INTERSPEECH 2009 BRIGHTON

Eye Tracking for the Online Evaluation of Prosody in Speech Synthesis: Not So Fast!

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

Department of Linguistics, The Ohio State University, USA

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

Abstract

This paper presents an eye-tracking experiment comparing the processing of different accent patterns in unit selection synthesis and human speech. The synthetic speech results failed to replicate the facilitative effect of contextually appropriate accent patterns found with human speech, while producing a more robust intonational garden-path effect with contextually inappropriate patterns, both of which could be due to processing delays seen with the synthetic speech. As the synthetic speech was of high quality, the results indicate that eye tracking holds promise as a highly sensitive and objective method for the online evaluation of prosody in speech synthesis.

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

10.21437/Interspeech.2009-665

1. Introduction

In evaluating speech synthesis, offline methods have remained predominant. For example, in the most recent editions of the Blizzard Challenge [1, 2], speech synthesizers have been evaluated via listening tests involving mean opinion scores of how natural or human a synthetic voice sounds, along with word error rates in transcribing semantically unpredictable sentences. As Swift et al. [3] and van Hooijdonk et al. [4] have noted, however, offline methods do not offer insight into how listeners actually process synthetic speech. As an alternative, online (and objective) method of evaluation, Swift et al. proposed using eye tracking to investigate how synthetic speech is processed incrementally in comparison to human speech. Swift et al. showed that human listeners process segmental information in synthetic speech incrementally at both the lexical and discourse levels, though with processing delays in comparison to human speech. They also found subtle differences in the online processing of two synthetic voices, demonstrating the potential of eye tracking to serve as a fine-grained evaluation measure.

Subsequently, van Hooijdonk et al. used eye tracking to investigate the impact of both segmental and supersegmental information on how human listeners process synthetic speech, comparing both a diphone voice and a unit selection voice to human speech. In their experiment, participants followed two consecutive instructions (in Dutch) to click on an object within a visual display. The first instruction mentioned an initial referent (e.g. *roze vork / pink fork*) with neutral intonation. The second instruction mentioned a target referent using a double accent pattern, the choice of which was forced by the unit selection synthesizer, as it typically produced these patterns and did not allow the accent pattern to be controlled. The target could either be of the same type but a different color (e.g. *blauwe vork / blue fork*), or a different type and color (e.g. *blauwe vos / blue fox*). In both cases, the other possible referent served as a competitor (there was also an unrelated distractor). When the target was of a different type, the double accent pattern was considered felicitous; in contrast, when the target was of the same type, the double accent pattern was considered infelicitous. The results of the experiment showed that the diphone voice induced significantly more fixations to the competitor than the unit selection voice or human speech, which could be explained by the relatively poor segmental intelligibility of diphone synthesis. Perhaps surprisingly though, in all three voice conditions, there were significantly more anticipatory looks to the competitor when the noun was of a different type, despite the expected felicity of the double accent pattern in this context. As the disambiguating segmental information in the noun arrived, looks to the competitor subsided more quickly with the human speech, echoing Swift et al.'s findings of processing delays with synthetic speech.

It is not entirely clear how unexpected the anticipatory looks to the competitor in the different type condition should be taken to be, as van Hooijdonk et al. did not report on whether listeners perceived the accents as contrastive, and did not provide an acoustic analysis. Moreover, as they did not compare different accent patterns, there was no way to observe the impact of felicitous and infelicitous tunes in the same context.

In this paper, we present an experiment that investigates whether different accent patterns in synthetic speech yield significant differences in anticipatory eye movements. The experiment replicates with synthetic speech Ito and Speer's [5, 6] 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*$_{!H*}$. The results 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 double accent pattern. Furthermore, in addition to this facilitative effect of L+H*, the study showed an intonational 'garden-path' effect in non-contrastive sequences, with increased looks to the contrastive competitor and delayed looks to the target. A subsequent experiment with size instead of color adjectives established the statistical significance of this effect.

Our experiment used a custom Festival Multisyn [7] unit selection voice with prosodic specifications given in APML [8]. A previous evaluation [9] of the limited domain Festival voice

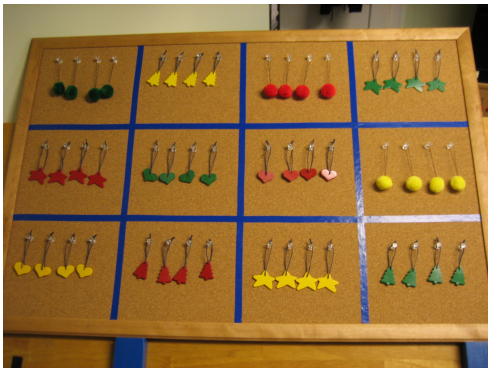

Figure 1: *Example ornament layout on a grid*

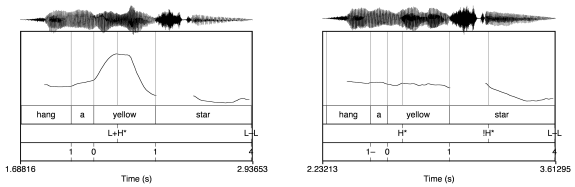

Figure 2: *Example F0 traces and ToBI annotations for [L+H* noaccent] (left) and [H* !H*] (right)*

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*

used in the FLIGHTS concept-to-speech system [10] has shown that such prosodic specifications yield significantly more natural synthetic speech in listening tests and in an expert evaluation. To our knowledge, this paper is the first to investigate the potential of using eye tracking for the online evaluation of varied intonation in synthetic speech.

As the quality of speech synthesized with a unit selection voice depends in large part on how well the speech database covers the target utterances, we were immediately confronted with the question of how to design the speech database for the voice used in our experiment. Since we did not know whether to expect to find similar facilitative and garden-path effects with varied intonation in synthetic speech as Ito and Speer found with human speech, we decided to construct a database that would enable high quality (though still non-trivial) synthesis for the current experiment.

2. Experiment

2.1. Design and materials

Participants decorated holiday trees following pre-synthesized auditory instructions. Each participant decorated three trees using the ornaments laid out on three separate grids. Four types of ornaments (3 targets: star, tree, ball, and 1 filler: heart) were painted in three colors (red, yellow and green), yielding 12 ornament sets that occupied 12 cells on each grid. (The size of ornaments was unified within a grid but altered across the three grids to distract participants from the experimental manipulation.) Each cell contained four identical ornaments. The three target ornaments in three colors were distributed to occupy nine out of ten peripheral cells surrounding the two central cells. The two central cells and the remaining one peripheral cell were occupied by the filler ornaments (i.e., hearts). The locations of ornaments were altered across the three boards. An example ornament grid is shown in Fig. 1. Each tree was decorated with 26 ornaments; as mentioned earlier, the orders of decoration were constructed to include locally contrastive and non-contrastive sequences.

In the original Ito and Speer [5, 6] 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 by an annotator blind to the experimental design. Example F0 traces and the ToBI transcriptions for the natural speech are given in Fig. 2 of [5]; 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.

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 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 the green*$_{L+H*}$ *ball*$_{\emptyset}$. Another trained ToBI annotator then marked F0, adjective & 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. Volumes across the segments was normalized according to the default settings of Adobe Audition.

2.2. Participants and eye-tracking procedure

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.

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. Trials in which a

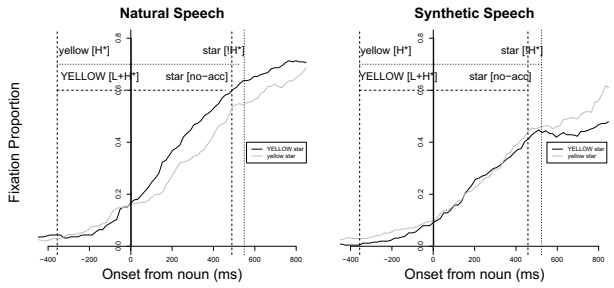

Figure 3: *Fixation proportions to the target in two contrastive sequences, e.g. red star → YELLOW/yellow star*

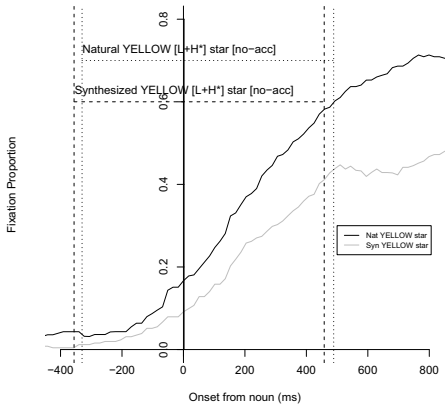

Figure 4: *Fixation proportions to the target due to contrastive accent in contrastive sequences with natural and synthetic speech*

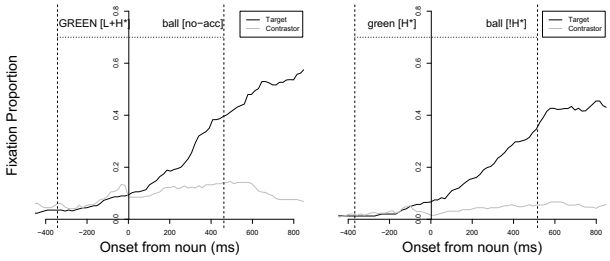

Figure 5: *Fixation proportions to the target and contrastive competitor in two non-contrastive sequences with synthetic speech, e.g. yellow tree → GREEN/green ball*

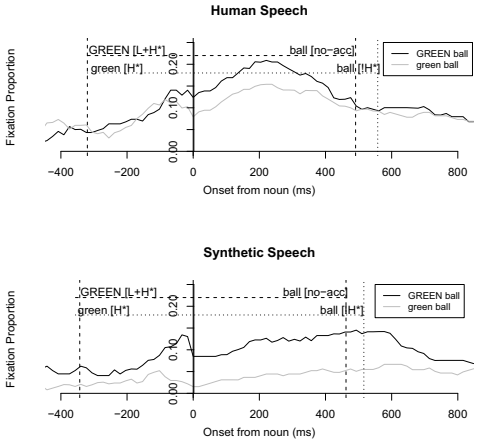

Figure 6: *Fixation proportions to the contrastive competitor in non-contrastive sequences with natural and synthetic speech*

participant was looking at the intended target at the onset of the adjective were eliminated. After this correction, we use repeated measures analysis of variance (ANOVA) for calculating the sigificance of the various effects we describe in this section. For eleven 100ms windows starting from the onset of the adjective, we did ANOVA calculations for both subjects and items.

As Fig. 3 shows, when synthetic speech stimuli are used in a contrastive discourse sequence, [L+H* no-accent] does not have any facilitative effect as compared to [H* !H*] in searching for a contrastive target. This is unlike the response of subjects to natural speech. For natural speech, the two lines diverged at the onset of the noun, and until about 300ms into the noun, fixation proportions to the target were significantly higher for [L+H* no-accent] than for [H* !H*] trials. But for synthetic speech, the lines are almost together throughout the entire length of the noun. Thus synthesized [L+H* no-accent] tunes did not have a significant effect in facilitating more looks towards the target. In addition, the synthesized speech is processed more slowly than the natural speech, as Fig. 4 shows.

The left panel of Fig. 5 shows that in non-contrastive sequences using synthetic speech, [L+H* no-accent] does evoke contrast, as looks to the contrastive competitor keep increasing past the end of the noun. By contrast, the right panel shows that [H* !H*] in non-contrastive sequences does not result in more looks to the contrastive competitor. This is similar to the behavior using natural speech (not shown here for space reasons). For both kinds of speech, a direct comparison of fixations to the contrastive competitor in the two non-contrastive sequences is shown in Fig. 6. For natural speech, the two lines diverge at around 50ms before the noun onset, with fixations to the contrastive competitor rising after that point. Ito & Speer reported the relative increase in looks to the competitor with [L+H* no-accent] as a trend in their experiment using color adjectives (the one we replicated), and a subsequent experiment using size adjectives—which involved a more difficult visual search task—established the statistical significance of the effect. In the current 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. Importantly, up to 100ms into the noun, this effect is statistically significant for both subjects as well as items (at $p < 0.05$). The effects for subjects are in fact very prominent: 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.

4. Discussion

It is remarkable that despite our best attempts to produce high quality synthetic stimuli, the eye movement monitoring method was able to clearly distinguish how listeners process these stimuli in comparison to human speech. Our informal listening tests suggested that the stimuli were all of excellent quality, with only quite subtle audible joins or other artefacts, and our ToBI annotator had no trouble identifying the intended tune in a blind test.[1] Additionally, few of the participants suspected that the experiment involved synthetic speech prior to the debriefing.

As our results failed to replicate the facilitative effect of contextually appropriate prosody with human speech, while producing a relatively stronger garden-path effect, one might be tempted to conclude that contextually appropriate prosody in synthetic speech does not help listeners, though contextually inappropriate prosody can certainly hamper their processing, and thus that one might be better off aiming for some sort of neutral intonation rather than risk getting the tune wrong for the context. Though we cannot prove that this glass-entirely-empty position is misguided, we can speculate on other possible reasons for the lack of facilitation in our experiment. Since our experiment confirms the presence of processing delays with synthetic speech that have been observed in earlier studies by Swift et al. and van Hooijdonk et al., one likely reason is that a processing delay in interpreting the segmental and supersegmental information in the adjective means that the disambiguating segmental information in the noun in a sense arrives too soon—that is, before there is time to see any anticipatory facilitative effect of a contrastive L+H* accent on the adjective. This suggests that designers of future eye-tracking experiments for online synthesis evaluation should take into account possible delays in processing, and accordingly lengthen the time before the disambiguating segmental information arrives, for example by using longer or extra adjectives in the stimuli. Another possibility is to use a more difficult visual search, which can lengthen the time window in which facilitation plays out, as in Ito and Speer's [6] related experiment using size rather than color adjectives, where the task of identifying relative size was found to be more difficult than identifying color.

With the relatively stronger garden path effect seen with the synthetic stimuli, we might also seek an explanation in terms of processing delays. If it takes listeners longer to process the contrastive adjective information, then listeners might be updating their referential domain for the target at the same time as the conflicting information from the noun is arriving, causing additional delays in identifying the correct referent. Another possibility is that with somewhat less intelligible segmental information, due to some imperfect joins, listeners are relying more heavily on prosody to guide their interpretation.

Returning to van Hooijdonk et al.'s somewhat surprising results, it is worth noting that they likewise demonstrated a garden path, rather than facilitative, effect, using a much simpler visual layout. In their case, the increased looks to the contrastive competitor could have been due to the first accent (on the adjective) receiving a contrastive interpretation by listeners. That this possibility was not investigated shows the importance of comparing the effects of different accent patterns, not just different contexts, in evaluating speech synthesis with eye tracking. Finally, we note that van Hooijdonk et al.'s finding that the diphone voice had more looks to the competitor than either the unit selection voice or natural speech is consistent with the hypothesis that listeners pay more attention to the prosody when the segmental intelligibility is lower.

[1] We intend to soon conduct a rating study comparing the synthetic stimuli to the natural ones.

5. Conclusions

In this paper, we have presented an experiment which indicates that eye tracking has the potential to be a highly sensitive and objective method for the online evaluation of prosody in synthetic speech, as even with high-quality unit selection synthesis, the results failed to replicate the facilitative effect of contextually appropriate accent patterns found with human speech, while producing a more robust intonational garden-path effect with contextually inappropriate patterns. As we observed processing delays with the synthetic speech that could explain the absence of facilitation, we suggest that experimental designs for eye-tracking evaluations should make allowances for such processing delays, thereby providing sufficient time for any facilitative effects to arise. In future work, we plan to investigate this possibility with unit selection voices of varying quality.

6. Acknowledgements

This work was supported in part by an OSU Arts & Humanities Innovation Grant. We thank Ping Bai for assistance with data analysis, Laurie Maynell for serving as our voice talent, Ross Metusalem for assistance with ToBI-annotating our spoken stimuli, and Rob Clark for help with Festival.

7. References

[1] 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.

[2] V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Proc. Blizzard Challenge 2008 Workshop (INTERSPEECH-08)*, 2008.

[3] 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.

[4] 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.

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

[6] ——, "Semantically-independent but contextually-dependent interpretation of contrastive accent," 2009, submitted.

[7] 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.

[8] B. de Carolis, C. Pelachaud, I. Poggi, and M. Steedman, "APML, a Mark-up Language for Believable Behavior Generation," in *Life-like Characters. Tools, Affective Functions and Applications*, H. Prendinger and M. Ishizuka, Eds. Berlin: Springer, 2004, pp. 65–85.

[9] M. White, R. A. J. Clark, and J. Moore, "Generating tailored, comparative descriptions with contextually appropriate intonation," 2009, submitted.

[10] J. Moore, M. E. Foster, O. Lemon, and M. White, "Generating tailored, comparative descriptions in spoken dialogue," in *Proc. FLAIRS-04*, 2004, pp. 917–922.