Using Eye Movements for Online Evaluation of Speech Synthesis

Charlotte van Hooijdonk, Edwin Commandeur, Reinier Cozijn, Emiel Krahmer & Erwin Marsi

Department of Communication & Information Sciences
Tilburg University, The Netherlands
{C.M.J.vanhooijdonk; E.Commandeur; R.Cozijn; E.J.Krahmer; E.C.Marsi}@uvt.nl

Abstract

This paper[*] describes an eye tracking experiment to study the processing of diphone synthesis, unit selection synthesis, and human speech taking segmental and suprasegmental speech quality into account. The results showed that both factors influenced the processing of human and synthetic speech, and confirmed that eye tracking is a promising albeit time consuming research method to evaluate synthetic speech.

Index Terms: eye movements, human speech, diphone synthesis, unit selection synthesis.

1. Introduction

The evaluation of synthetic speech in terms of intelligibility has primarily been done with offline research methods. For example, the Modified Rhyme Test (MRT) [1] has been used to investigate the segmental intelligibility of synthetic speech [2]. In this test, listeners are presented with spoken words and have to select the word they heard from a set of alternatives that differ only in one phoneme.

A disadvantage of offline research methods is that we obtain no insight in how listeners process synthetic speech. Online research methods, like eye tracking, give us a direct insight into how speech is processed incrementally. In the "visual world paradigm", participants are asked to follow spoken instructions to look up or pick up objects within a visual display (e.g., [3, 4]). The fixation pattern on the objects within the display is used to draw inferences about the processing of spoken instructions. Eye tracking might give us an idea of how similar the processing of synthetic speech is compared to the processing of human speech. This idea was first explored by Swift et al. [5] in a study concentrating on acoustically confusable words (e.g., beetle, beaker, and speaker) to see if the "disambiguation" point was processed at comparable time windows for two instances of synthetic speech and human speech.

The intelligibility of speech does not only depend on its segmental quality but also on the quality and the appropriateness of the prosodic information in the speech signal [6]. The visual word paradigm has more recently been used to investigate how humans process prosodic information. For example, Weber et al. [7] used eye tracking to investigate how prosodic information influences the processing of spoken referential expressions. In two experiments, participants followed two consecutive instructions to click on an object within a visual display. The first instruction mentioned the referent (e.g., purple scissors). The second instruction either mentioned a target of the same type but with a different colour (red scissors) or of a different type and a different colour (red vase). The instructions were either realised with an accent on the adjective (e.g., Click on the PURPLE scissors, Click now on the RED scissors) or the noun (e.g., Click on the purple SCISSORS, Click now on the red SCISSORS), and listeners were indeed shown to be sensitive for this prosodic difference.

Both segmental and suprasegmental quality are important factors in the intelligibility of synthetic speech. In this paper, we therefore extend on the work by Swift et al. by focussing on both segmental and suprasegmental aspects of speech. In our evaluation experiment, the participants were given two consecutive spoken instructions to look at a certain object within the visual display. These instructions were presented in three speech conditions: diphone synthesis, unit selection synthesis, and human speech. Diphone synthesis is based on concatenating prerecorded diphones (i.e., phoneme transitions), followed by signal processing to obtain the required pitch and duration. Unit synthesis is also based on concatenation, but on a much larger scale, where units are of variable size (e.g., sentences, constituents, words, morphemes, syllables, and diphones). As larger units of natural speech are exploited, requiring less concatenation, the segmental quality of unit synthesis is in general significantly higher than that of diphone synthesis. At the same time, the prosody may be inadequate, because the intended realisation of, for example, pitch accents, may not be available in the speech database. Thus, while quality of diphone synthesis is in general inferior to that of unit synthesis, it has the advantage that it can always produce contextually appropriate prosody (albeit by human intervention). In this experiment, we investigate this trade-off between segmental quality on the one hand and contextually appropriate prosody on the other from the perspective of humans processing synthetic speech. The human speech condition was added as an upper limit to compare processing of natural and synthetic speech.

2. Method

2.1. Participants

Thirty-eight native speakers of Dutch (13 male and 25 female, between 18 and 33 years old) were paid to participate. They had normal or corrected-to-normal vision and normal hearing. None of the participants were colour-blind and none had any involvement in speech synthesis research.

2.2. Stimuli

Fifteen pairs of Dutch monosyllabic picturable nouns were chosen as stimuli. These nouns shared the same initial phonemes (e.g., vork - vos, 'fork - fox'). Each experimental trial consisted of a 3x3 grid with four objects in the corner cells, see Figure 1. For every grid, the participants were given
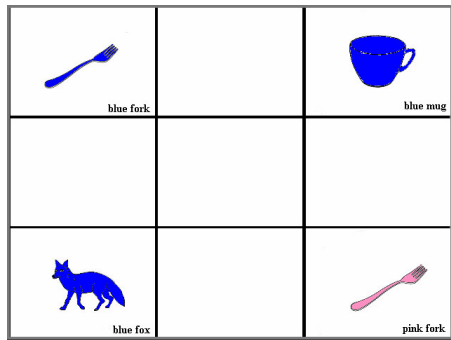
Figure 1: An example of the visual display

two consecutive spoken instructions each referring to a certain object within the grid. In both instructions, the nouns were modified with a colour adjective. The first instruction mentioned the **referent** (e.g., Kijk naar de roze vork, 'Look at the pink fork'). The second instruction mentioned the **target**. The target could either be of the **same type** as the referent modified with a different colour adjective (e.g., Kijk nu naar de blauwe vork, 'Now look at the blue fork'), or of a **different type** than the referent modified with a different colour adjective (e.g., Kijk nu naar de blauwe vos, 'Now look at the blue fox'). A fourth object was added as a **distractor** (e.g., blauwe mok, 'blue mug'). The distractor did not share the form of the other objects, but did share the colour with the two target objects. The distractor was never mentioned in the experimental trials. The colours blue and pink could occur in both instructions and were randomized across the trials.

The adjective and noun mentioned in the second instruction always had a double accent pattern (e.g., BLAUWE VOS, 'BLUE FOX'). In half of the cases the second instruction had a contextually appropriate double accent pattern while the other half had not. The second instruction had an appropriate accent pattern when it mentioned a different colour adjective and object type than mentioned in the first instruction. The second instruction had an inappropriate accent pattern when it mentioned a different colour adjective but the same object type than mentioned in the first instruction [8, 9]. Note that the choice of a double accent pattern was forced by the output of the unit selection synthesizer, as it typically produced these double accents.

The instructions were realised in three speech conditions, i.e., unit selection synthesis, diphone synthesis, and human speech. A female voice was used for all three speech conditions. For the unit selection synthesis a commercially available synthesizer was used. The instructions were obtained through an interactive web interface of the synthesizer. The output that was given by the interface was stored. Note that it was not possible to control the accent patterns of the instructions, as the synthesis was dependent on the intonation of the selected units in the database of the synthesizer. The diphone stimuli were produced using the Nextens[1] TTS system for Dutch, which is based on the Festival TTS system [10]. The input consisted of words and prosodic markup. Pitch accents were phonetically realised with a rule-based implementation of the Gussenhoven & Rietveld model for Dutch intonation [11]. The instructions in the human speech condition were recorded by a native speaker of Dutch (the first author) in a quiet room at Tilburg University. The instructions were digitally recorded, sampling at 44 kHz, using Sony Sound Forge™ and a Sennheiser™

---

[1] http://nextens.uvt.nl

microphone. The instructions were recorded multiple times and the best realisations were chosen. An independent intonation expert checked the utterances using PRAAT [12] to make sure that the intended accents in the second instructions were properly realised. We also checked whether there were significant durational differences between the target nouns in the various conditions, and this turned out not to be the case. All instructions in the three speech conditions were normalized at -16 dB using Sony Sound Forge™ and stored in stereo format.

In addition to the 90 experimental trials (15 stimuli × 3 speech conditions × 2 object types [same, different]), 20 filler trials were constructed to add variety to the visual display, and the accent pattern of the second instruction. In the filler trials, either the adjective or noun mentioned in the second instruction was accented (i.e., ROZE mok, 'PINK mug' or roze MOK, 'pink MUG'), and they were only realised in human speech and diphone synthesis. Moreover, all objects within the visual display had the same colour (pink or blue).

Three lists were constructed in a semi-Latin square design, each containing 90 experimental and 20 filler trials.

2.3. Procedure

Each participant was invited to an experimental laboratory, and was seated in front of a computer monitor. First, the participants were familiarised with the objects that occurred within the visual display during the experiment to ensure that they identified them as intended. This was done by asking them to describe the thirty depicted objects and their colour (pink or blue) aloud. The objects were shown in the middle of the computer screen. Participants could view each object at their own pace by clicking on a button, and they were corrected when an object was described incorrectly. This object was viewed again until it was described correctly.

Subsequently, instructions of the actual experiment were read to the participants, and the eye-tracking system was mounted and calibrated. Participants' eye movements were monitored using an SR Research EyeLink II eye-tracking system, sampling at 250 Hz. Only the right eye of the participant was tracked. The spoken instructions were presented to the participants binaurally through headphones. Next, the participants were presented with a practice session in which the procedure of the experiment was illustrated. This practice session consisted of six trials (3 speech conditions × 2 object types). The structure of a trial was as follows. First, participants saw a white screen with in the middle a little black cross, and they pressed a button to continue. Next, a white screen appeared with in the middle a central fixation point, and the participants were instructed to look at this point. The experimenter then initiated an automatic drift correction, after which the visual display appeared. The first instruction was given after 50 milliseconds. The participants had to look at the object that was mentioned, after which they pushed a button. Subsequently, a little black cross appeared in the centre of the grid and the participants were instructed to look at this cross. After 2000 milliseconds, the cross disappeared and the second instruction was given. Again, the participants had to look at the object that was mentioned, after which they pushed a button. Subsequently, a white screen appeared with in the middle a little black cross, and the participants pressed on a button again to continue to the next session. After completing the practice session, the actual experiment started, proceeding in the same way as during the practice session. During the experiment, there was no further interaction between participants and experiment leader.
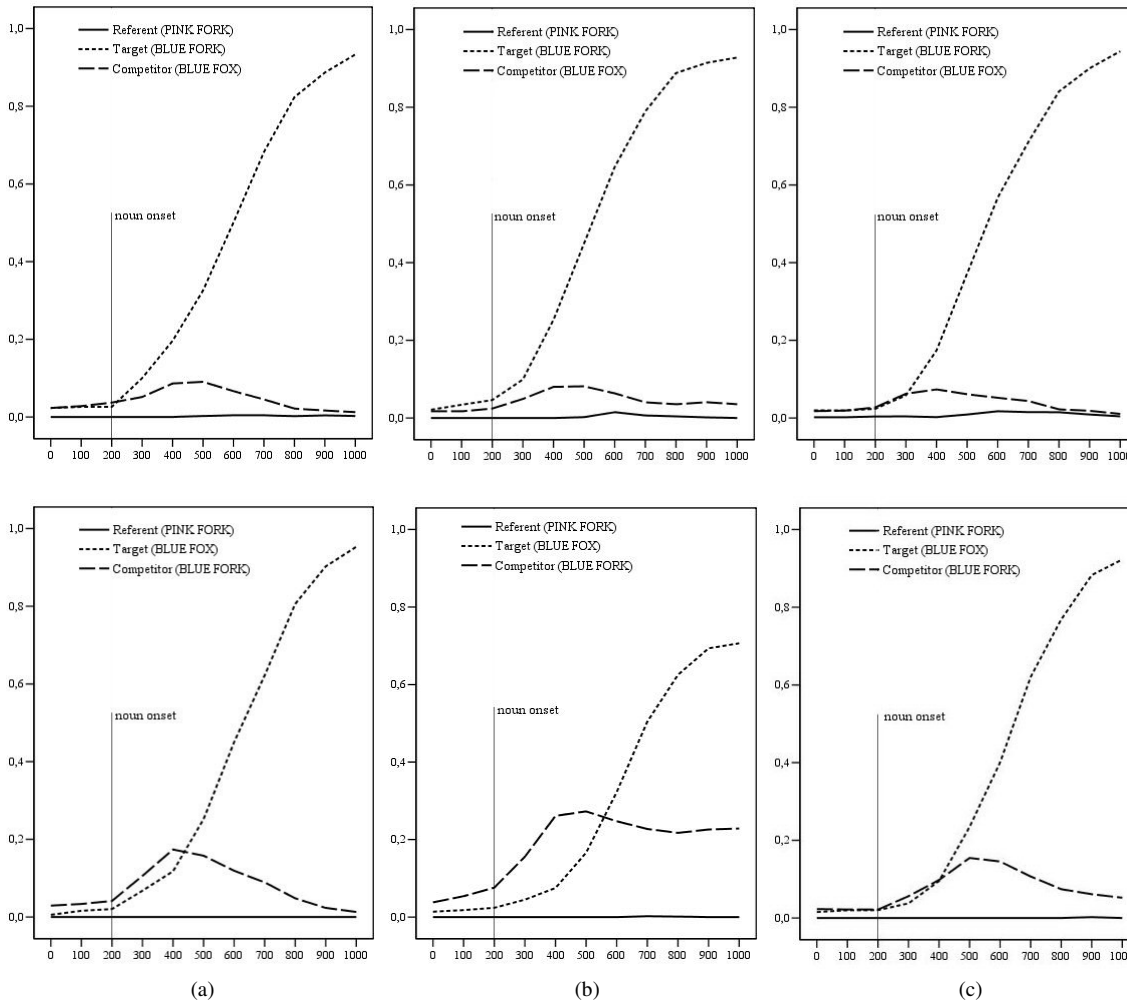
Figure 2: *Mean proportion of fixations to the referent, target, and competitor for (a) human speech, (b) diphone synthesis, and (c) unit selection synthesis for the second instruction mentioning a same object type (top row) and different object type (bottom row).*

2.4. Coding procedure and data processing

Eyelink software parsed the eye movement data into fixations, saccades, and blinks. Fixations were automatically mapped (i.e., using the program Fixation[2]) on the objects presented in each trial, and this was checked by hand. The fixations occurring in the first and second instruction of a trial were analysed. In the first instruction, trials in which less than 50% of the sample points after the onset of the referent noun belonged to fixations on the referent object were excluded from further analysis. In the second instruction, trials in which less than 50% of the sample points before the onset of the target noun belonged to fixations on the centre of the grid were excluded from further analysis. These trials were excluded because the instructions were not followed. The data of one participant was excluded, as she did not meet the above-mentioned criteria in any of the trials. The total amount of data that was excluded from further analysis was 7.7%, including the data discarded for the above-mentioned participant.

Fixation proportions were averaged over two time windows for each participant (F_1) and item (F_2) and were analysed with a 3 (speech condition) × 2 (object type) repeated measures analysis of variance (ANOVA), with a

[2] http://www.tilburguniversity.nl/faculties/humanities/people/cozijn/research

significance threshold of .05. The dependent variables were the mean proportion of fixations to the target and the competitor. The first time window extended from 200-600 ms after the target noun onset. The second time window extended from 600-1000 ms after the target noun onset.

3. Results

Figure 2 shows the mean proportion of fixations to objects within the visual display over time for the second instruction mentioning the same object and a different object type and for the three speech conditions. The statistics showed that for the time window 200-600 ms after the target noun onset, the mean proportion of fixations to the target did not differ significantly between three speech conditions (human speech: 21.3%, diphone synthesis: 21.5%, and unit selection synthesis: 20.0%; F_1 and F_2 < 1). However, there was a significant difference between the three speech conditions in the mean proportion of fixations to the competitor (human speech: 10.6%, diphone synthesis: 14.8%, unit selection synthesis: 8.5%; F_1 [2,72] = 20.68, p< .001, F_2 [2,28] = 10.13, p< .001). There was also a significant main effect of the object type mentioned in the second instruction in the mean proportion of fixations to the target (F_1 [1,36] = 48.82, p< .001, F_2 [1,14] = 34.08, p< .001). The proportion of fixations to the target was higher when the second instruction mentioned the same object type (26.3%) than when it

mentioned a different object type (15.5%). Conversely, the proportion of fixations to the competitor were higher when the second instruction mentioned a different object type (15.8%) than when it mentioned the same object type (6.8%) (F_1 [1,36] = 44.40, p< .001, F_2 [1,14] = 21.67, p< .001). Moreover, a significant interaction was found between speech condition and the object type mentioned in the second instruction for both the mean proportion of fixations to the target (F_1 [2,72] = 18.93, p< .001, F_2 [2,28] = 11.18, p< .001) and the competitor (F_1 [2,72] = 21.95, p< .001; F_2 [2,28] = 9.73, p< .005). For all three speech conditions, the mean proportion of fixations to the target was significantly higher when the second instruction mentioned the same object type than when it mentioned a different object type. Conversely, for all three speech conditions the mean proportion of fixations to the competitor was significantly higher when the second instruction mentioned a different object type than when it mentioned the same object.

For the time window 600-1000 ms after the target noun onset, the mean proportion of fixations to the target differed significantly between three speech conditions (human speech: 77.8%, diphone synthesis: 72.2%, unit selection synthesis: 78.1%; F_1 [2,27] = 12.70, p< .001, F_2 [2,28] = 1.32, p = .28). Also, a significant difference was found between the three speech conditions in the mean proportion of fixations to the competitor (human speech: 4.4%, diphone synthesis: 13.6%, unit selection synthesis: 5.5%; F_1 [2,72] = 57.16, p< .001, F_2 [2,28] = 5.28, p< .025). Furthermore, there was a significant main effect of the object type mentioned in the second instruction in the mean proportion of fixations to the target (same object type: 81.9%, different object type: 70.2%; F_1 [1,36] = 72.92, p< .001, F_2 [1,14] = 19.93 p< .005). The reverse was found for the mean proportion of fixations to competitor (same object type: 3.2%, different object type: 12.4%; F_1 [1,36] = 83.13, p< .001, F_2 [1,14] = 10.87, p< .01). Finally, a significant interaction was found between speech condition and the object type mentioned in the second instruction for both the mean proportion of fixations to the target (F_1 [2,72] = 57.20, p< .001; F_2 [2,28] = 5.96 p< .01) and the competitor (F_1 [2,72] = 53.45, p< .001; F_2 [2,28] = 3.70, p< .05). This interaction can be explained as follows: for the target it was the case that for the diphone synthesis (F_1 [1,36] = 129.89, p< .001; F_2 [1,14] = 13.18, p< .005) and unit selection synthesis (F_1 [1,36] = 17.12, p< .001; F_2 [1,14] = 7.26, p< .025), the mean proportion of fixations was significantly higher when the second instruction mentioned the same object type than when it mentioned a different object type. However, for human speech, no significant difference was found in the mean proportion of fixations to the target in the object type mentioned in the second instruction (F_1 [1,36] = 1.52, p = .27; F_2< 1). For the competitor it was the case that for all three speech conditions the mean proportion of fixations was significantly higher when the second instruction mentioned a different object type.

4. Conclusion and discussion

In this paper we described an experiment in which eye tracking was used to evaluate human speech, diphone synthesis, and unit selection synthesis having either contextually appropriate or inappropriate accent patterns. We found differences in the performance accuracy between the three speech conditions. In the time window 600 to 1000 ms, the mean proportion of fixations to the target was lowest for diphone synthesis and highest for unit selection synthesis and human speech. Also, in both time windows significant differences between the three speech conditions were found in the mean proportion of fixations to the competitor. The mean proportion of fixations to the competitor was highest for diphone synthesis. An explanation for these results could be the relatively poor segmental intelligibility of the diphone synthesis making it harder for the participants to process the disambiguation point of the acoustically confusable words. We also found that the participants anticipated the upcoming target. In both time windows, the mean proportion of fixations to the target was higher when the second instruction mentioned the same object type. Moreover a significant interaction was found between speech condition and the object type mentioned in the second instruction. In the time window 200-600 ms the mean proportion of fixations to the target was significantly higher for all three speech conditions when the second instruction mentioned the same object type. However, in the time window 600 - 1000 ms this interaction was only found for the diphone and unit selection synthesis. These results indicate that not only the segmental intelligibility of synthetic speech plays an important role in speech processing, but also listeners' anticipation based on the accent patterns within the speech.

Eye tracking seems to be a promising research method to evaluate synthetic speech. The results give us an insight in how similar the processing of synthetic speech is compared to the processing of human speech on a segmental and suprasegmental level. A disadvantage of this evaluation method is that is rather time consuming. It would be interesting to create a test bed environment in which it would be easy to compare the processing of a new speech synthesis system with reference fixation patterns.

5. References

[1] House, A.S., Williams, C.E., Hecker, M.H. and Kryter, K.D., "Articulation-testing methods: consonantal differentiation with a closed-response set." JASA, 37, 1965, pp 158-166.

[2] Schmidt-Nielsen, A., Intelligibility and acceptability testing for speech technology. In A. Syrdal, R. Bennett, and S. Greenspan (eds.), 1995. Applied Speech Technology. CRC: Boca Raton, pp. 194-231.

[3] Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., and Sedivy, J.E., "Integration of visual and linguistic information in spoken language comprehension". Science, 268, 1995, pp. 1632-1634.

[4] Altmann, G.T., and Kamide, Y., Now you see it, now you don't: Mediating the mapping between language and the visual world. In J. Henderson and F. Ferreira (eds.), 2004. The interface of language, vision, and action: Eye movements and the visual world. Psychology Press, New York, pp. 347-386.

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

[6] Sanderman, A.A., and Collier, R., "Prosodic phrasing and comprehension". Language and Speech, 40, 1997, pp 391-409.

[7] Weber, A., Braun, B., and Crocker, M. W., "Finding referents in time: eye-tracking evidence for the role of contrastive accents". Language and Speech, 49, 2006, pp. 367-392.

[8] Nooteboom, S.G., and Kruyt, J.G., "Accent, focus distribution, and perceived distribution of given and new information: An experiment". JASA, 82, 1987, pp. 1512–1524.

[9] Terken, J., and Nooteboom, S.G., "Opposite effects of accentuation and deaccentuation on verification latencies for Given and New information", Language and Cognitive Processes, 2, 1987, pp. 145-163.

[10] Black, A.W., Taylor, P., and Caley. R., The Festival Speech Synthesis System, System documentation. Centre for Speech Technology Research University of Edinburgh, Edition 1.4, for Festival Version 1.4.3, 2002.

[11] Gussenhoven, C., and Rietveld T., "A target-interpolation model for the intonation of Dutch". Proceedings of the ICSLP, Banff, Canada, pp. 1235-1238, 1992.

[12] Boersma, P. and Weenink, D., Praat, a system for doing phonetics by computer, version 3.4. Institute of Phonetic Sciences of the University of Amsterdam, Report 132., 1996.