

EVALUATING THE EFFECT OF PAUSES ON NUMBER RECOLLECTION IN SYNTHESIZED SPEECH

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

*Language Science and Technology, Saarland University
elmers@lst.uni-saarland.de*

Abstract: This study investigates the effects of an inserted pause on digit recollection for synthesized speech. Participants took part in a perception experiment which involved listening to a 7-digit random number that was rendered by a speech synthesis system. Some of the stimuli had pauses (200 ms or 500 ms in duration) inserted before one of the digits, while others did not include a pause. Immediately following each stimulus the participants were asked to provide a missing sequence of three adjacent digits. Results indicate that recall accuracy is improved immediately following a pause. Additionally, we found a significant effect for a pause duration of 500 ms but not for a pause duration of 200 ms. When investigating response time, we found that participants' response time increased when a pause was present. Overall, the results show that pauses have a role to play in synthesized speech. This research can be regarded in the context of investigating pauses and pause-internal particles (e.g. breath noises) in synthesized speech and the effects they have for human listeners.

1 Introduction

Speech synthesis systems have become ubiquitous in the banking and telephone industries. This progression has created situations where the average person is required to interact with synthesized recordings, often of strings of numbers (e.g. credit card and bank account). These exchanges are regularly complicated by a necessity for high accuracy and show no redundancy, in contrast to most other types of linguistic information. There is evidence that telephone numbers are grouped prosodically, which helps to recall those numbers [1]. This prosodic grouping is usually realized by rhythmic features such as alternations of accented and unaccented digits within a minor prosodic phrase. The boundaries of these minor prosodic phrases are sometimes marked by a short pause. Therefore, in this work, a perceptual experiment was conducted to investigate the effects of the presence of a pause on recollection accuracy for synthesized digits. More specifically, this experiment endeavoured to investigate the consequences of pauses on short-term digit recollection.

We previously conducted a pilot study with a similar focus using MaryTTS [2]. The pilot study focused on a single, short pause duration of approximately 200 ms and compared it against a no-pause condition. The results from the pilot study indicated that pauses in synthesized speech could improve digit recollection. For the current study, we aimed to elaborate on our previous findings and document the improvement of digit recollection with pause insertion for synthesized speech.

When researching TTS systems for the present study, multiple systems were considered, including MaryTTS [2], Festival [3], and Amazon Polly [4]. Interestingly, none of these systems created pauses automatically when generating synthesized digit sequences, and they all required some form of text markup. Both MaryTTS and Festival occasionally experienced problems

where part of the digit audio was truncated. To avoid any fractured audio we opted to use Polly to synthesize the audio clips. While all three TTS systems use voices created by concatenative synthesis, Polly was found to be superior in audio quality when compared to MaryTTS and Festival. This punctuated our decision to move forward using Polly and facilitated our intention to keep any audio quality discomfort as low as possible when listening to the audio clips. This in turn, allowed the participants to focus on the prompted digits rather than audio irregularities.

Another change that we made between the pilot and the present experiment was the addition of a second pause duration. In the pilot study, we observed a tendency towards the pause condition in the recollection accuracy between no pause and the 200 ms pause insertion. In this study we have added an additional pause duration of 500 ms. Our decisions regarding pause duration were based on a large multilingual study of silent pause durations [5]. By adding the longer pause to the experiment we hoped to see further exaggerations of the results indicated in the pilot study.

2 Method

2.1 Material

Participants listened to audio clips of synthesized speech that contained a randomized 7-digit number (e.g. 3852791). A 7-digit number was selected based upon Miller's Law [6], which states that the average person's short-term memory capacity is 7 ± 2 . Participants were asked to type a 3-digit sequence. There were 5 potential sequences (Figure 1).

```
Sequence 1:  {1 2 3} 4 5 6 7
Sequence 2:  1 {2 3 4} 5 6 7
Sequence 3:  1 2 {3 4 5} 6 7
Sequence 4:  1 2 3 {4 5 6} 7
Sequence 5:  1 2 3 4 {5 6 7}
```

Figure 1 – Five possible locations for the 3-digit target sequences within the 7-digit sequences.

In both pause conditions (200 and 500 ms), a pause was inserted prior to one of the digits participants were asked to type, i.e., the digits within the curly brackets in Fig. 1. A 3-digit graphical sequence was chosen to mask the critical digit, viz. the digit following the pause. The experiment included three pause durations: 0 ms, 200 ms, 500 ms. The durations of 200 and 500 ms were chosen to represent a short and a normal pause, respectively. The first and last digits were included as a baseline to confirm primacy and recency effects [7].

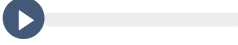
The stimuli were generated using Amazon Polly's TTS service with Joanna's voice, the standard TTS voice generated using concatenative synthesis. The pauses were inserted by using an instruction in the Speech Synthesis Markup Language (SSML) [8] indicating the pause duration in milliseconds.

2.2 Experiment

The material was uploaded to the online experiment platform Labvanced [9]. When beginning the experiment, participants were asked to use headphones and test their audio on the instruction screen. Then participants were instructed that they would hear a sound clip with a 7-digit number and that each clip would be played only once. After listening to the prompted clip, a box appeared on the screen and participants were asked to type a 3-digit sequence from the 7-digit

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.



Figure 2 – The instructions participants received before beginning the experiment.

Please write in the missing digits: 4 9 2 3 ___ !




Figure 3 – The screen participants saw after listening to the stimulus. Here they would enter a 3-digit sequence.

number. Figure 2 shows the instructions participants received before beginning the experiment. And Figure 3 shows the screen that appeared after listening to the audio clip, where participants entered the 3-digit sequence.

The experiment consisted of 35 audio clips, which included two trial runs that were not counted in the results, and a follow-up questionnaire. The experiment was designed so that each participant would experience every condition, including all pause locations, sequences, and durations. Total completion time was 10–20 minutes for each subject.

2.3 Participants

Participants were recruited from an online service using Prolific [10] and were offered payment for their time. There were a total of 15 subjects (10 F and 5 M, age range 25–60, mean age 36.2 years). All participants, except one, self-reported no form of hearing impairment. The participant who self-reported hearing impairment was excluded from the analysis. In order to determine familiarity with synthesized speech, subjects were asked, 'how often do you listen to text-to-speech audio?' Possible responses included were, "never", "monthly", "weekly", and "daily". Of the 15 participants, 8 (53%) indicated that they never listen to TTS audio, 4 (27%) indicated monthly, 1 (7%) indicated weekly, and 2 (13%) indicated daily usage.

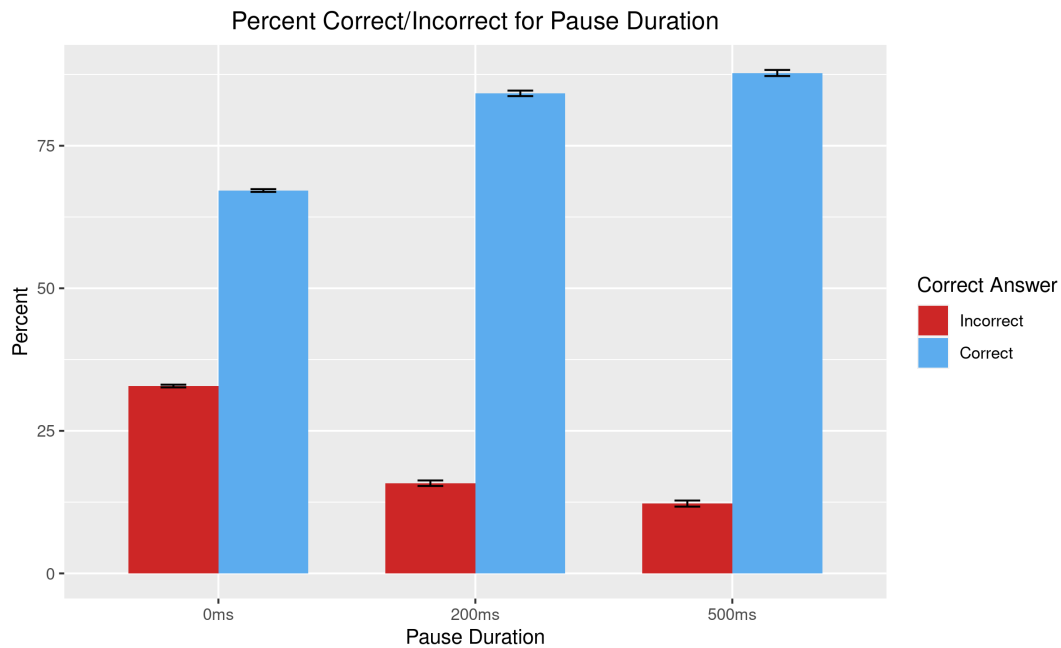


Figure 4 – Accuracy of the critical digit following the pause, or the sequence-central number in the no-pause condition, arranged by pause duration.

3 Results

The presence of a pause resulted in a higher recollection accuracy for the following digit than when the pause was absent. Similarly, Figure 4 shows that both the short (200 ms) and the normal (500 ms) pause durations caused a higher accuracy than the condition with no pause.

3.1 Accuracy Modeling

Multiple statistical models were analyzed for accuracy of the critical digit, i.e. the digit following the pause, and for response time (RT). The response variables were analyzed by using generalized linear mixed-effects models (GLMMs) from the lme4 package [11] in R [12].

Model decisions were made bottom-up, beginning with only random intercepts for *subject* and *item*, and progressively adding fixed effects. Random slopes for the fixed effects were added for *subject* assuming no issues from over-fitting or non-convergence. Models were compared via the Akaike information criterion (AIC) [13] to determine unexplained variance. If the AIC decreased by at least two points, then a factor was kept in the model.

The GLMMs were analyzed with *accuracy* of the critical digit (binary categorical variable, 0 for incorrect and 1 for correct) as the response variable. Models with the following predictor variables were evaluated: *pause occurrence* (binary categorical variable: 0 for absent, 1 for present), *pause duration* (factor with three levels: 0 ms, 200 ms, and 500 ms), *sequencing* (factor with 5 levels), and *digit position* (factor with 6 levels). For digit position, the first digit was not taken into account as it was never the critical digit. Due to collinearity effects, pause occurrence and pause duration were modeled separately.

Model 1 (Table 1), the model with the lowest AIC, included pause occurrence as the only fixed effect. Subject and item were included as random intercepts. The GLMM used a binomial family and logit link. This model shows that the presence of a pause is statistically significant and increases recollection accuracy (estimate (log-odds) = 1.6214, SE = 0.7475, $z = 2.169$, $p < 0.05$).

Model 2 (Table 2), the model with the lowest AIC, included pause duration as the only fixed effect. Subject and item were included as random intercepts. The GLMM used a binomial

Table 1 – 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 *

Table 2 – 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
PauseDur200ms	1.3019	0.7918	1.644	0.1001
PauseDur500ms	1.9911	0.8309	2.396	0.0166 *

family and logit link. This model indicates that the pause duration of 500 ms was statistically significant (estimate (log-odds) = 1.9911, SE = 0.8309, $z = 2.396$, $p < 0.05$) and is beneficial for recollection accuracy. However, the 200 ms pause duration was not statistically significant. Models were also analyzed for accuracy predicted by RT, yet none of the models achieved a lower AIC than the model with only random effects.

3.2 Reponse Time Modeling

The subject's response time (RT) was also recorded. Participants were only able to hear the clip once, and the RT timer started as soon as the audio clip ended. Upon submitting their answers the RT timer finished. The participants' RT had a highly positive skew, therefore, values that exceeded 3 standard deviations above the mean were excluded. Even with these values removed RT still skewed positive but was far less extreme. Even so, a gamma distribution was chosen. Additionally, while investigating RT, only correct answers were included in the models.

For Model 3 (Table 3), the model with the lowest AIC, included only pause occurrence as a fixed effect. Subject and item were included as random intercepts. A GLMM, with a gamma family and identity link, was chosen over a LMEM with a log-transformation of RT. This decision was made to prevent issues that can occur from seeking normality of a log-transformed RT [14].

Table 3 – 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 ***

Model 3 shows that pause occurrence is significant for RT (estimate = 363.40, SE = 28.09, $z = 12.94$, $p < .001$). Interestingly, the effect is an increase in RT. The coefficient value of 363.40 is similar to the average duration between the two pause durations, 200 and 500 ms. This duration might be representative of an abstract pause involved when the participants mentally recall the synthesized digits, before typing their answer. Models were also analyzed for RT predicted by pause duration, yet none of the models achieved a lower AIC than the model with only random effects.

4 Discussion and Summary

In this study, participants were tasked with listening to a 7-digit clip of synthesized speech to determine if a pause affected their recollection accuracy for the following digit. This study aimed at improving an effect found in our pilot study, specifically that a pause in synthesized speech aided in digit recollection. This study made improvements over the pilot by including: a higher quality concatenative TTS system, an additional 500 ms pause duration, and investigating RT. Using GLMM models, we have shown, generally, that the presence of a pause indeed affects recollection accuracy. Moreover, we also found that the 500 ms pause duration improved digit recollection. However, we were unable to confirm the results from our pilot study that a 200 ms pause duration improved digit recollection. These results emphasize the importance of further research on pauses in synthesized speech.

An important aspect of synthesized digit sequences is the prosodic structure, specifically how the number sequences are grouped and the number of groups. All stimuli in this study contained two prosodic groups. The first included all digits up to the pause, while the second consisted of all the digits following the pause. It is important to investigate these prosodic structures with more attention. Additionally, in the future we could include basic grouping strategies (e.g. 3-2-2) for 7-digit numbers [1] to evaluate different prosodic groups and their influence on digit recollection accuracy.

In the current study we have shown that the presence of a pause also influences response time, with RT increasing when a pause is present. Results indicate that the participant does not differentiate between pause durations while recalling the digits. The RT model showed that participants might be retaining some abstract pause duration in their mind during recollection. RT was measured after the sound clip finished and without a delay. Future research should evaluate whether the duration between when the clip finishes, and when the participant is able to respond, affects their accuracy. A promising next step in this research would be to investigate pause-internal particles, such as breath noises, for their effects on synthesized speech digit recollection.

References

- [1] BAUMANN, S. and J. TROUVAIN: *On the prosody of German telephone numbers*. In *Eurospeech*, pp. 557–560. 2001. doi:10.1215/00222909-2781749.
- [2] 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.
- [3] 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.
- [4] *Amazon Polly*. 2016. URL <https://aws.amazon.com/polly/>. Accessed: 10.01.2021.
- [5] 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*. Aix-en-Provence: Laboratoire Parole et Langage, pp. 199–202. 2002.
- [6] MILLER, G. A.: *The magical number seven plus or minus two: some limits on our capacity for processing information*. *Psychological Review*, 63 (2), pp. 81–97, 1956.

- [7] MCLEOD, S. A.: *Serial position effect*. 2008. URL <https://www.simplypsychology.org/primacy-recency.html>. Accessed: 10.12.2020.
- [8] *Speech synthesis markup language (ssml) version 1.1*. 2010. URL <https://www.w3.org/TR/speech-synthesis11>. Accessed: 07.01.2021.
- [9] 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.
- [10] *Prolific*. 2014. URL <https://www.prolific.co>. Accessed: 12.01.2021.
- [11] 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.
- [12] R CORE TEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL <https://www.R-project.org/>.
- [13] AKAIKE, H.: *Information theory and an extension of the maximum likelihood principle*. In *International Symposium on Information Theory*, pp. 267–281. 1973.
- [14] 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.