THE EFFECTS OF DURATION ON HUMAN PROCESSING OF REDUCED SPEECH

Dylan Bernhard[1] and Benjamin V. Tucker[2]

[1]Department of Linguistics, University of Alberta, Edmonton, AB, Canada T6G 2E7, dbernhar@ualberta.ca

[2]Department of Linguistics, University of Alberta, Edmonton, AB, Canada T6G 2E7, bvtucker@ualberta.ca

1 Introduction

For most listeners, parsing of reduced speech is required on a daily basis. When humans produce speech in a fast manner or casual context, there is a tendency to delete, assimilate, and generally weaken the phonemes that are being produced. The resulting acoustic signal can vary by a large degree from what speakers would consider a canonical pronunciation [3]. While studies have been performed to identify exactly how humans can translate from reduced forms to standardized internal representations, we are still far from a comprehensive model to explain all of the aspects of our understanding of how listeners parse a reduced speech input.

Past studies have shown that listeners use a combination of syntax, semantics, rate of speech, and proximal phonetic cues to aid in the processing of reduced speech [4, 5, 6, 9]. This present study investigates the degree to which duration affects human processing of reduced speech stimuli, with the hope of aiding in the building of a comprehensive model of how humans perceive reduced speech.

2 Method

Responses were gathered from 51 participants, 2 sets of responses had to be excluded from the analysis due to lost or damaged data. Participants in this study were students from introductory level Linguistics classes at the University of Alberta, and received partial course credit for participation.

This study consisted of cloze tasks [8] with an auditory component, where sections of reduced speech corresponding to between one and five words were removed from spontaneous utterances (frames) and replaced with silence. Seventy-two frames consisting of phrase length utterances with varying degrees of reduction were extracted from recorded audio of a young adult Western Canadian English speaker [7]. In each frame, a target of 1 to 5 words in length was extracted and replaced with silence to form the original gap stimuli. This set of 72 original gap length stimuli was then manipulated into two additional sets. In the first, the original stimuli were manipulated to have silent gaps with twice the original length. In a similar way, the second set was manipulated to consist of frames with half-length silent gaps. Two controls were also prepared: one where only a visual cloze was presented, and another where the participant heard the full audio from the original frame.

Participants were presented with a visual cloze where the target had been replaced with ten underscores, and a corresponding audio. They were then asked to type in the standard English orthography what word or words they thought best fit in the gap. Each participant was presented with the stimuli in five blocks. The first contained all of the visual cloze stimuli in a randomized order. Blocks two through four were then run, with each consisting of a set of seventy-two audible cloze stimuli with varying gap conditions (a balanced set of 24 short, original, and long gaps). Each of these blocks only contained one of each frame. The final block consisted of the full audio stimuli. This setup ensured that each participant was given an opportunity to respond to every stimulus that had been created for this study.

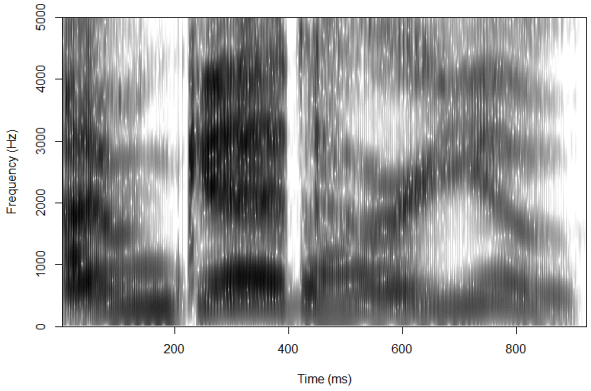

Figure 1: Example original stimulus

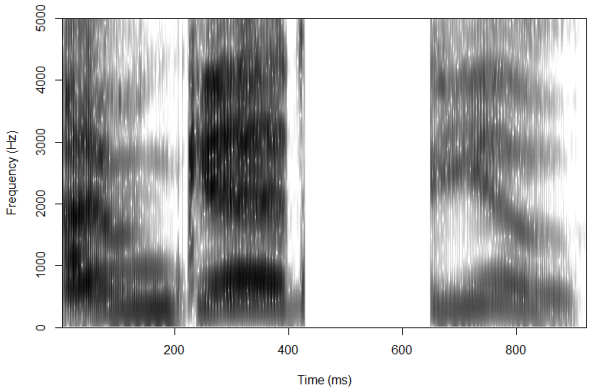

Figure 2: Example auditory cloze stimulus (original gap length)

3 Results

Responses were graded as correct or incorrect based upon what words were part of the original utterance. In trials where the participant did not give a response, the trial was marked as incorrect, as this implies that the information given was not enough to properly identify any possible response. Contractions were marked as their constituent words (e.g., "didn't" becomes "did not").

With only the visual information given in the Visual Cloze task, the rate of correct responses was low, at 8.4%. Given more auditory data, participants became approximately twice as good at the task, with the short gap, original gap, and long gap response rates being 15.8%, 16.7%, and 17.7% respectively. Finally, with the Full Audio task, participants' responses were 79.6% correct.

A logistic linear mixed-effects regression using the lme4 package [1] in R was performed to analyze the significance of the difference between the different tasks (Visual Cloze, Auditory Cloze and Full Audio). A first simple effects analysis with subjects and items as random-effects factors comparing the three different tasks finds that responses to the Visual Cloze task are significantly less accurate than the Auditory Cloze ($p<0.001$) and Full Audio ($p<0.001$). We also found that the responses to the Auditory Cloze were significantly less accurate then to the Full Audio ($p<0.001$). A second analysis was performed on the items for the Auditory Cloze task to investigate differences between the three manipulation types (Original, Short, Long). In this analysis, the responses from the Visual Cloze task were used as a way to measure the predictability of each item. To create this variable we calculated the average response for each item in the Visual Cloze task and used this value as a predictor (vcScore) in the second model. In this model Type of Manipulation and vcScore were our independent variables with subjects and items as random-effects factors. A summary of the model is found in Table 1.

	Estimate	S. Error	z value	Pr(>|z|)
Intercept	-4.47	0.213	-21.02	< 2e-16 ***
Type: Long	0.22	0.082	2.71	0.00672 **
Type: Orig	0.10	0.083	1.22	0.22235
vcScore	11.78	0.748	15.75	< 2e-16 ***

Table 1: Results of linear mixed-effects regression for the auditory cloze manipulations

In this analysis we find that vcScore is a significant predictor, indicating that as the predictability increases the responses are more accurate. We also see a significant different between the Long and the Short manipulations, indication that participants were significantly more accurate for the Long manipulation than the Short manipulation.

4 Discussion

We find that the increase in correct response rate between the visual cloze control and the auditory cloze conditions supports the claim that there are proximal phonetic cues that humans use when parsing speech [7].

With regard to the control condition which presented participants with the full audio frame and with no silent gap, the rate of correct responses is lower than might be expected. At approximately 80%, it would seem that some factor from our stimuli is hindering complete recognition of the target. This is interesting because there was no indication in the original recording the other interlocutor had any difficulty understanding the conversation. This is likely an effect of degree of reduction; more reduced targets are harder to parse than those closer to their canonical pronunciation. This combined with the fact that the participants don't have the full conversational context make it a more difficult task.

The results of this study also indicate that there is an effect of duration on human speech processing of spontaneous speech. However, there does appear to be a trend in the data that the Long stimuli were easier to predict than the Original duration, and that Short stimuli were the most difficult, even though the only significant comparison is between the Long and Short items. This difference suggests that the increased processing time offered by a larger gap is aiding in word prediction.

It is likely that the listeners make predictions with regard to the expected duration when processing speech [2]. This would support a speech processing model that contains activation and competition in the mental lexicon. The reason for shorter gaps being harder to predict correctly is that the correct response is removed from the pool of possible candidates due to the gap being too short to fit it.

The current experimental results suggest that the inclusion of a measure of reduction would benefit the analysis. Measures of reduction have proved only marginally predictive in previous research [6] but may be useful here. It may also be fruitful to investigate the response length. Items that may have been marked incorrect might indicate a correlation between the duration manipulation and response length, such that participants provide shorter responses for the shorter duration and longer responses for the longer silence duration manipulation.

Overall, the results of this study indicate that listeners are sensitive to manipulations of duration in the reconstruction of spontaneous speech. In combination with the results of other studies investigating speech perception, we may yet discover aspects which assist speech recognition technologies to recognize more spontaneous speech.

References

[1] Bates D, Maechler M, Bolker B and Walker S (2014). _lme4: Linear mixed-effects models using Eigen and S4_. R package version 1.1-7, <URL: http://CRAN.R-project.org/package=lme4>.

[2] Dilley, L. C., & Pitt, M. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science*, 21(11). 1664-1670.

[3] Dilts, P. (2013). *Modeling phonetic reduction in a corpus of spoken english using random forests and mixed-effects regression*. Thesis, Department of Linguistics, University of Alberta, Edmonton, AB.

[4] Ernestus, M., Baayen, H., & Schreuder, R. (2002). The recognition of reduced word forms. *Brain and Language, 81*(1), 162-173.

[5] Mehta, G., & Cutler, A. (1988). Detection of target phonemes in spontaneous and read speech. *Language and Speech. 31*(2), 135-156.

[6] Pickett, J., & Pollack, I. (1963). Intelligibility of excerpts from fluent speech: effects of rate of utterance and duration of excerpt. *Language and Speech 6*(3), 151-164.

[7] Podlubny, R. (2013). *Acoustic Decomposition: The Roles of Duration and Intensity in Spoken Language Processing*. Honors Thesis, Department of Linguistics, University of Alberta, Edmonton, AB.

[8] Taylor, W. L. (1953). "Cloze Procedure": a new tool for measuring readability. *Journalism Quarterly, 30*(4), 415-33.

[9] van de Ven, M., Tucker, B. V., & Ernestus, M. (2011). Semantic context effects in the comprehension of reduced pronunciation variants. *Mem. & Cog., 39*(7), 1301-1316.