Evaluating Prosody in Synthetic Speech with Online (Eye-Tracking) and Offline (Rating) Methods

Rajakrishnan Rajkumar, Michael White, Shari R. Speer, Kiwako Ito

Department of Linguistics, The Ohio State University, USA

{raja,mwhite,speer,ito}@ling.osu.edu

Abstract

This study examines the relationship between online processing effects observed in earlier eye-tracking experiments [1, 2] and offline quality ratings gathered for the synthetic and natural speech stimuli used in these experiments, along with their acoustic-prosodic properties. White et al. [2] reported that even high-quality synthetic speech failed to replicate the facilitative effect of contextually appropriate accent patterns found with human speech, while it produced a more robust intonational garden-path effect with contextually inappropriate patterns. They conjectured that both of these effects could be due to processing delays observed with the synthetic speech. In this paper, we present an acoustic analysis of the stimuli used in the eye-tracking experiments and an offline stimuli rating task, which was designed to investigate whether a context-independent measure of utterance quality could predict processing-based effects. The analysis reveals that for synthetic speech, longer adjectives—which provide more processing time—do facilitate anticipatory looks to the target. Larger values of F0 drop (difference between the F0 values of the adjective and following noun) also negatively influenced looks to the target and were negatively correlated with offline ratings, suggesting that this may be a specific acoustic factor that merits attention in future work on improving synthesis quality. Finally, the study shows that online measures of unconscious processing and offline measures of conscious judgments, taken together, can provide a more comprehensive evaluation of synthetic speech than either method alone.

Index Terms: speech synthesis, evaluation, prosody, eye tracking, unit selection

1. Introduction

Traditionally, synthetic speech has been evaluated by offline methods such as listener ratings of speech quality and error rates in transcribing semantically unpredictable sentences [3, 4]. As Swift et al. [5] and van Hooijdonk et al. [6] have noted, however, offline methods do not offer insights into how listeners actually process synthetic speech. These studies used eye tracking to investigate the impact of segmental and supersegmental information on how human listeners process synthetic speech. Subsequently, to study the processing of varied intonation in synthetic speech, White et al. [2] investigated whether different accent patterns in synthetic speech yielded significant differences in anticipatory eye movements. They replicated with synthetic speech Ito and Speer's [1] eye-tracking experiment, where participants followed recorded instructions to decorate holiday trees with ornaments laid out on a grid. The decoration sequences were carefully constructed to include contrasts between consecutively-mentioned ornaments (e.g., *Hang a red star. Next, hang a yellow star.*), as well as locally non-contrastive sequences (e.g., *Hang a yellow tree. Next, hang a green ball.*) The noun phrases in these critical utterances had one of two pitch accent patterns: (1) a contrastive L+H* accent on the adjective, and no accent on the noun, e.g. *hang a YELLOW*$_{L+H*}$ *star*$_{\emptyset}$; (2) H* on the adjective and !H* on the noun, e.g. *hang a yellow*$_{H*}$ *star*$_{!H*}$. Ito and Speer's results using natural speech demonstrated a robust effect of the contrastive L+H* accent together with the prosodically attenuated noun, which produced very early looks to the target cell in contrastive sequences, significantly faster than with the H* accent. Furthermore, in addition to this facilitative effect of L+H*, the study showed an intonational 'garden-path' effect in non-contrastive sequences, where listeners directed looks to the contrastive competitor and delayed looks to the target. In the replication study, however, synthetic speech resulted in a more robust garden pathing effect, but no facilitation was observed. White et al. conjectured that processing delays observed with synthetic speech (similar to delays observed in the earlier studies cited above) may have caused both of these effects. No facilitation might have been found because by the time listeners finished processing the L+H* accent on the adjective, information about the correct referent was already available to them. Stronger garden pathing might have been observed because listeners took longer to process the adjective and as a consequence they might have been updating their referential domain for the target at the same time as the conflicting information from the noun was arriving, causing additional delays in identifying the correct referent. To allow for possible processing delays in future experiments, they suggested giving more time to the listener before the arrival of the disambiguating segmental information (e.g., by using longer or extra adjectives in the stimuli).

In this paper, we present an acoustic analysis of the stimuli used in these eye-tracking experiments, with the aim of identifying specific acoustic factors influencing the processing of synthetic speech. We also present the results of an offline rating task, which was designed to ascertain whether a context-independent measure of quality can predict online effects. The analysis revealed that larger F0 drops between the adjective and noun hinder online processing and negatively influence offline judgements. We also found that longer adjectives do facilitate more looks to the target, consistent with the processing time explanation, even in the case of synthetic speech; however, duration has little influence on offline perceptual judgements. Our analysis also shows that stimuli ratings by themselves do not predict all the effects seen in the eye-tracking experiments. Thus to make a statement about the quality of synthetic speech, the results of both the experiments as well as the findings from the acoustic analysis are needed.

The rest of the paper is structured as follows. Section 2 re-

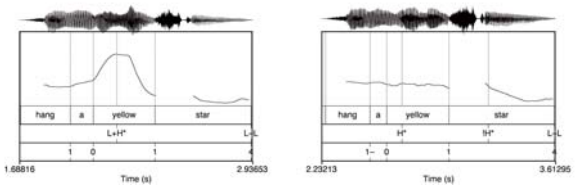

Figure 1: *Example F0 traces and ToBI annotations for [L+H* noaccent] (left) and [H* !H*] (right)*

Contr? / Tune	Adj Dur (ms)	Adj F0 (Hz)	N dur (ms)	N F0 (Hz)
Y / L+H* ∅	356 *330*	332 *299*	458 *489*	148 *148*
Y / H* !H*	366 *332*	223 *207*	524 *549*	192 *164*
N / L+H* ∅	343 *320*	332 *300*	462 *491*	152 *150*
N / H* !H*	368 *316*	223 *208*	516 *558*	197 *163*

Table 1: *Mean duration and F0 of target NPs across conditions; corresponding natural speech values are in italics*

views White et al.'s [2] replication with synthetic speech of Ito and Speer's [1] eye-tracking experiment. Section 3 describes the offline rating task we designed to gather naturalness ratings for the stimuli used in the eye-tracking experiments. Section 4 presents the acoustic analysis of the stimuli, and Section 5 discusses the implications of our findings for speech synthesis evaluation. Section 6 summarizes the main findings of the study and concludes with directions for future work.

2. Experiment I: Eye-tracking

2.1. Design and materials

Participants decorated holiday trees following pre-synthesized auditory instructions. Each participant decorated three trees using ornaments (3 targets: star, tree, ball, and 1 filler: heart) painted in three colors (red, yellow and green). Each tree was decorated with 26 ornaments and the decoration sequences were carefully constructed to include locally contrastive and non-contrastive sequences. In the original Ito and Speer [7, 1] experiment, the auditory instructions were recorded by a trained female phonetician who maintained her overall pitch range and speech rate within and across conditions. All the instruction utterances were ToBI transcribed [8] by an annotator blind to the experimental design. Example F0 traces and the ToBI transcriptions for the natural speech are given in Fig. 1 of [1]; the F0 traces for the synthesized speech are very similar. Table 1 shows the mean durations and F0 values for the adjectives and nouns across conditions, for both the synthesized and natural speech. Compared to natural speech, the mean adjective duration for synthetic speech was longer but the nouns were shorter.

To produce the synthetic stimuli, 192 pseudo-instructions (with ToBI tune annotations) were recorded by the same speaker as in the original experiment, and used to construct a Festival [9] unit selection speech database. The pseudo-instructions—e.g., *hang a greedy*$_{H*}$ *ball*$_{!H*}$—were designed to ensure that the stimuli would require at least two joins, while otherwise providing excellent coverage of diphones in context. Festival was then used to generate critical phrases like *hang a green*$_{L+H*}$ *ball*$_{\emptyset}$. Another trained ToBI annotator then marked F0, adjective and noun durations and certified that the tunes were clear in all the items. Synthesized critical phrases were spliced in at the end of the natural speech stimuli of the original experiments.

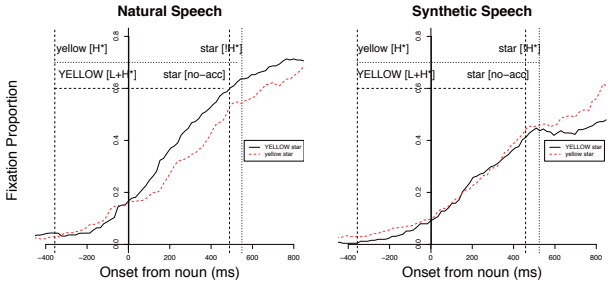

Figure 2: *Fixation proportions to the target in two contrastive sequences, e.g. red star → YELLOW/yellow star*

Volume levels across the segments were normalized according to the default settings of Adobe Audition.

2.2. Participants and eye-tracking procedure

In both experiments, 33 undergraduate students at the Ohio State University participated in partial fulfillment of a course requirement. Data from 29 native speakers of American English are analyzed below. Participants sat in front of a drafting table with the top tilted at 35 degrees to support the ornament display board. They wore lightweight headgear fitted with an eye-camera and a magnetic transmitter that functioned to correct measured eye positions for head movement. Participants followed instructions to choose an ornament from the grid and place it on a small tree located to their right. The x and y coordinates of eye-fixations on the board were recorded at 60 Hz using ASL Eye-Trac 6 data-collection system. The experimenter monitored the participant's eye locations and body orientations via a ceiling-mounted camera, and pressed a key to play each instruction when the participant had finished hanging an ornament and had faced back to the board.

2.3. Results

Participants had nine trials in each of the four critical conditions. The dependent variables were the mean proportion of fixations to the target and to the competitor. The fixation proportion was calculated for each time point by dividing the total number of actual fixations to the target/contrastive competitor by the total number of possible fixations. We used repeated measures analysis of variance (ANOVA) to calculate the significance of the various effects we describe in this section, for eleven 100ms windows starting from the onset of the adjective for both subjects and items.

Fig. 2 shows fixation proportions over time for both types of speech in a contrastive discourse sequence. For natural speech, fixation proportions to the target were significantly higher for [L+H* no-accent] than for [H* !H*] trials (the two lines diverge at the onset of the noun, and remain apart until at least 300ms into the noun). In contrast, for synthetic speech, [L+H* no-accent] does not have any significant facilitative effect as compared to [H* !H*] in searching for a contrastive target (the lines are almost together throughout the entire length of the noun). In addition, the synthesized speech is processed more slowly than the natural speech, as Fig. 3 shows. For both kinds of speech, a direct comparison of fixations to the contrastive competitor in the two non-contrastive sequences is shown in Fig. 4. For natural speech, Ito & Speer reported the relative increase in looks to the competitor with [L+H* no-accent] as a trend in

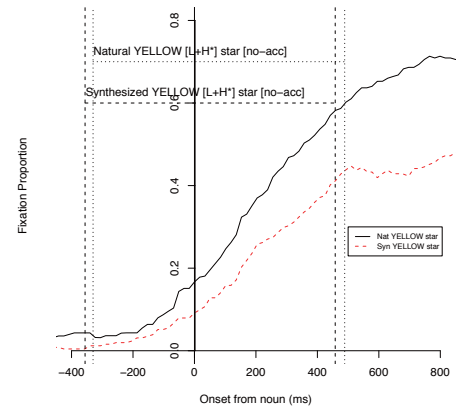

Figure 3: *Fixation proportions to the target due to contrastive accent in contrastive sequences with natural and synthetic speech*

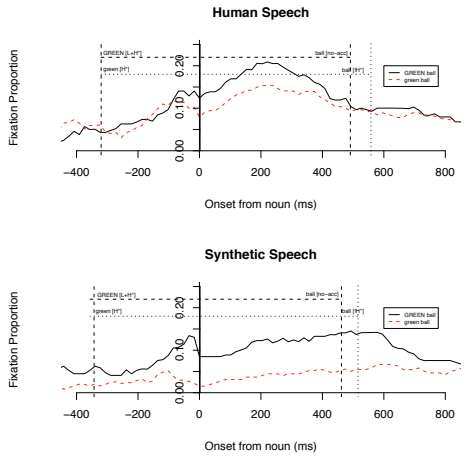

Figure 4: *Fixation proportions to the contrastive competitor in non-contrastive sequences with natural and synthetic speech*

their experiment (the two lines diverge at around 50ms before the noun onset, with fixations to the contrastive competitor rising after that point). In the experiment with synthetic speech, from just before the noun onset there was a clear separation between the fixation proportions to the contrastive competitor in the two conditions. For all time windows up to 300ms into the noun, significance with $p < 0.0001$ is observed, and the remaining two windows of the noun show significant results with $p < 0.001$. But for items (of which there are fewer), this effect is significant at $p < 0.05$ for just the two windows surrounding the onset of the noun. Thus the experiment confirms the effect of synthesized L+H* on the adjective in evoking a contrast set for the upcoming noun, in contrast to H* accents; indeed, the results show a relative garden-path effect that is more robust than the one found with natural speech.

3. Experiment II: Offline Rating Task

3.1. Design and Materials

Quality ratings were collected by means of an online interface. For each item, two types of responses were elicited from the participants: (1) a naturalness rating on a scale ranging from 1 (bad) to 7 (excellent); and (2) a forced-choice judgment as to whether the item was natural or synthetic speech. The stimuli for the rating experiment consisted of 144 items, presented as 2 lists. Each list contained 18 synthetic items from the White et al. [2] eye-tracking experiment and 18 natural items from the Ito and Speer [1] experiment and 36 filler items (18 synthetic and 18 natural items) from a directions-giving domain.[1] Each list was rated by 10 native speakers of American English between the ages of 18 and 40. The filler items were introduced to ensure that each item was presented without any notion of discourse context. We made a conscious decision to collect context-independent naturalness ratings in order to see whether some notion of absolute or context-independent quality had any relationship with the eye-tracking effects observed. Inside each list, domains were strictly interleaved (items from the directions domain alternated with tree domain items) so that subjects did not infer (even if inadvertently) contextual dependencies between successive items presented. Inside each list, sound files were arranged in a pseudo-random order. To this end we ensured that there were no cases where 2 consecutive files in the same domain had the same word sequence repeated. We also ensured that in the same domain, the same tune did not appear more than three times in a row.

3.2. Summary of Results of the Offline Rating Task

Domain	AvrRating	%Acc	Correlation	%Syn Guesses
Overall	4.92	72.00	0.90	48.00
Synthetic Tree	4.64	65.83	-0.83	65.80
Natural Tree	5.47	66.95	0.80	33.08
Synthetic Directions	3.82	74.44	-0.90	74.40
Natural Directions	5.74	80.83	0.86	19.16

Table 2: Domain wise split-up of item ratings

Table 2 summarizes the results of the experiment for each domain. In both the tree and directions domains, the synthetic items received a lower average rating compared to the natural items in that domain. We also calculated identification accuracy (percentage of times that the subjects were actually right in their classification decision) and the percentage of times they classified items as synthetic (irrespective of whether the outcome was right or wrong). There was a positive correlation between ratings and identification accuracy of natural speech items and a negative correlation in the case of the synthetic speech items (the overall correlation adjusts for this inversion). In the tree domain, the identification accuracy was lower for both natural and synthetic speech items in comparison to the directions domain. This could be attributed to the perceivable quality difference in the case of items in the directions domain as compared to the tree domain. The ratings in Table 2 also directly illustrate that the synthetic speech used in the tree experiment was of higher quality than typical unit selection speech, of which the speech from the directions domain is fairly representative; this was an expected result, given that the tree speech corpus was

[1] The directions domain data had L*-L*-L*-L*-L* and H*-!H*-!H*-!H*-!H* tunes.

carefully constructed for excellent coverage and that the phones were hand-aligned. Table 3 illustrates the results of the experiment for each tune. Here we observe that L+H* items were rated lower than the H* !H* items even for the natural speech items. This could be explained by the fact that a contrastive L+H* tune presented out-of-context can sound odd for listeners.

Tone	Type	Avr-Rating	%Acc	Correlation	%SynGuesses
L+H*	Overall	4.34	63.61	0.68	62.00
H* !H*		5.77	69.17	0.92	40.00
L+H*	Synthetic	3.96	75.56	-0.65	75.50
H* !H*	Tree	5.32	56.11	-0.88	56.00
L+H*	Natural	4.72	51.67	0.089	48.3
H* !H*	Tree	6.22	82.22	0.83	17.8

Table 3: Tone-wise split-up of item ratings

4. Acoustic Analysis using Generalized Linear Mixed Models (GLMMs)

Speech/Model	Condition	Chi-Sq
Synthetic	Contrastive sequence	7.168, $p < 0.01$
	Non-contrastive sequence	0.1539
Natural	Contrastive sequence	0.1194
	Non-contrastive sequence	4.687e-08
Model 1 (D.f 3)	looks $\sim$ cond + (1 | subj)	
Model 2 (D.f 4)	looks $\sim$ cond + (1 | subj) + (1 | item)	

Table 4: Model comparison using the log-likelihood test (degree of freedom is 1)

The motivation for conducting an acoustic analysis of the stimuli used in the eye-tracking experiments was to isolate specific acoustic properties potentially influencing listeners' processing of synthetic speech and explain the findings reported by White et al. [2]. As the results described in the Section 2 show, L+H*-accented adjectives in natural speech induce both facilitation and garden pathing effects, while synthetic speech causes only the latter effect. The first step was to examine the extent of item-wise variability in the stimuli of the two experiments in the various experimental conditions. For this analysis (and for all the statistical analyses reported in this paper hereon), Generalized Linear Mixed Models (GLMMs) were used. For these analyses, we adopted a binary choice dependent variable encoding the presence or absence of looks to the desired area of interest within the first 300ms of the noun. This particular time window was postulated on the basis of the ANOVA analyses which demonstrated significant subjects and item effects till this time. Once listeners encountered adjectival information, after a gap of 100-150ms (time for planning and executing saccades), they start fixating on a particular nominal referent. The eye-tracker provides data separated by 17ms and thus in this case, this involved considering 17 data points. Table 4 presents a comparison between two models—with and without item random effects—for both types of speech using a log-likelihood test, where the two models are fitted to the data and their log-likelihoods were compared. The models are presented in R GLMM format.[2] The results indicate that a GLMM with an item intercept was significantly different from a model without an item intercept only in the facilitation case with synthetic speech, the only experimental condition which did not induce an effect akin to natural speech. Thus there is item-wise variation in the eye-tracking results with the synthetic speech stimuli not seen with the natural data. So for all the subsequent analyses, item effects were considered only in the case of those involving synthetic speech facilitation.

[2]The dependent variable occurs to the left of '$\sim$ ' and independent variables occur to the right; '*' denotes an interaction between two fixed effects; random effects are represented after the '|' symbol.

4.1. Acoustic Factors and Looks to the Area of Interest

As predictors of looks to the desired area of interest, we examined several acoustic factors including **Duration of the adjective and noun**, **F0 drop** (difference in F0 between the adjective and noun), **Adjective F0 latency** (time from the onset of the adjective to the adjective F0 peak time-point), and Festival's **Join cost** of the synthetic items. Various acoustic factors were treated as the fixed effects of the model and subjects and items were considered to be the random effects. The basic regression equations (in R GLMM format) for predicting looks to the desired area of interest are:

Synthetic Speech looks $\sim$ acoustic-factor*cond + (1|subj) + (1|item)

Natural Speech looks $\sim$ acoustic-factor*cond + (1|subj)

These equations were used to answer the question: *Is a given acoustic factor a significant predictor of looks to the desired area of interest?* We built models where each factor was considered separately, instead of a single model with all these factors considered together. This was because individual factors were not strongly correlated with each other. All the analyses were done by considering the interaction between acoustic factor and condition as predictors of looks to the target. Such an interaction term was considered because many of the acoustic factors were very related to the experimental condition in which they were used (see Table 1); for example, L+H* items have consistently higher adjective F0 peaks compared to the H* items.

Factor	Coeff	z-value
adjDur	0.001633	0.5702, $p < 0.05$
condL+H*	-3.591332	-2.3068, $p < 0.05$
adjDur:condL+H*	**0.010052**	2.4045, $p < 0.05$
F0Drop	-0.006239	-0.4679
condL+H*	10.170012	2.7537, $p < 0.01$
F0Drop:condL+H*	**-0.050244**	-2.1107, $p < 0.05$

Table 5: GLMMs testing the effect of acoustic factors in the synthetic speech facilitation case

Tables 5 and 6 show the results of GLMM-based acoustic analyses for synthetic speech facilitation and natural speech garden pathing, respectively. For synthetic speech contrastive sequences, longer L+H*-accented adjectives induced more looks to the area of interest. This lends credence to the hypothesis about synthetic speech processing offered in White et al. [2] that given more processing time (as in longer adjectives), listeners would be able to integrate adjectival information more effectively to constrain upcoming nominal referent choices and importantly early looks to the target would be visible at the onset of the noun itself. A preliminary item analysis revealed that duration was related to adjective identity. So, synthetic speech items with longer adjectives *yellow* (411 ms) and *green* (368 ms) did exhibit a facilitation trend, as opposed to *red* (287 ms) which did not exhibit this trend (Figure 5).

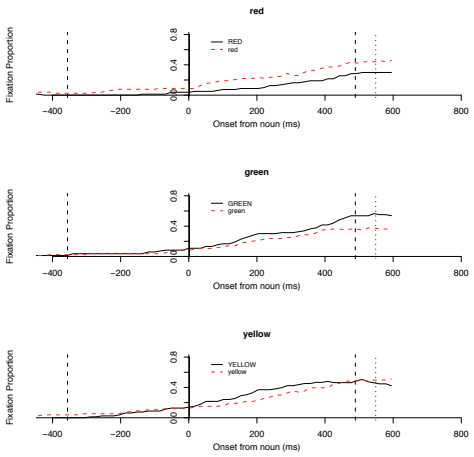

Figure 5: *Colour-wise fixation proportions to the target for synthetic speech facilitation*

The analysis also shows that for synthetic speech, smaller F0 drop induced more looks to the target in contrastive sequences. This leads to the possibility that items with large F0 drop values were considered to be unnatural and diverted listeners from focusing on the desired nominal referent. For natural speech garden pathing, Table 6 also shows that longer adjectives induced more looks to the contrastive competitor, but here greater F0 drop induced more looks to this area of interest. Thus, as might be expected for natural speech, L+H* items with longer adjectives and large F0 drop values in infelicitous sequences induced more looks to the competitor. Similar analyses revealed that none of the other acoustic factors we looked at was significant in predicting natural speech facilitation or synthetic speech garden pathing. In particular, it is worth noting that for synthetic speech, join cost, the one automatic measure of acoustic fit we considered, was not a significant predictor of any of the effects discussed above.

Factor	Coeff	z-value
adjDur	-0.013232	-3.392, $p < 0.001$
condL+H*	-4.778460	-2.963, $p < 0.01$
adjDur:condL+H*	**0.017399**	3.348, $p < 0.001$
F0Drop	-0.05591	-2.291, $p < 0.05$
condL+H*	0.41848	0.231
F0Drop:condL+H*	**0.05436**	2.089, $p < 0.05$

Table 6: GLMMs testing the effect of acoustic factors in the natural speech garden pathing case

4.2. Relationship between Ratings and Looks

This analysis connects the eye-tracking experiment described in Section 2 and the offline rating task described in Section 3 by investigating the relationship between the quality ratings and looks to the relevant areas of interest. The motivation was to find out whether a context-independent measure of quality could predict processing-based effects. Table 7 shows the results of the GLMM analyses for items in a contrastive sequence. Note that for both kinds of speech, the first row shows the analysis of rating by itself, while the second row shows the analysis of rating in its interaction with the tune condition. For synthetic speech, rating by itself was not a significant predictor of looks to the target. This is probably because of the high correlation between rating and condition (condition does not predict looks to the target for synthetic speech facilitation), as seen in Table 3, where L+H* items are rated lower across the board. But the rating-condition interaction actually predicts looks to the target. The positive coefficient estimated by the regression model indicates that higher rated items induced more looks to the target in the L+H* condition. Since the independent quality ratings for the class of items rated as least natural predicted facilitatory looks to the target during the processing of synthetic speech, we might have expected the ratings to predict facilitation in the case of the better-rated natural speech items as well. However for natural speech neither the rating nor the rating-condition interaction was a significant predictor of looks to the target, suggesting that the natural speech items all sound natural in context. In the case of garden pathing, Table 8 shows the corresponding GLMM analyses. Here, the analysis which considers the interaction between rating and condition does not significantly predict looks to the target for either type of speech.

Speech	Factor	Coeff	z-value
Synthetic	rating	-0.21359	-1.30
	rating	-1.3554	-2.806, $p < 0.001$
	condL+H*	-7.2991	-2.288, $p < 0.05$
	rating:condL+H*	**1.4334**	2.109, $p < 0.05$
Natural	rating	-0.1783	-1.6023
	rating	-0.04947	-0.1798
	condL+H*	-1.09159	-0.4253
	rating:condL+H*	0.29348	0.6000

Table 7: GLMMs testing the effect of rating on looks to the target in the facilitation case

Speech	Factor	Coeff	z-value
Synthetic	**rating**	**-0.6577**	-3.639, $p < 0.001$
	rating	-0.5529	-1.5310
	condL+H*	-0.9457	-0.3996
	rating:condL+H*	0.3700	0.6972
Natural	**rating**	**-0.4779**	-3.441, $p < 0.001$
	rating	-0.7097	-1.5487
	condL+H*	1.0430	0.2908
	rating:condL+H*	-0.3258	-0.4959

Table 8: GLMMs testing the effect of rating on looks to the competitor for garden pathing

4.3. Correlation between Rating and Acoustic Factors

In the light of the findings about F0 drop and adjective duration influencing processing of synthetic speech (described in Section 4.1), we examined whether specific acoustic factors also influenced the offline ratings, in order to see whether the same factors were at work in both situations. The quality ratings were highly correlated with F0 drop ($r = -0.76$ for synthetic speech and $r = -0.87$ for natural speech; both at $p < 0.001$), while adjective duration was not correlated with the ratings ($r = 0.07$ for synthetic speech and $r = 0.01$ for natural speech; both at $p < 0.001$). Thus F0 drop affects both online processing and the offline naturalness ratings, while adjective duration was a factor specific to online processing. A possible criticism of this finding is that H*-accented items (which have smaller values of F0 drop) may have been rated relatively higher because the corresponding L+H*-items were rated lower, due to the absence of context as mentioned before. But, as the results of the eye-tracking experiment show, in the facilitation case of syn-

thetic speech, L+H*-accented adjectives with smaller F0 drop induced more looks to the target. Thus to confirm that F0 drop is a relevant factor influencing the quality of items as revealed by the offline rating task, we also need the results from the eye-tracking experiment. On the other hand, adjective duration (or for that matter any of the other acoustic factors we looked at, except F0 drop) was not correlated with the ratings in any of the conditions, while it is a factor which is relevant to the processing of synthetic speech. Thus, while adjective duration has an important effect on the time course of processing, it does not appear to be a factor related to the quality of the speech items.

5. Discussion

A close look at the results of the two experiments presented in the previous sections suggests that the online measures of unconscious processing and the offline measures of conscious judgments together contribute to a more comprehensive evaluation of synthetic speech than is possible using either method alone. The acoustic analysis of the eye-tracking stimuli revealed that larger values of F0 drop in a [L+H*-no accent] contrastive discourse sequence hinders processing of synthetic speech. The ratings study confirms the effect of F0 drop, since synthetic items with larger F0 drop values were given lower ratings (as were natural items). The fact that both studies demonstrate the influence of F0 drop point to the conclusion that this is a factor which merits attention in future speech synthesis work.

The analysis also brings to light the fact that longer adjectives facilitate more looks to the target for synthetic speech in [L+H*-no accent] contrastive sequences. Consistent with this finding, we noted in an item-specific analysis of the synthetic speech stimuli that longer color adjectives tended to facilitate more looks to the target. However, adjective duration was not seen to be a factor having any bearing on the offline ratings, suggesting that longer adjectives are simply providing more time for listeners to exhibit anticipatory effects of contrastive prosody.

The ratings study indicates that the synthetic stimuli used in the eye-tracking experiment are of higher quality than typical unit selection speech, as is evident from the difference in ratings with the directions domain data. The analysis which looks at both studies reveals that higher synthetic speech ratings do indicate increased looks to the target in the facilitation case. It is interesting to note, however, that the ratings are not otherwise predictive of eye-tracking effects, at least when their interaction with condition is taken into account. This could be because context-independent ratings do not provide an accurate picture of the quality of items, especially in the case of contrastive prosody presented out of context. Thus the studies demonstrate that to make a statement about the quality of synthetic speech, one can obtain a more complete picture by relying on the results of the eye-tracking experiment as well as the offline rating study.

6. Conclusions

Our study revealed that for synthetic speech, adjectives with longer durations were associated with an increase in looks to the target in felicitous trials. This lends credence to White et al.'s [2] conjecture that more processing time may simply be needed to see facilitations effects with synthetic speech, in order to offset the observed delays in processing synthetic speech compared to natural speech. The offline rating task illustrates that synthetic items that were rated as more natural also produced more facilitation in looks to the target in the online eye-movement monitoring experiment, suggesting that the less natural items were hindering the effect of condition observed with the natural speech data. We also found that the eye-tracking effects and acoustic analysis help us interpret the context-independent ratings better. Larger drops in F0 between the accented adjective and following noun in the case of synthetic speech resulted in reduced looks to the target. This suggests that unnaturally large F0 drops may be a specific acoustic factor that hinders online processing, and thus merits specific attention in future work on improving synthesis quality. Interestingly, the offline ratings correlated highly with F0 drop, but not with adjective duration (or with any of the other acoustic measures we considered). As future work it would be interesting to design an eye-tracking experiment with more items and specifically controlling for F0 drop and adjective duration, the two acoustic factors which the present study brought to light, and examine the eye-tracking effects observed.

7. Acknowledgements

We thank Cynthia Clopper and the OSU Speerlab group for providing useful feedback, Dominic Espinosa for providing the filler stimuli of the offline rating task, Laurie Maynell for serving as our voice talent, and Rob Clark for help with Festival. This work was supported in part by an OSU Arts & Humanities Innovation Grant.

8. References

[1] K. Ito and S. R. Speer, "Semantically-independent but contextually-dependent interpretation of contrastive accent," in *Prosodic categories: production, perception and comprehension*, P. Prieto, S. Frota, and G. Elordieta, Eds. Springer, to appear.

[2] M. White, R. Rajkumar, K. Ito, and S. R. Speer, "Eye tracking for the online evaluation of prosody in speech synthesis: Not so fast!" in *Proc. of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH-09)*, 2009.

[3] M. Fraser and S. King, "The Blizzard Challenge 2007," in *Proc. Blizzard Workshop (in Proc. of the 6th ISCA Workshop on Speech Synthesis)*, 2007, pp. 7–12.

[4] R. A. J. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *Proc. Blizzard Workshop (in Proc. of the 6th ISCA Workshop on Speech Synthesis)*, 2007, pp. 1–6.

[5] M. D. Swift, E. Campana, J. F. Allen, and M. K. Tanenhaus, "Monitoring eye movements as an evaluation of synthesized speech," in *Proc. of the IEEE 2002 Workshop on Speech Synthesis*, 2002.

[6] C. van Hooijdonk, E. Commandeur, R. Cozijn, E. Krahmer, and E. Marsi, "The online evaluation of speech synthesis using eye movements," in *Proc. of the 6th ISCA Workshop on Speech Synthesis (SSW-6)*, 2007.

[7] K. Ito and S. R. Speer, "Use of L+H* in immediate contrast resolution," in *Proc. of Speech Prosody 2008*, 2008.

[8] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," *Proceedings of ICSLP92*, vol. 2, pp. 867–870, 1992.

[9] R. A. J. Clark, K. Richmond, and S. King, "Multisyn: Open-domain unit selection for the Festival speech synthesis system," *Speech Communication*, vol. 49, no. 4, pp. 317–330, 2007.