*University of Utah*
UNDERGRADUATE RESEARCH JOURNAL

# THE EFFECTS OF TEXT CAPTIONING ON NON-NATIVE ENGLISH SPEAKERS'  IMMEDIATE RECALL AND DELAYED RECOGNITION OF SPEECH

**Vanessa Srivastava (Brennan Payne)**
**Department of Psychology**

Studies based on Rabbitt's Effortfulness Hypothesis suggest that speech perceived in challenging perceptual situations comes with a cost of speech memory and comprehension because more cognitive resources are being allocated to speech perception. However, research shows that cognitive losses can be offset by text captioning in monolingual English speakers. Research also suggests that non-native speakers have more difficulty with speech perception. The purpose of this study is to determine whether or not text captioning effects immediate recall and delayed recognition of speech stimuli presented to non-native English speakers in varying levels of background noise. The participants were non-native English speaking young adults. Participants were presented with complex sentences in varying levels of background noise. In one condition, these auditory stimuli were paired with text captioning, and in the other condition, no text captioning was present. Immediate recall of these sentences was measured immediately after stimuli were presented, and delayed recall was tested after a block of forty-five trials. Text captioning was found to have a significant effect on both immediate recall and delayed recognition of speech stimuli in every background noise condition. Background noise was found to have an effect on immediate recall of speech stimuli. Correlation between language proficiency and immediate recall was found to be extremely high in every captioning X background condition. Moderate correlations were found between language proficiency and delayed recognition in most captioning X background noise condition. Such results imply that text captioning helps non-native English speakers with immediate recall and delayed recognition of speech stimuli and that background noise decreases immediate recall of speech stimuli in non-native English speakers.

## ACKNOWLEDGEMENTS

This work was supported by funding from the Undergraduate Research Opportunities Program at the University of Utah awarded to Vanessa Srivastava.

## TABLE OF CONTENTS

# INTRODUCTION

The world we live in is changing in significant ways due to globalization. According to the Associated Press, at least two-thirds of children in the world are raised speaking more than one language (Marian and Shook, 2012). Therefore, it is essential that we think about the effects of learning multiple languages on our cognitive abilities. The majority of research in psychology is conducted on American university students (Heinrich, Heine, & Norenzayan, 2010). There is a lot of research being conducted in other Western countries as well (Heinrich, Heine, & Norenzayan, 2010). However, much of the literature on non-native language speakers is research into bilingualism, not L2 language learning, which is the time during which people are acquiring a non-native language. There is some evidence that language comprehension in an L2 language is more difficult than a native language (Song and Iverson, 2018). One obvious but important component of language is listening comprehension, which is relatively understudied in the context of second language abilities. Therefore, it is important to study listening comprehension in L2 language learners and identify potential ways to increase comprehension. In the following study, we examined whether text captioning affects immediate recall and delayed recognition of auditory stimuli presented to non-native speakers in varying background noise conditions.

The task of speech comprehension for a non-native language is cognitively demanding for several reasons, and background noise non-surprisingly impairs performance. Mayo et al. (1997) and Rogers et al. (2006) conclude that non-native speakers never match the abilities of native speakers when there is background noise

(Lecumberri et al., 2010). Non-native language listeners have been found to be at a disadvantage when compared to native listeners in a few aspects. Several studies find that non-native listeners struggle more with word or sentence processing than do native listeners in the presence of noise, whether it be in fixed or variable noise levels (Lecumberri et al., 2010). Furthermore, as described in more detail later in the paper, noise increases the cognitive load required to attend to and comprehend stimuli. Non-native speech adds another layer of unintelligibility, thus further increasing the cognitive load (Lecumberri et al., 2010). There are many potential reasons for this additional unintelligibility. For example, non-native speech could be harder to comprehend because non-native speakers may be less familiar with the phonemes associated with the stimuli language. Because of the relatively small number of phonemes and large number of words seen in every language, many words are phonemically similar. There are different phonemes associated with different languages, and due to familiarity, it is more difficult to distinguish between phonemes that are not as common in a native language (Lecumberri et. al, 2010). Additional explanations can come from limitations in vocabulary and multiple characteristics of one's native language (Lecumberri et. al, 2010). Finally, non-native listeners have a harder time attending to one stimuli in the presence of multiple stimuli than native listeners when those stimuli interfere with low-level processing information, such as consonant sounds that aid in word predictability, which it seems that non-native speakers rely upon more heavily than their native speaking counterparts (Lecumberri et al., 2010). Taken together, research suggests that non-native English speakers have a more difficult time with the perception of auditory stimuli.

The comprehension of auditory stimuli is a complex process. It requires the detection of sound, the separation of the background noise from the meaningful information, understanding that the input is language, and processing the speech. Such a process requires significant cognitive demand, which is affected by effortfulness. Rabbitt's (1968) effortfulness hypothesis proposes that an increase in the effort required to perceive sensory input, such as speech, leads to depletion in cognitive processing resources, which, in turn, causes a decline in stimulus memory (Wingfield, Tun, & McCoy, 2005). Rabbitt conducted an experiment on young adults in which he confirmed this hypothesis. In the experiment, Rabbitt presented verbal information to participants in clear speech, then in background noise. He found that participants' memory for the first list of clear speech was worse when followed by unclear noise stimuli (Wingfield, Tun, & McCoy, 2005). This suggests that background noise interfered with the participant's ability to encode memory of the vocal stimuli. A number of studies have since found that increased effort in the perception of stimuli leads to weaker memory, including in adults with hearing loss (Wingfield, Tun, & McCoy, 2005). These studies show that, even if speech can be perceived in challenging perceptual situations, the perception comes with a cost of speech memory and comprehension because those cognitive resources are being allocated to speech perception.

However, several studies have shown that text captioning, written words on a screen that accompany speech, can improve the recognition of speech in background noise and hearing loss. Krull and Humes (2016) measured the effects of text captioning on speech recognition in monolingual, English speaking younger and older adults. They found that participants were best able to comprehend the stimuli when presented with both speech and text information (Krull & Hulmes, 2016). This implies that text captioning can help to offset the cognitive load caused by more effortful speech perception. In a study of monolingual adults, Payne et al. (2018) found that as signal-to-noise ratio declines (meaning that there is more background noise in relationship to the speech stimuli), so does immediate recall and delayed recognition of speech. However, text captioning was revealed to offset these declines (Payne et. al, 2018). These results suggest that text captioning alleviates the cognitive load of noisy speech comprehension even in relatively simple speech perception tasks (i.e., listening to one's own primary language). Additionally, the effects of SNR and text captioning were found to vary as a function of individual differences in adults with hearing loss (Payne et. al, 2018). Hearing loss would increase effortfulness in speech processing, as would non-native listening, suggesting that these effects may also vary along the lines of language proficiency.

The effects of text-captioning on understanding speech in noise have only been tested in monolingual adults, and these effects are not understood in terms of second language captioning. Research suggests that when non-native speakers use one language, the other language is simultaneously active (Marian and Shook, 2012), which could mean that bilingual speakers are applying a higher cognitive effort than their monolingual counterparts when using either language. More recent research further implies that speakers must work harder to listen to speech in a non-native language. This suggests that encoding speech in a non-native language is more effortful (Song and Iverson, 2018).

This information led to the following hypotheses: First, I hypothesized that immediate recall and delayed recognition would decline with lower signal to noise ratios

(SNR) in non-native English speakers. There were two potential effects that text captioning may have had on this decline. It was possible that text-captioning would improve speech memory in non-native English speakers, as seen in monolingual English speakers (Krull & Hulmes, 2016). If this was the case, we would expect the text-captioning to offset the declines caused by the lower SNR. It was also possible that the cognitive effort required to both listen and read in a non-native language would be too overwhelming for non-native English speakers, in which case captioning would not help to offset the decline.  Finally, I hypothesized a positive correlation between language proficiency and recall accuracy, as well as language proficiency and recognition accuracy.

## METHODS

### *Participants*

Nineteen non-native English speaking participants were recruited for this study. Of those nineteen, eighteen participants completed the study, with one stopping midway due to frustration with the difficulty of the study.  The average age of the participants was twenty-two years old, with the youngest participant being eighteen years of age and the oldest participant being twenty-seven years of age. All of the participants identified as male or female gender (8 male, 11 female). Participants varied significantly in the number of years they had been speaking English (M = 13.63 years, range = 4-27 years). The majority of participants were students at the University of Utah and therefore spoke English at a college level.  Participants were recruited both through the University of Utah psychology participant pool and by word of mouth.  Participants recruited through the participant pool were given class credit for their time spent participating in the study. Those participants who were not enrolled in psychology courses were compensated at $10.00 per hour for their time spent in the study.

*Materials and Design:* Participants were presented with two blocks of forty-five trials of speech stimuli in background noise conditions. In one block, they were presented with only auditory stimuli, and in the other they were presented with auditory stimuli and text captioning. Speech stimuli were complex, propositionally dense eighteen word sentences taken from National Geographic and similar magazines. Speech stimuli were taken from previous experiments conducted by Payne and Stine-Morrow (Stine-Morrow et al., 2001; Payne & Stine-Morrow, 2017). The sentences were recorded by a female native English speaker using a cardioid USB microphone (frequency response: 20Hz - 20kHz; sample rate 48 kHz, Max SPL: 120 dB) in a quiet environment. The auditory stimuli were recorded and digitized at a sampling rate of 44.1 kHz. The silent periods preceding and following the target sentence were removed from the recordings. Auditory stimuli were equated on root mean square amplitude. The noise masker was a stationary speech-shaped noise with the long-term frequency spectrum of the speech.

There were three different noise conditions: Quiet, 7 dB SNR (mild background noise), and 3 dB SNR (moderate background noise). These noise levels were chosen because they were determined to increase listening effort while maintaining intelligibility during pilot studies. The noise conditions were randomly assigned to an even number of trials in each block, so that there were six categories of caption/no caption X Quiet/7 dB SNR/3 dB SNR for a two by three within-subjects design.

*Procedures, Measures, and Apparatus:* Before beginning the main experiment, participants gave informed consent and filled out a demographics sheet asking about their age, ethnicity, vision and hearing problems, language background, and mental health history. Participants were administered two language tests. The first was the Extended Range Vocabulary test created by the United States Educational Test Services (ETS), which takes up to six minutes. The second is the Michigan English Language Institute College English Test (MELICET), which is untimed. In order to quantify language proficiency for the correlational models, we calculated participants' z-scores on each the MELICET grammar, MELICET cloze (semantics), and ETS Extended Range Vocabulary test, and we averaged them together. Additionally, eye tracking data was collected from the participant, but we did not use it in this study, so we will not discuss it further. Auditory stimuli were presented to participants at 60 dB through a MA-41 audiometer via the auxiliary input. The participants heard the stimuli through IP-30 insert air-conduction earphones. During the text captioning condition, captions were presented in segments designed to simulate the experience of Internet Protocol Captioned Telephone Service users. In order to keep this consistent, speech was presented in three word chunks with all text remaining on the screen until 1000 ms after the end of the spoken sentence. The rate of each block was based on the timing of the auditory stimuli so that the text could not precede the auditory stimuli. After that 1000 ms, a cross appeared on the screen for 5000 ms, after which a recall cue was presented. This recall cue prompted participants to recall as much of the sentence as they could remember out loud. The accuracy of this response was considered the immediate recall task. Their response was recorded with a microphone and saved for future transcription. At the end of the forty-five trial block, the participant was given a survey to determine their memory for the sentences and a survey indicating the difficulty of the task. The accuracy of this response was considered the delayed recall task. After an optional break, the computer task and the following surveys were repeated in a second block, varying only in that the sentences were different and whether or not text captioning appeared with the speech stimuli.

## RESULTS

In order to analyze the effects of text captioning on immediate recall and delayed recognition of auditory stimuli, we conducted two 2x3 repeated measures factorial Analyses of Variance (ANOVAs). The first ANOVA was designed to determine the effects of text captioning on immediate recall of auditory stimuli in three different noise conditions; no background noise, mild background noise (7 dB SNR), and severe background noise (3 dB SNR). The second ANOVA was designed to determine the effects of text captioning on delayed recognition of stimuli in those same three noise conditions. For each of the two ANOVAs, we looked at the effect of captions, the effect of SNR, and the interaction effect between the two variables. We used a significance level of $p$=0.05.

The first ANOVA determined the effects of text captioning and SNR on immediate recall of speech stimuli. Captioning was found to have a significant main effect on immediate recall with $F(1, 17) = 28.78$, $p < .05$, *partial* $\eta^2 = .63$. Posthoc pairwise comparisons allowed us to determine that the mean difference between the caption and no caption condition was $t\ (17) = 5.79$, $p<.05$. SNR was also found to have a significant main effect on immediate recall with $F(2,17) = 10.63$, $p < .05$, *partial* $\eta^2 =$ .57. Posthoc pairwise comparisons allowed us to identify significant differences between each of the noise conditions. The mean difference between the Quiet and the 7 dB SNR was $t$ (17) =4.35, $p$<.05.  The mean difference between the Quiet and the 3 dB SNR was $t$ (17)=1.89, $p$<.05.  The mean difference between the 7dB SNR and the 3dB SNR conditions was $t$ (17) =-2.46, $p$<.05.  The interaction between captioning and SNR was not found to have a significant effect on immediate recall with $F$(5,17) = 1.43, $p$>.05, *partial* $\eta^2$ = .15.

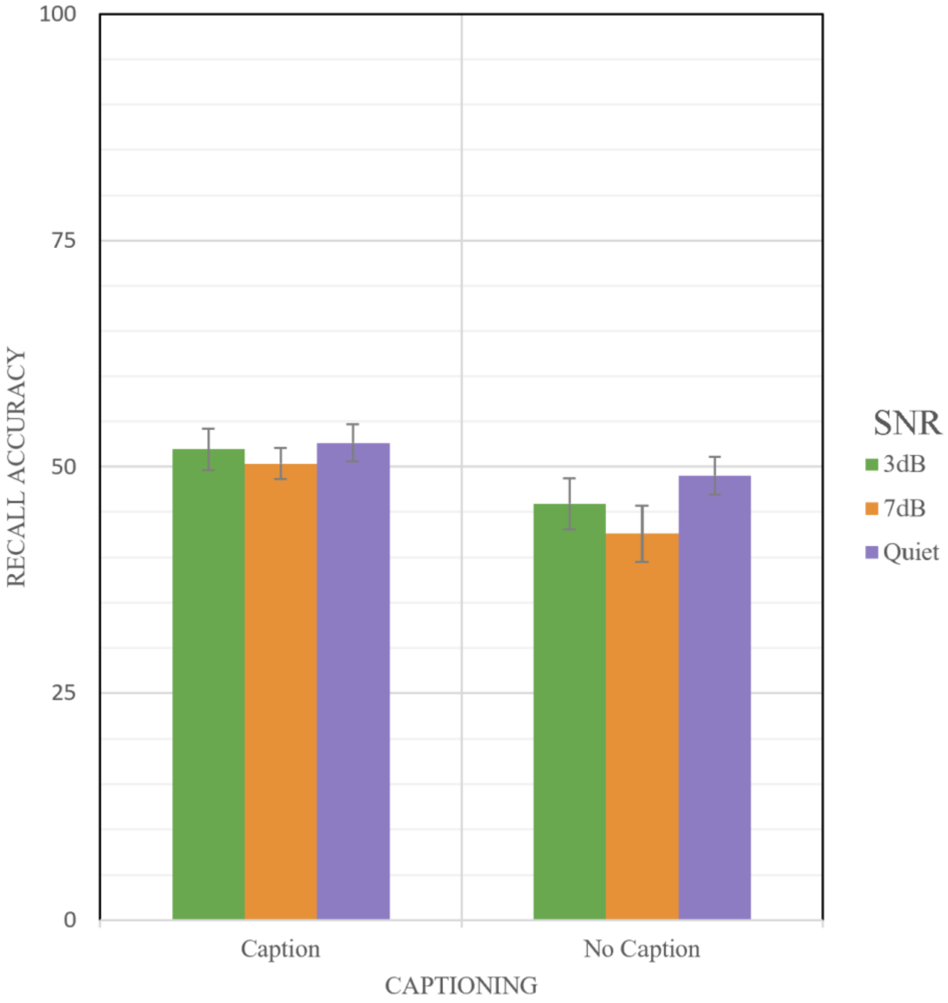

*Figure 1: Immediate Recall Accuracy as a function of captioning and SNR*

The second ANOVA determined the effects of text captioning and SNR on delayed recognition of speech stimuli.  Captioning was found to have a significant effect on delayed recognition of stimuli with $F($1,17) = 10.36, $p$ <.05, *partial* $\eta^2$ = .38. Posthoc pairwise comparisons allowed us to determine that the mean difference between the caption and no caption condition was $t$ (17) = 4.1, $p$<.05. SNR was not found to have a significant effect on delayed recognition of stimuli with $F$(2,17) = 2.21, $p$>.05, *partial* $\eta^2$ =.22. Additionally, the interaction between captioning and SNR was not found to have a significant effect of delayed recognition of stimuli with $F$(5,17) = .08, $p$>.05, *partial* $\eta^2$ = .01.

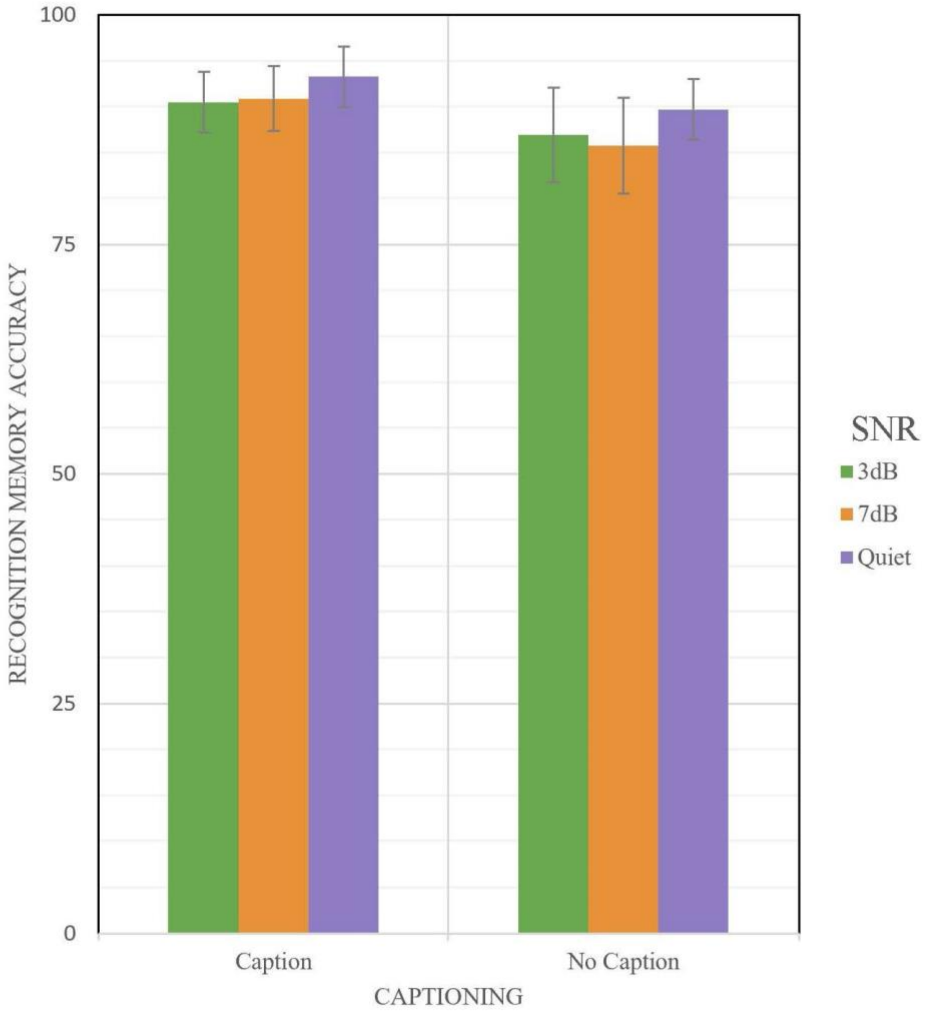

*Figure 2: Delayed Recognition Accuracy as a function of captioning and SNR*

In order to analyze the effects of language proficiency on immediate recall and delayed recognition of auditory stimuli, we have looked at the Pearson's two-tailed correlations of the average z-score score of the three language tests with each immediate recall scores and delayed recognition scores in each of the six captioning X SNR categories.

**Table 1**

***Correlation between Language Proficiency and Immediate Recall***

| **Captioning X SNR** | **r** | **p** |
|---|---|---|
| **Cap X Quiet** | .83 | <.05 |
| **Cap X 7dB SNR** | .85 | <.05 |
| **Cap X 3dB SNR** | .80 | <.05 |
| **No Cap X Quiet** | .87 | <.05 |
| **No Cap X 7dB SNR** | .88 | <.05 |
| **No Cap X 3 dB SNR** | .86 | <.05 |

Table 1

There is a significant strong correlation between language proficiency and immediate recall in all six of the captioning X SNR conditions tested in this study.

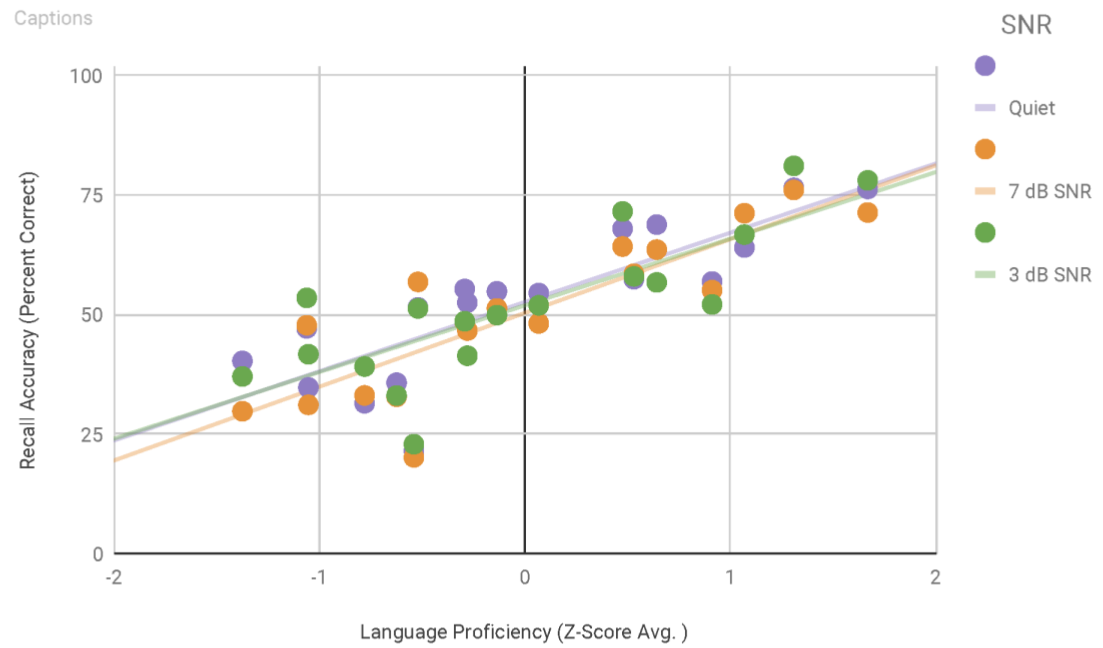

*Figure 3: Correlation between Language Proficiency and Immediate Recall of Speech with Captions*

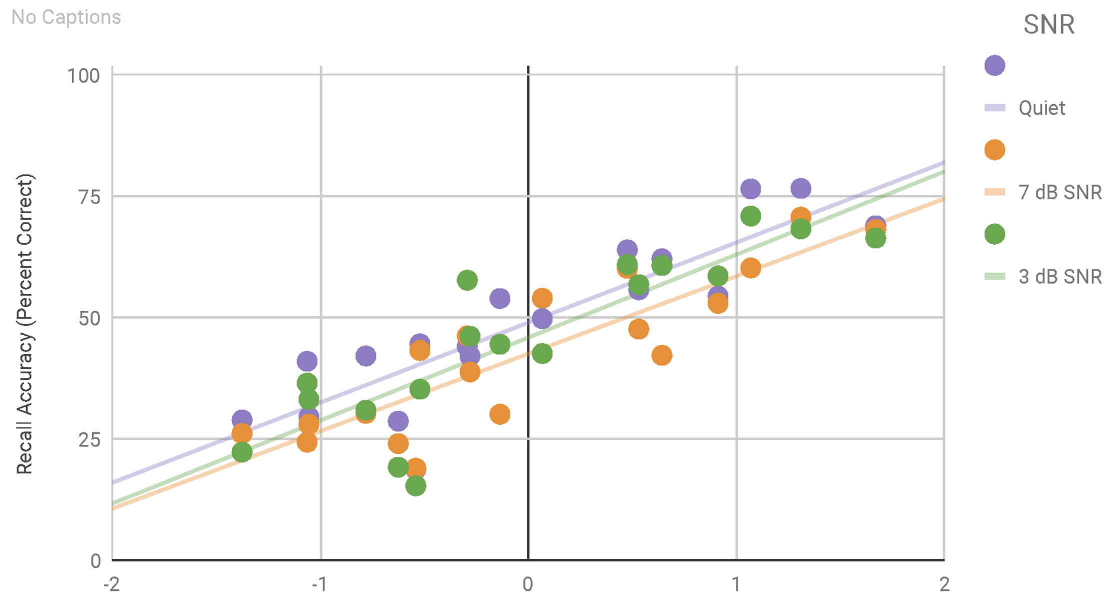

*Figure 4: Correlation between Language Proficiency and Immediate Recall of Speech with No Captions*

**Table 2**
***Correlation between Language Proficiency and Delayed Recognition***

| **Captioning X SNR** | **r** | **p** |
|---|---|---|
| **Cap X Quiet** | .65 | <.05 |
| **Cap X 7dB SNR** | .42 | >.05 |
| **Cap X 3dB SNR** | .28 | >.05 |
| **No Cap X Quiet** | .48 | <.05 |
| **No Cap X 7dB SNR** | .66 | <.05 |
| **No Cap X 3 dB SNR** | .54 | <.05 |

Table 2

There are significant moderate correlations in the Caption X Quiet condition, No Caption X Quiet condition, No Caption X 7dB SNR, and No Caption X 3dB SNR. There are no significant correlations between language proficiency and delayed recognition of speech stimuli in the categories that have both captions and background noise.

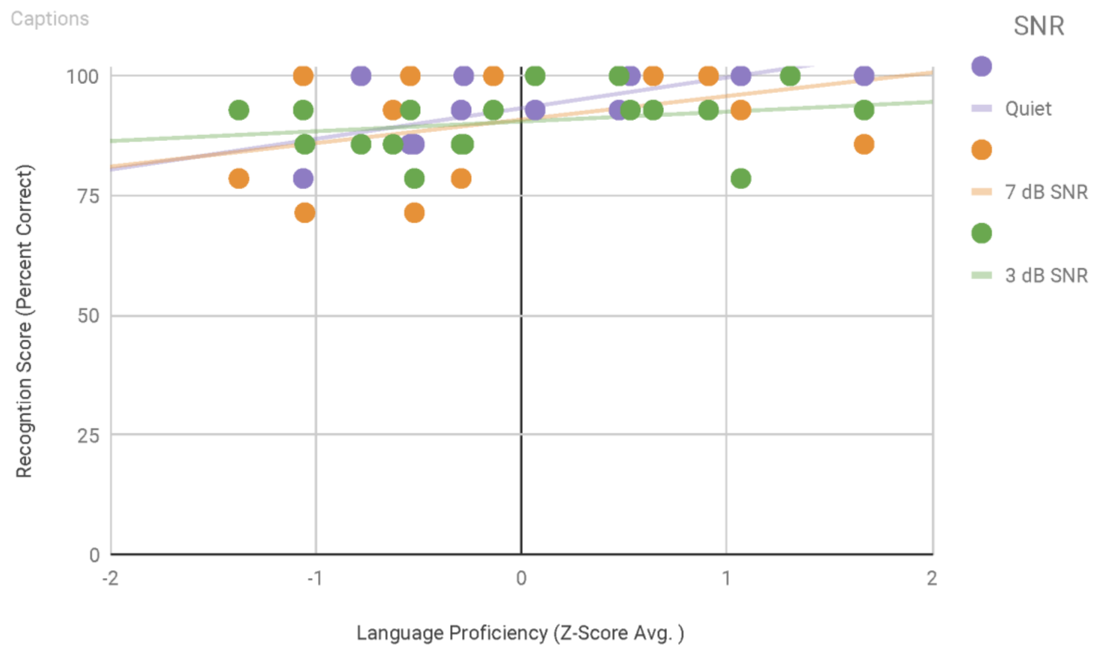

*Figure 5: Correlation between Language Proficiency and Delayed Recognition of Speech with Captions*

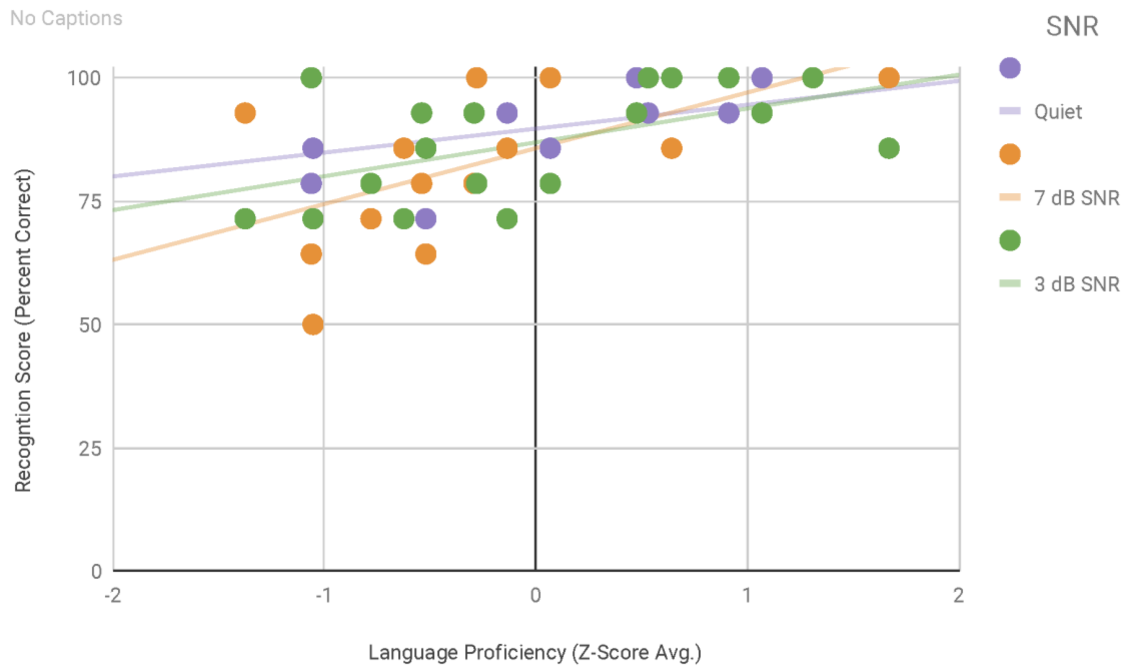

*Figure 6: Correlation between Language Proficiency and Delayed Recognition of Speech with No Captions*

# DISCUSSION

Much of the underlying research for this study was based on Rabbitt's (1968) effortfulness hypothesis, which suggests that the more difficult it is to perceive sensory input, such as speech, the fewer cognitive resources there are available to process the sensory input, leading to a decline in stimulus memory (Wingfield, Tun, & McCoy, 2005). Based on this effortfulness hypothesis and results from former studies, we hypothesized that non-native English speakers' immediate recall and delayed recognition would both decline with lower SNRs. We know, based on former research, that speech perception is already more difficult for participants who are non-native language speakers (Song & Iverson, 2018; Lecumberri et. al, 2010). Prior research conducted on monolingual adults also suggests that background noise makes auditory perception more difficult (Wingfield, Tun, & McCoy, 2005). In fact, we even know that background noise makes speech perception more difficult in non-native English speakers (Lecumberri 1. et. al, 2010). A similar study conducted on monolingual adults yielded results in line with this; in Payne's study, participants' immediate recall and delayed recognition of auditory stimuli was hindered in background noise (Payne et. al, 2018). The results in this study, however, did not fall in line with the results of former research. Background noise was found to affect immediate recall as predicted, but was not found to have a significant effect on delayed recognition. This result was unexpected and was possibly due to the relatively small sample size of our study. However, it is possible that the effects of background noise are more detrimental to immediate recall than delayed recognition in non-native English speakers. This could be due to the levels of difficulty for each task. As described in the methods of this study, immediate recall was calculated by the percentage of words participants replicated correctly when asked to repeat each sentence verbatim almost immediately after the target sentence was presented. The delayed recognition task only required participants to distinguish between old sentences and semantically similar new sentences. Such a task would require less resources from working memory, and therefore could be less impacted by increased effortfulness.

We also hypothesized that text captioning could either attenuate the difficulties associated with the presence of background noise or that the cognitive effort associated with both reading and listening in a non-native language would be overwhelming for participants, which would lead to a decline in performance. The results suggested that text captioning had a positive effect on speech perception. In this case, we expected that text captioning would improve both immediate recall and delayed recognition accuracy because such results were seen in former studies conducted on monolingual adults (Krull & Holmes, 2016; Payne et. al, 2018). Text captioning was found to have a statistically significant positive effect on both immediate recall and delayed recognition of speech stimuli. However, there was not a statistically significant interaction effect between text captioning and SNR as observed in monolingual samples. This could be due to the fact that non-native speakers benefitted significantly from text captioning, even in the quiet background noise condition.

We decided to test for correlation between language proficiency and performance in both immediate recall and delayed recognition of speech stimuli because of the wide range of language proficiency scores found amongst participants. There were some interesting trends in the correlations between language proficiency and recall of speech stimuli, as well as language proficiency and recognition memory. It is interesting to note how high the correlation between language proficiency and immediate recall accuracy is regardless of SNR or text captioning. Prior research has shown a positive correlation between both L2 language proficiency and L2 language comprehension and between L2 language proficiency and working memory (Jeon and Yamashita 2014). The nature of the immediate recall task, described earlier in the paper, clearly relates strongly to both language comprehension and to working memory. These relationships could explain the high correlation between language proficiency and working memory seen in our results. In both immediate recall and delayed recognition, we see a trend of lower correlations between language proficiency and accuracy with text captioning than without it. This could suggest that language proficiency is less important when captioning is present for the recall and recognition of speech stimuli. This very clearly supports our hypothesis that text captioning could help with the perception of auditory stimuli and refutes our hypothesis that text captioning could overwhelm non-native speakers and detriment their perception of auditory stimuli. It is possible that text captioning helps to relieve the difficulties of speech perception in non-native speakers because some of the added difficulties of non-native speech perception come from auditory difficulties, such as accent or phonemic similarities, and the ability to read the same stimuli reduces that confusion.

It is important to note that the target sentences used for both immediate recall and recognition memory may have been culturally biased in favor of those individuals who had been in the United States for the longest amount of time. I noticed often when coding the data that those participants who did not do as well overall would also struggle to remember popular historical figures like George Washington, Teddy Roosevelt, or Jane Goodall. It would likely be significantly harder for an individual to remember a name he or she had not grown up hearing than one they knew very well. Although only six of the sentences have names of famous figures discussed in the American school system, it is possible that there are additional cultural biases that are less obvious.

Despite the many positive results, there are important limitations to this study that must be considered. There were considerable individual differences in language proficiency amongst our sample that contributed to our results in significant ways. However, with a larger sample size, it would be possible to get a clearer understanding of how L2 language learners process auditory stimuli as a more collective group. It is possible that the differences between our results and the expected results based on prior research could be due to the relatively small sample size of this study. Additionally, every participant in the study was either a student, an employee or a former student of the University of Utah. This likely limited the diversity of the participants based on their education level, and therefore could impact the generalizability of the study. Some of the participants in this study were personal friends of the experimenter. This could have made participants either more nervous (because they were performing the study in front of someone they knew) or more comfortable for the same reason. This could also have affected some of the participants' scores in positive ways (feeling comfortable) or negative ways (feeling nervous).

There are many directions in which this research may extend in the future. Future research could benefit by examining a larger section of the population in this study. A larger sample size could allow researchers to separate the effects of individual differences from the ways in which SNR and text captioning affect L2 English speakers overall. Along similar lines, it could be interesting to conduct a similar experiment on groups of individuals in different periods of the language learning process. For example, L2 language learners could be separated from balanced bilingual speakers and non-balanced bilingual speakers who are even stronger in their non-native language. Finally, it could be fascinating to directly compare the ways in which non-native English speakers' perception of speech stimuli differ from the ways in which native English speakers perceive the same stimuli in various background noise conditions, as well as with and without text captioning.

The results of this study suggest that text captioning improves immediate recall and delayed recognition of speech regardless of whether or not background noise is present. Additionally, results suggest that the presence of background noise decreases immediate recall of speech stimuli. Finally, language proficiency correlates extremely highly with immediate recall of speech stimuli and can have moderate correlation with delayed recognition of speech stimuli depending on the captioning X SNR condition.

These results have important real world applications. In a classroom setting, for example, it may be helpful for non-native English speaking students to have word-heavy PowerPoint presentations to help them follow along with a lecture. Additionally, it would be helpful to minimize background noise as much as possible in classroom and workplace settings. In a shared cultural setting, such as a movie theater, it could be beneficial for L2 language learners to have the option of seeing a movie with text captioning. We have seen the use of text captioning help monolingual English speakers with hearing loss in both research and in real world settings (Payne et.al, 2018; Hoffman

& Warner, 2016). The results of this study may encourage us to expand text captioning to non-native English speaking populations as well in an attempt to improve L2 learning outcomes.

# REFERENCES

Heinrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 61-135.

Hoffman, C., Warner, M. (2016) *United States Patent No. USD775653S1*

Jeon, Eun Hee, and Junko Yamashita (2014). L2 Reading Comprehension and Its Correlates: A Meta-Analysis. *Language Learning 64(1)*, 160–212.

Krull, V., & Hulmes, L. E. (2016). Text as a Supplement to Speech in Young and Older Adults. *Ear and Hearing*, 164-176.

Lecumberri, Maria, Garcia, Luisa, Cooke, Martin, & Cutler, Anne (2010). Non-Native Speech Perception in Adverse Conditions: A Review. *Speech Communication 52(11). Non-Native Speech Perception in Adverse Conditions*, 864–886.

Marian, V., & Shook, A. (2012, September-October). The cognitive benefits of being bilingual. *Cerebrum*.

Payne, B.R., Silcox, J., Lash, A., Ferguson, S.H., & Lohani, M. (2018). Assistive text captioning offsets the effects of background noise on speech memory. *Academy of Rehabilitative Audiology Institute*.

Payne, B.R., Stine-Morrow, E. A. L., (2017) The Effects of Home-Based Cognitive Training on Verbal Working Memory and Language Comprehension in Older Adults. *Frontiers in Aging Neroscience (9)*, 1-20.

Song, Jieun & Iverson, Paul (2018) Listening Effort during Speech Perception Enhances Auditory and Lexical Processing for Non-Native Listeners and Accents. *Cognition 179*, 163–170.

Wingfield, A., Tun, P. A., & McCoy, S. L. (2005). Hearing Loss in Older Adulthood: What it is and How it Interacts with Cognitive Performance. *American Psychological Society*, 144-148.