CHAPTER 1

Human Speech Perception of Ear-Recorded speech

1.1 Introduction

In order to judge the usefulness and intelligibility of the modified sounds recorded at the ear, it is necessary to run a human perception task on the recorded sounds. Of primary interest is whether the speech recorded at the ear in a noisy condition a) has resulted in a sufficient reduction in the ambient noise level, and b) is markedly more intelligible that speech recorded at the *mouth* in a noisy condition.

1.2 Background

The inability for human listeners to understand speech occurs from an information loss that can be due to a host of factors. For example, information can be lost in the domain of time (eg. an intermittent signal), or from loss of intensity (eg. due to distance of the source), as well as the distortion of the source itself, such as a speech deficit (**?**). Of particular interest to the present study is the difficulty for human listeners to perfectly understand a speech signal due to additive background noise from different sources other than the desired speech signal.

The ability of the human auditory system to hear and differentiate multiple sources of sound from a single pressure wave is often given the term "auditory scene analysis" (**?**). The term "scene analysis" is borrowed from the visual domain, implying the separation of an "scene" (be it auditory or visual) into its component objects (again, be they auditory or visual). For the purposes of this study, we will be discussing human auditory ability to find multiple sound sources from a temporal stream of air pressure fluctuations (ie. sound) reaching the tympanic membrane.

To visualize the auditory scene, note the waveform (ie. the graph of air pressure fluctuations) that reaches the tympanic membrane, eg.

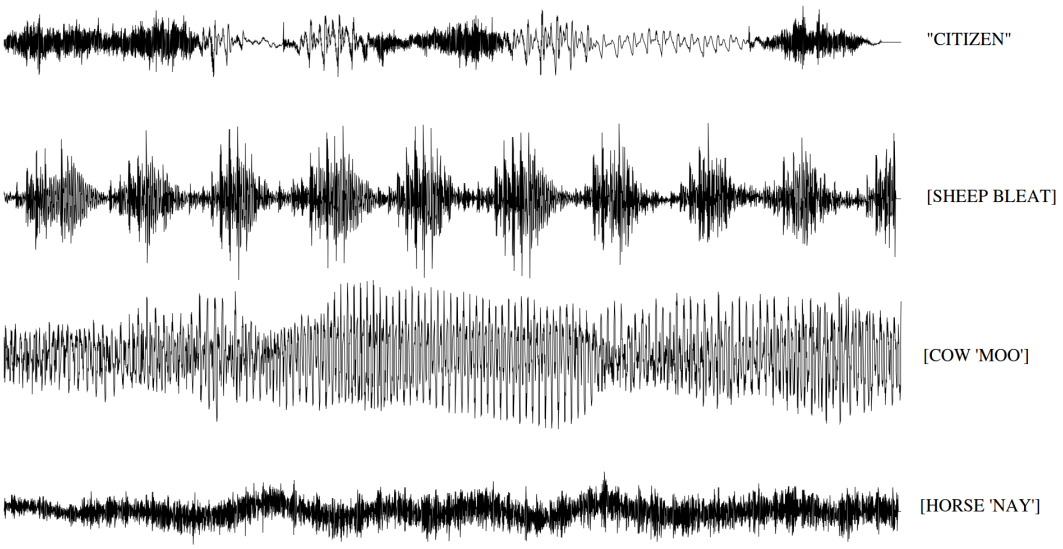


Figure 1.2: The four component waveforms (human speech, sheep, cow, horse), of the combined waveform seen in Figure 1.1.

see Figure 1.1. It is composed of all environmental sounds contributing to the air pressure fluctuation at the tympanic membrane[1].

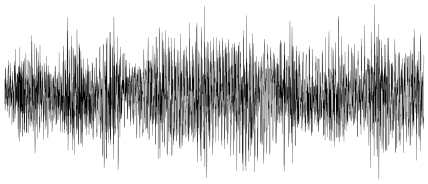


Figure 1.1: A waveform composed of multiple sound sources (cf. Fig. 1.2.

However, under the framework of auditory scene analysis, the human auditory system is able to separate this input signal into its various sources, or "auditory objects". In effect, this would separate the above waveform into its actual component sources of human speech, and the sounds of a sheep, cow, and horse, seen in Figure 1.2; "The normal auditory system exhibits a remarkable ability to parse these complex scenes" (**?**, 2). Of course, there reaches a point at which the auditory system fails and can no longer differentiate all sources, or, more relevant to this paper, recognize the information in a human speech signal when embedded

[1]Note that this is for illustration purposes; the waveform will obviously look different when shaped by a given environment and the human ear canal before reaching the tympanic membrane.

with background noise from one or more additional sources. The following section will describe in more depth the acoustics of speech in noise.

1.2.1 Acoustics of Speech in Noise

Speech in noise can be intuitively grouped into two components, the speech (more specifically the voice one is intending to hear) and the noise, called the "masking" element. Broadly, masking can be defined as "the process by which the threshold of hearing for one sound is raised by the presence of another" (**?**, 61). This masking element is anything *but* the voice[2] (speech signal) that one is interested in.

The masking process can be broken down into two forms: energetic masking and informational masking. Energetic masking occurs when the masking element shares the same temporal and frequency elements of the voice. It can be thought of as if the masked element and the voice are competing for "space" along the basilar membrane and then the auditory nerve (**?**), but can also be considered to be competing for the listener's attention (ie. the listener must concentrate on ignoring the mask, and exclusively listening to the target, **?**). Energetic masking is normally thought to occur primarily in the "lower" auditory processes, eg. the cochlea and auditory nerve, though this is not always the case, as described further below.

Informational masking can be broadly thought of as difficulties relating to memory, linguistic processing, and the like, oftentimes generalized to speech-on-speech noise. This can even be restricted to cases in which the masking speech is intelligible speech, as **?** failed to find informational masking in a cross-linguistic task. This is thought to occur primarily in the "higher auditory processes" in the brain.

This can be visualized in a diagram of overlapping speech presented in **?**, as seen in Figure 1.3. Say that utterance (C) in the figure is the desired "voice", leaving utterance (B) the masking element. In (D), one can see the voice (red), the areas of direct frequency and temporal overlap in green, and in blue the remainder of the masking speech (B). This could primarily be viewed as a form of energetic masking

[2]For the purposes of this paper, the term "voice" will be used throughout to refer to the singular speech source the listener desires to hear out of the masked signal.

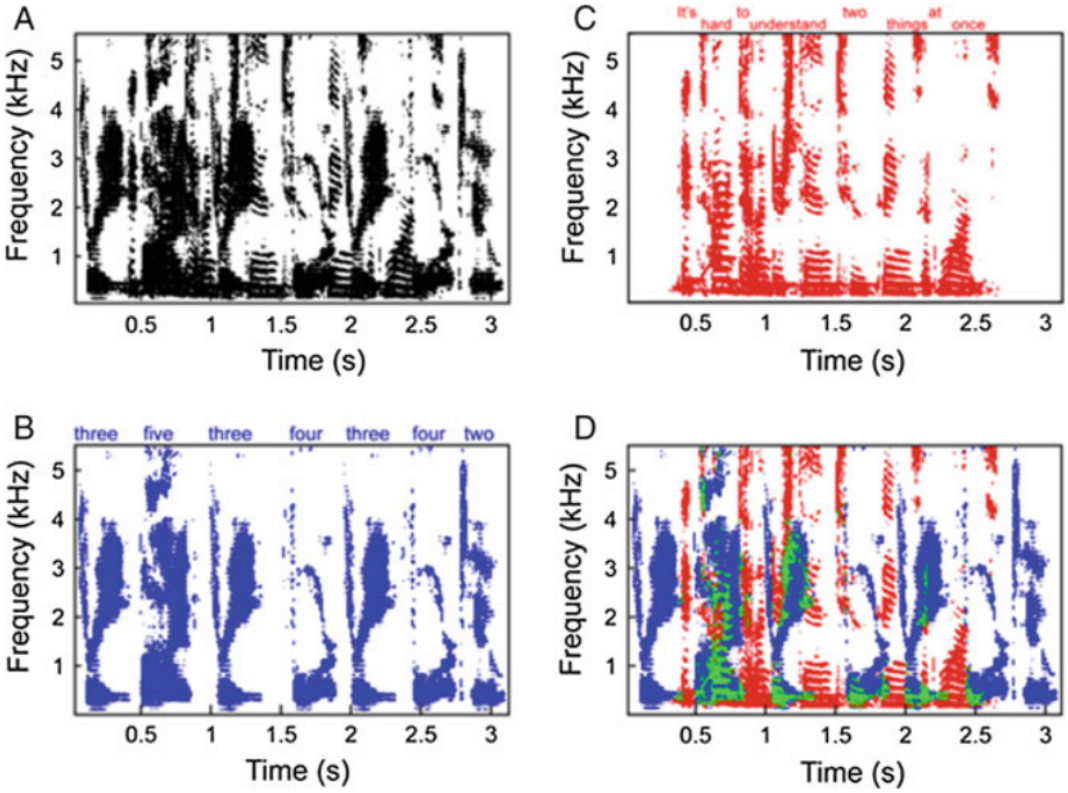


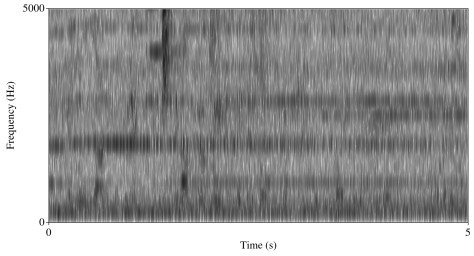
Figure 1.3: Diagrams of different spectograms. (A) The spectogram of two temporally overlapping spoken utterances. (B) The spectogram of the utterance "three five three four three four two." colored in blue (C) The spectogram of the sentence "It's hard to understand two things at once." colored in red. (D) The overlap of the two spectograms (B) and (C), with the color green highlighting the areas of energy in frequency and time that overlap.

(competition for lower-level processing), though upper level processing is required to take meaning from the desired voice, which is masked informationally by the other, competing voice carrying it's own information.

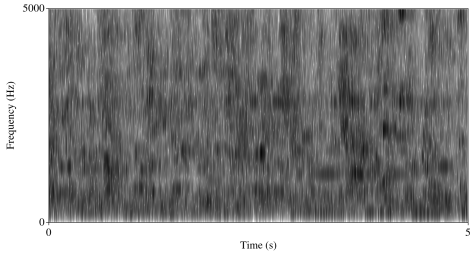
The five different background noises used in the study described in Chapter 2?? primarily serve the purpose of energetic masking of the voice in the signal. A small (5 second) portion of the spectrogram of each sound can be seen in Figure ??. These sounds don't produce any informational content themselves which mask the desired voice (the 'cafe' noise, seen in Figure ??, does contain speech babble, none of it intelligible), and so masking occurs by producing energy at the same time as - and in the same frequency range as - the recorded voice.

Yet simply because a sound may be "masked" does not necessarily imply that the voice is not heard or understood. There are a number of methods used by the auditory system to overcome the masking and understand the voice; this process is termed "release from masking" (?). One such method, binaural hearing, uses both ears to tease apart the different sources, due to the very small temporal difference that occurs when different sound sources reach each ear. In this example, it is easy to see that energetic and informational masking are not strictly limited to masking separate "lower" and "higher" processes (?). The use of binaural hearing is an example of a "higher" process used for a release from energetic masking, as it necessarily requires signals from both ears to be interpreted (?). Binaural hearing is basically making use of the spatial directionality of the noise(s) from the listener to separate the different sources (?).

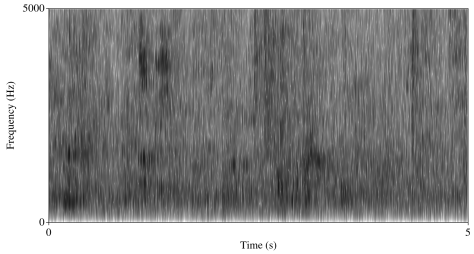
Other methods of release from energetic masking include recognizing the difference between the fundamental frequency and timbre of two different sources, as well as making note of acoustic transitions: "when [a] sound...changes its properties gradually, [it] is likely to be heard as a single changing sound. However, when [it] changes...abruptly, [it] tends to be treated as a newly arriving sound, this tendency increasing with the abruptness of the change." (?, 5). Fundamental frequency (F0) has been shown to be an effective tool to presumably inpterpret the location of harmonics, and parse apart different sources (eg. two separate, simultaneous vowels



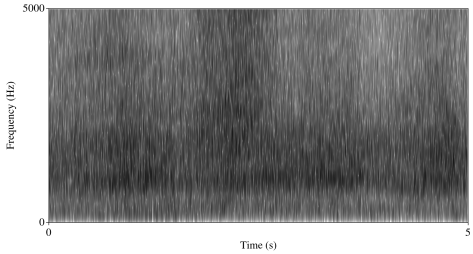
(a) Bus background noise.



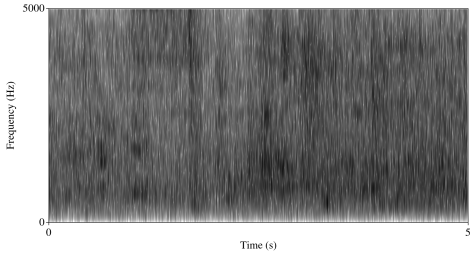
(b) Cafe background noise.



(c) Pedestrian background noise.



(d) Street background noise.



(e) Factory background noise.

Figure 1.4: Example spectograms of the first five seconds of the background noise tracks. Most recorded sentences occurred within these stretches of noise.

with different F0s, ?). There are many other proposed methods to release masking that the auditory system uses, particularly among informational masking (?), but these are beyond the scope of this project.

1.2.2 Human Recognition Performance of Speech in Noise

Eventually, however, with enough background noise and masking, the methods listed above for releasing the masking will fail and recognition will begin to break down. Under the most simple conditions to measure, steady-state noise, (?) report that, when listening to speech in noise, human self reported intelligibility ratings don't drop significantly until the SNR reaches approximately -3 dB, where intelligibility drops to about 55%, and it doesn't hit near floor level (0%) until -9 db SNR.

This subjective measure is backed by a study performed by ?, who used the PRESTO corpus (?) to test sentence intelligibility among 121 native English speakers. ? found that - similar to ? - that the median score (at the 50th percentile) of speech with a -3dB SNR had about 55% accuracy. At +3 dB SNR, the 50th percentile increased to approximately 88% accuracy.

? does however comment that there was great inter-speaker variation among the subjective measures of their perceived intelligibility of an utterance, and this is also supported by ?'s results. They showed that, averaging over all SNR conditions (-5, -3, 0, +3), the variability between speakers had a range of almost 36%; a retest performed with a subgroup of the original participants on the same dataset yielded a similar (34%) range of variability in accuracy.

? discusses about how listening to speech in background noise places extra demands on working memory, as does listening to degraded speech (?). The diversion of working memory to acoustic processing can be particularly detrimental to performance when simultaneously working on other computation, e.g. syntactic and semantic parsing (?), such as in a phrase or sentence recognition. ? tested a group of high-performance against a separate group of low-performance hearers of speech in noise on several different working and short term memory tasks. Not surprisingly, the group of listeners who are able to better hear speech in noise also perform statis-

tically better on the working memory tasks. This is by no means the only indicator of perceptual performance.

? briefly discusses the concept of perceptual learning, which asserts that one can learn to accomodate a particular adverse condition (eg. background noise, signal distortion, etc.) with practice in that area. Learning will be less effective in cases which the degradation is variable or unpredicatble between trials, such as with unpredictable background noise. The present study is designed with this in mind, that the speech with the background noises will be presented in a random order, and since the background noise varies, it cannot be assumed or predicted from one sentence to the next.

Although there is expected to be a great amount of variability between subjects, the results from the **?** and **?** studies indicate that the average SNR for the highest noise condition from the data collected and described in Chapter 2**??**[3] is over 9 dB SNR above the 50% intelligibility threshold at -3 db SNR given by these two studies. It is unlikely that listeners will encounter much masking in the collected noisy speech that won't be overcome.

Due to this, two additional participants were rerun with lower SNRs, explained more in Section 1.3.

1.3 Experiment 2: Human Speech Perception in Noise

After analysing the results from a few pilot runs, it was deemed that the speech collected at the mouth in the noisy background described in Chapter **??** was too easily recognizeable for human listeners; i.e. the SNR ratio was not low enough, as explained in Section **??**. After additional stimuli were gathered (explained below in Section 1.3.1), a human speech perception experiment was run on the data collected in the first experiment in order to better understand and compare the ability of the auditory system to accurately comprehend the speech with a noisy background, and the speech distorted by passage through the speaker's head, and then modified

[3]The 80 dB noise condition yielded approximately +6dB SNR

to become more intelligible. To act as a control, participants would also listen to normal, clean speech.

1.3.1 Stimuli Generation

To remedy the problem of the noisy speech being *too* intelligible and having a high SNR, two additional participants (one male, one female) were recorded following the procedure in the first task (cf. Chapter **??**). The list of stimuli was increased to 80 sentences (eight Harvard Sentence lists[4]) to provide more reaction data from this present experiment.

To increase the SNR, the directional microphone was pointed away from the mouth of the participant, and directed toward the loudspeaker (see fig. 1.5). This, of course, results in some of the limitations outlined in section **??** in Chapter 2; for example, simply pointing the mouth microphone toward the loudspeaker, rather than increasing the noise, ignores the fact that the noise level inside the ear might increase as well with an increase in ambient noise. Given the alternatives outlined in section **??**, this was seen as the best available option. Figures 1.6a and 1.6b show the new noisy and ear recorded speech, respectively.

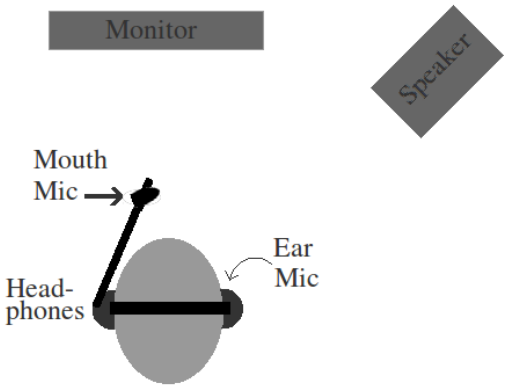
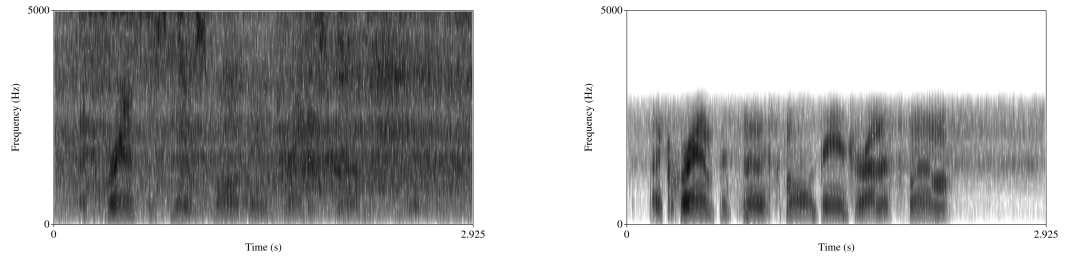


Figure 1.5: This is the same setup as described in Chapter **??**, except that the microphone is facing the loudspeaker, rather than the participant's mouth.

Furthermore, due to the issue of the lack of noise, only one noise level condition was used - the highest available noise level (80 dB). All previously used background noises were used for this data collection

[4]This included the previous three lists, 14, 28, and 57, as well as lists 21, 29, 37, 53, and 68. These additional lists were pseudo-randomly chosen, as were the original three lists, to contain words that would be readily recognizeable by the participant population.



(a) New recording at the mouth with the microphone pointed toward the loudspeaker noise source.

(b) New recording at the ear; all recording conditions for the ear were the same as in the first group of recordings.

Figure 1.6: The sentence "A cramp is no small danger on a swim" spoken by the male speaker, recorded at the mouth (Fig. 1.6a) with "cafe" noise, and simultaneously at the ear (Fig. 1.6b)

task as well.

1.3.2 Design

The experiment had three factors - gender of speaker **x** microphone location **x** noise type - resulting in a 2x2x6 experiment. There were two genders, two mic locations (recording at the ear, and at the mouth), and six noise types (bus, cafe, pedestrian, street, factory, and no noise (clean)). Since the ability to understand speech in noise is quite variable between individuals, the design of this experiment was a within-subjects experiment. This meant that each of the 2x2x6 (ie. 24) conditions needed to be seen by each participant. The sessions that were re-recorded constituted 80 sentences, which allowed for three sentences in each of the 24 conditions, totalling 72 sentences used in the experiment. The eight remaining sentences were used as a "training" set, intended to get the participants used to the task itself, rather than to acclimate them the type of speech that they would hear.

Since any given speaker could not hear the same sentence twice without introducing a confound, and since each sentence occurred in each of the 24 conditions, this necessitated the use of 24 co-balanced lists to ensure that each sentence was heard in every condition. For example, sentences 1, 2, and 3 would occur in Factor #1 (e.g. female speaker, mic at the mouth, with bus background noise) in co-balanced

group #1 for Participant #1. Participant #2 would see co-balanced group #2, which placed sentences 1, 2, and 3 in Factor #2 (e.g. female speaker, mic at the mouth, with cafe background noise). For simplicity's sake, each grouping of three sentences would appear together in a given condition, and were not mixed up between the different co-balanced groups. However, once the sentences are assigned to a particular condition, the order of presentation to the participants is randomized.

1.3.3 Participants

Twenty-four native speakers of English with self-reported normal hearing participated in the experiment. Each participant was placed into a separate co-balanced group, as specified in 1.3.2.

1.3.4 Equipment

The experiment was conducted in a soundbooth with a pair of over-the-ear headphones. The experiment used in-house developed software, and participants answers were typed into a computer whose monitor could be seen from inside the soundbooth.

1.3.5 Procedure

The participant was seated in the sound booth in front of a keyboard and computer monitor, with a pair of headphones, and were given a set of instructions. They were told that they would hear each utterance only once, and what they would hear was comprised of real English words, but may not constitute a "complete" sentence. They were forewarned that many of the sentences they heard would be noisy and difficult to understand. They were instructed to write all words they heard, even if what was heard did not make syntactic sense, or if the words were not adjacent (e.g. if only the first and last word of the sentence was heard). They would be timed, with 18 seconds to type their response starting from the beginning of the soundfile and that their answer would be saved as is if they ran into the time limit, preventing them frrom typing more.

The participant was told that the first set of eight utterances they heard were part of a "training" set of eight utterances. These were given to introduce the participants to the experiment and their task[5]. None of the utterances from the training set were used in the analysis. Once completed with this initial set, participants were asked if they had any questions. Afterwards they began the task in the soundbooth. They would hear one of the sentences, and type their answer in a text box. When finished with their answer, they would either click to advance to the next sentence, or, if they ran into the time limit, were prevented from modifying their answer, and were prompted to click another button to advance. When finished with all 72 stimuli, the participant was given a brief questionnaire to fill out.

After the experiment, the researcher would double check participant answers for correct spelling. Only obvious errors were modified (e.g. 'teh' to 'the', 'crakers' to 'crackers', 'mantle' to 'mantel'), while ambiguous errors were left as-is (e.g. 'blo' was not changed to 'block', 'finde' was not changed to 'fine'). Numbers were also lexicalized (e.g. '30' to 'thirty'). Punctuation was removed for ease of analysis and calculation of word error rate. Exact responses given by each participant can be found in Appendix F**??**.

1.4 Results

The word error rate (WER) for each (spell-checked) response for each participant was calculated. The code for the WER calculation can be found in Appendix F**??**

A 3-way, within-subjects ANOVA was performed with the collected data, 72 sentences from each of the 24 participants. Factors included the gender of the speaker (of the stimulus) with two levels - male and female - the location of the recording microphone with two levels - at the mouth and at the ear - and the background noise type with six levels -no noise, bus noise, cafe noise, pedestrian noise, street noise, and factory noise. There was no significant 3-way interaction between speaker gender, noise type, and mic location, as can be seen in the by-

[5]The same eight sentences were heard by every participant

subjects ANOVA (Table 1.1) and the by-items ANOVA (Table 1.2). Two, two-way interactions were significant, speaker gender by mic location, and noise-type by mic location. The two way interaction between speaker gender and noise type was not significant. The main effects of all three factors were also significant (cf. Tables 1.1 and 1.2).

Effect	DFn	DFd	F	p	p<.05
speaker_gender	1.00	23.00	6.69	0.02	*
noise_type	5.00	115.00	84.83	0.00	*
mic_location	1.00	23.00	155.07	0.00	*
speaker_gender:noise_type	5.00	115.00	0.55	0.74	
speaker_gender:mic_location	1.00	23.00	18.53	0.00	*
noise_type:mic_location	5.00	115.00	55.53	0.00	*
speaker_gender:noise_type:mic_location	5.00	115.00	1.61	0.16	

Table 1.1: ANOVA for by-subjects analysis of the three-factor, within-subjects experiment.

Effect	DFn	DFd	F	p	p<.05
speaker_gender	1.00	71.00	4.01	0.05	*
noise_type	5.00	355.00	103.50	0.00	*
mic_location	1.00	71.00	354.53	0.00	*
speaker_gender:noise_type	5.00	355.00	0.52	0.76	
speaker_gender:mic_location	1.00	71.00	21.36	0.00	*
noise_type:mic_location	5.00	355.00	71.03	0.00	*
speaker_gender:noise_type:mic_location	5.00	355.00	1.86	0.10	

Table 1.2: ANOVA for by-items analysis of the three-factor, within-subjects experiment.

Mauchley's Test for Sphericity was conducted, both for the by-subjects and by-items ANOVAs. Significant sphericity violations were found for the main effect of noise type, and the interaction of speaker gender and noise type, as can be seen in Tables 1.3 and 1.4.

The corrections for sphericity were performed using a Greenhouse-Geisser test, for both by-subjects and by-items ANOVAs. Both resulted in a change to the two way interaction between speaker gender and noise; it is no longer a significant

Effect	W	p	p<.05
noise_type	0.21	0.00	*
speaker_gender:noise_type	0.61	0.73	
noise_type:mic_location	0.39	0.13	
speaker_gender:noise_type:mic_location	0.67	0.87	

Table 1.3: Sphericity test for the by-subjects ANOVA.

Effect	W	p	p<.05
noise_type	0.78	0.26	
speaker_gender:noise_type	0.64	0.01	*
noise_type:mic_location	0.85	0.66	
speaker_gender:noise_type:mic_location	0.86	0.70	

Table 1.4: Sphericity test for the by-items ANOVA.

interaction.

Effect	GGe	p[GG]	p[GG]<.05
noise_type	0.61	0.00	*
speaker_gender:noise_type	0.86	0.71	
noise_type:mic_location	0.77	0.00	*
speaker_gender:noise_type:mic_location	0.87	0.17	

Table 1.5: Sphericity corrections for the by-subjects ANOVA.

Noting the sphericity violations involving the noise type condition, the data was viewed in the box plots in Figures 1.7, 1.8, and 1.9. When viewing the simple effects of noise in Figure 1.7, it is apparent that the no-noise condition differs distinctly from the other noise types.

This holds true when viewing the interaction of speaker gender and noise type in Figure 1.8 and the interaction of noise type and mic location in Figure 1.9. The conditions in which there is no noise present differs noticeably from those with noise; this can visually seen even in the condition in which the speech was recorded at the ear. This is likely the root of the sphericity violations.

Effect	GGe	p[GG]	p[GG]<.05
noise_type	0.91	0.00	*
speaker_gender:noise_type	0.87	0.73	
noise_type:mic_location	0.94	0.00	*
speaker_gender:noise_type:mic_location	0.95	0.11	

Table 1.6: Sphericity corrections for the by-subjects ANOVA.

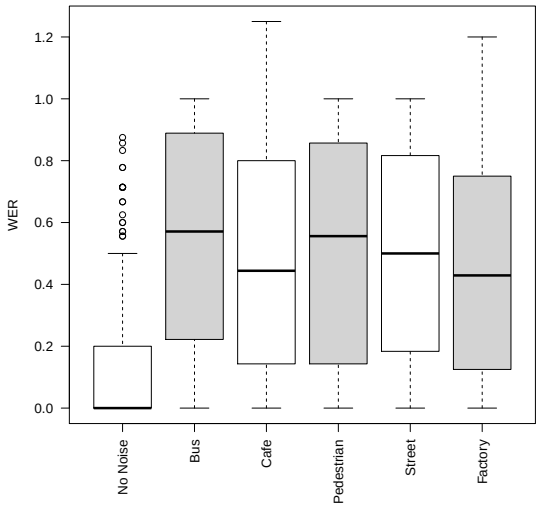


Figure 1.7: Boxplot displaying the average word error rate (WER) averaged over each participant for every noise type. WER is the variable on the y-axis, and noise type is on the x-axis.

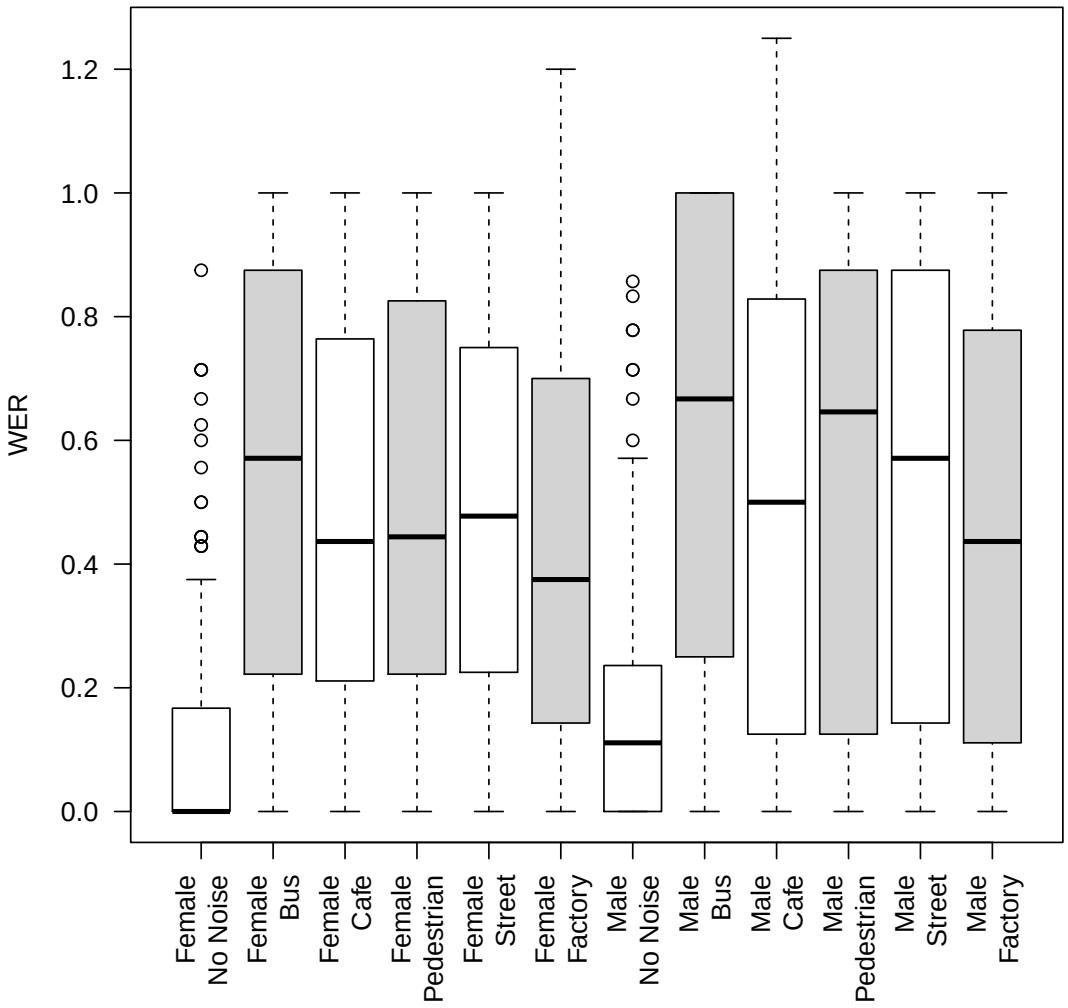


Figure 1.8: Boxplot displaying the average word error rate (WER) averaged over each participant for the interaction of every noise type by the speaker gender. WER is the variable on the y-axis, and noise type by speaker gender is on the x-axis.

Since it there is a main effect of statistical difference in noise, and since the "no noise" condition is very apparently different from the rest another ANOVA was calculated with the "no noise" level of nose type removed. This was to test for statistical difference only in noise conditions containing actual noise (and any resulting interaction).

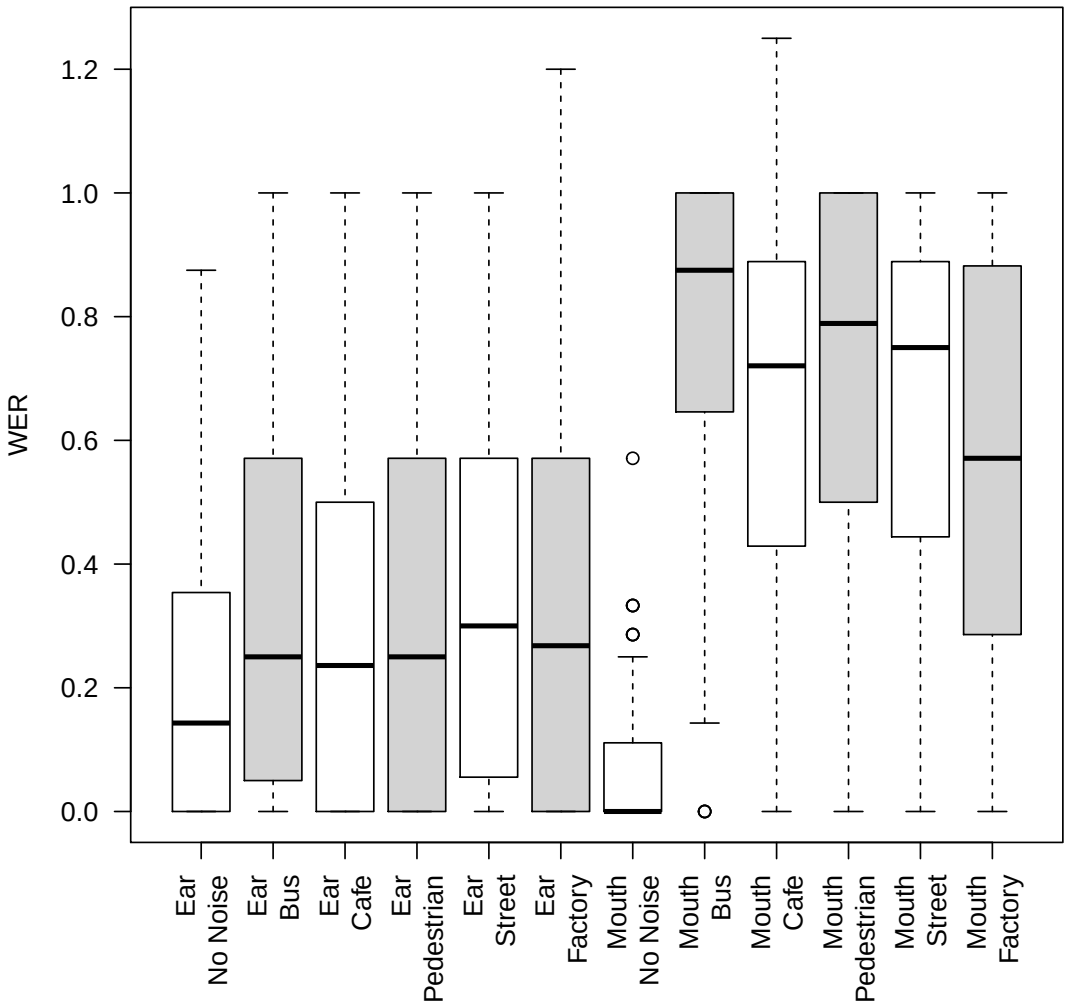


Figure 1.9: Boxplot displaying the average word error rate (WER) averaged over each participant for the interaction of every noise type by the mic location. WER is the variable on the y-axis, and noise type by mic location is on the x-axis.

This modified ANOVA is a two by two by five design, with the noise type factor only having 5 levels (with the "no noise" condition removed from the noise type factor).

The results are similar, with no significant three-way interaction, but - as before - two significant two-way interactions of speaker gender by mic location and noise-type by mic location (cf. Tables 1.7 and **??**). There are main effects of noise type

Effect	DFn	DFd	F	p	p<.05
speaker_gender	1.00	23.00	5.26	0.03	*
noise_type	4.00	92.00	5.95	0.00	*
mic_location	1.00	23.00	215.08	0.00	*
speaker_gender:noise_type	4.00	92.00	0.60	0.66	
speaker_gender:mic_location	1.00	23.00	16.68	0.00	*
noise_type:mic_location	4.00	92.00	6.00	0.00	*
speaker_gender:noise_type:mic_location	4.00	92.00	1.14	0.34	

Table 1.7: ANOVA for by-subjects analysis of the three-factor, within-subjects experiment. The "no noise" condition was removed from the noise type factor, resulting in a 2x2x5 design.

Effect	DFn	DFd	F	p	p<.05
speaker_gender	1.00	71.00	3.46	0.07	
noise_type	4.00	284.00	6.92	0.00	*
mic_location	1.00	71.00	515.73	0.00	*
speaker_gender:noise_type	4.00	284.00	0.55	0.70	
speaker_gender:mic_location	1.00	71.00	20.62	0.00	*
noise_type:mic_location	4.00	284.00	7.14	0.00	*
speaker_gender:noise_type:mic_location	4.00	284.00	1.35	0.25	

Table 1.8: ANOVA for by-items analysis of the three-factor, within-subjects experiment. The "no noise" condition was removed from the noise type factor, resulting in a 2x2x5 design.

and mic location. Speaker gender only significant in the by-subjects anova, but is not significant in the by-items anovas.

Again, Mauchley's Test for Sphericity was conducted, resulting only in a significant sphericity violation in the by-subjects ANOVA for noise (cf. Tables 1.9 and 1.10). This was corrected with a Greenhouse-Geisser test, which resulted in the interaction between speaker gender and mic location no longer being significant, but the interaction between noise type and mic location, and the main effects of noise type and mic location, remaining significant (cf. Tables 1.11 and 1.12).

Effect	W	p	p<.05
noise_type	0.26	0.00	*
speaker_gender:noise_type	0.76	0.74	
noise_type:mic_location	0.58	0.23	
speaker_gender:noise_type:mic_location	0.82	0.89	

Table 1.9: Sphericity test for the by-subjects ANOVA with the "no noise" condition removed.

Effect	W	p	p<.05
noise_type	0.87	0.36	
speaker_gender:noise_type	0.88	0.43	
noise_type:mic_location	0.93	0.86	
speaker_gender:noise_type:mic_location	0.98	1.00	

Table 1.10: Sphericity test for the by-items ANOVA with the "no noise" condition removed.

1.5 Initial Discussion

The primary hypotheses from Chapter 2**??** included a) that the signal recorded from the ear, premphasized, filtered, and preemphasized again, would be intelligible by human listeners, and b) that it would be more intelligible than speech with a noisy background. The results in Section 1.4 above show a statistical difference between the WERs of the sentence transcriptions of the speech recorded from the ear canal and the speech recorded in front of the mouth. This can be seen more clearly in the graph of the simple effects of microphone location, in Figure 1.10. The speech recorded at the ear has a significantly lower transcripiton word error rate than the speech recorded at the mouth, over all noise conditions (including clean speech). These primary hypotheses seem to have been validated.

A "near" main effect of speaker gender can be observed, in that speaker gender there was a statistical main effect of speaker gender for the by-subjects ANOVA, but not the by-items. To dissuade the notion that there is an effect of gender, it should also be noted that, in the instance of these two particular speakers, gender is also confounded with SNR. The average female speaker's SNR was higher than the

Effect	GGe	p[GG]	p[GG]<.05
noise_type	0.57	0.00	*
speaker_gender:noise_type	0.90	0.65	
noise_type:mic_location	0.82	0.00	*
speaker_gender:noise_type:mic_location	0.91	0.34	

Table 1.11: Sphericity corrections for the by-subjects ANOVA with the "no noise" condition removed.

Effect	GGe	p[GG]	p[GG]<.05
noise_type	0.93	0.00	*
speaker_gender:noise_type	0.94	0.69	
noise_type:mic_location	0.97	0.00	*
speaker_gender:noise_type:mic_location	0.99	0.25	

Table 1.12: Sphericity corrections for the by-items ANOVA with the "no noise" condition removed.

male speaker's SNR[6]. This likely contributed to the observed 'near' significance.

There is also a definite main effect of noise. It is obvious that the speech recorded with no background noise (particularly at the mouth) would be easier to transcribe and recognize than the speech recorded with background noise. However, when the level of 'no noise' within the noise-type factor is removed, the statistical difference within this condition remains.

A closer look at the main effect of noise-type, excluding the level of 'no noise', can be seen in Figure 1.11.

[6]**These two SNR values are on a computer in the US - will add later**

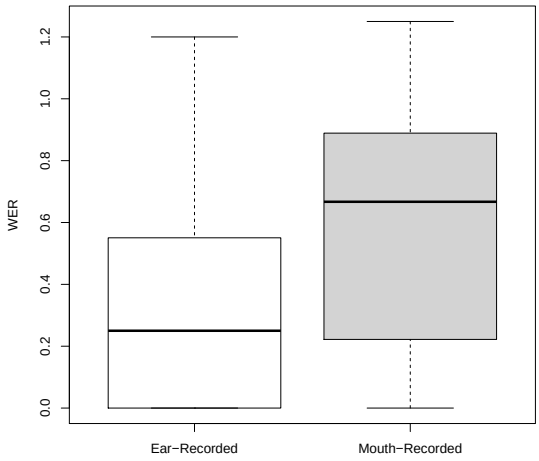


Figure 1.10: Simple effects of Microphone Location.

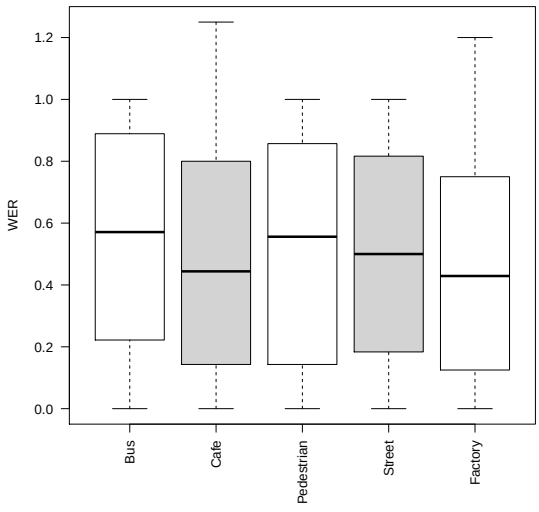


Figure 1.11: Simple effects of Noise Type, excluding the level of 'no noise'.

The difference here is not nearly as stark as that seen within the microphone location distinction in Figure 1.10. Even still, there is a noticeable difference, particularly between the bus background noise (with the highest relative WER) and the factory background noise (with the lowest relative WER).

Referring back to Figure 1.4, containing the spectograms of the background noises in Section **??**, it is certain appears that the bus noise (cf. Figure 1.4e) contains bands in the frequency spectrum that contain higher amplitude. This may adversely affect a person's ability to parse the harmonics and/or formants from the desired speech.

It is unclear, however, why the cafe background noise, among the other noises,

has a relatively lower WER, since it also contains many more bands of frequency in this area from speaker babble. It is possible that the more prominent frequency bands in the bus noise play a role. Similarly, it is difficult to observe much difference between the factory background noise and the pedestrian background noise, despite the pedestrian noise having a higher upper bound.

To see if any more apparent difference can be found, Figure 1.12 shows the noise-type factor split between the two levels of mic-location, displaying what was originally seen above in Figure 1.9. When only looking at the noise as it occurs in the mouth-recorded condition, the differences between the levels of noise become much more stark. In particular, the difference between the bus noise (high WER) and factory noise (relatively low WER) greatly expands.

This distinction reaches as low as a median of approximately 60% WER for the factory background noise conditions, and nearly a median of 90% WER for the bus background noise. The variance within noise conditions also differs, with the bus noise containing less variance (upper quartile 100% WER, lower quartile 68% WER) than the factory noise condition (upper quartile 90% WER, lower quartile 30% WER). The remaining noise conditions are more similar to one another, and fall in between the two, with median WERs hovering near 80%.

The differences between the transcription WERs of the bus and cafe noises and factory and pedestrian noises, despite the seemingly similar spectograms, are also heightened. As stated earlier, the bus noise has more distinct frequency bands, which might offer an explanation for it's divergence from the WER of the cafe noise, but no explanation is proposed for the apparent difference between the factory and pedestrian noises.

Returning again to Figure 1.12, the WER frontrunner is quite clearly the speech recorded at the mouth with no background noise. There was never any doubt that this would be the case, as speech with relatively little background noise is the sort of speech from which learners acquire their language model, and whereby most communication occurs. The median WER is, unsuprisingly, 0%, though there is some variance from perfect perception; some errors do occasionally occur. The

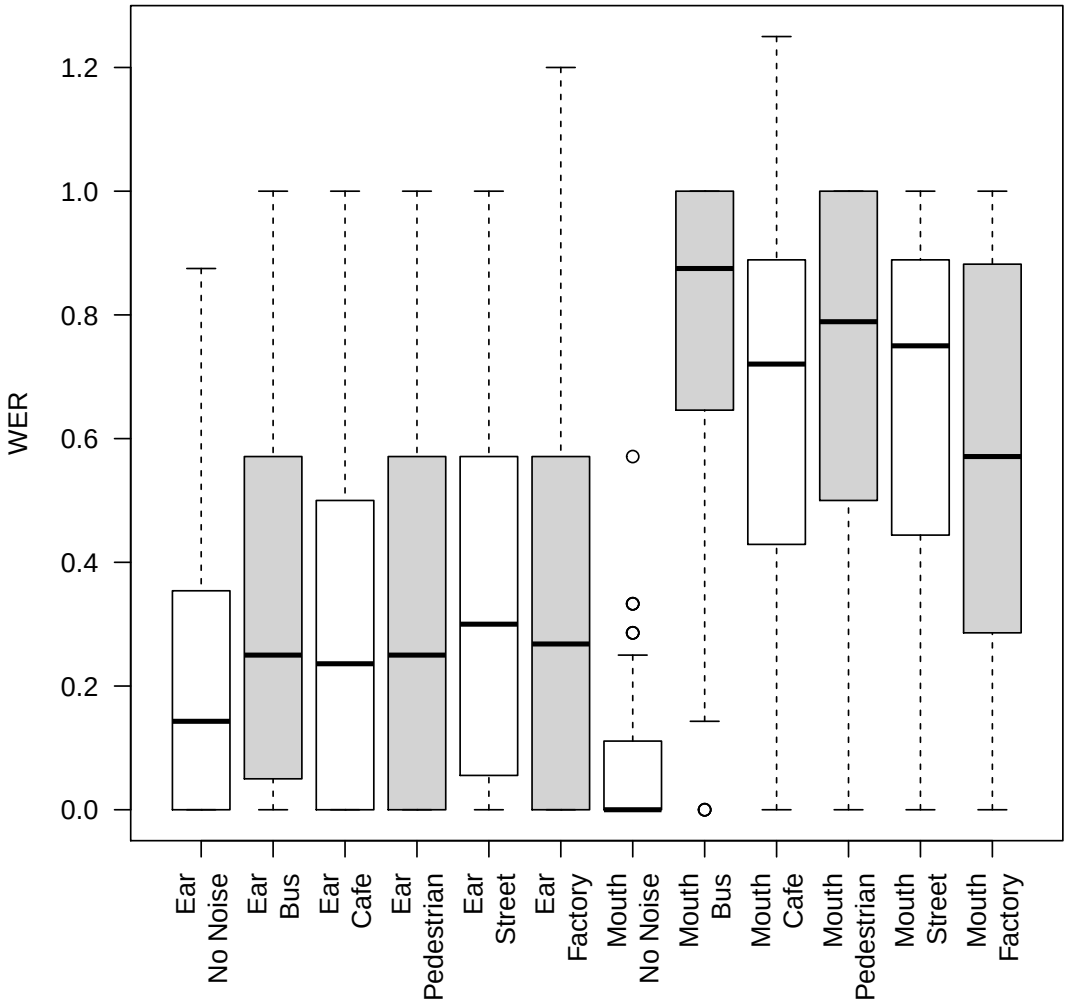


Figure 1.12: Boxplot displaying the average word error rate (WER) averaged over each participant for the interaction of every noise type by the mic location. WER is the variable on the y-axis, and noise type by mic location is on the x-axis.

ear-recorded speech in a clean environment manages to also acheive a respectable transcription WER median of approximately 15%, with a lower quartile boundary at 0% WER and an upper quartile boundary of 35% WER.

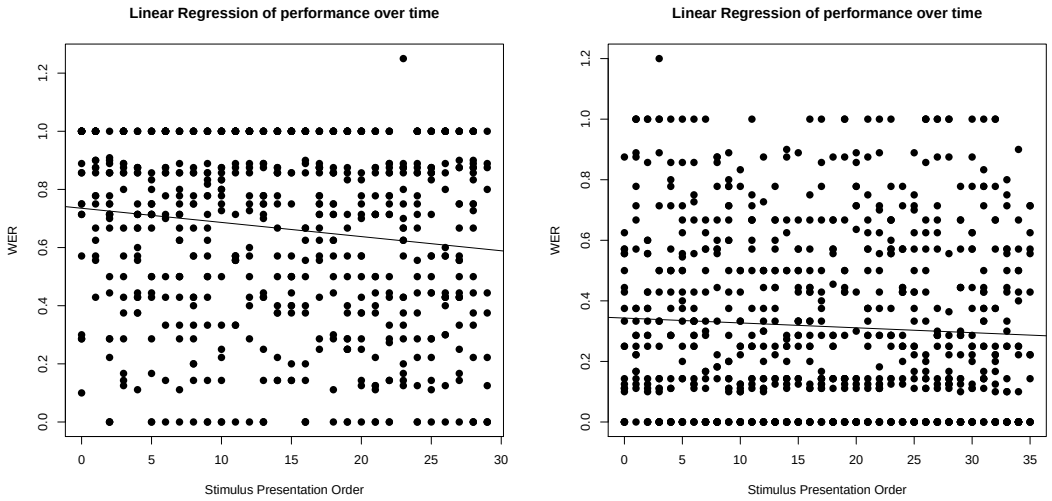
Despite being more similar than their mouth recorded counterparts, there is still an noticeable difference between the ear recorded speech with no noise, and ear recorded speech under noise conditions. Most noise conditions recorded from the ear acheive a near 0% lower quartile boundary, but the upper quartile boundary for most nears 60% WER. The median WER for noise conditions generally falls at or slightly below 30%.

Even though a higher WER than the no-noise condition, the ear-recorded speech in noise is quite consistent across noise categories. This is very different from the mouth-recorded speech in noise, which vary greatly between noise conditions. This indicates that while a noise presence hampers transcription ability and increases WER in ear-recorded conditions, the varying qualities of the different background noises were dampened to the point of having no effect resembling that which occurs in the mouth-recorded speech.

Since the ear-recorded speech, even in noise, appears to be quite consistent, the conditions are right for the participants to use perceptual learning to improve their recognition performance of the ear-recorded speech.

To visualize whether the performance of participants generally improves after exposure to ear-recorded speech, a scatterplot graph following participant's chronological performance with mouth-recorded speech over the course of the experiment can be seen in Figure 1.13a. A linear regression model was fit onto the data, with a slope of -0.0048728. A similar graph, containing the participant's performance with ear-recorded speech can be seen in Figure 1.13b. A linear regression model was also fit to this data, having a slope of -0.0016076.

It is important to note the linear scales both axes, particularly the y-axis, as an explanation for why the slope values themselves are so small. The take away from both graphs and both fitted regression models in Figures 1.13a and 1.13b is that participants do improve over the course of the experiment for both mouth recorded



(a) Scatterplot of all participants' WER values for responses to speech recorded at the mouth **and** in noise.

(b) Scatterplot of all participants' WER values for responses to speech recorded at the ear.

Figure 1.13: The x-axis is the order of the responses; eg. "1" on the x-axis is the first response given by the participants. The x-axis only corresponds to order of response, and does not indicate the specific noise type or gender of the speaker. A line was fitted to the data using linear regression.

speech and ear recorded speech.

1.5.1 Limitations

Normalized amplitude.

*Some participants noted that the computer screen was a distraction to their ability to perceive the utterances. Cite **?** regarding "Reduced Attentional Capacity", briefly touched on in the Background/lit review.

Number of participants is low.

1.6 Follow-up Pilot Experiments

To expound on the previous study, two additional pilot studies were performed to give insights into possible future directions. The impetus for the first study was the study performed by **?**, which demonstrated that the fundamental frequency (F0) was a tool used by the auditory system to separate a desired source from masking noise. This follow up proposes to recombine the very clear, lower frequencies with the "noisy" upper frequencies recorded at the mouth. The hypothesis is that the auditory system will use the clear fundamental frequency harmonic information in the lower frequencies to extract the upper harmonics out of the noise. Accuracy is predicted to improve over that of the low-pass filtered, "muffled", ear speech (which many participants subjectively observed to be annoying and difficult to understand), as more high frequency information will be present and available for participants. This speech will sacrifice the advantage of being completely "noise-free", to sound more natural. Additionally, since the ear-recorded speech consists of very clean harmonics, it is hypothesized that this speech, combined with the higher frequency mouth recorded speech in the noise-free condition, will perform equal to the plain "clean" speech condition. This will be referred to as the "F0" study.

The second follow-up experiment was based on the concept of "perceptual learning" discussed earlier. This presumes that the auditory system can learn to adapt to understand speech in a degraded signal better over time. According to **?**, significant

learning can occur with even a small number of training trials. While also intuitive, **?** demonstrate that during training, successful recognition of a degraded signal will help one recognize a similar signal more than unsuccessful recognition of a degraded signal. Based on these findings, a short story was read and recorded from inside the ear canal for the participants in this follow up study to listen to prior to completing the experiment itself. Since the type of distortion from the ear recorded signal is regular and predictable, it is hypothesized that perceptual learning will take place (unlike with the unpredictable speech in noise c.f. **?**), and those who have listened to the training story will perform better on ear-recorded speech than those who had not. This will be referred to as the "perceptual learning" study.

1.6.1 "F0" Study Methods

The stimuli used for this study consisted of the exact same sentences produced by the exact same speakers. No modification was performed to the sentences recorded at the mouth. For the sentences recorded at the ear, the same modifications as before (pre-emphasis, lowpass filtering[7], and a second pre-emphasis), but this time the simultaneously recorded speech from the mouth was filtered and combined with the ear-recorded speech. The speech from the mouth was bandpass filtered between 3000Hz and 8000Hz, with a 500Hz slope. This allowed for an overlap of the frequencies from the mouth-recorded speech and the lowpass filtered ear-recorded speech. The two signals were converted to a stereo signal, and then combined into a mono signal. This resulted in clean speech below approximately 2700Hz, and noisy speech above approximately 2700Hz, as seen in Figure 1.14.

Four native speakers of English with self-reported normal hearing participated in this study. The design and procedure of this experiment was exactly the same as the initial perception ex-

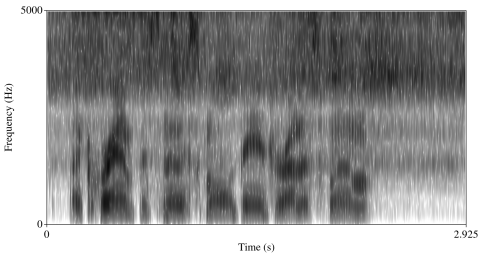


Figure 1.14: A spectrogram of the sentence "A cramp is no small danger on a swim". The low-pass filtered ear-recorded signal was combined with the simultenous

[7]Lowpass filtered allowing 0-2500Hz, with a 500 Hz slope

periment, save the alteration in modifications performed on the ear-recorded stimuli.

1.6.2 "F0" Study Results

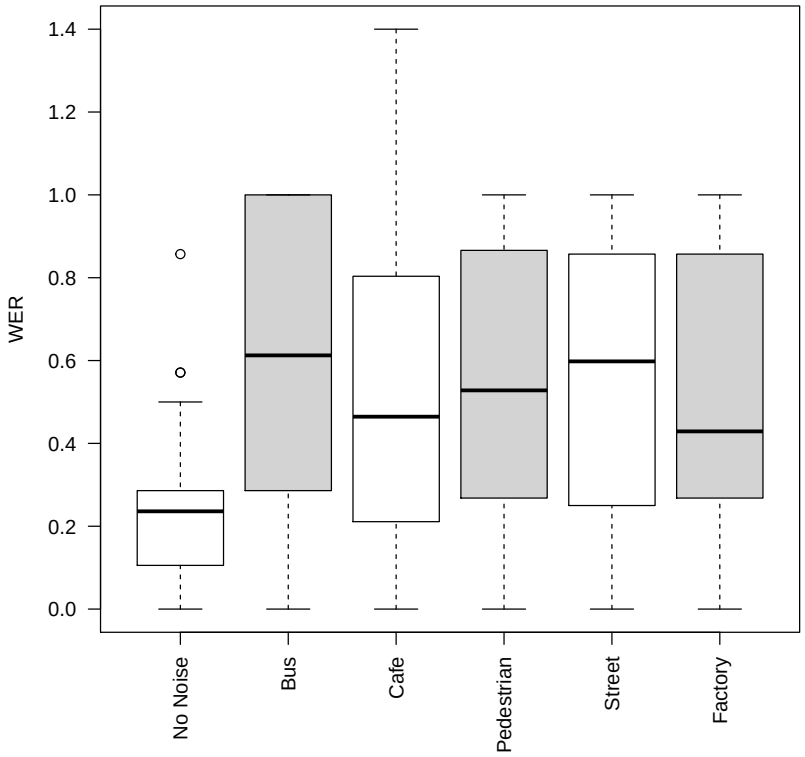


Figure 1.15: Using the data from the five participants who performed the experiment using the speech in which the higher frequencies were added back in from the noisy mouth-recorded speech. Boxplot displaying the average word error rate (WER) averaged over each participant for every noise type. WER is the variable on the y-axis, and noise type is on the x-axis.

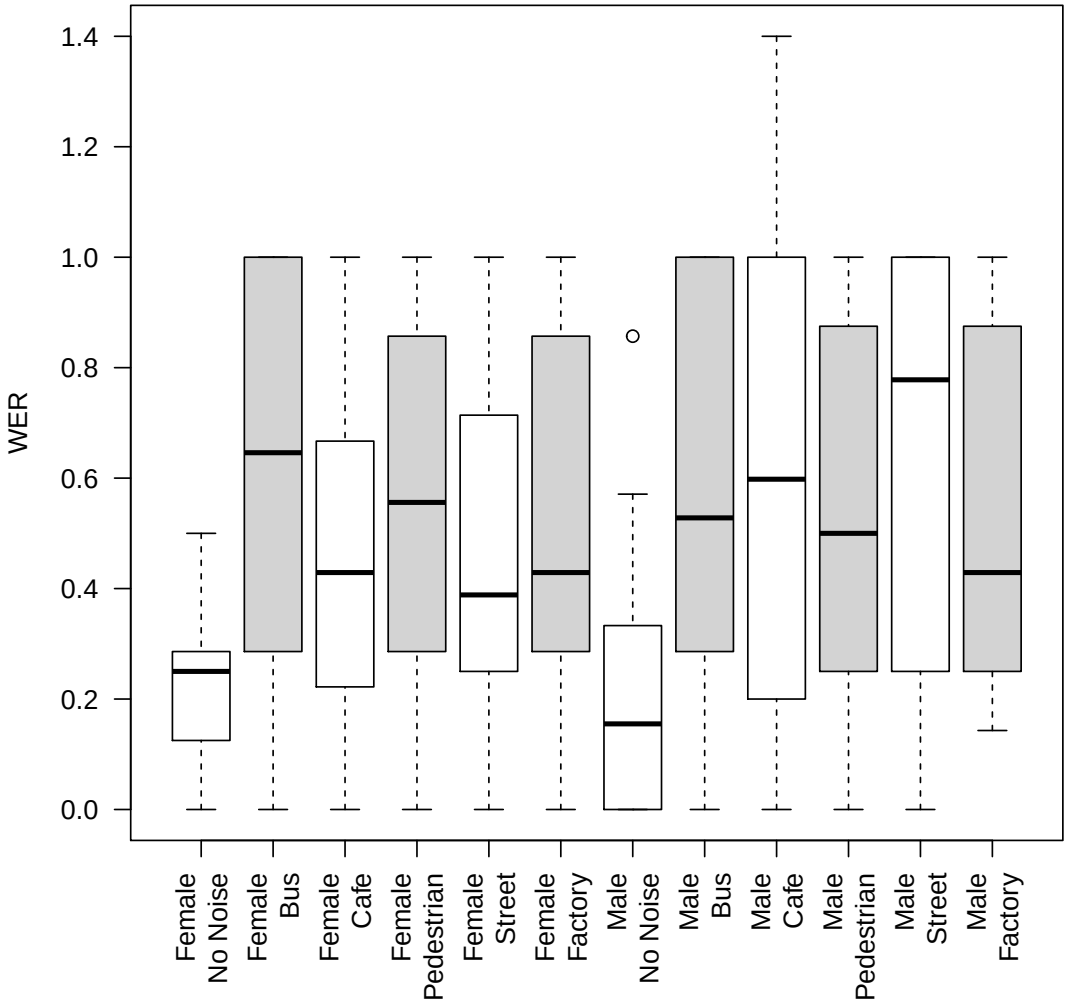


Figure 1.16: Using the data from the five participants who performed the experiment using the speech in which the higher frequencies were added back in from the noisy mouth-recorded speech. Boxplot displaying the average word error rate (WER) averaged over each participant for the interaction of every noise type by the speaker gender. WER is the variable on the y-axis, and noise type by speaker gender is on the x-axis.

1.6.3 "Perceptual Learning" Methods

For this study, a new speaker was recorded from the ear canal (with the same set-up as all previous recordings) reciting the short story "Peter Rabbit", by Beatrix Potter. The recorded had a total length of approximately 5 minutes and 13 seconds. The recorded story underwent the same transformations as the ear-recorded stimuli (pre-emphasis, lowpass filtering[8], and pre-emphasis).

There were four native speakers of English with self-reported normal hearing who participated in this follow up study. They were first presented with a transcript of the story, and asked to listen to the audio and read along. This offers ample chance for "successful" recognition of the degraded ear-recorded signal (cf. **?**). After the reading session, the participants conducted the experiment as was done in the other studies mentioned in Sections 1.3.5 and 1.6.1.

[8]Lowpass filter of 0-2500Hz with a 500Hz slope.

1.6.4 "Perceptual Learning" Results

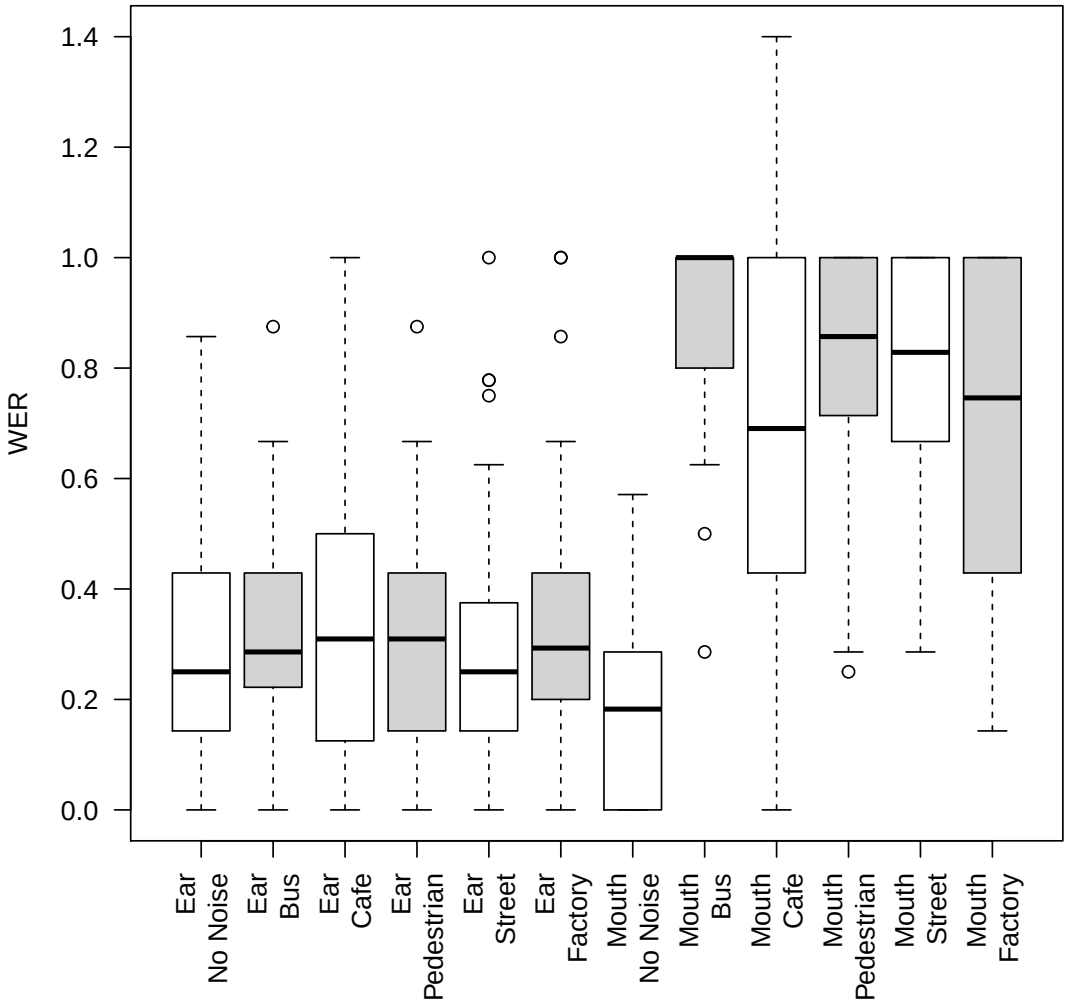


Figure 1.17: Using the data from the five participants who performed the experiment using the speech in which the higher frequencies were added back in from the noisy mouth-recorded speech. Boxplot displaying the average word error rate (WER) averaged over each participant for the interaction of every noise type by the mic location. WER is the variable on the y-axis, and noise type by mic location is on the x-axis.

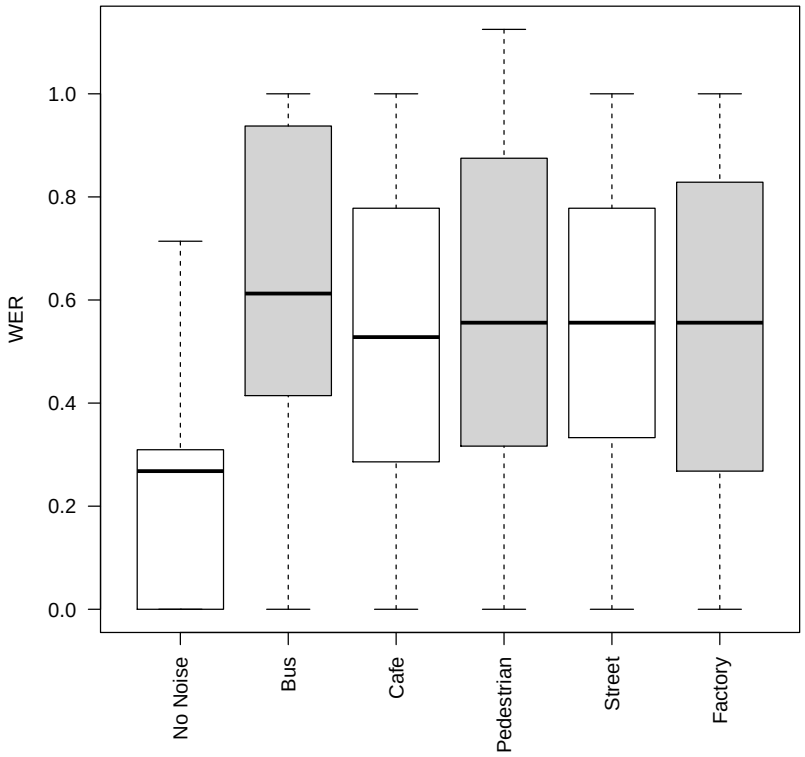


Figure 1.18: Using the data from the four participants who performed the training task in which they listened and read along to a story prior to the experiment. Boxplot displaying the average word error rate (WER) averaged over each participant for every noise type. WER is the variable on the y-axis, and noise type is on the x-axis.
blah

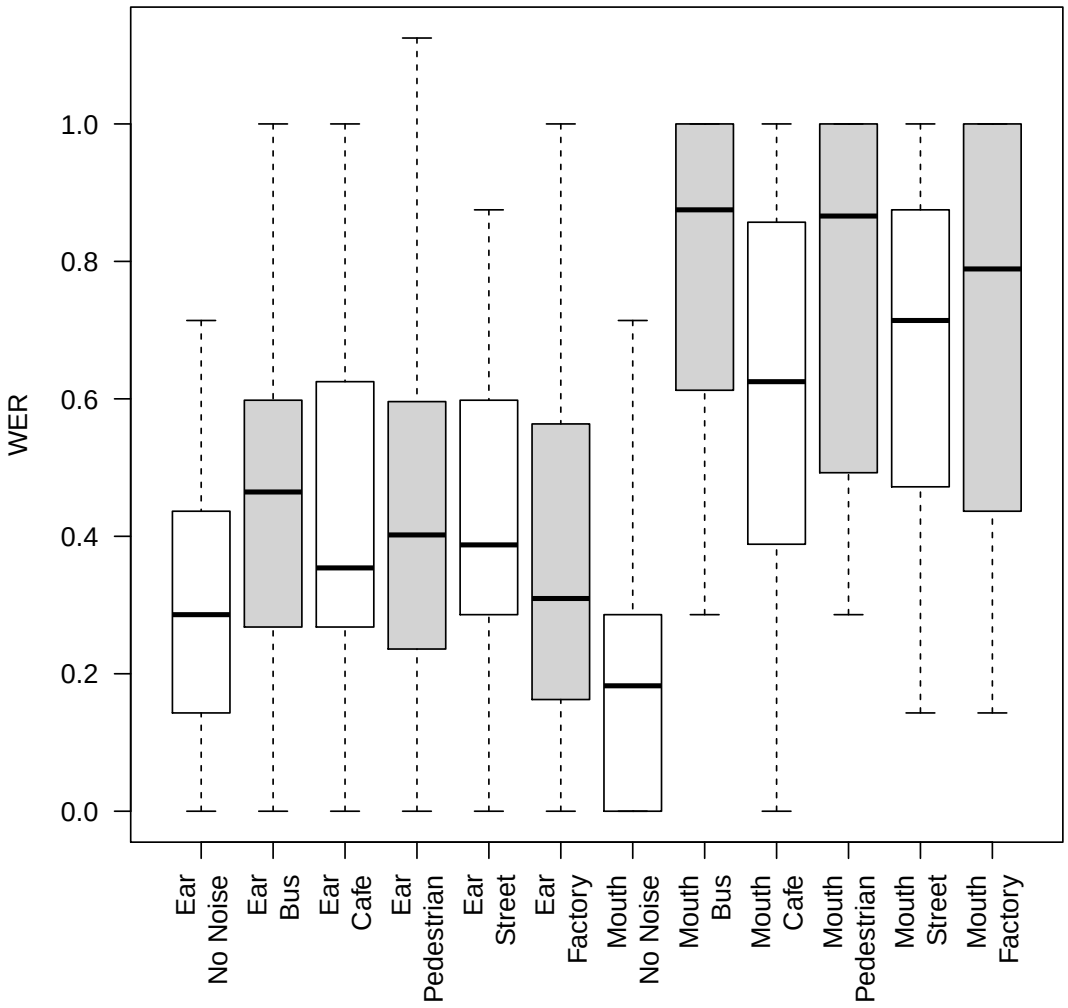


Figure 1.20: Using the data from the four participants who performed the training task in which they listened and read along to a story prior to the experiment. Boxplot displaying the average word error rate (WER) averaged over each participant for the interaction of every noise type by the mic location. WER is the variable on the y-axis, and noise type by mic location is on the x-axis.

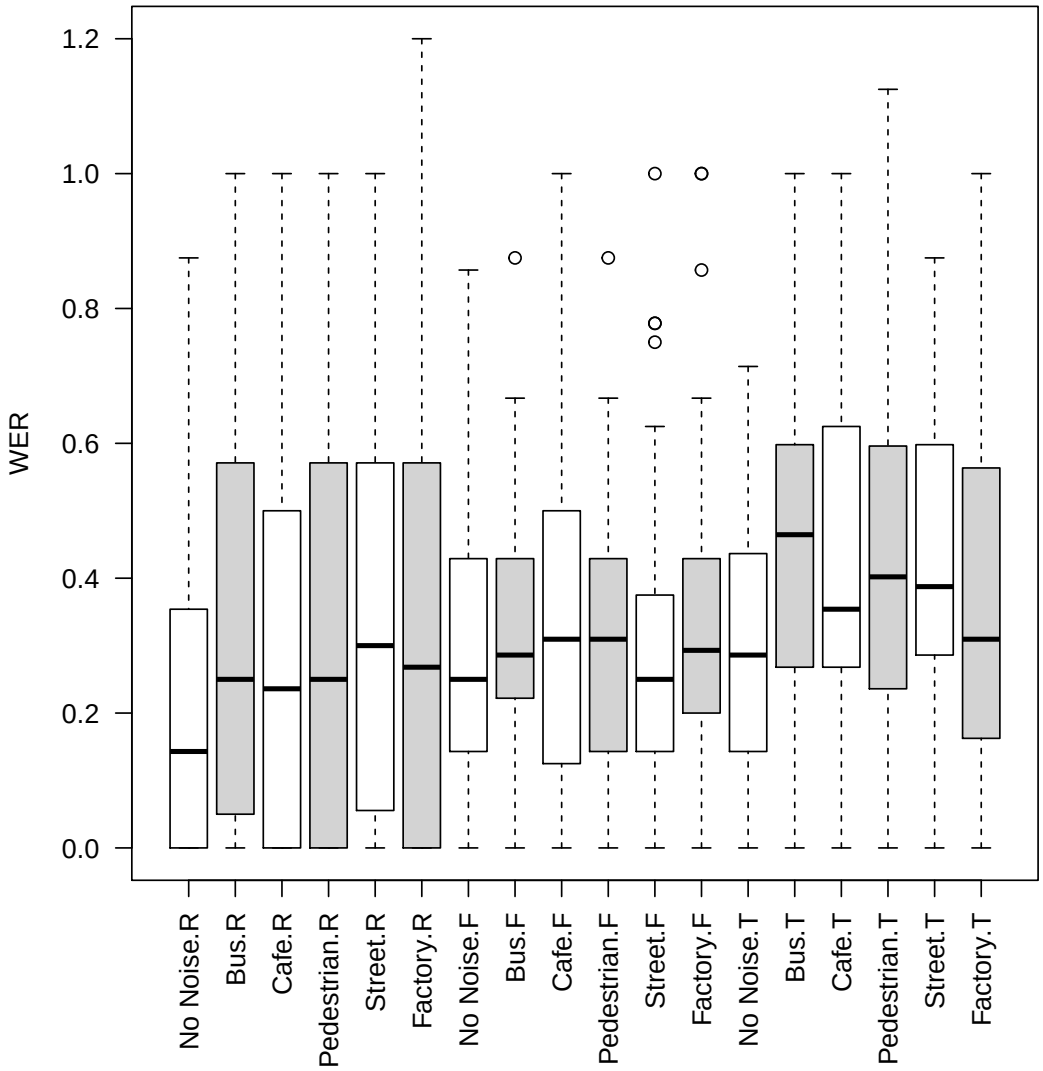


Figure 1.21: AllData-ear. Boxplot displaying the average word error rate (WER) averaged over each participant for the interaction of every noise type by the mic location. WER is the variable on the y-axis, and noise type by mic location is on the x-axis.

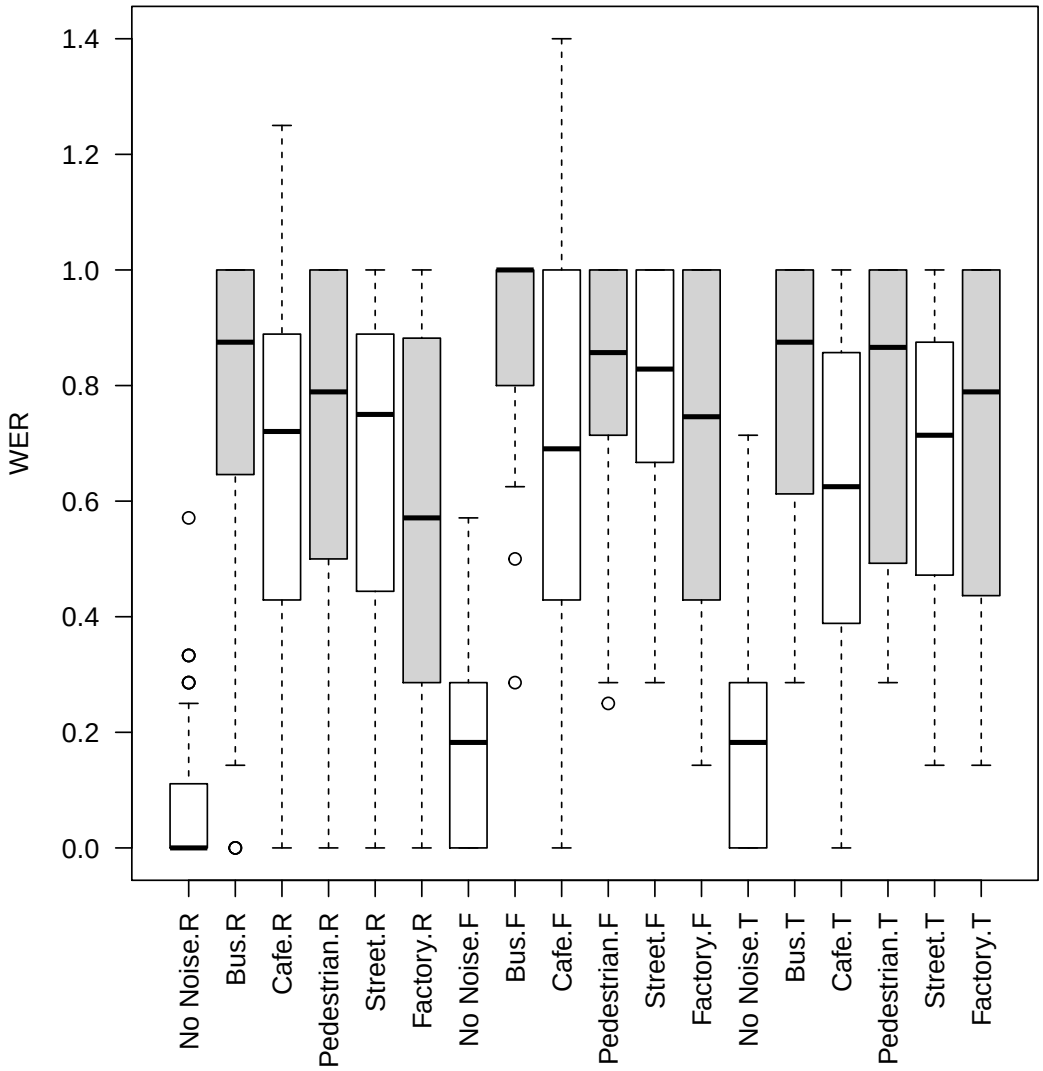


Figure 1.22: AllData-mouth. Boxplot displaying the average word error rate (WER) averaged over each participant for the interaction of every noise type by the mic location. WER is the variable on the y-axis, and noise type by mic location is on the x-axis.

1.6.5 Global Discussion

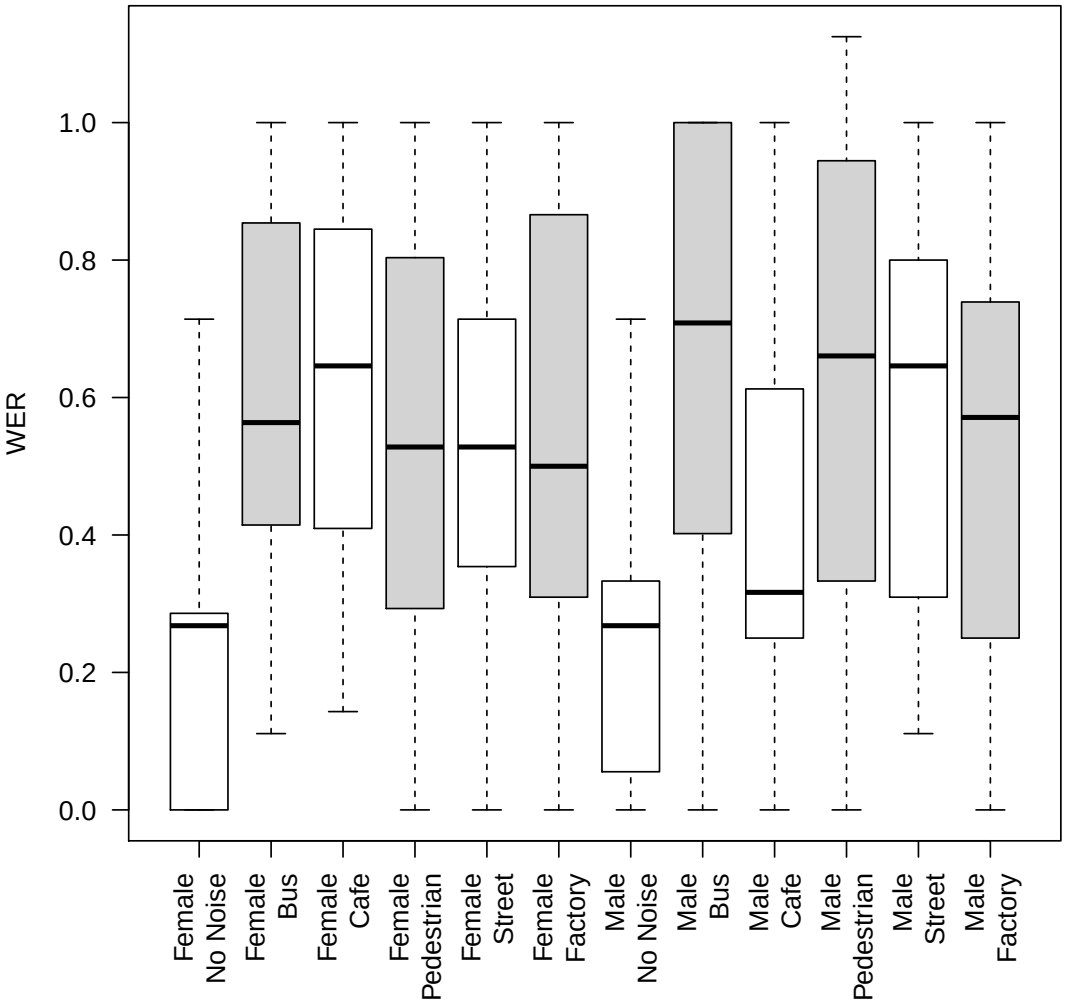


Figure 1.19: Using the data from the four participants who performed the training task in which they listened and read along to a story prior to the experiment. Boxplot displaying the average word error rate (WER) averaged over each participant for the interaction of every noise type by the speaker gender. WER is the variable on the y-axis, and noise type by speaker gender is on the x-axis.

REFERENCES