INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Take a breath: Respiratory sounds improve recollection in synthetic speech

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

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

Abstract

This study revisits Whalen et al. (1995, JASA) by evaluating English speaking participants in a perception experiment to determine if their recollection is affected by including breath noises in sentences generated by a speech synthesis system. Whalen found an improvement in recollection for sentences that were preceded by a breath noise compared to sentences without one. While Whalen and colleagues used formant synthesis to render the English sentences, we use a modern concatenative synthesis system. The present study uses inhalations of three different lengths: 0 ms (no breath noise), 300 ms (short breath noise), and 600 ms (long breath noise). Our results are consistent with Whalen and colleagues for the 600 ms condition, but not for the 300 ms condition, indicating that not all inhalations improved recollection. The present study also found a significant effect for sentence length, illustrating that shorter sentences have higher accuracy for recollection than longer sentences. Overall, the present study indicates that respiratory sounds are important to the recollection of synthesized speech and that researchers should focus on longer and more complex types of speech, such as paragraphs or dialogues, for future studies.
Index Terms: speech synthesis, pause particles, breath noises, inhalation, memory

1. Introduction

In the present study, we examined pause particles in synthesized speech. Pauses include stretches of silence and often particles, such as breath noises and sometimes clicks, which are an underresearched aspect of speech synthesis. As the general segmental quality of text-to-speech synthesis (TTS) systems has improved, the focus has shifted towards suprasegmental elements, with the intent of creating natural and expressive sounding speech. Breath noises, for example, can mark speaker individuality [1] or indicate formality in Korean [2]. The intensity and duration of breath noises are influenced by speech planning [3] and can help the listener predict the amount of upcoming material. While evaluating single sentences, [4] found a positive correlation between the duration of an inhalation and the length of the upcoming sentence. The inclusion of breath noises in speech synthesis may improve the naturalness and expressiveness desired in audiobooks, conversational assistants, and characters for movies and games.

Speech synthesis systems display large amounts of variation in how they handle pause particles. For example, pauses are often handled haphazardly, applying rudimentary punctuation-based heuristics for determining their location, frequency, and duration. Most modern TTS systems do not implement pauses with appropriate placement and duration [5], and fail to include any breath noises whatsoever. See [6, 7] for notable exceptions.

Previous work by Whalen et al. [8] (henceforth Whalen) found that English speaking participants' recollection, sometimes referred to as recall (as in Whalen), was better for sentences preceded by a breath noise than those not preceded by a breath noise. Whalen's study was conducted using a formant synthesizer, KLATTALK [9]. In contrast, [10] used concatenative synthesis to evaluate the perception of telephone numbers preceded by an inhalation. They found that the majority of subjects did not have a preference. The results from [8] and [10] offer conflicting interpretations of the effect of breath noises in synthesized speech, which called for further investigation. In a recent study [11], we found that the insertion of a silent pause in a 7-digit sequence improved the recollection of the following digit. An appropriate next step was to evaluate breath noises and revisit Whalen's study.

The primary objective for this experiment is to clarify the conflicting interpretations between [8] and [10]. Therefore, we endeavoured to examine if breath noises aid in recollection. In an effort to investigate this question, the present study closely mirrors the Whalen study, with some updates concerning technology. Specifically, we used Amazon Polly [12] to generate our stimuli and a web-based platform to conduct the experiment. By including these modifications, and other nuanced updates, we intend to contribute research to pauses and pause particles in synthesized speech.

2. Comparison of the present study and Whalen

The present study is a partial replication of Whalen, combining ideas from their experiments 1, 3 and 4. In each experiment participants listened to synthesized audio and, afterwards, wrote down what they heard. Experiment 1 focused on the effect between a breath noise and a no breath noise condition, with each condition separated into a single block. For example, the participants would hear a block of 20 sentences each preceded by a breath noise, followed by a second block of 20 sentences not preceded by a breath noise. The opposite ordering of blocks was also included. They found a significant effect for breath noises on the improvement of recollection. Moreover, the no breath noise condition did not have a significant effect on recollection improvement. Lastly, they found an improvement due to practice.

Experiment 3 and 4 maintained the breath noise/no breath noise conditions from experiment 1, but with more specificity. In experiment 3, rather than using the same block system from experiment 1, the breath noises were inserted randomly before sentences. Once again practice was found to be significant, but breath noises were not significant. While experiment 3 focused on random distribution of the breath noises, experiment 4 focused on appropriateness. In their earlier experiments they had maintained the appropriateness of the breath noise. In other words, short sentences were only preceded by the short (mean duration $\sim$ 600 ms) breath noise and long sentences were only preceded by the long (mean duration $\sim$ 740 ms) breath noise. In experiment 4, they tested appropriateness in a way that both short and long breath noises appeared before both short and long sentences. They found appropriateness was not significant but

http://dx.doi.org/10.21437/Interspeech.2021-1496

Table 1: *Mean (SD) for breath and sentence lengths reported here compared to Whalen (SD was not reported in Whalen)*

	Present Study	Whalen
Short breath duration (in ms)	300	597
Long breath duration (in ms)	600	738
Short sentence length (in words)	8.5 (2.0)	8.1
Long sentence length (in words)	16.2 (3.6)	15.2

they indicated this may be due to the small range of sentence lengths.

The present experiment synergizes many of the aforementioned ideas from Whalen. We incorporated the breath noise vs no breath noise conditions from experiment 1. We assigned breath noises randomly before sentences, rather than in blocks (like experiment 3). Lastly, we evaluated appropriateness by inserting short breath noises before long sentences, and vice versa (like experiment 4). This experiment examines the following durational conditions: a 0 ms no breath noise (henceforth NO-brn), a 300 ms breath noise (henceforth SHORT-brn), and a 600 ms breath noise (henceforth LONG-brn). Table 1 contains a comparison between the present study and Whalen for breath noise durations and sentence lengths. Participants in both experiments heard synthesized audio and recollected what they heard. However, in the present experiment participants typed their responses after each stimulus rather than writing them by hand. The experimental design in Whalen is easily converted into a web-based study like we did here.

3. Method

3.1. Creating the stimuli

For this experiment we used Amazon Polly, which Amazon describes as a "Text-to-Speech service that uses advanced deep learning technologies to synthesize speech that sounds like a human voice" [12]. The documentation for Polly does not provide further information beyond "advanced deep learning technologies" to clarify how the breath noises were created or the amount of breath noise variation. Polly's breath feature announcement claims that Polly can parrot the sounds of both inhalation and exhalation for normal speech. However, in our time working with Polly only inhalations could be identified. Additionally, the breath tags required for synthesizing respiratory sounds are currently only available for the standard voices, which use concatenative synthesis, not for the neural voices.

Polly includes an automated mode which allows the user to indicate (using preset values) the volume, frequency and duration for the synthesized breaths. The current experiment uses the manual mode to specify exact locations, and to customize the duration and volume. The breath noises (for both automated and manual mode) must be indicated using text mark-up, specifically Speech Synthesis Markup Language (SSML) [13].

Whalen's stimuli were created with KLATTALK [9], a formant synthesizer. Their breath noises were made from recordings of a person with a similar vocal tract to the voice model of their synthesizer. They recorded a total of six breath noises (3 short and 3 long) to add variety and factor out any oddities. Additionally, they indicated that their sentences were *not* completely comprehensible, but every sentence was answered correctly by at least one participant.

With the goal of creating more natural and expressive

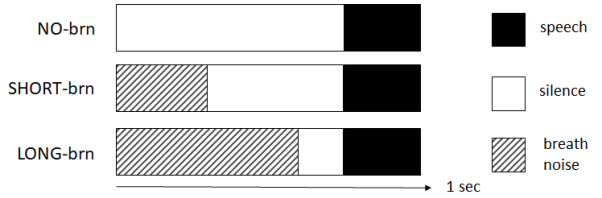

Figure 1: *Schematic for the first second of the stimuli in the three conditions.*

speech, we used Polly to generate inhalation sounds. The three conditions for this experiment were: 1) NO-brn (i.e. 0 ms), SHORT-brn (mean duration: 300 ms), and LONG-brn (mean duration: 600 ms). Our justification for the short and long inhalation durations are from a study on phrase-initial inhalation noises [14], which differ from phrase-internal inhalations.

We chose to use Polly's "default" breath noise since it is ∼300 ms in duration. We also chose Polly's "default" intensity value since it is ∼40 dB, which is consistent with what [14] found as a median intensity for phrase initial inhalations. [14] also found a median value of ∼140 ms for the right-edge pause between the phrase-initial inhalation and the onset of speech. Polly naturally inserts a ∼50 ms right-edge pause between the inhalation and the speech, so we increased this to a total of 150 ms by adding an additional 100 ms of silence (Fig. 1).

The present study uses a total of 28 different sentences (24 experimental, 3 practice, and 1 for instruction). For the 24 experimental stimuli, 12 were short sentences (mean number of words = 8.5, SD = 2.0, range: 5–12) and 12 were long (mean number of words = 16.2, SD = 3.6, range: 13–26). Some sentences included simple numbers, but none included complex or alphanumeric expressions. The sentences were created with Polly using the aforementioned methodology and consisted of situations that are typically discussed with conversational assistants such as weather, schedule information, restaurant bookings, etc.[1]

Three versions of each sentence were created using our three conditions (NO-brn, SHORT-brn, and LONG-brn), resulting in a total of 72 tokens. These 72 tokens were evenly divided into three lists, designed in such a way that each sentence appeared only once per list. Additionally, the lists were balanced to achieve an equal number of each breath noise condition. The tokens in each list were randomized, so that different lists had a different ordering of sentences. However, participants who saw the same list encountered the sentences in the same order.

3.2. Participants

We created our web-based experiment using Labvanced [15] to present the audio stimuli to the participants and collect their typed answers, questionnaire information, and response time (RT). Participants were recruited with Prolific [16] and consisted of 63 monolingual English participants (mean age 36.92 years; age range 18–70 years; 29 females, 33 males, 1 nonbinary; 59 British accented, 2 American accented, 2 Australian accented) who were paid for their participation. One participant indicated hearing impairment and was excluded from the results. For the experiment, subjects were instructed to type what they heard exactly as they heard it. Subjects were presented with one of three lists. Each list consisted of the same 24

[1]The sentences are available on pauseparticles.org.

sentences. However, they varied in which breath noise condition (NO-brn, SHORT-brn, LONG-brn) preceded the sentence, and in the overall ordering of the stimuli. Participants listened to one audio clip during the instruction screen which was followed by three practice sentences (not included in the results). The practice sentences included examples that were preceded by a breath noise and some not preceded by a breath noise. After completing the listening portion they filled out a questionnaire.

3.3. Scoring and data processing

After collecting the participants' results, we standardized the data by tokenizing, removing punctuation and extra whitespace, converting words to lowercase, and correcting some spelling errors. For example, if a participant typed "appoxiately", we corrected it to "approximately", and counted it as correct during the scoring. However, homophones or words that did not preserve the intended meaning of the sentence were not corrected. For example, if a participant wrote "weight" instead of "wait" in the context of waiting for a table at a restaurant then their word was not corrected, and consequently, not scored positively.

After standardizing the data, participants were scored based on how many of the correct words they had included in their response. They were awarded 1 point for each correct word. In the present study we focused on whether the correct word was included, not on the order. Whalen's scoring method provided one point for a correct word in the correct location. A mostly correct word was worth 0.5. A correct word in the incorrect location provided 0.5. Whalen scored homophones as correct and did not encounter semantically related words. In the present study, the scoring system was simplified so that participants were awarded 1 point if the word in their submission was found in the canonical version (i.e., the correct version). The present study and Whalen, counted function and content words equally when scoring, since the TTS systems used in the two studies did not reduce function words as in human connected speech. Scores were normalized by dividing the participant's score by the length (i.e., number of words) of the canonical version of the sentence. Normalized scores ranged from 0 to 1. The differences in scoring methods might affect differences between the two studies. However, within the study, since all stimuli were scored using the same method, there is a level of consistency when comparing the scores.

4. Results

The mean and standard deviation for the different breath noise conditions can be seen in Table 2. When looking at the mean scores for all conditions, it is clear that participants are already scoring near the normalized score ceiling, which can also be seen in Fig. 2. When looking at the individual breath noise conditions, we find higher scores for the LONG-brn condition compared to the NO-brn and SHORT-brn conditions. As for length, we also find a score difference between short and long sentences.

Statistical models were analyzed with linear mixed-effects models (LMEM) from the lme4 [17] package (Version 1.1.25) and the lmerTest [18] package (Version 3.1.3) in R [19] (Version 3.6.3). Models were made using backwards selection, i.e., starting with the maximal model for fixed and random effects and gradually reducing (starting with random slopes) in the case of over-fitting or non-convergence. Models were compared with the Akaike information criterion (AIC) [20], which calculates unexplained variance, and the model with the lowest AIC was considered as the model with the best fit.

Table 2: *Scores normalized by number of words for different conditions.*

Condition	Mean	SD
All Conditions	0.909	0.154
NO-brn	0.902	0.164
SHORT-brn	0.902	0.160
LONG-brn	0.923	0.136
Length Short	0.959	0.110
Length Long	0.860	0.175

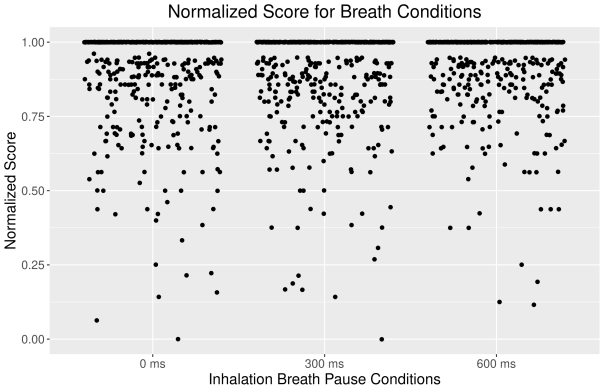

Figure 2: *Scatterplot for score normalized by number of words for each of the breath noise conditions.*

The final model was: $lmer(NormalizedScore \sim BreathNoise + Length + (1 \mid Subject) + (1 \mid Sentence), REML = FALSE)$. This model includes breath noise duration and sentence length as fixed effects (without an interaction term). As random effects, intercepts were included for both the subject and the individual sentences. Visual inspection for the residual plot revealed deviations from homoscedasticity and a violation of normality (partly caused by the ceiling effects). However, [21] has shown that linear mixed-effects models are robust to these types of violations. Our analysis revealed a main effect for the LONG-brn condition (*Estimate* = 0.02077, SE = 0.00722, t = 2.877, $p < 0.01$) and the short sentence length (*Estimate* = 0.09894, SE = 0.02986, t = 3.314, $p < 0.01$). These main effects indicate an increase in recollection of the sentence. We found that shorter sentences are recalled better than longer sentences, and that sentences immediately preceded by a LONG-brn are recalled better than sentences preceded by the NO-brn or the SHORT-brn.

5. Discussion

The present study replicated one of the major findings from Whalen, namely that the LONG-brn condition improves recollection. With these results in mind, future research can investigate the following: duration, learning effects, sentence length, and measuring recollection.

5.1. Duration

When designing the experiment, the first author found the SHORT-brn to be most natural, while the LONG-brn appeared abnormally long. However, the SHORT-brn condition was not

significant while the LONG-brn condition was significant. The short and long breath noises used by Whalen were longer than the versions used in the present study, and found to improve recollection. Importantly, the present study's LONG-brn was approximately the same duration as Whalen's short condition. This finding may indicate that exaggerated breath noises, and possibly other pause particles, are more suitable for synthesized speech with respect to recollection.

There are many hypotheses that could explain the improvement in recollection caused by various particles in speech [22], including breath noises. While we describe these options, we do not position one as the primary rationale for recollection improvement. Three possibles hypotheses are: 1) processing-time hypothesis, i.e., the breath noises are providing more time for the listener to process what they hear, 2) attention orienting hypothesis, i.e., the breath noises are drawing the listener's focus, and 3) predictive processing hypothesis, i.e., participants use the breath noises to predict upcoming speech content. Future work should further investigate the specific mechanisms for improving recollection in synthesized speech.

5.2. Learning effects

Whalen found that participants performed better during the second half of the stimuli than during the first half (i.e. learning effect). The present study did not find any kind of learning effect, possibly due to improvements in audio quality for modern TTS systems. Another possibility is that listeners have become more acclimated to hearing synthesized audio. In a follow-up questionnaire, participants were asked how often they listen to computer-generated audio, such as conversational assistants or in-car navigation. Only 11 of the 63 participants reported never listening to computer-generated speech; however, this number might be inaccurate if participants misunderstood potential situations in which they hear computer-generated audio, such as robocalls or online videos.

5.3. Sentence Length

Whalen measured sentence length in number of words. Consequently, the present study also measured length via number of words, in order to maintain parity with Whalen. Ideally, length would be evaluated using a more stable metric such as a speech timing unit, e.g., number of syllables. This would alleviate the problem that arises when two sentences share the same number of words but vary greatly in their number of syllables.

The present study found high recollection scores for short and long sentences. Therefore, future work should include longer material lengths, such as paragraphs or fragments of dialogue. In the present study, short sentences (mean length = 8.5 words) had a mean accuracy of 0.959, whereas the long sentences (mean length = 16.2 words) had a mean accuracy of 0.860. The high quality of the synthesizer allows participants to not only understand the material, but repeat it verbatim, with near perfect accuracy. While we see an accuracy drop in the longer sentences, future experiments should investigate both longer and more complex sentences and discourses. In fact, [6] concluded that paragraphs and longer sentences are important and might improve naturalness for the listener by reducing the monotony and improving the prosody of speech synthesis. Interesting examples would be paragraphs of material, such as audiobooks, or dialogic conversation between humans and conversational agents. Finally, it would be interesting to look into semantically unpredictable sentences to see if these results for recollection hold.

5.4. Measuring Recollection

Both the present study and Whalen tested the participants' ability to recollect the exact message they had heard. While typing or writing their answer, participants are required to focus on spelling, potentially reducing the amount of effort they can give to the general content. It is important to think about what metrics and constructs are used to measure participant recollection, since there are many different ways to measure understanding and memorization. One possible alternative could have participants listen to an audio clip and record a summary in their own words, similar to [22], so that a participant's score would be dependent on overall comprehension rather than a word-for-word memorization. Another alternative could provide participants with multiple-choice questions. Future work should focus on a particular format to evaluate specific details with more nuance.

6. Conclusions

The present study investigated the effect of an inserted breath noise on recollection of synthesized speech, similar to Whalen et al. [8]. Our results are comparable to the results found by Whalen and colleagues. Three breath noise conditions were evaluated, a NO-brn (i.e. 0 ms) condition, a SHORT-brn (mean duration: 300 ms) condition, and a LONG-brn (mean duration: 600 ms) condition. Participants displayed a high level of recollection overall, even in the NO-brn condition. The LONG-brn improved recollection, whereas the SHORT-brn did not. We also found a significant effect for sentence length, which indicates that recollection is better in shorter sentences.

This experiment evaluated breath noises in single sentence contexts, avoiding connected speech due to difficulties in determining whether the breath noise influences the planning of the upcoming sentence or is a consequence of the preceding speech. Therefore, we chose to investigate breath noises in a smaller, more manageable context before looking towards longer and more complex forms of discourse in the future.

This work on breath noises is a component of a larger project, investigating pause-internal particles. Beyond investigating recollection abilities as a function of breath noises, future work will also view this phenomenon from the perspective of naturalness, which is important for maintaining expressiveness without sacrificing the pleasantness of synthetic speech.

7. Acknowledgements

This research was funded in part by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID MO 597/10-1. The authors would like to extend their thanks to Vera Demberg and Heiner Drenhaus for their assistance with regards to statistical analyses and experimental design. Additionally, we thank our student assistant Hanna Zimmermann for her support.

8. References

[1] M. Kienast and F. Glitza, "Respiratory sounds as an idiosyncratic feature in speaker recognition," in *Proc. 15th International Congress of Phonetic Sciences*, Barcelona, 2003, p. 1607–1610.

[2] B. Winter and S. Grawunder, "The phonetic profile of Korean formal and informal speech registers," *Journal of Phonetics*, vol. 40, no. 6, pp. 808–815, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447012000666

[3] S. Fuchs, C. Petrone, J. Krivokapic, and P. Hoole, "Acoustic and respiratory evidence for utterance planning in German," *Journal of Phonetics*, 01 2013.

[4] D. Whalen and J. Kinsella-Shaw, "Exploring the relationship of inspiration duration to utterance duration," *Phonetica*, vol. 54, pp. 138–52, 02 1997.

[5] J. Trouvain and B. Möbius, "Zu Mustern der Pausengestaltung in natürlicher und synthetischer Lesesprache," in *Proc. 29th Conference Elektronische Sprachsignalverarbeitung (ESSV '18)*, Ulm, 2018, pp. 334–341.

[6] N. Braunschweiler and L. Chen, "Automatic detection of inhalation breath pauses for improved pause modelling in HMM-TTS," in *Proc. 8th Speech Synthesis Workshop*, Barcelona, 2013, pp. 1–6.

[7] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "Breathing and speech planning in spontaneous speech synthesis," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7649–7653.

[8] D. H. Whalen, C. E. Hoequist, and S. M. Sheffert, "The effects of breath sounds on the perception of synthetic speech," *The Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3147–3153, 1995. [Online]. Available: https://doi.org/10.1121/1.411875

[9] D. Klatt, "The KLATTALK text-to-speech conversion system," in *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 7, 1982, pp. 1589–1592.

[10] J. Trouvain and B. Möbius, "Einatmungsgeräusche vor synthetisch erzeugten Sätzen — eine Pilotstudie," in *Proc. 24th Conference Elektronische Sprachsignalverarbeitung (ESSV '13)*, Bielefeld, 2013, pp. 50–55.

[11] M. Elmers, R. Werner, B. Muhlack, B. Möbius, and J. Trouvain, "Evaluating the effect of pauses on number recollection in synthesized speech," in *Elektronische Sprachsignalverarbeitung 2021, Tagungsband der 32. Konferenz*, ser. Studientexte zur Sprachkommunikation. Berlin: TUD Press, 2021, pp. 289–295.

[12] "Amazon Polly," 2016, accessed: 22.02.2021. [Online]. Available: https://aws.amazon.com/polly/

[13] "Speech Synthesis Markup Language (SSML) version 1.1," 2010, accessed: 15.02.2021. [Online]. Available: https://www.w3.org/TR/speech-synthesis11

[14] R. Werner, S. Fuchs, J. Trouvain, and B. Möbius, "Acoustic and physiological characteristics of breath noises," *Submitted to INTERSPEECH 2021*, 2021.

[15] H. Finger, 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.

[16] "Prolific," Oxford, UK, 2014, accessed: 03.03.2021. [Online]. Available: https://www.prolific.co

[17] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[18] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[19] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[20] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *International Symposium on Information Theory*, 1973, pp. 267–281.

[21] H. Schielzeth, N. J. Dingemanse, S. Nakagawa, D. F. Westneat, H. Allegue, C. Teplitsky, D. Réale, N. A. Dochtermann, L. Z. Garamszegi, and Y. G. Araya-Ajoy, "Robustness of linear mixed-effects models to violations of distributional assumptions," *Methods in Ecology and Evolution*, vol. 11, no. 9, pp. 1141–1152, 2020. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13434

[22] S. H. Fraundorf and D. G. Watson, "The disfluent discourse: Effects of filled pauses on recall," *Journal of Memory and Language*, vol. 65, no. 2, pp. 161–175, 2011.