ADJUSTEMENTS OF SYNTHETIC SPEECH TO OPTIMISE THE INTELLIGIBILITY FOR HEARING-IMPAIRED USERS

Anne-Marie Öster
Department of Speech Communication and Music Acoustics, KTH,
Box 700 14, S-100 44 Stockholm, Sweden

ABSTRACT

Results are reported of perceived intelligibility of synthetic speech, produced by the Infovox-system at three different rates, by eight hearing-impaired elderly users and by eight younger normal-hearing persons with a simulated hearing-loss. The speech material consisted of 17 Swedish consonants embedded in /a-a/ context, words and sentences. Preferred rate of some written sentences was investigated through a synthesis production experiment.

INTRODUCTION

Nowadays, synthetic speech is used to a great extent in public information systems and in various applications for visually handicapped persons or persons with speech disorders. The Infovox-system used in Sweden is a text-to-speech synthesis that was originally developed at KTH in the seventies. The system transforms text to synthetic speech by means of rules implemented in the system and it has multilingual capabilities. Continuing improvements have been introduced over time [1]. Engstrand [2] reported that the users describe the synthesis as a fantastic aid that facilitate the exercise of a profession, studies or communication in general. The comprehension of the synthetic speech was good with some effort, free from interference, in a familiar context and after some learning time. However, a higher degree of concentration was needed to profit by synthetic speech than to natural speech.

The segmental intelligibility of the KTH- and the present Infovox-system was evaluated by normally hearing listeners by Goldstein and Till [3]. Eighteen Swedish consonants were used in a VCV-test in the context of the vowels /a, i, o/. The overall error rate of the KTH-synthesis was 8.7 % and the Infovox-system obtained slightly above 12 %. Neovius and Raghavendra [4] evaluated the comprehension of natural and synthetic speech in the Swedish and American English KTH-system. These studies show that the segmental intelligibility is very good as well as the general comprehension of the Swedish text-to-speech systems.

However, many users with a visual defect are elderly people and many persons in the surrounding of a user with a speech disorder are elderly people and consequently it could be assumed that many of those also have a deteriorating hearing according to age. Until now, synthetic speech in general has never been evaluated by hearing-impaired listeners in Sweden.

In a preliminary study by Granström & Öster [5] some experiments were made relating to talking rate vs. articulation speed with only one very co-operative severely hearing-impaired subject. In the synthesis production experiment a feature of the system was utilised, that makes it possible to interactively change the synthesis by moving a point on the computer screen. The adjustment of segmental duration was allowed on the X-axis and the speed of formant transitions on the Y-axis. The settings were made while listening to a paragraph speech, one sentence at a time. Pure intelligibility was, intentionally not part of the test, since the text of the story was also given to our subject. His settings did actually not deviate a lot from the settings by normally hearing listeners, except that he preferred slightly slower speech and more rapid formant transitions.

This paper reports of eight elderly hearing-impaired persons' ability to understand synthetic speech at a natural rate, a low rate as well as at a high rate. Eight younger normal-hearing persons were tested with the same speech material at the three different rates with and without a simulated hearing-loss at high frequencies.

METHOD

Subjects

Eight hearing-impaired adults; five men and three women between 59 and 76 years of age and eight normal-hearing persons below 30 years of age participated in the study. The hearing-impaired subjects were listening to the synthetic speech at their individually selected most comfortable level (67-107 dBSPL) and their best ear pure-tone averages at .5, 1, 2, 4 Hz varied between 35-78 dB.

Stimuli

The material consisted of a segmental intelligibility test, a word recognition test, comprehension of simple questions and subjective adjustments of preferred reading-rate of written text.

The segmental test consisted of 17 Swedish consonants which were embedded in an /a-a/-context. Three random lists were constructed and presented to the subjects at three different rates: natural, low and high. Each consonant appeared twice in every test list. In the beginning of each list five extra stimuli were presented to familiarise the subjects with the synthetic speech. The word recognition test was made up of 50 low-predictable words, based on which sentences of five words each were constructed. Three lists of ten sentences each were chosen for this test to be presented at different speaking rates. The lists were made by choosing the words at random from the original list. Thus all lists had the same content of words but in different combinations. Hence, the three lists were equally difficult and the linguistic ability of the subjects had low influence on the result. However, this word test, with five words in sequence, loaded the short-term memory more than single word-tests. The simple questions read by synthetic speech were so-called Helen-questions of the type "What colour is a lemon?" The listeners' task was to answer the question with one word only. Each list contained 20 questions and every list started with three training-questions to familiarise the subjects with the synthetic speech condition. The subjective adjustments of preferred reading-rate of written text were performed in a similar way as in the experiment cited above [5]. By moving a point along the X-axis on the computer screen the articulation rate was adjusted from very low to very high rate. The settings were made while listening to a paragraph speech, one sentence at a time.

Speech rate conditions

All test material was presented to the subjects at three different rates: natural, low and high. The natural rate was set to a somewhat lower value than the default value of the system. The low rate was selected as the rate an experienced person used when talking clearly to hearing-impaired persons. The high rate was settled according to a normal conversation rate with normal-hearing people. The time elapsed at the low rate condition was increased with 58 % of the time at the natural rate condition and decreased at the high rate condition with 37 % of the time at the natural rate.

Testing conditions

The hearing-impaired subjects (HI) were listening to all test lists binaurally through headphones at their comfortable level. The subjects wrote down which consonant they heard on the segmental test and echoed the perceived words at the word recognition test immediately after presentation. The test-leader wrote down all words. The comprehension test of simple questions was made in the same way as the word recognition test.

The natural rate conditions were always presented first to all subjects while the high rate conditions were presented to half of the subjects before the low rate conditions and to half of the subjects in opposite order. The normal-hearing subjects were tested with the segmental VCV-test with (NHSIM) and without (NH) a hearing-loss simulation at high frequencies. They were also tested with the word recognition test and the comprehension test of Helen-questions with the hearing-loss simulation. The testing conditions were the same as for the hearing-impaired subjects. The test material was delivered binaurally through headphones at the subjects' most comfortable level.

Description of the hearing-loss simulation

It has been shown that the intelligibility of synthetic speech is more impaired in masking noise compared to natural speech, because of the small number of acoustic cues used to specify phonetic segments in synthetic speech. This was the reason for choosing a recently developed simulation method of a hearing-loss, instead of masking noise. The simulation method used [6] is an approximation of threshold shifts and loudness functions of sensorineurally hearing-impaired individuals in real time. Based on audiogram data, processing is done binaurally in octave bands with the use of DSP programming tool ALADDIN. This simulation method does not include loss of temporal and spectral resolution. In the experiment with the normal hearing-listeners, one type of audiogram was chosen based on the audiograms of the eight hearing-impaired subjects. The synthetic speech samples were played through the simulation model and presented to the normal-hearing listeners through headphones at preferred level.

RESULTS AND DISCUSSION

The segmental test

The mean result for the 8 NH at natural rate was 94 % correctly identified consonants in an /a-a/-context. This can be compared to the result that was obtained of the Infovox-system by Goldstein & Till [3]. In their study the mean result of correctly identified consonants for normal-hearing listeners was 87.7 %. However, in their study the contexts were both /a-a/, /i-i/ and /o-o/ and it seemed that the error rates were lowest in the /a-a/ context. The mean percent-correct perceived synthetic consonants at natural rate for the HI-group was 51 % and 53 % for the NHSIM-group. The results of low and high rate indicate that changing the speech rate did not increase or decrease the mean percent intelligibility of synthesised consonants surrounded by the vowel /a/ for hearing-impaired or normal-hearing listeners with a simulated hearing-loss.

The HI-group perceived all voiced consonants as voiced at all rates but the unvoiced sounds were perceived as voiced in roughly 19 % in all conditions. There were fewer confusions of manner of articulation than of place of articulation. The confusions made of manner of articulation occurred for unvoiced stops. Unvoiced stops were perceived as unvoiced fricatives in approximate 45 % at all rates. A common feature was that /t/ and /k/ was understood as /s/ to a high degree at all rates and /p/ was perceived as /f/ at a low rate and equally much as /f/ and /s/ at a high rate.

Percent-correct place of articulation within each manner of articulation was highest for the lateral, the tremulant and voiced fricatives at all rates and lowest for unvoiced stops at all rates.

There were practically no confusion made for voiced/unvoiced consonants by the normal-hearing group with or without the hearing-loss simulation. Confusions of the manner of articulation by the NH-group were negligible. With the hearing-loss simulation the normal-hearing group extremely often perceived unvoiced stops as unvoiced fricatives at all rates, but especially at high rate. Most unvoiced consonants were heard as /s/. Unvoiced fricatives were at almost all rates understood as unvoiced stops to about 28 % of the cases. Moreover, voiced fricatives were frequently perceived as /l/ at high rate; 59 % and also the tremulant was heard as /l/ in 38 % of the events at the low rate.

The consonants best perceived by the hearing-impaired elderly adults within each manner of articulation were the voiced stop /b/, the nasal /n/, the lateral /l/, the tremulant /r/, the voiced fricative /v/, the unvoiced fricative /sj/ and the voiced stop /k/. The results of over-all correct-perceived consonants were almost identical for the HI-group and the NHSIM-group but there were some interesting perceptual differences between the groups. Some of the confusions made by the NHSIM-group could be due to the approximation of the sensorineural loudness function that is part of the simulation procedure.

As a rule, /g, ng, tj, p, t/ were the consonants that the HI-subjects had difficulties to perceive correctly in the context with /a/.

The word recognition test

The result showed that the HI-group understood 65 % of the words at natural rate, that was always presented first, compared to 60 % correct words by the NHSIM-group. The NH-listeners understood all words at all rates to almost 100 %. The result of the NHSIM-group was

lower than expected probably due to their unfamiliarity with perceptual consequences of the simulated hearing-loss. The hearing-impaired listeners were used to listen to a reduced speech signal and did not give up the task as easily as the NHSIM-group did. Both groups performed best at the low rate-condition probably because of a longer processing time available.

There was a learning effect for both groups and for low and high rate when presented last. However, both groups performed best at low rate irrespective of order of presentation.

Comprehension of simple questions

Both groups performed very well at all rates but the low rate seemed to be a bit superior in intelligibility to the high rate. The HI-group got slightly better results than the NHSIM-group at all rates. Their was no apparent learning effect according to order of presentation. However, the HI-group got better scores at low rate when presented last and the NHSIM-group did slightly better when high rate was presented last. The spontaneous utterance of most subjects was that they found the synthetic speech thick, drawling and unnatural at low rate even though they got more time to process the meaning of it.

Subjective adjustments of preferred reading-rate of written text

The preferred speech rates by the HI-subjects were almost the same for all of them, except for one of the subjects, who preferred a higher rate than the others. In general, the result showed that the HI-subjects preferred almost the same rate as the natural rate condition of the listening tests reported in this study. The natural rate condition was set to a somewhat lower value than the default value of the system.

From the above results it is shown that changing the speech rate of the synthesis did not significantly affect the overall intelligibility or the intelligibility of Swedish consonants for either of the listener groups. This is in accordance with the discussion by Engstrand [2] where he reports that many users found that a high rate did not improve the naturalness and a low rate did not improve the intelligibility of the synthesis. Engstrand assumed that a manipulation of the speech rate of synthetic speech disagrees with the manipulation of the rate of natural speech due to the fact that changes of the duration of natural speech are not linear.

A possible approach to improve the intelligibility of synthetic speech to hearing-impaired users could be to manipulate it with regard to the unlinearity changes of duration mentioned above. Revoile et al. [7] have shown that manipulation of both consonant and vowel durations can improve consonant recognition in cases of severe hearing impairment.

ACKNOWLEDGEMENT

The work has been supported by grants from the Swedish Handicap Institute.

REFERENCES

[1] Carlson, R., Granström, B., and Hunnicutt, S. (1990). "Multilingual text-to-speech development and applications", *Advances in Speech Learning and Language Processing*, (A.W. Ainsworth, ed), JAI Press, London, 269-296.

[2] Engstrand, O. (1993) "Behovsanalys av talarstilar i text-till-talsystem för handikapptillämpningar" *Handikappinstitutet*.

[3] Goldstein, M. and Till, O., (1992). "Is % overall error rate a valid measure of speech synthesiser and natural speech performance at the segmental level?", *Int. Conf. on Spoken Language Processing, ICSLP 92* Proceedings, Vol.2, 1131-1134.

[4] Neovius, L. and Raghavendra, P. (1993). "Comprehension of KTH text-to-speech with "listening speed" paradigm". *Proceedings of Eurospeech 93*, 1687-1690.

[5] Granström, B., Öster, A-M. (1994). "Speech synthesis for hearing-impaired persons - in research, training and communication". *STL-QPSR* 2-3, 93-112.

[6] Öhngren, G., & Dahlquist, M. (1995). "To hear hearing-impaired; simulation of hearing impairments". *Proceedings of European Conf. on Audiology*, Noordwijkerhout, Netherlands.

[7] Revoile, S. G., Holden-Pitt, L., Pickett, J., and Brandt, F. (1986). "Speech cue enhancement for the hearing impaired: Altered vowel durations for perception of final fricative voicing". *JSHR*, 29, 240-255.