Running head: IMPACT OF TALKER VARIATION

Impact of Talker Variation in Following Context on Resolving Lexical Ambiguities in Speech Recognition

Grace Park

Department of Psychology, Ohio State University

Abstract

Studies have shown that when a word in a sentence is hard to recognize, a listener relies on context occurring subsequent to the word in order to accurately identify the word. For example, when someone talking on the phone said *The wing had feathers* but the listener only heard *The *ing had feathers* due to cell-phone interference, the listener will tend to focus on following context (e.g., *feathers*) to understand it was the "wing" that "had feathers" rather than "thing" or "ring". The current study explores whether a listener will rely on following context if the target word (e.g., *wing*) was spoken by one talker and the following word (e.g., *feathers*) was spoken by another talker. Results show that the talker change neither weakens nor strengthens the influence of subsequent context on ambiguity resolution.

Keywords talker variation, subsequent context, ambiguity resolution, phonemic restoration, spoken word recognition

Impact of Talker Variation in Following Context on Resolving Lexical Ambiguities in Speech Recognition

Lexical ambiguity refers to a speech string that has more than one possible interpretation and therefore, requires the appropriate meaning to be selected. For example, use of *bank* in the sentence *My father went to the bank* can be interpreted as a financial service institution or the edge of a river. Such lexical ambiguities occur frequently in speech and various noises can easily mask a critical phoneme (e.g., /k/, /b/, /t/). Yet, listeners can resolve ambiguities immediately. How do listeners resolve such lexical ambiguity? One method to ambiguity resolution is utilizing information that surrounding context provides. For instance, if a talker said "He was driving a nice car" and noise interference caused a listener to hear the sentence as "He was driving a nice *ar (* indicating a phoneme covered by noise)", despite the fact that there are several lexical competitors (i.e., car, bar, far) for an ambiguous word, the listener almost immediately chooses the word "car", which is semantically congruent with "driving".

Studies demonstrate that contextual information, especially preceding information, has a significant influence on ambiguity resolution (Marslen-Wilson, 1987; Samuel, 1981; Van Alphen & McQueen, 2001). For instance, Samuel (1981) showed that preceding context biases a listener in the decision making of an appropriate word. For example, listeners could recognize an ambiguous word more accurately in the sentence "the soldier's thoughts of the dangerous *battle* made him very nervous" rather than in the sentence "the soldier's thoughts of the dangerous *batter* made him very nervous".

Compared to prior context, the influence of subsequent context (e.g., "the *ar was towed yesterday") has been studied far less. When disambiguating information follows the ambiguous word, a listener has to either wait until additional information becomes available or resolve the ambiguity when it occurs. If the perceptual system waits until disambiguating information is provided, then the chance of identifying an ambiguous word correctly increases. On the other hand, if the perceptual system does not wait and identifies an ambiguous word before the disambiguating information is available, processing would take longer if the selection turns out to be incorrect (e.g., *bar* instead of *car*) because then the listener would have to analyze the sentence once more in order to find the correct word (Szostak & Pitt, 2010).

Warren and Sherman (1974) support the first solution and further suggest that the perceptual system stores the available phonetic information from the ambiguous word (e.g., *ar* from **ar*), delaying the resolution process until additional information is provided to identify the correct word. However, in such a case, several factors (e.g., number of syllables between the ambiguous word and the following disambiguating information) could increase the memory load of a listener and could interrupt subsequent disambiguating information from helping ambiguity resolution. For instance, Connine, Blasko, and Hall (1991) suggested that there is a time limit of approximately one second for subsequent disambiguating information to influence the resolution process. In other words, if the disambiguating information is not provided within one second after the ambiguity is encountered, then the perceptual system processes the resolution without the additional information.

Building on the study by Connine et al. (1991), Szostak and Pitt (2010) studied the influence of subsequent context in resolving ambiguous words. Their study explored whether increasing amount of information between an ambiguous word and the disambiguating word

would weaken the strength of subsequent context on ambiguity resolution. They presented the sentences in which the initial phoneme of the target words were either superimposed with noise or replaced by noise. Each sentence contained a target word, the ambiguous word (e. g., *ing; which * represents the phoneme covered by the noise), followed by a disambiguating word. The disambiguating word was either congruently related to the ambiguous word (e.g., *the *ing had feathers*) or congruently related to a competing word that only differed in the initial phoneme from the target word (e.g., *the *ing had diamond*). Then they manipulated the number of syllables between an ambiguous word and a disambiguating word, either one syllable (near condition) or six to eight syllables (far condition) (e.g., near condition: *the *ing had feathers* versus far condition: *the *ing had an exquisite set of feathers*). Participants had to determine whether the initial phoneme was covered by the noise or replaced by it.

Szostak and Pitt (2010) predicted that if there is a time limit for the perceptual system to wait until disambiguating information becomes available in order to resolve the ambiguity, the influence of subsequent context in the near condition should be larger than that of the far condition. The study found as they predicted. As number of syllables between an ambiguous word and a disambiguating word increased the influence of the context on ambiguity resolution decreased. Szostak and Pitt (2010) concluded that the reason subsequent context effect diminished is that additional information between an ambiguous word and a disambiguating word produced memory interference that prevented listeners' utilization of the subsequent context.

The results of the above study suggest that memory interference reduces the processing capacity availability. In other words, the perceptual system had to work harder to find the appropriate meaning when the distance between an ambiguous word and a disambiguating word

increased. A possible factor that is as influential as distance is a change in talker. Newman and Sawusch (2009) suggest that generally, the perceptual system perceives the spoken language sources and assigns them to different streams according to the talker that produced the source, because identifying the talker helps a listener to separate one sentence from another. By this account, changing the talker of the disambiguating word is predicted to have a similar effect as increasing the number of syllables between the ambiguous word and disambiguating word. In other words, changing the talker will decrease the influence of subsequent context on ambiguity resolution.

The current experiment investigated how much participants rely on the subsequent context when the talker changes. More specifically, the experiment studied the influence of subsequent context on ambiguity resolution when an ambiguous word and a disambiguating word in a sentence were produce by different talkers. In the experiment, the sentence "The wing had feathers" was divided and produced by two talkers: "The wing had" was spoken by a female talker and "feathers" was spoken by a male talker. Gender was the only difference between two talkers, because it is the most notable property that distinguishes one talker from another among several acoustic attributes (e.g., age, dialect, speed of speech) (Creel, Aslin & Tanenhaus, 2008).

I used the phonemic restoration paradigm (Samuel, 1981) in order to learn how well participants can discriminate the phoneme from the white noise without response biases. Phonemic restoration refers to an auditory illusion that a listener hears the sound that is not present. The greater restoration would lead to the better ambiguity resolution. I predicted that talker variation would reduce the strength of subsequent context on ambiguity resolution. In other words, the subsequent context information would be less likely to be used in the two talker condition than in the single talker condition. If this result was found, it would suggest that talker

variation increases memory load to the cognitive system and interrupts the identification process of an ambiguous word.

Method

Participants. Twenty-four undergraduate students from the Ohio State University participated in the study and received course credit in an introductory psychology class. Prior to the study, they reported being native English speakers with no hearing disorders or impairments.

Stimuli. Stimuli consisted of 24 monosyllabic target words (e.g., *wing*, *lip*, *mole*) that were consonant –vowel –consonant (CVC). All initial phonemes of target words were either masked by or replaced with the white noise to create ambiguity. The reason initial phonemes were manipulated was that any disruption at the beginning of a word is more noticeable than any other position (Samuel, 1981). For the initial phoneme of the target word, three phonetic categories, liquids (e.g., /r/ and /l/), nasals (e.g., /n/ and /m/), and glides (e.g., /w/ and /y/) were used. A study of Samuel (1981) showed that out of five phoneme classes (stop, fricative, liquid, nasal, and glide), listeners could not differentiate stop (e.g., /p/ and /t/) and fricative (e.g., /f/ and /th/) from white noise. The target words had at least one rhyme competitor (e.g., *ring* is a rhyme competitor of *wing*). Also the target words without the initial phoneme could not be a word; therefore, participants would not be able to identify an ambiguous word by itself (e.g., *ing* cannot be a word without *w*). All sentences began with a determiner (e.g., the) followed by the target word (e.g., *wing*) and verb (e.g., *had*) and ended with the disambiguating word (e.g., *feathers*). Based on the results of Szostak and Pitt (2010), all sentences had a single syllable between the target word and the disambiguating word in order to strengthen the influence of subsequent

context. The sentences were created as either congruent, where a disambiguating word was semantically congruent with the target word (e. g., *the wing had feathers*), or incongruent, where the disambiguating word was semantically incongruent with the competitor of the target word (e. g., *the wing had diamonds*). Appendix A contains the full list of the stimuli used in this experiment.

Each of the 24 target words was placed in eight different types of sentence that varied in number of talkers (single talker versus two talkers), congruency (congruent versus incongruent), and condition (added versus replaced) for a total of 192 items. Two talkers, female and male, recorded the sentences. They both were American English speakers born and raised in Central Ohio and similar in age. Prior to recording, talkers were instructed to articulate each sentence slowly and properly. Each talker recorded the sentences in a sound-dampened room.

The sentences recorded by a female talker were created and white noise was either added to the initial phoneme of the target word to create the "added" stimuli or replaced to create the "replaced" stimuli. Once editing was finished, two copies of each sentence were created: one was used as the single talker condition and the other was further manipulated to be used as the two talker condition. The male talker also created the same sentences and the disambiguating words (e.g., *feathers*) from each sentence were spliced out and replaced with the disambiguating words of the female talker (e. g., *The wing had* spoken by a female talker combined with *feathers* spoken by a male talker). This cross-splicing technique was performed to ensure that information in the sentence fragment, including the target word, was identical across conditions. Care was taken to ensure the portion altered in the target word was identical across the four conditions. For instance, the length and speech speed of the disambiguating word "feathers" from the sentence *the wing had feathers* was the same in the "added" and "replaced" sentence.

Procedure. The experimental design was within subjects, meaning each subject participated in all four conditions. The experiment took place in a sound-attenuated room that can hold up to four participants at a time. Each participant was provided with headphones attached to a computer as well as a box with two buttons labeled "added" and "replaced". Participants were asked to carefully listen to each sentence over headphones and decide whether the initial phoneme of the ambiguous word was present or not. After the decision, they were instructed to press the "added" button for the initial phoneme present sentence and "replaced" button for initial phoneme not-present sentence. It was emphasized that the participants should wait until the sentence had completely finished before providing a response.

Prior to the experimental trials, 24 practice trials were presented for participants to become familiar with the experiment, and a short break was offered after 96 trials. The experiment lasted approximately 40 minutes.

Results

I predicted that if the talker variation affected the strength of influence of subsequent context in ambiguity resolution, accuracy in performance will decrease in the case of a sentence spoken by two talkers compared to sentences spoken by a single talker. One explanation for this is that alternation of talker adds an extra load on perceptual system, because listeners have to adjust to a different talker to resolve sentence ambiguity.

Three of twenty four participants were removed from analysis because in at least two conditions, they were suspected that they misunderstood the instruction and responded "replaced" to added stimuli and "added" to replaced stimuli. The results suggest that the

prediction was incorrect. Figure 1 shows the hit rates and false alarm rate across the number of talkers and congruency conditions. *Hit rate* indicates the proportion of added response divided by the total number of added stimuli presented [hit counts / (hit counts + miss counts)]. The *False alarm rate* indicates the proportion of false alarm responses divided by the total number of replaced stimuli presented [false alarm counts/(false alarm counts + correct rejection counts)]. The hit rate and false alarm rate were analyzed independently. If the prediction was correct, hit rate of the two talker congruent condition would be as equally low as the single talker incongruent and the two talker incongruent condition, indicating the talker change effect prevent the subsequent context from helping ambiguity resolution. In contrast, hit rate of the two talker congruent condition (0.885) is as high as that of the single talker congruent condition (0.891). Appendix B provides the mean counts of hit, miss, false alarm, and correct rejection across the conditions.

The mean hit rates in the single talker condition in Figure 1 (left side of bars) show strong evidence of a congruency effect. The hit rate in the two talker congruent condition is larger than that of the two talker incongruent condition, yielding a congruency effect of 0.19. A congruency effect was also found between the single talker congruent condition and the single talker incongruent condition, which was 0.16. As the results show, the congruency effect in the two talker condition was 0.03 larger than that in the single talker condition. An Analysis of Variance (ANOVA) with number of talkers and congruency as the two factors yielded a reliable main effect of congruency, $F(1, 20) = 13.264$, $p < .05$. However, a reliable main effect was not found for talker, $F(1, 20) = 1.565$, $p = 0.225$. The interaction did not approached statistical significance, $F(1, 20) = 0.964$, $p = 0.338$. The drop in mean hit rate from the single talker

incongruent condition to the two talker incongruent condition (0.03) was slightly bigger than that in the single talker congruent condition and the two talker congruent condition (0.01).

On the other hand, inspection of the false alarm rates showed hardly any congruency effect regardless of whether there were two talkers or one. A two-way ANOVA yielded no main effect of congruency ($F(1, 20) = 0.001$, $p = 0.976$), main effect of talker ($F(1, 20) = 0.315$, $p = 0.581$), nor an interaction between two ($F(1, 20) = 1.79$, $p = 0.196$).

Discussion

This experiment explored whether a change in talker will weaken the effects of subsequent context in resolving ambiguous words. The result showed comparable congruency effects across the single talker condition and the two talker condition. However, the current experiment does not show that the talker variation effect as reliable.

There are several possible causes that lead to the unreliable result. One hypothesis is that listeners opted for a strategy that ignores the talker later in the experiment because they found that the change of talker does not provide helpful information in the resolution. This strategy has been found in other studies. For example, Clopper and Pisoni (2004) found that listeners begin to opt for the strategy that ignores talker change. In the study they separated participants into two groups and trained them with six American English dialects. One group received the sentences spoken by a single talker for each dialect (total of six talkers) and the other group the sentences spoken by three talkers for each dialect (total of eighteen talkers). Then participants were asked to categorize novel sentences according to their dialect. The results showed that the single talker group performed better with a sentence spoken by previously introduced talker, while the three talker group outperformed on a sentence spoken by a new talker. Clopper and

Pisoni (2004) concluded that the three -talker group focused on the distinction among six dialects rather than the talkers. They suggest that throughout the trials, the participants in the three-talker group developed strategies that help them to encode the common properties among the three talkers and ignore the properties of individual talkers. The result suggests that when participants are primed with the multiple talkers, they may develop a strategy to ignore difference properties among talkers and focus more on important factors such as context. In order to examine whether the above strategy was used in the current experiment, the results of the current experiment were divided into the beginning and last sections and further analyzed. If the listeners developed the strategy, the main effect of the talker change effect would be found in the first half of the experiment, while no main effect would be present in the second half of the experiment. A three-way ANOVA yielded a reliable main effect of congruency, $F(1, 20) = 13.433$, $p < .02$, but reliable main effect was not found for the first versus second sections, $F(1, 20) = 0.538$, $p = 0.472$. The result indicates that listeners did not develop any strategy that ignores the change of talker.

Another possibility is that the participants did not have sufficient time to be influenced by the talker change effect during ambiguity resolution. At the beginning of the current experiment, participants were instructed to respond as soon as a sentence ended. Requiring a response before a listener recognizes the change of talker might lead to an unreliable talker effect. McLennan and Luce (2005) showed that duration between the stimuli and response influences the talker effect. In their study, listeners heard numerous stimulus words and were asked to repeat the stimuli. Some of the words were presented in the same voice several times, while others were produced by different voices. When the listeners repeated the stimuli words immediately, no talker specific effect was found. However, when the listeners delayed response by 150

milliseconds after the word was given, they repeated faster and more accurately for the words that were produced by one voice several times. The result indicates that time is required for listeners to be primed by talker change effect.

The above hypothesis leads to another possibility that the sentences were too short for a listener to be influenced by the talker change effect during ambiguity resolution. All of the sentences were approximately 1.4 to 1.7 seconds long. Especially disambiguating words, where talker change occurs, were no longer than .6 second in the sentences. In such a short time, participants might not notice that the talker changed.

A different hypothesis is that the change of talker only influences the stage that is indirectly related to the influence of subsequent context on identifying an ambiguous word. The process of understanding the spoken language consists of two stages: identifying the word and retrieving the meaning (Gaskell & Marslen-Willson, 2001). The perceptual system identifies a word or a sentence, and then accesses the associated meaning. The influence of subsequent context on lexical ambiguity resolution is related to the second stage in a way that the appropriate meaning must be selected in order to resolve a previous ambiguity (Gaskell & Marslen-Willson, 2001). On the other hand, the change of talker generally influences the first stage, identification. If this is the case, the effect of the talker can be found in the influence of prior context, because the perceptual system is affected not only by contextual information but also word by word processing.

In conclusion, the present study has provided further insight into the influence of subsequent context on ambiguity resolution. Although the results do not show the talker variation effect to be reliable, possible causes were proposed: listeners opted for a strategy that

ignores the effect of talker change, the change of talker only influences the ambiguity resolution indirectly, or the sentences were too short for a listener to be influenced by the talker change effect. Further studies can eliminate possible variables and examine the effect of talker change can be found in different setting of the experiment. For instance, presuppose the stimuli were not long enough for talker change to be effective. In this case, the sentences can be improved by increasing the speech proportion of the second talker such as increasing the number of syllables (e.g., "The wing had feathers"). If the above study shows that the effect of talker change decrease the influence of subsequent context on lexical ambiguity, it would indicates that talker change actually increase the memory load of perceptual system.

References

Connine, C. M., Blasko, D., & Hall, M. (1991). Effect of subsequent sentence context in auditory word recognition: Temporal and linguistic constraints. *Journal of Memory and Language,* 30, 234-250

Clopper, C.G. & Pisoni, D.B. (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. Journal of Phonetics, 32, 111-140.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. Cognition, 106(2), 633-664.

Gaskell, G. M., & Marslen-Wilson, W. D. (2001). Lexical Ambiguity Resolution and Spoken word Recognition: Bridging the Gap. *Journal of Memory and Language*, 44, 325-349.

Marslen-Wilson, W, D. (1987). Functional parallelism in spoken word-recognition. Spoken word recognition, 71-102.

McLennan, C.T., & Luce, P. A. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory & Cognition, 31,* 306-321.

Newman, R. S., & Sawusch, J. R., (2009). Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another. *Journal of Phonetics,* 38, 46-65

Samuel, A. G. (1981). Phonemic restoration: Insights from a new methodology. *Journal of Experimental Psychology: General,* 110, 474-494

Szostak, C. M. & Pitt, M. A. (2010). *Using subsequent context to resolve prior lexical ambiguities in spoken word recognition.* Unpublished manuscript, Ohio state University, OH.

Van Alphen, P & McQueen, J. M. (2001). The Time-Limited Influence of Sentential Context on Function Word Identification. *Journal of Experiment Psychology: Human Perception and Performance*, 27(5), 1057-1071

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science,* 167, 392-393

Warren, R. M. & Sherman, G. L. (1974). Phonemic restoration based on subsequent context. *Perception & Psychophysics,* 16, 150-156

Appendix A

The list of stimuli used in experiment. The word in italics was the word congruent with the incongruent context. The final two words in each sentence are the congruent and incongruent disambiguating words respectively.

The knock/*lock* was heard/snapped.

The leak/*beak* was fixed/sharp.

The lip/*rip* was chapped/stitched.

The list/*wrist* was numbered/injured.

The log/*dog* was rotting/barking

The look/*book* was modeled/written.

The match/*latch* was struck/closed.

The math/*bath* had fractions/bubbles.

The mole/*bowl* was digging/assigned.

The mug/*rug* was filled/swept.

The nap/*wrap* was restful/tasty.

The nest/*guest* had sparrows/luggage.

The niece/*lease* was married/cosigned.

The night/*light* was starless/shining.

The roast/*ghost* was salty/spooky.

The rum/*gum* was poured/chewed.

The well/*bell* was dry/rung.

The wick/*nick* was burning/repaired.

The wife/*knife* was pregnant/pointed.

The wig/*rig* was shampooed/driven.

The wing/*ring* had feathers/diamonds.

The wish/*dish* was granted/shattered.

The wool/*bull* was knitted/angered.

The yarn/*barn* was woven/painted.

Appendix B

Mean number of hits, misses, false alarms, and correct rejections in the experiment. Values are rounded to the nearest whole number.

Number of talker	congruency	hits	misses	false alarms	correct rejection
one	congruent	20	2	6	17
one	incongruent	16	6	5	18
two	congruent	20	2	5	17
two	incongruent	16	6	6	16

Figure 1

Hit rate and false alarm rates in four conditions: single-talker congruent, single-talker incongruent, two-talker congruent and two-talker incongruent.