor et al., 1998)
  - Amazon Polly (Amazon, 2016)

- **None** of the systems automatically created pauses

- Instead, they all required some form of text markup

Stimuli Creation

- Stimuli created using Amazon Polly‘s TTS service with Joanna‘s voice
- Voice generated using concatenative synthesis
- Pauses (duration indicated in ms) were inserted with the Speech Synthesis Markup Language (SSML)

Experiment

- 35 audio clips, including two trial runs (not included in the results)
- Follow-up questionnaire
- Experiment was conducted in English
- Each participant encountered every condition, including all pause locations, sequences, and durations

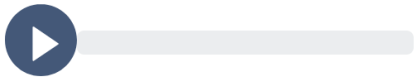
Experiment

- Total completion time was 10-20 minutes for each subject
- Experiment was conducted with Labvanced (Finger et al., 2017)
- Participants were recruited via Prolific (Prolific, 2014)

Instructions for Participants

Welcome and thank you for taking the time to participate in this study!

You will hear a 7-digit number. Afterwards, you will be asked to enter a 3-digit grouping. You will hear each audio clip only ***once***. Please put in headphones and test your audio with the example before clicking the "Next" button.



Ex. You hear 1 7 6 2 5 9 0

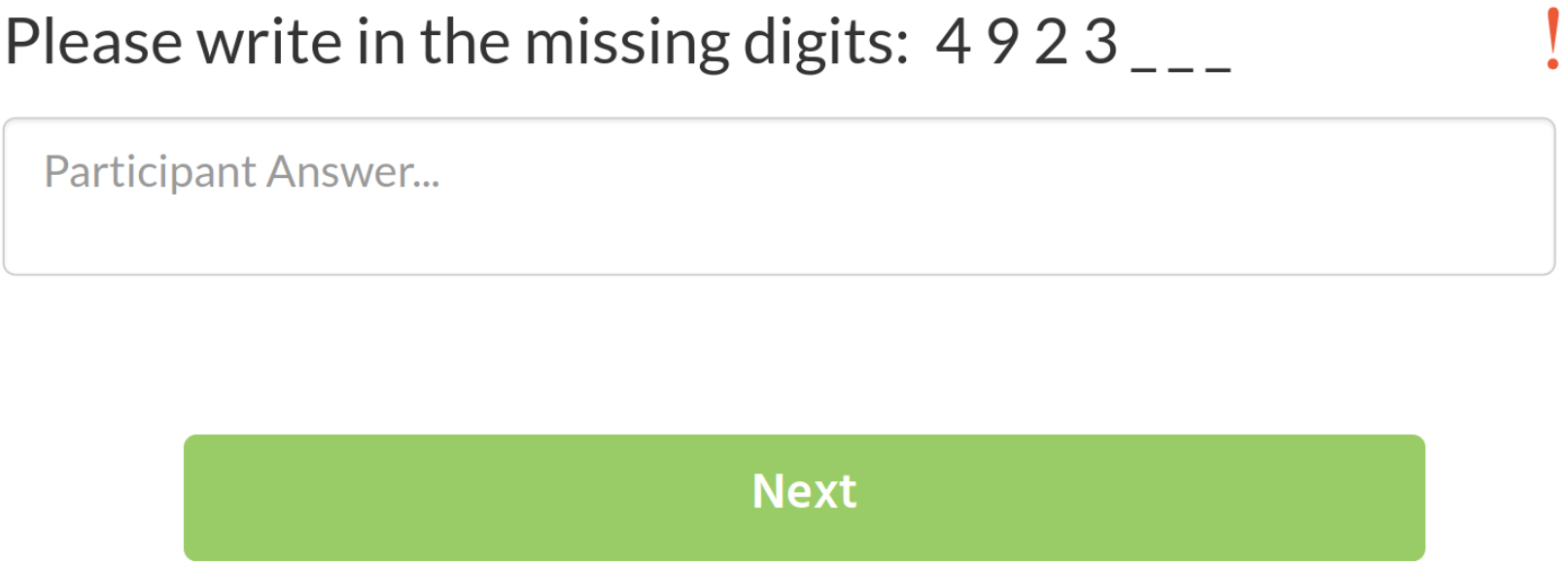
You are asked to fill in the blanks 1 7 6 _ _ _ 0

You should answer 259 (please write without spaces)

The experiment consists of 35 audio clips and a follow-up questionnaire. Please ***do not*** make notes while listening. Total completion time is 10-20 minutes.

Next

Example Participant Screen



Participant Background Information

- 15 Subjects (10F & 5M, age range: 25-60 years, mean age = 36.2 years)
- Monolingual speakers of English
- One participant self-reported hearing impairment and was excluded from the analysis

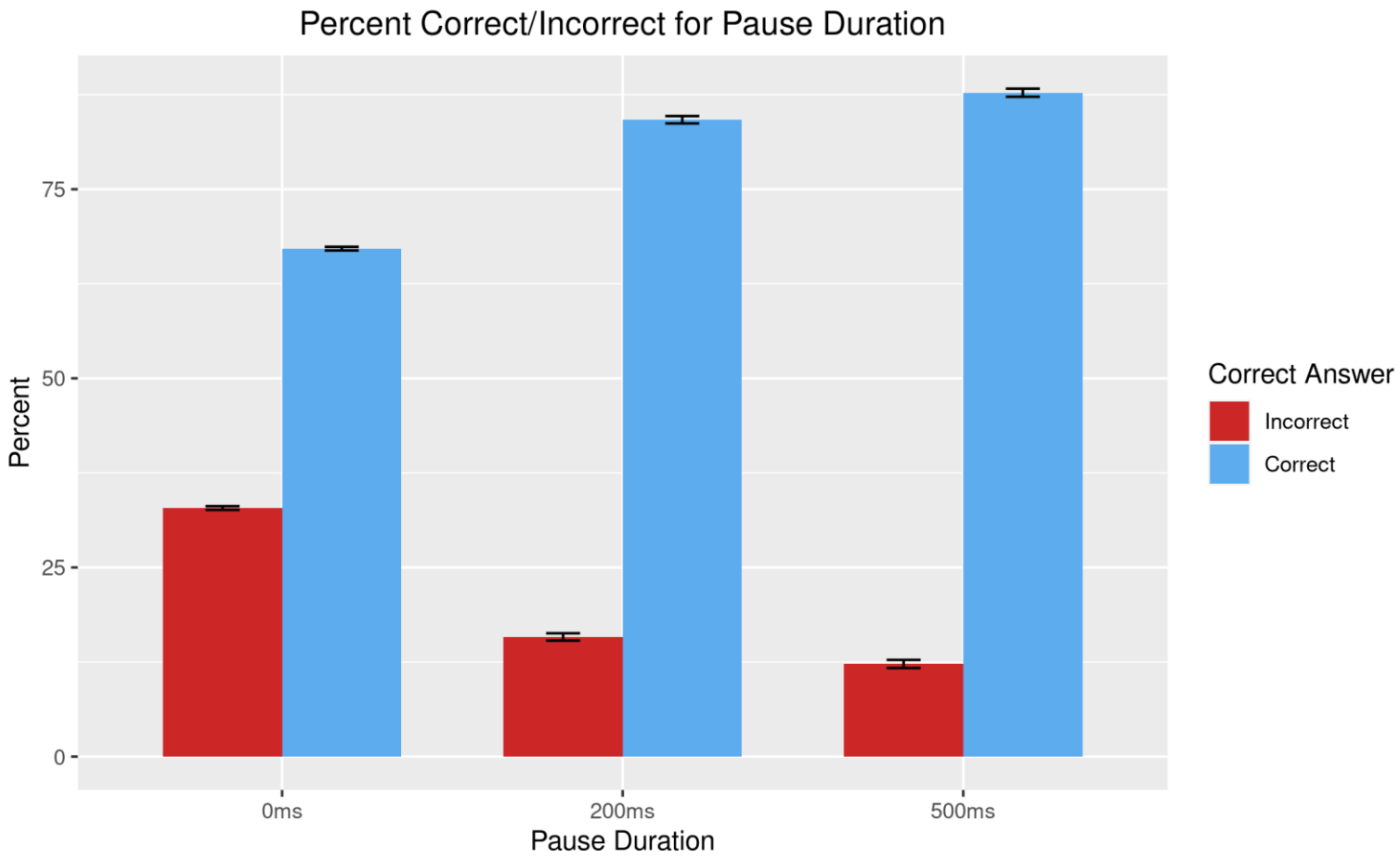
Participant TTS Familiarity

- Participants were asked, “how often do your listen to text-to-speech audio?“
- Possible responses included, “never“, “monthly“, “weekly“, and “daily“

Participant TTS Familiarity

- 8 (53%) indicated “never“
- 4 (27%) indicated “monthly“
- 1 (7%) indicated “weekly“
- 2 (13%) indicated “daily“

Results


Percent Correct/Incorrect for Pause Duration
Percent
75
50
25
0
0ms
200ms
500ms
Pause Duration
Correct Answer
Incorrect
Correct

Modeling

- Models were analyzed for accuracy (critical digit) and response time (RT)
- Analyzed with generalized linear mixed-effects models (GLMMs) from lme4 package (Bates et al., 2015) in R (R Core Team, 2020)

Accuracy Modeling

- *Accuracy:* binary categorical variable (0 for incorrect, 1 for correct)
- *Pause occurrence:* binary categorical variable (0 for absent, 1 for present)
- *Pause duration:* factor with 3 levels (0 ms, 200 ms, and 500 ms)

Accuracy Modeling

- *Sequencing:* factor with 5 levels
- *Digit position:* factor with 6 levels
- The first position (i.e. the first digit) was not considered since it can‘t be the critical digit
- Due to collinearity effects, *pause occurrence* and and *pause duration* were modeled seperately

Accuracy Modeling – Pause Occurrence

Model 1: GLMM Results Accuracy~Pause Occurrence + (1 | Subject) + (1 | Item)

Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.8947	0.6915	1.294	0.1957
PauseOccur1	1.6214	0.7475	2.169	0.0301 *

- Presence of pause is statistically significant and increases recollection accuracy

Accuracy Modeling – Pause Duration

Model 2: GLMM Results Accuracy~Pause Duration + (1 | Subject) + (1 | Item)
Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.8940	0.6871	1.301	0.1932
200ms	1.3019	0.7918	1.644	0.1001
500ms	1.9911	0.8309	2.396	0.0166 *

- Pause duration of 500ms is statistically significant and increases recollection accuracy (200ms was not)

Response Time Modeling

- Recorded the subjects RT
- Participants heard each clip only once
- RT started as soon as the audio clip finished
- RT ended as soon as they submitted their answer

Response Time Modeling

Model 3: GLMM Results RT~Pause Occurrence + (1 | Subject) + (1 | Item)

Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4930.62	26.48	186.19	<2e-16 ***
PauseOccur1	363.40	28.09	12.94	<2e-16 ***

- Pause occurrence is statistically significant with an increase in RT

Response Time Modeling

- The coefficient value 363.40 is similar to the average duration of 200 and 500 ms
- Might represent an abstract pause involved when participants mentally recall synthesized digits, before typing their answer

Summary

- This study investigated whether a pause in synthesized speech aided in digit recollection
- None of the TTS systems investigated automatically generate pauses
- We found, generally, that the presence of a pause improved digit recollection
- Unable to confirm that 200 ms pause significantly improved digit recollection

Summary

- RT is influenced by presence of pause
- RT increases when pause is present
- RT model indicated that participants might retain some abstract pause duration in their mind when recollecting

Discussion

- Investigate if a delay between hearing the clip and responding affects their accuracy and/or RT
- Exploraton of pause-internal particles and their effects
  - Ex. breath noises

Discussion

- Presently, all stimuli contained two prosodic groups
  - Group 1: all digits before the pause
  - Group 2: all digit following the pause

- Prosodic structure, specifically how numbers are grouped and the number of groups, are an important aspect of synthesized speech

Discussion

- Investigate prosodic structures further
- Basic grouping strategies for 7-digit numbers
  - Ex. 3-2-2 (Baumann & Trouvain, 2001)
- Evaluate different prosodic groups and their influence on digit recollection accuracy

References

Baumann, S. and J. Trouvain: On the prosody of German telephone numbers. In Eurospeech, pp. 557–560. 2001. doi:10.1215/00222909-2781749.

Schröder, M. and J. Trouvain: The German text-to-speech synthesis system mary: A tool for research, development and teaching. International Journal of Speech Technology,6, pp. 365–377, 2003. doi:10.1023/A:1025708916924.

Taylor, P., A. W. Black, and R. Caley: The architecture of the Festival speech synthesis system. In Third ESCA Workshop in Speech Synthesis, pp. 147–151. 1998.

Amazon Polly. 2016. URL`https://aws.amazon.com/polly/`. Accessed: 10.01.2021.

Campione, E. and J. Véronis: A large-scale multilingual study of silent pause duration. In B. Bel & I. Marlien (Eds.), Proceedings of the Speech Prosody Conference. Aixen Provence: Laboratoire Parole et. Langage, pp. 199–202. 2002.

Miller, G. A.: The magical number seven plus or minus two: some limits on our capac-ity for processing information. Psychological Review, 63 (2), pp. 81–97, 1956.

McLeod, S. A.: Serial position effect. 2008. URL `https://www.simplypsychology.org/primacy-recency.html`. Accessed: 10.12.2020.

Speech synthesis markup language (ssml) version 1.1. 2010. URL `https://www.w3.org/TR/speech-synthesis11`. Accessed: 07.01.2021.

Finger, H., C. Goeke, D. Diekamp, K. Standvoss, and P. König: Labvanced: a unified JavaScript framework for online studies. In International Conference on Computational Social Science (Cologne). 2017.

Prolific. 2014. URL `https://www.prolific.co`. Accessed: 12.01.2021.

Bates, D., M. Mächler, B. Bolker, and S. Walker: Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67(1), pp. 1–48, 2015.doi:10.18637/jss.v067.i01.

R Coreteam: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

Akaike, H.: Information theory and an extension of the maximum likelihood principle. In International Symposium on Information Theory, pp. 267–281. 1973.

Lo, S. and S. Andrews: To transform or not to transform: using generalized linear mixed models to analyse reaction time data. Frontiers in psychology, 6, p. 1171, 2015.doi:10.3389/fpsyg.2015.01171