

MONITORING EYE MOVEMENTS AS AN EVALUATION OF SYNTHESIZED SPEECH

Mary D. Swift,¹ Ellen Campana,² James F. Allen,³ Michael K. Tanenhaus²

¹Department of Linguistics, ²Department of Brain and Cognitive Sciences

³Department of Computer Science

University of Rochester

Rochester, NY 14627

ABSTRACT

We describe a novel empirical method for synthesized speech evaluation that relies on monitoring participant eye movements as they respond to spoken instructions in a visual workspace. We use instruction sets generated with two different text-to-speech synthesizers and a human voice for baseline comparison. We replicate findings demonstrating incremental processing for natural speech and at the same time demonstrate incremental processing for synthesized speech for both lexical and discourse-level processing. Results reveal differences in the time course of processing between natural and synthesized voices as well as between the two synthesized voices. These results, including the detection of subtle processing differences between voices, demonstrate the potential for eye movement monitoring as a promising new methodology that provides an on-line measure of synthesized speech processing more fine-grained than possible with techniques used to date.

1. INTRODUCTION

The evaluation of text-to-speech (TTS) systems has been an active research area since the inception of speech synthesis technology. System evaluations serve a variety of goals, from establishing a basis for comparison with other systems or with previous versions of the same system, to determining the appropriateness of a given system for a given application, to diagnostic tests such as detecting implementation problems and identifying areas for improving system performance. This multiplicity of evaluation levels has spurred the development of a comparable variety of evaluation tools and methods. Segmental intelligibility has traditionally been primary among the criteria for synthesized speech evaluation and assessment. However, as intelligibility improves and the range of synthesized speech applications increases, in particular the use of spoken dialogue systems, naturalness of synthesized speech has become increasingly important.

In spite of continued development in TTS system evaluation techniques (e.g. [1]), areas for improvement remain. In particular, we focus on the role of human comprehenders in speech synthesis development, a factor emphasized by Benoit along with the need for collaboration between the psychology and speech engineering communities to work to improve understanding of mechanisms of speech communication in general [2]. Evaluation techniques for the most part rely on human listener responses. In some cases, listeners must be trained to provide the specific feedback required, especially at the level of prosody and discourse evaluation.

We propose a novel method for synthesized speech evaluation using the technique of monitoring naive listener eye movements as they respond to spoken instructions. The overarching goal of this work is to investigate the feasibility of this method by demonstrating that 1) processing of synthesized speech is similar to the processing of human speech, and 2) the eye-tracking measure is sensitive enough to reveal subtle differences between the processing of synthesized speech and human speech, and even differences in the processing of speech synthesized using different packages. We begin with well-established psycholinguistic results related to the incremental processing of human speech. We duplicate the experimental conditions that were used in the psycholinguistic research, but critically include synthesized speech conditions. We verify that the relevant results from the original psycholinguistic study were replicated in the human voice condition so it can form an effective baseline. Finally, we compare eye movements from the synthesized voice conditions to the human baseline, and to each other.

Psycholinguistic research has shown that people comprehend human speech incrementally, as utterances unfold in time. This incremental processing of speech gives rise to temporary ambiguities at all levels of linguistic processing, including choice of lexical candidate [3], commitment to syntactic structure, and the establishment of referential domains [4, 5]. For example, as the instruction *Click on the beaker* unfolds, the word *beaker* is briefly consistent with multiple lexical candidates that share initial phonemes, such as *beetle* and *beacon*. Monitoring eye movements in response to spoken instructions in a visual workspace provides an on-line measure that tracks the temporary commitments listeners make as speech unfolds. The probability of fixating on items in the visual workspace (e.g. Figure 1 below) is determined in part by aspects of language processing such as lexical activation [6] and referential domain circumscription [7], as shown in the experiments below. These eye movements occur as early as 200ms after the onset of the referring expression. Thus the fixation probabilities over time provide a continuous measure of processing that is closely time-locked to the spoken instruction. Eye-tracking measurements are sensitive to language processing distinctions from subtle acoustic variation at the sub-segmental level (e.g. [8]) to higher levels of discourse processing such as the effect of prosodic emphasis in reference resolution [9].

If synthesized speech is also processed incrementally, then the eye movement monitoring technique can be used to provide an on-line processing measure more fine-grained than possible with techniques used to date, such as word monitoring and self-paced listening tasks [10]. Thus comparison of synthesized speech processing measures with a natural speech processing baseline can

In *Proceedings of the IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, CA. This research was supported by NIH HD-27206 to MKT

provide valuable information with respect to naturalness.

In the remainder of the paper, we present experimental evidence from studies following the evaluation method described above. We monitor eye movements as listeners respond to pre-recorded instructions generated by a human speaker and by two TTS synthesizers. We replicate psycholinguistic findings demonstrating that people process spoken language incrementally during lexical access as well as higher-level discourse processing. The on-line eye-tracking measure shows that listeners make partial commitments as the instruction unfolds. Specifically, listeners consider multiple lexical alternatives and they establish different referential domains on the fly based on whether a definite or indefinite article is used. Importantly, incremental understanding is observed for both natural speech and synthesized speech. These results, including some suggestive differences in responses with the two TTS voices, establish the potential for using such techniques for synthesized speech evaluation as well as real-time evaluation of spoken dialogue systems.

2. THE EXPERIMENTS

We describe two experiments intended to establish the potential of monitoring eye fixations to evaluate listener responses to synthesized speech. The first experiment addresses incremental processing at the lexical level, while the second experiment addresses incremental processing at the discourse level in terms of use of definiteness information. In both experiments, 15 participants followed spoken instructions directing them to click on objects on a display. The instructions included three voice conditions which were generated using two TTS synthesizers and a digitally recorded human voice. Throughout the session, eye movements were monitored using a lightweight head-mounted pupil/corneal reflection tracking system (ISCAN, model RK-726PCI).

2.1. Experiment 1: Lexical Access

This experiment addressed the question of whether listeners would process words incrementally, entertaining multiple lexical candidates, by analyzing looks to pictures of items that had the same initial phonemes as the name of a target item such as *beaker*, e.g. *beetle*, a picture that rhymed with the target, e.g. *speaker*, and an unrelated picture, e.g. *dolphin*, as in Figure 1.

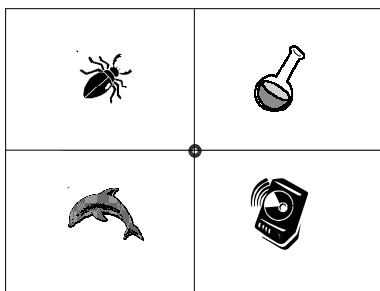


Fig. 1. Sample experimental display for Experiment 1.

2.2. Results

Results show clear evidence for incremental lexical processing for the human voice instructions (as in previous psycholinguistic stud-

ies [6]), as well as for the synthesized instructions. Participants were more likely to look at a competitor when the initial portion of its name overlapped with the target (e.g., *beetle* (target *beaker*)) than when the initial segments differed (*speaker*) or when there was no overlap (*dolphin*). An ANOVA revealed a significant main effect of item type, $F(2,14)=5.02$, $MSE=5.02$, $p < .01$, with no main effects or interaction of voice condition.

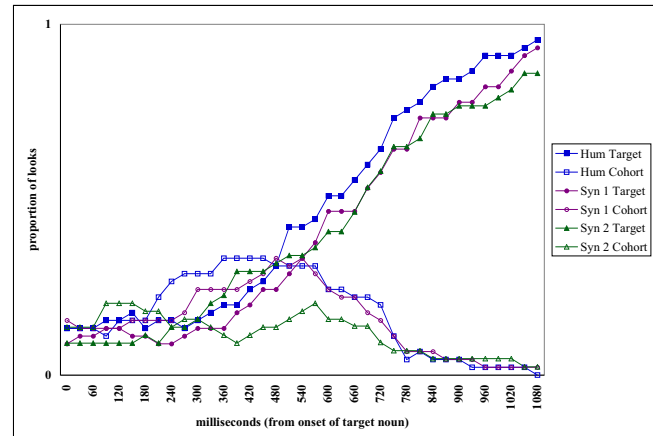


Fig. 2. Time course of looks to target and cohort for all voices.

Figure 2 shows the proportion of trials containing looks to the target and the cohort competitor for all voice conditions. At target word onset, participants look to both target and cohort items. Looks to the cohort subside when participants hear the disambiguating information from the second syllable of the word. Although the pattern of looks to the target is similar for all voices, participants identify the target more rapidly in the human voice condition.

There are differences in the patterns of looks to the cohort across voice conditions, although there is no difference in performance accuracy. This variation could indicate that lexical competition may contribute to the processing load of synthesized speech. The differences in the time course and proportion of looks to the cohort may be evidence for graded effects of incrementality across voices.

2.3. Experiment 2: Referential Domain Circumscription

This experiment addressed the question of whether listeners would use the presence of a definite or indefinite article to differentially circumscribe potential referents as a spoken instruction unfolds. Definiteness cues such as the definite article *the* and indefinite article *a* are monosyllabic and typically unstressed words. These function words often appear in strongly reduced contexts, which are difficult to synthesize. We examined eye movements within displays containing four shapes (e.g., square, circle, triangle, etc.), two unique and a pair of identical shapes, in a grid such as that shown in Figure 1, together with instructions such as *Click on the/a square*.

2.4. Results

Results show that participants use the definiteness information conveyed in the spoken instructions for all voice conditions. For

definite instructions, e.g., *Click on the heart*, where one of the unique objects was the target, participants were more likely to look at the unique distractor than either of the duplicated distractors ($F(2,14)=310.38$, $MSE=7.34$, $p < .01$), and there was no interaction with voice type. For indefinite instructions, e.g., *Click on a square*, either of the duplicated objects could serve as target, and participants were more likely to look at either of the duplicated items than at the definite distractors ($F(2,14)=117.52$, $MSE=10.29$, $p < .01$). Again, there was no interaction with voice type.

First, consider the trials with instructions containing definite articles. Figure 3 shows the proportion of looks over time to the target, the unique unrelated item, and the two duplicate unrelated items in the trials with a definite article for the human voice condition. Results replicate previous findings with human speech [7]. The zero point on the x-axis corresponds to the onset of the noun phrase, e.g., *the heart*. Participants clearly use the definiteness information carried by the article because looks to the duplicate unrelated items subside approximately 100ms before the target is distinguished from the unique unrelated item. Thus the items that are consistent with the definite article are first disambiguated from the items that are not consistent with the definite article, and then the target is disambiguated from the unrelated item.

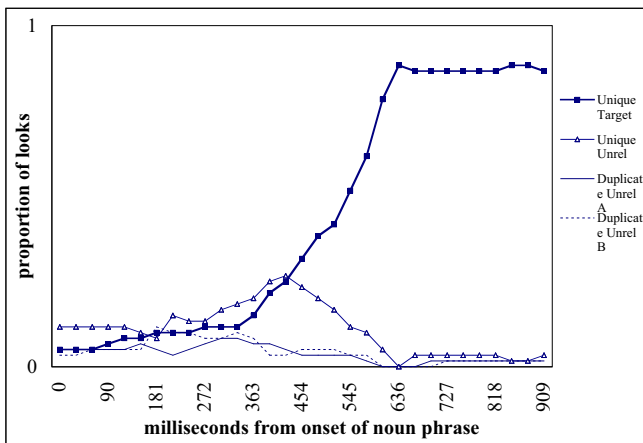


Fig. 3. Time course of looks to all items for definite instructions for the human voice.

The data from the two synthesized voice conditions follow the same general pattern. Specifically, the disambiguation between unique and duplicated items (i.e., definite vs. indefinite) occurs approximately 100ms before the two unique items (i.e., definite target vs. definite unrelated) are disambiguated in the synthesized voice conditions. There is, however, an important difference between the human and synthesized voice conditions - the disambiguation points occur later for the synthesized voices than for the human voice (Figure 4). As there is no significant difference in length of the articles across voice conditions, this cannot be the reason for the delay in disambiguation.

One explanation for this difference could be the increased processing load associated with synthesized voices. We tested this hypothesis by conducting a voice judgment survey. We found no differences in accuracy between voice conditions for the definite instructions - performance was at ceiling for all definite instruction auditory stimuli except one. In contrast, we observed large differences in accuracy for the indefinite instructions ($F(2,14)=6.43$,

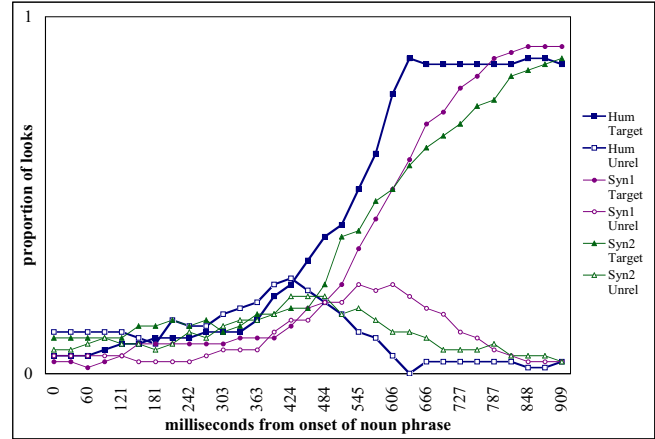


Fig. 4. Time course of looks to target (unique) and unique unrelated items for definite instructions.

$p < .01$). The average accuracy for the human voice was 80%, while it was 65% for SYN1 and 31% for SYN 2, suggesting that the delay in reference resolution for synthesized definite instructions may be due to distributional characteristics of the voices over the course of the interaction.

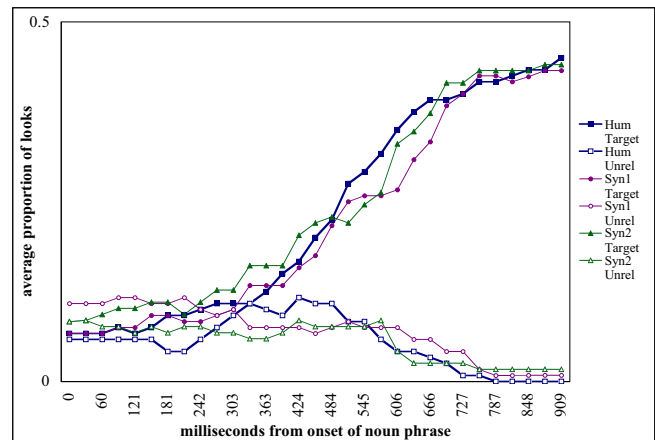


Fig. 5. Time course of looks to all items for indefinite instructions.

Figure 5 shows the average proportion of looks over time to the target (duplicate) items and the unrelated (unique) items for each of the three voice conditions during the trials with indefinite instructions. In these cases, either duplicated item is an appropriate target. Looks to either of the indefinite targets are summed together and represented as a single line for each voice condition in Figure 5. Similarly, looks to either unrelated (unique) item are summed together in a single line for each voice. Again, the zero point on the x-axis corresponds to the onset of the noun phrase, e.g., *a square*.

For all voice conditions, looks to the duplicated items diverge from looks to the unique items at roughly the same point. We cannot tell from this data whether these eye movements are due to processing of the indefinite article or whether they are due to processing of the noun. Given the differences in accuracy for the voice judgment survey, it is surprising that we do not see differ-

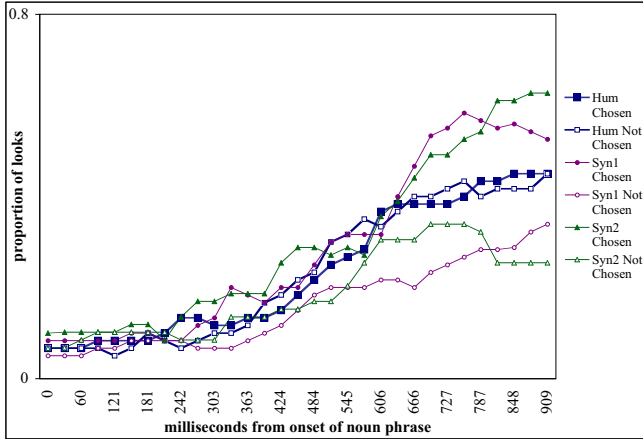


Fig. 6. Time course of looks to target (duplicate) items chosen and not chosen for indefinite instructions.

ences in the time course of looks between the voice conditions, but examining looks to the two duplicated items may provide an explanation.

Figure 6 shows the proportion of looks over time to the two possible target (duplicated) items. For each voice condition, the item identified as “chosen” is the target item that the participant clicked on, and the item identified as “not chosen” is the other duplicated item. For instructions in the human voice condition, participants considered each of the duplicated items before clicking on one, reflecting the expected circumscription of referential domain in the indefinite condition. For instructions in the synthesized voice conditions, however, participants tended to click on the first of the duplicated items that came to their attention, reflecting a more restricted referential domain than expected - one more appropriate to a definite article interpretation.

These results show that participants make different assumptions about the felicity of article use for the synthesized speech instructions than for the human speech instructions, due to global differences in how well the indefinite articles could be understood in the three voice conditions. This could also explain the delays in disambiguation for the definite article instructions.

3. IMPLICATIONS

The experiments described in this paper represent the first attempt to apply a novel method of using eye tracking to evaluate synthesized speech. Critically, our results support the use of this methodology by demonstrating that 1) processing of synthesized speech is similar enough to processing of human speech to allow comparison, and 2) the eye tracking measure is sensitive enough to reveal subtle differences between processing of human speech and synthesized speech, and between synthesized speech produced by different synthesis packages. We have also shown that the method is effective at the lexeme level and at the discourse level. This method of evaluation is promising because it yields more fine-grained measures than possible with techniques used to date. Moreover, this technique works well with naive listeners and can be adapted to specific applications and natural tasks.

Future work using this method may be used to detect subtle differences in synthesized speech processing that can reveal prob-

lems and areas for improvement at different levels in a TTS system, from grapheme-to-phoneme conversion to the implementation of prosody. The detection of on-line speech processing differences allows direct and objective comparisons between different synthesis algorithms that can provide valuable data regarding specific implementation issues. This methodology can be used to target specific linguistic and psycholinguistic processing issues for synthesized speech. Moreover, the use of this methodology in psycholinguistic research on natural speech processing provides an existing body of empirical evidence for comparison and on which to build. A direct comparison of listener eye movements while processing natural and synthesized speech may offer insights into specific factors involved in issues that have been difficult to quantify, such as ‘naturalness’ and the additional cognitive load that synthesized speech imposes on processing.

Adapting the eye-tracking technique as a new method of synthesized speech evaluation represents a step towards bringing together the psycholinguistic and speech engineering communities that promises to advance our understanding of speech processing. There is an extensive body of psycholinguistic research using eye tracking that documents many facets of human language processing. By performing fine-grained comparisons of synthesized speech processing to this evidence, we come closer to achieving truly natural speech.

4. REFERENCES

- [1] J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, editors. *Progress in Speech Synthesis*. Springer-Verlag, New York, 1997.
- [2] C. Benoît. Evaluation inside or assessment outside? In J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, editors, *Progress in Speech Synthesis*, pages 513–517. Springer-Verlag, New York, 1997.
- [3] W. Marslen Wilson. Functional parallelism in spoken word recognition. *Cognition*, 25:71–102, 1987.
- [4] G. Altmann. Ambiguity in sentence processing. *Trends in Cognitive Sciences*, 2, 1998.
- [5] M. K. Tanenhaus and J. Trueswell. Sentence comprehension. In J. Miller and P. Eimas, editors, *Handbook of Perception and Cognition*. Academic Press, San Diego, 1995.
- [6] P. Allopenna, J. Magnuson, and M. K. Tanenhaus. Tracking the time course of spoken word recognition using eye-movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 1998.
- [7] J. Hanna. *The effects of linguistic form, common ground, and perspective on domains of referential interpretation*. PhD thesis, University of Rochester, 2001.
- [8] D. Dahan, J. S. Magnuson, M. K. Tanenhaus, and E. M. Hogan. Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, 2001.
- [9] D. Dahan, M. K. Tanenhaus, and C. G. Chambers. Accent and reference resolution in spoken-language comprehension. *Journal of Memory and Language*, in press.
- [10] D. B. Pisoni. Perception of synthetic speech. In J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, editors, *Progress in Speech Synthesis*, pages 541–560. Springer-Verlag, New York, 1997.