The Online Evaluation of Speech Synthesis Using Eye Movements

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.

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 [1] has been used to investigate the segmental intelligibility of synthetic speech [2]. In this test, listeners are presented with spoken words and are instructed to select the word they heard from a set of alternatives that differ only in one phoneme. Another example is the Mean Opinion Score [3] in which listeners have to rate the quality of spoken sentences on scales (i.e., excellent - bad).

A disadvantage of offline research methods is that no insight is obtained in how listeners process synthetic speech. Online research methods, like eye tracking, give a direct insight in 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., [4, 5]). The fixation patterns on the objects within the display are used to draw inferences about the processing of spoken instructions. Eye tracking might give 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. [6] 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 results showed that both human speech instructions and synthetic speech instructions were indeed processed incrementally. Moreover, when hearing the onset of the target noun (e.g., beaker), the listeners were more likely to look at the cohort competitor (e.g., beetle) than at the rhyme competitor (e.g., speaker). Finally, the listeners identified the target more rapidly in the human speech condition than in the two synthetic speech conditions.

*The current research is performed within the IMIX-IMOGEN (Interactive Multimodal Output GENeration) project sponsored by the Netherlands Organisation for Scientific Research (NWO). The authors would like to thank Lennard van der Laar, Pascal Marcelis, and Marie Nilsenova for their help in setting up the experiment, Marc Swerts for discussing the findings of the experiment and Carel van Wijk for his statistical advice.

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 (i.e., suprasegmental quality) [7]. The visual word paradigm has more recently been used to investigate how humans process prosodic information. For example, Weber et al. [8] 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 on the noun (e.g., Click on the purple SCISSORS, Click now on the red SCISSORS). The results showed that the listeners were affected by this prosodic difference. When the first instruction was realized with an accent on the adjective (e.g., Click on the PURPLE scissors), listeners anticipated the upcoming target, i.e., before the onset of the target noun, listeners looked at the target of the same type as the referent but with a different colour (red scissors). When both instructions were realized with an accent on the adjective (e.g., Click on the PURPLE scissors, Click now on the RED scissors) this anticipation only increased. However, when the instructions were realized with an accent on the noun (e.g., Click on the purple SCISSORS, Click now on the red SCISSORS), listeners did not anticipate the upcoming target.

Both segmental and suprasegmental quality are important factors in processing synthetic speech. In this paper, we therefore extend on the work by Swift et al. by focusing 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 investigated this trade-off between segmental quality on the one hand and contextually appropriate prosody (i.e., suprasegmental quality) on the other from the perspective of humans processing synthetic speech. The human speech condition was added as a baseline 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[1].

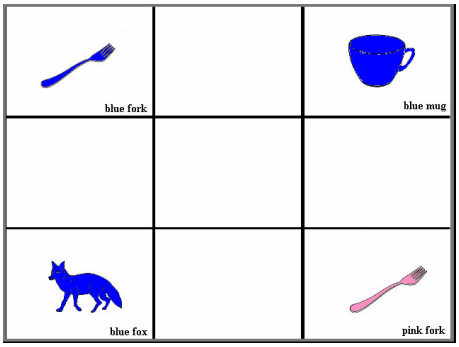

Figure 1: An example of a visual display

For every grid, the participants were given two consecutive spoken instructions each referring to a certain object within the grid. In both instructions, the nouns were modified with a colour adjective (blue or pink). 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** as 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 targets. 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.

[1] The textual descriptions in figure 1 are only added for illustrative purposes, they did not occur in the actual experiment.

The first instruction was realised with a standard, neutral intonation contour. In the second instruction, the adjective and noun were both accented (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, see Table 1. The second instruction had an appropriate accent pattern when it mentioned a different colour adjective and a different object type as the referent in the first instruction. The second instruction had an inappropriate accent pattern when it mentioned a different colour adjective but had the same object type as the referent in the first instruction [9, 10]. Note that the choice of a double accent pattern was forced by the output of the unit selection synthesizer, as it typically produces these double accents.

Table 1: Example of the instructions

First instruction	*Kijk naar de roze vork* Look at the pink fork
Second instruction contextually appropriate double accent pattern	*Kijk nu naar de BLAUWE VOS* Now look at the BLUE FOX
Second instruction contextually inappropriate double accent pattern	*Kijk nu naar de BLAUWE VORK* Now look at the BLUE FORK

The instructions were realised in three speech conditions, i.e., diphone synthesis, unit selection synthesis, and human speech. A female voice was used for all three speech conditions. The diphone stimuli were produced using the Nextens[2] TTS system for Dutch, which is based on the Festival TTS system [11]. 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 [12]. 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 this type of synthesis is dependent on the intonation of the selected units in the database of the synthesizer. 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™ microphone (type SKM 135 G2). The instructions were recorded multiple times and the best realisations were chosen. An independent intonation expert checked the utterances using PRAAT [13] to make sure that the intended accents in the second instructions were properly realised. All instructions in the three speech conditions were normalized at -16 dB, using Sony Sound Forge™, and stored in stereo format.

We checked whether there were durational differences between the target nouns mentioned in the second instruction between the various conditions. It turned out that speech condition did not affect the duration ($F < 1$). Also, the target object type (same object type vs. different object type) mentioned in the second instruction did not affect its duration

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

(F< 1). Finally, there was no interaction for duration between speech condition and target object type (F< 1).

In addition to the 90 experimental trials (15 stimuli × 3 speech conditions × 2 target object types), 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 the 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.

Three lists were constructed in a semi-Latin square design, each containing 90 experimental and 20 filler trials. In each list, the stimuli were mixed up and presented as one block to the participants. Thus, the participants were presented with all three speech conditions and both target object types during the experiment.

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, the 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 target 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 to correct for shifts of the head-mounted tracking system. After the automatic drift correction, 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, the white screen with in the middle a little black cross appeared again and the participants pressed on a button indicating the start of the next trial. After completing the practice session, the actual experiment started, proceeding in the same way as the practice session. During the experiment, there was no interaction between the participant and the experiment leader.

After the participants completed the experiment, they were asked to listen to an instruction (i.e., *Kijk nu naar de BLAUWE BLOEM,* Now look at the BLUE FLOWER) realised in diphone synthesis, unit selection synthesis, and human speech. Next, they were asked to fill out a questionnaire. This questionnaire consisted of four items about the intelligibility (i.e., audibility, comprehensibility, perceptibility, and distinctness) and four items about the naturalness (i.e., intonation, pleasantness to listen, speech rate, and naturalness) of the three speech conditions. Each question was accompanied with a 7-point Likert scale on which the participants could indicate how much they agreed or disagreed with the content of each item.

2.4. Coding procedure and data processing

Eyelink software parsed the eye-movement data into fixations, saccades, and blinks. Fixations were automatically mapped (using the program Fixation[3]) on the objects presented in each trial, and this mapping 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 analysed with a 3 (diphone synthesis, unit selection synthesis, human speech) × 2 (same target object type, different target object type) repeated measures analysis of variance (ANOVA)[4], with a significance threshold of .05. For post hoc tests, the Bonferroni method was used. The dependent variables were the mean proportions of fixations to the target and to the competitor. The first time window began 200 ms after the onset of the target noun, because this is the earliest point at which fixations driven by information from the target noun were expected [5, 14]. The time window extended over 400 ms, which roughly corresponded to the mean duration of the target noun. The second time window extended from 600 to 1000 ms after the target noun onset.

The results of the questionnaire were processed by mapping the items to which the participants disagreed to 1 and agreed to 7 and were analysed with a 3 (speech condition) × 4 (items) repeated measures analysis of variance (ANOVA), with a significance threshold of .05. For post hoc tests, the Bonferroni method was used.

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

[4]Mauchly's test of sphericity was significant for some main effects and interactions. For these cases, we looked both at Greenhouse-Geisser and Huynh-Feldt corrections on the degrees of freedom, which gave similar results. For the sake of transparency, we report on the normal degrees of freedom.

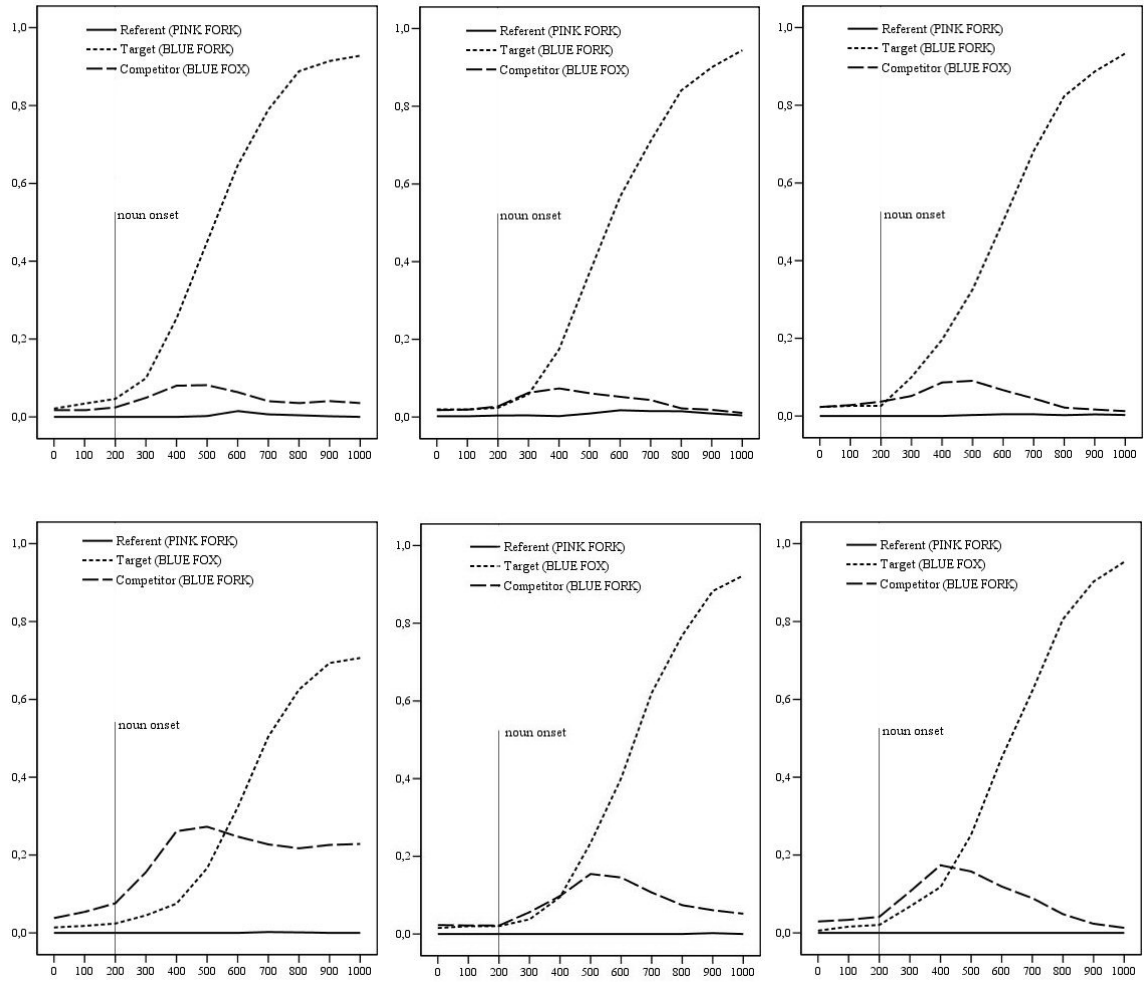

Figure 2: Proportions of fixations to the referent, the target, and the competitor for diphone synthesis, unit selection synthesis, and human speech for the second instruction mentioning a same target object type (top row) and different target object type (bottom row).

Table 2: Mean proportions of fixations to the target and competitor for the two time windows in relation to the speech condition

	Diphone synthesis		Unit selection synthesis		Human speech	
	Target	Competitor	Target	Competitor	Target	Competitor
Time window 200 – 600 ms	.22	.15	.20	.09	.21	.11
Time window 600 – 1000 ms	.72	.14	.78	.06	.78	.04

Table 3: Mean proportions of fixations to the target and competitor for the two time windows in relation to the target object types mentioned in the second instruction

	Target object of the same type		Target object of a different type	
	Target	Competitor	Target	Competitor
Time window 200 – 600 ms	.26	.07	.16	.16
Time window 600 – 1000 ms	.82	.03	.70	.12

3. Results

3.1. Eye movement data

Table 2 summarizes the mean proportions of fixations found within the time window 200 to 600 ms for the three speech conditions. The statistics showed that the mean proportions of fixations to the target did not differ significantly between the three speech conditions F_1 and $F_2 < 1$. In all three speech conditions, the mean proportions of fixations to the target were approximately 20%. However, there was a significant difference between the three speech conditions in the mean proportions of fixations to the competitor: F_1 [2,72] = 20.68, p< .001, partial eta squared = .37; F_2 [2,28] = 10.13, p< .001, partial eta squared = .42. The mean proportions of fixations to the competitor were the highest in the diphone synthesis condition and the lowest in the unit selection synthesis condition. The mean proportions of fixations to the competitor in the human speech condition fell between these two. Table 3 reveals that within the time window 200 to 600 ms, the mean proportions of fixations to the target were higher when the second instruction mentioned a target object of the same type than when it mentioned a target object of a different type: F_1 [1,36] = 48.82, p< .001, partial eta squared = .58; F_2 [1,14] = 34.08, p< .001, partial eta squared = .71. Conversely, the mean proportions of fixations to the competitor were higher when the second instruction mentioned a target object of a different type than when it mentioned a target object of the same type: F_1 [1,36] = 44.40, p< .001, partial eta squared = .55; F_2 [1,14] = 21.67, p< .001, partial eta squared = .61. Finally, within the time window 200 to 600 ms, an interaction was found between speech condition and target object type mentioned in the second instruction for both the mean proportions of fixations to the target: F_1 [2,72] = 18.93, p< .001, partial eta squared = .35; F_2 [2,28] = 11.18, p< .001, partial eta squared = .44, and to the competitor: F_1 [2,72] = 21.95, p< .001, partial eta squared = .38; F_2 [2,28] = 9.73, p< .005, partial eta squared = .41. For all three speech conditions, the mean proportions of fixations to the target were significantly higher when the second instruction mentioned a target object of the same type than when it mentioned a target object of a different type. Conversely, for all three speech conditions the mean proportions of fixations to the competitor were significantly higher when the second instruction mentioned a target object of a different type than when it mentioned a target object of the same type.

In the time window 600 to 1000 ms, an significant effect was found of speech condition in the mean proportions of fixations to the target, although not by items: F_1 [2,27] = 12.70, p< .001, partial eta squared = .26; F_2 [2,28] = 1.32, p = .28. The mean proportions of fixations to the target were the highest for unit selection synthesis and human speech and low for diphone synthesis. The results found for speech condition in the mean proportions of fixations to the competitor were similar to those found in the time window 200 to 600 ms, F_1 [2,72] = 57.16, p< .001, partial eta squared = .61; F_2 [2,28] = 5.28, p< .025, partial eta squared = .27. Also, similar results were found for the target object types mentioned in the mean proportions of fixations to the target: F_1 [1,36] = 72.92, p< .001, partial eta squared = .67; F_2 [1,14] = 19.93 p< .005, partial eta squared = .59, and to the competitor: F_1 [1,36] = 83.13, p< .001, partial eta squared = .7; F_2 [1,14] = 10.87, p< .01, partial eta squared = .44. Finally, a similar interaction was found between speech condition and target object type in the mean proportions of fixations to the competitor: F_1 [2,72] = 53.45, p< .001; partial eta squared = .60; F_2 [2,28] = 3.70, p< .05, partial eta squared = .21. The interaction found between speech condition and target object type in the mean proportions of fixations to the target was different from the results found in the previous time window, F_1 [2,72] = 57.20, p< .001; partial eta squared = .61; F_2 [2,28] = 5.96 p< .01, partial eta squared = .30. Only in the diphone synthesis condition: F_1 [1,36] = 129.89, p< .001; partial eta squared = .78; F_2 [1,14] = 13.18, p< .005, partial eta squared = .49, and in the unit selection synthesis condition: F_1 [1,36] = 17.12, p< .001; partial eta squared = .32; F_2 [1,14] = 7.26, p< .025; partial eta squared = .34, the mean proportions of fixations were higher when the second instruction mentioned a target object of the same type than when it mentioned a target object of a different type. For human speech, no difference between the speech conditions was found: F_1 [1,36] = 1.52, p = .27; $F_2 < 1$.

3.2. Intelligibility and naturalness of the three speech conditions

Figure 3 illustrates the results found for the questionnaire on the intelligibility and the naturalness of the three speech conditions.

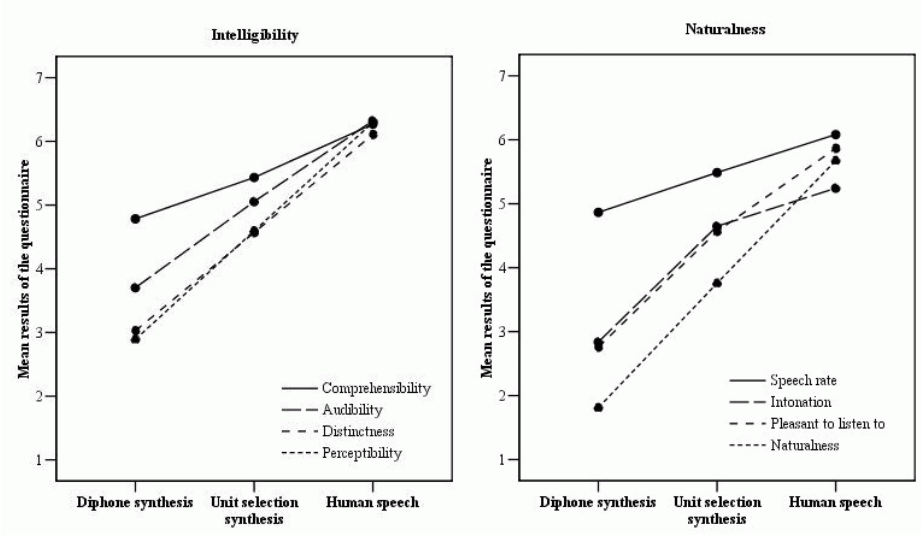

Figure 3: Mean results of the questionnaire on the intelligibility and naturalness of the three speech conditions

A significant effect was found of speech condition for both intelligibility: F [2,72] = 42.52, p< .001, η_p^2 = .54 and naturalness: F [2,72] = 49.83, p< .001, η_p^2 = .58. Post-hoc tests showed that all pairwise comparisons were significant at p< .001. For intelligibility and naturalness diphone synthesis was rated lowest followed by unit selection synthesis. Human speech was rated highest. Finally, Figure 3 shows that the participants were homogeneous in their ratings on the intelligibility and the naturalness of human speech, but they were heterogeneous synthesis.

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 proportions of fixations to the target were 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 proportions of fixations to the competitor. The mean proportions of fixations to the competitor were highest for diphone synthesis. An explanation for these results could be that the relatively poor segmental intelligibility of the diphone synthesis makes it harder for the participants to determine the disambiguation point of the acoustically confusable words. We also found that the participants anticipated the upcoming target. In both time windows, the mean proportions of fixations to the target were higher when the second instruction mentioned a target object of the same type. Moreover, interactions were found between speech condition and the target object type mentioned in the second instruction. In the time window 200 to 600 ms, the mean proportions of fixations to the target were significantly higher for all three speech conditions when the second instruction mentioned the same object type. However, in the time window 600 to 1000 ms, this interaction was only found for diphone synthesis 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' anticipations based on the accent patterns within the speech.

The results of the questionnaire showed that for both intelligibility and naturalness of the three speech conditions, diphone synthesis was rated lowest followed by unit selection synthesis. These subjective measures correspond with the results found in our eye-movement data. The combination of these offline subjective measures and online objective measures give a detailed insight in the perception and the processing of synthetic speech.

The experiment shows that eye tracking is a promising research method to evaluate synthetic speech. The results give an insight in how similar the processing of synthetic speech is compared to the processing of human speech on a segmental and a suprasegmental level. The complexity of the method could be reduced if a test bed environment would be created that enables an easy comparison of the processing of new speech synthesis systems. That way new speech synthesis methods could be tested in a standardised way.

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] Pisoni, D.B., *Some measures of intelligibility and comprehension*. In J. Allen, M.S. Hunnicutt, and D.H. Klatt (eds.), 1987. From Text to Speech: the MITalk System. Cambridge University Press, Cambridge, pp.151-171.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] Matin, E., Shao, K. and Boff, K., "Saccadic overhead: information processing time with and without saccades", *Perceptual Psychophysics*, 53, 1993, pp. 372-380.