Measuring Speech Recognition With a Matrix Test Using Synthetic Speech

Theresa Nuesse1,2, Bianca Wiercinski1, Thomas Brand2,3, and Inga Holube1,2

Abstract
Speech audiometry is an essential part of audiological diagnostics and clinical measurements. Development times of speech recognition tests are rather long, depending on the size of speech corpus and optimization necessity. The aim of this study was to examine whether this development effort could be reduced by using synthetic speech in speech audiometry, especially in a matrix test for speech recognition. For this purpose, the speech material of the German matrix test was replicated using a preselected commercial system to generate the synthetic speech files. In contrast to the conventional matrix test, no level adjustments or optimization tests were performed while producing the synthetic speech material. Evaluation measurements were conducted by presenting both versions of the German matrix test (with natural or synthetic speech), alternately and at three different signal-to-noise ratios, to 48 young, normal-hearing participants. Psychometric functions were fitted to the empirical data. Speech recognition thresholds were 0.5 dB signal-to-noise ratio higher (worse) for the synthetic speech, while slopes were equal for both speech types. Nevertheless, speech recognition scores were comparable with the literature and the threshold difference lay within the same range as recordings of two different natural speakers. Although no optimization was applied, the synthetic-speech signals led to equivalent recognition of the different test lists and word categories. The outcomes of this study indicate that the application of synthetic speech in speech recognition tests could considerably reduce the development costs and evaluation time. This offers the opportunity to increase the speech corpus for speech recognition tests with acceptable effort.

Keywords
speech audiometry, speech reception threshold, Oldenburg sentence test, text-to-speech, synthetic speech

Date received: 23 January 2019; revised: 13 June 2019; accepted: 19 June 2019

Introduction
Understanding speech is crucial for communication in everyday life. Therefore, measurements dealing with the ability to understand speech in different listening situations are frequently applied in audiological research and clinical practice. Normally, the test procedures involve a listening and a repetition part; participants listen to a speech snippet like phonemes, words, or whole sentences and repeat what they heard. The measurements take place in quiet or in noise at different signal-to-noise ratios (SNRs). They are implemented with fixed or adaptive level adjustments, yielding the speech recognition threshold (SRT) at an intended percentage of correct answers, which is typically 50% correct. The measured speech recognition scores can be described as a function of the SNR, resulting in psychometric functions. These allow for an estimate of the SRT and, therefore, for a comparison with the outcome of adaptive procedures. To fit psychometric functions, a logistic function can be utilized (Knoblauch & Maloney, 2012). A steep slope of the psychometric function is needed for high reliability when estimating the SRT with an adaptive procedure. To ensure such a steep slope, the speech recognition score for the words

1Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, Oldenburg, Germany
2Cluster of Excellence “Hearing4All”, Oldenburg, Germany
3Medizinische Physik, Universität Oldenburg, Oldenburg, Germany

Corresponding Author:
Theresa Nuesse, Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, Ofener Str. 16/19, D-26121 Oldenburg, Germany.
Email: Theresa.Nuesse@jade-hs.de

Trends in Hearing
Volume 23: 1–14
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2331216519862982
journals.sagepub.com/home/tia

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (http://www.creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).
within each sentence should be as similar as possible (Kollmeier, 1990).

Conventional speech recognition tests have one thing in common: The presented speech material was spoken by a natural speaker and recorded in a studio. In many cases, elaborate adjustments in postprocessing were necessary to ensure equivalent test lists with respect to their speech recognition score (perceptual balance, ISO 8253-3). This procedure is very time-consuming and thus often restricts the number of test lists or speech samples. An opportunity to shorten the development time of speech tests is to create a so-called matrix test. The structure of most of the available matrix tests is based on the Hagerman sentences (Hagerman, 1982), consisting of 50 well-known words. All test sentences follow the same syntactical structure, but are semantically unpredictable. The Oldenburg sentence test (OLSA) is a German matrix test in which the speech material is based on German sentence structure (name, verb, numeral, adjective, object). It was developed by Wagener and colleagues (Wagener, Brand, & Kollmeier, 1999a, 1999b; Wagener, Kühnel, & Kollmeier, 1999) using a male speaker and was later supplemented by a second version with a female speaker (Ahrlich, 2013; Wagener, Hochmuth, Ahrlich, Zokoll, & Kollmeier, 2014). To generate a variety of test lists with preferably low effort, a total of 100 sentences were recorded to generate the speech material. These recordings were cut into components, adjusted in intensity, and concatenated into new combinations. With this technique, coarticulation effects are maintained in the recordings. Level adjustments were necessary to obtain similar recognition for every word, to maximize the slope of the psychometric function and thus to minimize the uncertainty in the measured SRTs. Although the female and the original male version of the OLSA use the same base matrix, the measured SRTs differ by about 2 dB (−9.4 dB SNR for the female vs. −7.1 dB SNR for the male voice). This deviation is mainly due to different speech speeds and spectral properties. Matrix sentence tests are also available in many other languages (Kollmeier et al., 2015; Warzybok et al., 2015; Zokoll et al., 2015) and are applied in quiet as well as in noise. Due to the special recording and cutting procedure, the development of matrix tests can be achieved in less time compared to other speech recognition tests. Nevertheless, evaluation measurements are still necessary to determine the level adjustments.

Regarding progress in speech technology, the question arises whether synthetic speech could be an alternative to time-consuming recordings of natural speakers, whereas early so-called text-to-speech (TTS) systems struggled with intelligibility issues that led to error rates of up to more than 30%, depending on the system (Greene, Logan, & Pisoni, 1986); under certain circumstances, (simple wording and low background noise), more recent systems provide perfectly understandable speech (Benoit, Grice, & Hazan, 1996). Following the results of the Blizzard Challenge over more than one decade, modern, high-quality synthetic speech does not statistically differ in intelligibility (King, 2014). Most naturally sounding TTS systems are based on unit selection (King, 2014), which means that speech signals are created by concatenating pieces of recordings stored in a library. Complex linguistic analysis precedes this process to decompose the written text into prosodic units. Subsequent to the selection of appropriate phonemes, the database entries are adjusted in pitch, duration, and intensity before they are concatenated to new phonetic sequences.

Furthermore, applications in everyday life suggest robustness, to some extent, in background noise. To evaluate the quality of different synthesizers, in various studies measurements with participants listening to synthetic speech in noise and quiet were carried out (e.g., Benoit et al., 1996; Hazan & Shi, 1993; Isaac, 2015). Hazan and Shi (1993) found a high, within-subject variability for synthetic and natural speech in noise, covering the differences between the two speech types. Although present TTS systems do not differ from natural speech in intelligibility, synthetic speech leads to higher variability in recognition scores compared to degraded natural speech (Hazan & Shi, 1993). This suggests that listening to synthetic speech induces additional difficulties to the participants depending on individual perceptual strategies (Hazan & Shi, 1993). Differences in language acquisition or cognitive effects might underlie the increased difficulty, but are elusive. Recently, Govender and King (2018a, 2018b) examined the cognitive load and listening effort of different TTS systems in relation to natural speech. Although the cognitive load measured in a dual-task paradigm was not influenced by the speech synthesis (Govender & King, 2018a), the listening effort represented in pupil dilation tends to be higher for synthetic speech compared to natural speech, albeit without statistical significance (Govender & King, 2018b).

The overarching aim of this study was to examine whether synthetic speech is suitable for application in speech audiometry. In the long run, creating speech materials using TTS systems instead of natural speech may allow for the generation of large speech corpora more quickly and with less effort. Furthermore, specific properties of the speech such as fundamental frequency and speech rate can be adapted to the needs of the study or the patient population quite easily in TTS systems. To analyze the recognition of synthetic speech and to show possible differences to natural speech, a listening test with young, normal-hearing participants was performed, to choose a TTS system with the best possible natural sound. Using the selected system, the speech material of
the female OLSA was replicated. As a proof of concept, the matrix speech test material was chosen for this first approach utilizing synthetic speech in a recognition test in order to directly match the results of the two speech types. In addition, the female OLSA is the only common speech test in German with a female speaker, which was one requirement for the synthetic voice with regard to a possible application in, for example, hearing-aid evaluation. To allow for the direct comparison, the properties of the new, synthetic, female OLSA were fitted as far as possible to its natural equivalent. Subsequently, SRTs and slopes of the psychometric functions, as well as training effects, were evaluated, focusing on differences between the naturally spoken and the synthetic speech. Matrix tests exhibit a significant training effect caused by the simple sentence structure and the recurrently presented words. Wagener et al. (1999b) and Schluter, Lemke, Kollmeier, and Holube (2016) recommend a training phase of one to four test lists before data collection, depending on the required measurement accuracy. For clinical practice, at least 40 sentences of training are recommended when using matrix speech tests. The training effect of six test lists on the SRT is estimated to be approximately 2 dB for the male OLSA (Schluter et al., 2016; Wagener et al., 1999b) and 2.2 dB for the female OLSA (Ahrlich, 2013). The greatest training effect was observed from the first to the second test list, with a shift in SRT of approximately 1 dB (Schluter et al., 2016).

Overall, two main research questions were addressed by this study. The first question is: Can test results (SRT, slope, training effect) and characteristics (perceptual balance) of a matrix test using natural speech be reproduced using synthetic speech? The differences between natural speaker and TTS system are expected to lie within a range comparable to the variation between natural speakers, that is, among male and female OLSA. The number of recommended training lists is assumed to be the same.

The second research question is: Does the usage of synthetic speech reduce the development effort of speech test material? The production time of the synthetic speech material is expected to be shorter than for natural speech. Due to the consistent speech properties, no necessity for optimization is assumed (e.g., level adjustments), which reduces the development effort considerably.

Methods

The study design was similar to evaluations of the female OLSA (Ahrlich, 2013; Wagener et al., 2014) and matrix tests in other languages (Puglisi et al., 2015; Warzybok et al., 2015; Zokoll et al., 2015). The major difference was the alternating presentation of natural and synthetic speech.

Subjects

The speech recognition tests were performed monaurally on the better ear of 48 normal-hearing listeners (31 females and 17 males). They were between 18 and 25 years old, with an average of 21.8 years (SD: 2.1 years). All were students of Jade University of Applied Sciences or Carl von Ossietzky University Oldenburg, Germany. Before the speech tests, air-conduction pure-tone audiometry was administered using an audiometer (Auritec AT 700; Hamburg, Germany) and headphones (Sennheiser HDA 300; Wedemark, Germany). The pure-tone hearing thresholds were 10 dB HL or below at all audiometer frequencies from 125 Hz to 8 kHz. At up to two frequencies, hearing thresholds of 15 dB HL were allowed. If both ears fulfilled these criteria, the better ear was selected, based on a comparison of the average pure tone thresholds at 0.5, 1, 2, and 4 kHz (PTA4). All participants were unfamiliar with the OLSA and received 10 Euros per hour as reimbursement. They gave their informed consent prior to inclusion in the study. The experiment was approved by the ethics committee (“Kommission für Forschungsfolgenabschätzung und Ethik”) of the Carl von Ossietzky University in Oldenburg, Germany (Drs. 34/2017).

Speech Material

The speech material contained 150 sentences composed from the word matrix of the OLSA (Wagener et al., 1999). Each sentence included five words, one of each word category (name, verb, numeral, adjective, and object). For each word category, 10 different words were available, that is, 50 words in total. The 150 sentences were grouped pseudorandomly in 15 blocks of 10 sentences each. Each word appeared once in each block. Since the measurement procedure expects lists of 20 sentences, two blocks were merged to one list, thus each word appeared twice in each test list. One block was used twice. This procedure resulted in eight identical test lists with 20 sentences each. The speech material of the female OLSA was developed by Ahrlich (2013) and Wagener et al. (2014), as described in the “Introduction” section. For the measurements of this study, the same 150 sentences of the original test were arranged following the scheme described earlier, resulting in eight identical test lists for the two speech types. Since these sentences were already included in the measurement software, no further adjustments were necessary.

Selection of TTS System

For the development of a matrix test with synthetic speech, a TTS system had to be selected among the many available commercial and research systems for TTS. As a first step, a list of 36 TTS systems was
compiled based on a web search. For each TTS system, several words and sentences of different speech materials were generated with available demo versions and without further optimizations. The subjective impression of speech quality (mainly naturalness of accent and completeness of synthesis) was informally evaluated by one listener (a student employed as research assistant performing the web search) and assigning German school grades. The ratings were reviewed by one other listener (second author). Three TTS systems were preselected based on the subjective evaluation and the manufacturer’s information with regard to the following criteria:

- Assigned rating
- Terms of use (e.g., open-source usage)
- Adjustable speech rate
- German female voice available

The three preselected TTS systems were compared to each other and to an original natural speech recording using 12 sentences taken from existing speech recognition tests:

- Six (three short and three long) everyday sentences of the Göttingen sentence test (Kollmeier & Wesselkamp, 1997), spoken by a male speaker.
- Three sentences from the OLACS corpus (Oldenburg linguistically and audiologically controlled sentences; Uslar et al., 2013), spoken by a female speaker.
- Three sentences of the OLSA spoken by the same female speaker as for OLACS (Wagener et al., 1999, 2014).

These speech stimuli included eight statements with a simple sentence structure, two questions, and two sentences with a relative clause. In addition, a sequence of two grammatically more complex sentences of the German version of the tale “North Wind and Sun” and spoken by a female speaker (Holube, Fredelake, Vlaming, & Kollmeier, 2010) were used. The sentences were generated by all three TTS systems without further optimizations, using female speech. All stimuli were calibrated to the same root mean square level and presented at a level of 65 dB sound pressure level (SPL).

Twelve young, normal-hearing listeners (average age: 24 years, nine females and three males) evaluated the quality of the three TTS systems. Based on the universal perceptual quality dimensions described by Hinterleitner, Norrenbrock, and Möller (2013), they rated naturalness, prosody (stress and intonation), fluency, and intelligibility. The listeners ranked the audio files for the three quality dimensions relative to each other, with support of an absolute scale between “not at all” (“gar nicht” in German), “somewhat” (“wenig” in German), “middle-rate” (“mittelmäßig” in German), “very” (“sehr” in German), and “completely” (“vollkommen” in German), while switching between the synthetic speech and the original natural speech. The speech types were randomly arranged on the computer screen. Auditory presentations were selected by the listeners and were repeatable as often as necessary. In addition, the overall impression for the three synthetic speech types was rated, based on the question “If you would have to decide for one voice, which one would you select?” (“Wenn Sie sich für eine Stimme entscheiden müssten, welche würden Sie wählen?” in German).

On average, the TTS system of Acapela group (Acapela Group Babel Technologies SA) was rated the best in the relative comparisons. For the highest overall impression, 8 of the 12 listeners selected this TTS system, two selected each of the other systems. Therefore, the TTS system of Acapela group was used in this study. The technology employed in the chosen system is a nonuniform unit selection that allows synthesis of very natural-sounding speech, compared to other techniques (King, 2014).

**Procedure for Synthesis**

The speech material was generated with the German voice Claudia of the TTS system “Virtual Speaker” (Acapela Group, Mons, Belgium). The text of each of the 150 sentences was entered into the software as input. The TTS system generated 10 alternatives for each of the five words of each sentence. The 10 alternatives differed in stress. The selection of the most natural word combination was based on subjective auditory evaluation of the second author following three criteria:

- Avoidance of discontinuities (sudden changes in fundamental frequency, perception of intersections).
- Correct stress of each word as expected for standard German, and
- Natural intonation of each sentence.

Using this selection procedure, each of the 150 sentences was generated separately. The subjective evaluation described earlier led to speech signals with correct German pronunciation. However, the proper names at the beginning of the sentences seem to be difficult to synthesize. For every sentence, the best alternative was chosen, which sometimes did not sound as natural as the other parts of the sentences. The output of the TTS system was mono audio files with a sampling rate of 22.05 kHz, 16 bit. Each sentence was adjusted to the same root mean square level as its natural counterpart from the female OLSA to avoid any possible deviations in level. To be applicable in the measurement software and similar to the natural speech, the mono
audio files were doubled to stereo audio files and resampled to a sampling rate of 44.1 kHz.

**Speech Characteristics**

As additional parameters of the TTS system, the speech rate and the fundamental frequency of the speaker were adjustable. Those parameters were changed to match the characteristics of the female OLSA with natural speech as close as possible, while avoiding audible distortions of the synthesis procedure.

The speech rate was decreased compared to the standard setting of the TTS system. Figure 1(a) shows the speech rate of the synthetic speech, natural female, and natural male speech of the OLSA in syllables/min. The average speech rate of the natural female OLSA was 167 syllables/min, whereas the synthetic female speech was a little faster at 175 syllables/min. Nevertheless, both female versions were considerably slower than the natural male OLSA (208 syllables/min on average).

The fundamental frequency was increased compared to the standard setting of the TTS system. Figure 1(b) shows the fundamental frequency of the synthetic female speech and the natural female speech. The fundamental frequency of the synthetic speech is still on average somewhat lower than the fundamental frequency of the natural speech and shows a broader distribution. Nevertheless, the long-term average speech spectra of both speech types are very similar (Figure 1(c)).

**Noise**

For both natural and synthetic female speech, a stationary noise was used as masker. The noise was generated by superimposing the respective speech material 30 times with randomized time shifts, similar to the noise generation of the OLSA with a male speaker (Wagener et al., 1999). This procedure yielded the same noise spectrum as the long-term average speech spectrum. The noise for the female OLSA was available from Wagener et al. (2014) and implemented in the measurement software. The noise for the synthetic speech was generated using the same software. Each noise was always presented with a constant level of 65 dB sound pressure level. The noise masker started 500 ms before the beginning of each sentence and ended 500 ms after the end of each sentence.

**Equipment**

The measurements were conducted with a research version of the Oldenburg Measurement Application (OMA, Version 2.0) of Hö rTech gGmbH (Oldenburg, Germany) in a sound-attenuating booth. The stimuli were presented monaurally to the better ear via Sennheiser HDA 200 headphones (Wedemark, Germany) driven by a sound card RME Fireface UC (Audio AG, Heimhausen, Germany) and a headphone amplifier HB7 (Tucker-Davis Technologies, Alachua, USA). The listeners were asked to repeat the sentences, and the investigator entered the correctly repeated words on a touch screen.
Measurement Procedure

Figure 2 gives a schematic overview of the experiment. It was divided into two sessions within a maximum time span of 2 weeks for each participant. During each session, 18 test lists of 20 sentences each were presented to the listeners. The sessions were divided into a training phase and a measurement phase.

The goal of the training phase was to familiarize the listeners with the speech material as well as the procedure of the speech test and to study the improvement of the results over time. Within the training phase of the first session, eight test lists were presented. During the training phase of the second session, four test lists were presented. For each test list of the training phases, the SRT was estimated using the adaptive procedure of Brand and Kollmeier (2002). Within the adaptive procedure, the noise level was kept constant at 65 dB SPL, and the speech level was adjusted according to the responses of the listeners, starting with an SNR of 0 dB. To familiarize the listeners with both speech types, the type was changed after six test lists in the first session and after two test lists in the second session.

The measurement phase of the first session consisted of 10 test lists and the second session consisted of 14 test lists. Within the measurement phases of both sessions, the speech types were alternated from test list to test list. The SNR was kept constant during each test list of the measurement phases. To enable a reliable fit of the psychometric function, tests should result in speech recognition scores of about 20%, 50%, or 80%. A pilot experiment with five listeners using −12, −9, and −6 dB SNR according to Ahrlich (2013) revealed that some of the listeners scored below 10% at the lowest SNR. These results were discarded and the SNRs modified to −11, −8.5, and −6 dB SNR. Each of the eight test lists were presented to each listener at each of the three SNRs during the two measurement phases.

Latin squares were used to balance the test list number and the order of the SNRs. In addition, the speech-type selection for the first test list of the first session was balanced, and both speech types were presented alternately in the following measurement phase.

Results

Speech Recognition at Different SNRs

The speech recognition scores of the measurement phase were analyzed separately for the natural and synthetic speech and compared to each other. In a first step, the results of the eight test lists for each listener at each SNR were averaged. Figure 3 shows the average values as boxplots. Results were normally distributed (Shapiro–Wilk test, p > .05 for all SNRs, resulting SRTs and slopes s for both types of speech), wherefore no linearization

Figure 3. Speech recognition scores for the three SNRs and the two speech types, natural (gray and open box) and synthetic (black and filled box) for all listeners. The circles tag the mean values. Boxplots denote median (horizontal line within the box), interquartile range (length of the box), whiskers up to the highest or lowest value within 1.5 times the interquartile range and drawn from the upper quartile to the lower quartile, and outliers (+). OLSA = Oldenburg sentence test; TTS = text-to-speech; SNR = signal-to-noise ratio.

Figure 2. Overview of the experiments divided in two sessions separated into training and measurement phase. The numbers 1 to 18 indicate the quantity of measured test lists (not the test list number). The color (white or gray) indicates the natural or the synthetic speech, respectively.
(e.g., rationalized arcsine unit transform, Studebaker, 1985) was applied. Paired sample t tests with Bonferroni correction for three comparisons ($\alpha = .0167$) revealed statistically significant differences for $\text{SNR} = 11\text{ dB}$ ($t_{47} = 5.90, p < .001$), for $\text{SNR} = 8.5\text{ dB}$ ($t_{47} = 8.28, p < .001$), and for $\text{SNR} = 6\text{ dB}$ ($t_{47} = 4.44, p < .001$).

Speech recognition scores were better for the natural speech than for the synthetic speech, but the differences between the mean values were only 4.1% (at $\text{SNR} = 11\text{ dB}$), 7.1% (at $\text{SNR} = 8.5\text{ dB}$), and 2.4% (at $\text{SNR} = 6\text{ dB}$).

**Psychometric Functions**

Individual psychometric functions were fitted according to equation (1) in Brand and Kollmeier (2002) for every listener and each speech type (Figure 4), resulting in individual values for SRT and $s$. The mean values are given in Table 1 in comparison to the results of Ahrlich (2013). The SRTs for the two speech types were significantly different ($t_{47} = 9.78, p < .001$), but the slopes $s$ were not ($t_{47} = 1.55, p = .217$). The difference in mean SRT between the synthetic speech and the natural speech (0.5 dB) was in the same range as the difference between the natural speech measured in this study and the results for natural speech measured by Ahrlich (2013) (0.3 dB).

**Training Effects**

Figure 5 shows the SRTs of the training phase of the two sessions together with the mean SRTs of all test lists presented during the measurement phase of both sessions. The numbers on the x-axis give the presentation order in time (not the number of the test list). The two subfigures show the two different configurations: the first starting with synthetic speech for one half of the subjects and the second starting with natural speech for the other half of the subjects. A difference of 0.6 dB SNR was observed for the first test list of the training phase.

As expected, the improvement in SRT was largest from the first to the second test list, being larger for the synthetic speech (1.1 dB SNR) than for the natural speech (0.9 dB SNR). The training effect from the first to the sixth test list was 1.8 dB SNR for the synthetic speech and 1.7 dB SNR for the natural speech. A repeated-measures between-subjects analysis of variance (ANOVA) showed a statistically significant effect of training on SRTs, $F(7, 322) = 60.159, p < .001$, which according to Cohen (1988) is a strong effect (effect size: $f = 1.14$). There was no significant difference between subject effect of training sequence, $F(1, 46) = 1.128, p = .294$, for the two different training sequences in the first session. Therefore, post hoc comparisons ($t$ test) were carried out on the whole data set (for details, see Table 2). SRTs of the first test list differ statistically significant from all following test lists. Furthermore, $t$ test outcomes indicate that there were SRT improvements from test list to test list, until a plateau was reached following the fourth test list or after 60 sentences of training. Nevertheless, the mean SRTs of the plateau (test lists

![Figure 4](image-url)
4–6) differed statistically significant from the overall mean SRT estimates of the measurement phases for all participants, by 0.9 dB for the synthetic speech ($t_{23} = 5.08$, $p < .001$ for training with TTS; $t_{46} = 5.48$, $p < .001$ for training with natural speech) and by 0.7 dB for the natural speech ($t_{23} = 6.50$, $p < .001$ for training with natural speech; $t_{46} = 4.69$, $p < .001$ for training with TTS). The SRTs of the training phase of the second session revealed smaller differences to the SRTs of the measurement phases. For the second test list of the second training phase, a difference of 0.4 dB was observed for the synthetic speech and a difference of 0.2 dB was observed for the natural speech compared with the estimated SRTs in the measurement phase. This reduction indicates that the training was not finalized after the training phase in the first session, but continued during the measurement phase and the training of the second session.

In addition, the SRT changes after switching from one speech type to the other are noticeable. When switching from the synthetic speech to the natural speech during the training phase (upper panels in Figure 5), the training effect proceeds with a statistically significant mean improvement of 0.3 dB ($t_{23} = 2.74$, $p = .012$). When switching from the natural speech to the synthetic speech during the training phase (lower panels in Figure 5), the SRTs deteriorated by 0.7 dB on average (statistically significant, $t_{23} = -4.77$, $p < .001$) before training proceeded. As described earlier, the SRTs are lower for the synthetic speech compared to the natural

---

**Figure 5.** SRTs for the eight test lists of the training phase in the first session and four test lists of the training phase in the second session using the adaptive procedure and the means of all test lists presented in the measurement phase of the first and the second sessions. Training was conducted randomized and alternating in terms of start speech type. Boxplots of second training session do not necessarily contain data of the same participants as in the first session presented in the same row. The colors indicate the speech type (black lines and filled boxes: synthetic speech; gray lines and open boxes: natural speech). OLSA = Oldenburg sentence test; TTS = text-to-speech; SNR = signal-to-noise ratio; SRT = speech recognition threshold.

**Table 2.** $p$ Values (Bonferroni-Corrected for Multiple Comparisons) of the Post Hoc Pairwise Comparisons Using $t$ Test of the SRTs Measured in the Training Phase of the First Session.

| Training list | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---------------|---|---|---|---|---|---|---|---|
| 1             | <.001 | | | | | | | |
| 2             | <.001 | 1.000 | | | | | | |
| 3             | 0.288 | | | | | | | |
| 4             | <.001 | <.001 | | | | | | |
| 5             | <.001 | <.001 | <.003 | .937 | | | | |
| 6             | <.001 | <.001 | <.001 | .352 | .095 | | | |
| 7             | <.001 | <.001 | <.001 | | | | | |
| 8             | <.001 | <.001 | <.001 | .116 | .105 | 1.000 | .246 | \|

*Note: Statistically significant effects in bold font.*
speech. This was taken into account by adding the mean SRT difference of 0.5 dB to the SRTs measures with natural speech. After doing so, no statistically significant difference between training lists 6 and 7 was found for either sequence (t_{23} = -1.22, p = .235 for the switch from synthetic speech to natural speech and t_{23} = -1.52, p = .143 for the switch from natural to synthetic speech).

**Comparison of Test Lists**

To ensure reproducibility of speech test results, test lists should be perceptually balanced, and therefore result in similar speech recognition scores (ISO 8253-3). During the development process of the OLSA with natural speech, an extensive optimization phase was integrated, to homogenize the recognition scores of the speech corpus. Based on empirical data collected with normal-hearing listeners, words with high speech recognition scores were decreased in level and words with low speech recognition scores were increased in level. This procedure was omitted for the synthetic speech. Nevertheless, both speech types show similar deviations between the psychometric functions of the eight test lists (see Table 3). The maximum difference in SRT for the eight test lists is 0.5 dB for the synthetic speech and 0.4 dB for the natural speech.

**Comparison of Word Categories**

Speech recognition scores for each word category and both speech types were analyzed (see Figure 6). Ceiling and floor effects occur to some extent in all word categories for the highest (-6 dB SNR) and lowest SNR (-11 dB SNR).

A repeated-measures ANOVA was conducted to examine the differences between the word categories and the effects of speech type and SNR. Mauchly’s test indicated that the assumption of sphericity had been violated, \( \chi^2(9) = 46.522, p < .001; \) therefore, Greenhouse-Geisser corrected tests are reported (\( \epsilon = 0.72 \)). There was a

### Table 3. SRT and Slopes for the Psychometric Functions of All Eight Test Lists.

| Test list | 1   | 2   | 3   | 4   | 5   | 6   | 7   | 8   |
|-----------|-----|-----|-----|-----|-----|-----|-----|-----|
| SRT in dB SNR |     |     |     |     |     |     |     |     |
| Synth     | -8.7| -8.7| -8.4| -8.5| -8.9| -8.5| -8.5| -8.3|
| Nat       | -9.2| -9.0| -9.0| -8.9| -9.1| -9.3| -9.0| -9.1|
| \( s \) in \%/dB |     |     |     |     |     |     |     |     |
| Synth     | 12.7| 13.3| 11.6| 12.3| 13.4| 13.0| 13.4| 12.8|
| Nat       | 12.2| 11.3| 12.3| 13.2| 12.7| 12.7| 12.6| 11.9|

**Note.** SNR = signal-to-noise ratio; SRT = speech recognition threshold.

**Figure 6.** Speech recognition scores of all listeners for the different word categories at the three SNRs for the synthetic and the natural speech. OLSA = Oldenburg sentence test; TTS = text-to-speech; SNR = signal-to-noise ratio.
significant main effect of word category on speech recognition scores, \(F(2.88, 155.24) = 7.17, p < .001\). No significant interaction effects were found between the word category and SNR, \(F(5.75, 155.24) = 0.56, p = .742\), or between the word category and type of speech, \(F(2.88, 155.24) = 2.39, p = .073\). As no interaction effects were found, post hoc comparisons were made for results of all SNRs and speech types together. \(T\) tests with Bonferroni correction showed statistically significant differences between names and verbs, numerals and verbs, names and objects, as well as for adjectives and numerals (for details, see Table 4).

Psychometric functions were fitted to the speech recognition scores at the three SNRs for each word category for the speech types separately, to account for the SRT offset between them, as described earlier. The SRTs indicate that the speech recognition of names is better than for the other word categories when using natural speech. In addition, numerals were better understood than other categories for both speech types (for details, see Table 5). The range of SRTs is larger for natural speech (1.9 dB) than for synthetic speech (1.1 dB).

To compare the distribution of the speech recognition scores of the two speech types, a \(t\) test for variance homogeneity was conducted. Because of floor and ceiling effects occurring at \(-6\) dB SNR and \(-11\) dB SNR, the statistical test was conducted only for \(-8.5\) dB SNR. The assumption of variance homogeneity between TTS and female OLSA data was rejected for the word types name (\(p = .001\)), verb (\(p < .001\)), and object (\(p = .003\)) at the Bonferroni-corrected significance level (\(\alpha = .01\)). The variance of speech recognition scores for the synthetic speech was higher than for natural speech for the word types name and object, whereas it was lower for the word type verb. Further variance analysis regarding specific word comprehension showed no systematic effects. In word-by-word comparisons, differences in variance between the two speech types occurred in both directions, in that the distribution of speech recognition scores was wider for some words for the synthetic speech and narrower for others. These differences were inconsistent within the word categories. Only variances of two words differed statistically significant in a \(t\) test for variance homogeneity (\(\alpha = .005\), Bonferroni-corrected), namely Nina (name, \(p = .0004\)) and Blumen (object, \(p = .0002\)). Even without such conservative \(p\) value correction (using \(\alpha = .05\)), SRTs of three words differed statistically significant between natural and synthetic speech for the word type name, and only two words differed at this significance level for each of the word types verb and object.

### Discussion

This study aimed to examine whether synthetic speech is suitable for application in speech audiometry. The results of the listening tests clearly show the effectiveness of the new approach. Measuring speech recognition with synthetic speech signals leads to similar results to the traditional use of natural speech signals. Nevertheless, certain conditions should be fulfilled; especially the hand sorting of the appropriate word implementations by an experienced listener seems to be crucial for the TTS system used here. The slopes of the fitted psychometric functions were comparable in both versions, while the SRT was slightly higher for TTS than for natural speech. Apparently, the two speech types show only small differences in objective measures (i.e., spectrum and speed). It is questionable whether using another TTS system or another fine adjustment would reach a different outcome. Nevertheless, the measured differences between natural and synthetic speech lie within the same range as for two different natural speakers (in particular male and female OLSA), as hypothesized earlier.

Altogether, the reproduction of Ahrlich's (2013) results exceeded the expectations. This holds for both evaluations: the comparison with new data of the natural speaker as well as with the synthetic speech material. The listening test described earlier led to SRT estimates that were only slightly different from the reference, while the slopes of the fitted psychometric functions were the same. A possible reason for deviation when using the natural speaker may lie in the modification of the test setup: Two of the three measured SNRs were modified to provide a better estimate of speech recognition scores of
about 20% and 50%. In addition, the uncommon alternating presentation of two different speech types (or speakers) during the same session might have influenced the listeners’ adaptation to the signal and thereby might have caused changes in the speech recognition scores.

With regard to training effects, natural and synthetic speech again showed similar results, which are comparable to those described by Ahrlich (2013) for the female speaker. The improvement from training list one to six was about 2 dB, which is also in line with the findings for the male matrix test (Schlueter et al., 2016; Wagener et al., 1999b). A training phase of about three training lists or 60 sentences is recommended, reproducing the results of Schlueter et al. (2016). Regarding the intersession training, that is, the training effect between different sessions, the data indicated that training proceeds to a limited extent after the first session. The data described here lack deeper insights into intersession learning beyond the second session. Because of the strong analogy to the male matrix test, similar intersession training effects as described in Schlueter et al. (2016) can be expected. It is noticeable that there is a change in speech recognition when switching from the synthetic speech to the natural speaker (between sixth and seventh test list in the training session), although this trend was not statistically significant. It could be due to an adaptation to speaker-dependent speech characteristics.

In addition, speech recognition was analyzed in each word category after the ANOVA showing a main effect of word category. No significant interaction effects were found, but there was a tendency for an interaction between word category and speech type \((p = .073)\). Therefore, and because of the SRT offset between the two speech types, 50% word recognition scores were assessed separately (Table 5). For natural speech, larger absolute SRT differences between the word categories were found than for synthetic speech. Especially the word type name showed higher recognition scores compared to the other word types, when listening to natural speech. In the optimization process of the male and female OLSA (Ahrlich, 2013; Wagener et al., 1999a), better speech recognition scores for names than for the other word categories were also found, resulting in level adjustments of single words. Nevertheless, these adjustments did not entirely equalize word intelligibility, resulting in a better speech recognition of the word category name. This effect might be reduced for the synthetic speech because of the rather unfamiliar pronunciation of proper names in the TTS system. Here, instead, numerals had the highest intelligibility. Regarding the distributions of word recognition for each word type, variance homogeneity between the two speech types was only found for numerals and adjectives. Nevertheless, word-by-word comparisons of variance did not provide reasons for a large mismatch or unbalance. Despite the diverse representations of each word included in the synthetic speech (as stated in the “Methods” section), the speech production of the TTS system was at least as homogeneous as natural speech.

The findings of this study suggest that in this particular setup, synthetic speech could replace recordings of natural speakers for testing speech recognition in noise. This was evaluated with normal-hearing listeners, using one specific set of speech material and one specific TTS system. In this study, participants were chosen to be homogeneous in age and hearing loss according to the standard (ISO 8253-3). As supra-threshold auditory processes and cognitive abilities might also influence SRT results, future evaluations dealing with effects of (age-specific) deterioration of sensory processing on TTS signal perception is necessary for application in heterogeneous participant groups. Furthermore, it cannot be ruled out that other speech material or TTS system would lead to different outcomes. Prior to the experiments described here, the TTS system was chosen to synthesize a female speech for the speech recognition tests. This decision was based on a better comparability to international matrix tests that are mostly spoken by females (Kollmeier et al., 2015). Bearing in mind that the gender of the synthetic speech might affect the recognition and perceived naturalness (Stevens, Lees, Vonwiller, & Burnham, 2005), speech recognition might be different for male speech. In addition, the selection of the TTS system was based on the subjective evaluation of two expert listeners in a first step and of 12 naïve participants in a second step. In both processes, synthetic speech samples were assessed in quiet, while the speech recognition tests took place in noise. It is unknown whether the selection would have differed if the evaluation had been carried out for noisy speech samples. Furthermore, an expert listener subjectively chose the individual speech snippets supplied by the TTS system building the final test sentences. This procedure was much less effortful than the recording, postprocessing, and optimization of matrix tests with natural speech. In this study, one expert (second author) spent approximately 4 hours choosing the most natural-sounding sentence representation in the synthesizing software and postprocess the wav files (see earlier, “Procedure for Synthesis”) for use in the speech test. For natural speech, the recording and editing processes are more effortful, and in particular the optimization requires time-consuming measurements with participants (typically \(n = 12\); Hochmuth et al., 2012; Zokoll et al., 2015).

Here, the matrix test was chosen as a first approach to examine the usability of synthetic speech in speech recognition tests. This type of test is well documented, and the original speech corpus could be included in the study as a baseline with reasonable effort. As matrix tests are optimized to require less recording time than other
speech tests using more natural speech material (e.g., Hearing In Noise Test [HINT], Nilsson, Soli, & Sullivan, 1994; Speech Perception in Noise [SPIN] test Kalikow, Stevens, & Elliott, 1977), the reduction of effort would be even higher for these tests. However, the effort-reduction potential of synthetic speech is limited to the production of the speech material. The effort for evaluating the speech material for equivalent intelligibility of test lists cannot be reduced. For traditional speech recognition tests, the detailed evaluation and standardization is crucial for usage in clinical settings. In research requiring reproducible speech material without the need for standardization, the use of synthetic speech material might be even more efficient. Especially in virtual listening conditions, that is, mimicking real-world situations, higher variability of speech and large speech corpora could be preferable, depending on the research question. Nevertheless, it would be desirable to choose the optimal synthetic word representation for the particular sentence objectively in future applications. In theory, automatic speech recognition systems could provide this information. These systems were recently applied in the objective evaluation of synthetic speech (Chang, 2011; Tang, Cooke, & Valentini-Botinhao, 2016) as well as in the prediction of speech recognition test outcomes (Hülsmeyer, Buhl, Wardenga, Warzybok, & Schädler, 2017; Schädler, Warzybok, Hochmuth, & Kollmeier, 2015). However, the state of the art in automatic speech recognition and objective evaluation is currently unable to determine naturalness as precisely as a human expert listener. More research is necessary to provide an automatic procedure for the synthetization of appropriate speech signals applicable in speech recognition measurements.

Conclusion

In summary, it was shown that synthetic speech can be applied in speech audiometry. The results indicate that thresholds for synthetic speech are similar to natural speech if speaker properties are matched. In this study, SRTs were only 0.5 dB SNR higher for synthetic speech compared to the reference test, which has no relevance in practice. Synthetic speech reduced the development time considerably, assuming that the TTS system had already been selected. Replicating earlier findings in natural speech matrix tests, 40 to 60 sentences (2–3 lists) of training are recommended when using synthetic speech, based on the acquired measurement accuracy.

Acknowledgment

The authors would like to thank Jule Pohlhausen for the review and preselection of available TTS systems. English language support was provided by http://www.stels-ol.de/.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was supported by the governmental funding initiative Niedersächsisches Vorab of the Lower Saxony Ministry for Science and Culture, research focus Hören im Alltag Oldenburg (HALLO) and by the European Regional Development Fund (ERDF-Project Innovation network for integrated, binaural hearing system technology [VIBHear]), together with funds from the State of Lower Saxony.

ORCID iD

Theresa Nuesse https://orcid.org/0000-0002-6202-8952

References

Ahrlich, M. (2013). Optimierung und Evaluation des Oldenburger Satztests mit weiblicher Sprecherin und Untersuchung des Effekts des Sprechers auf die Sprachverständlichkeit [The influence of different background noise situations on speech intelligibility in different languages] (Bachelor thesis). Carl von Ossietzky Universität Oldenburg, Oldenburg.

Benoit, C., Grice, M., & Hazan, V. (1996). The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. Speech Communication, 18, 381–392. doi:10.1016/0167-6393(96)00026-X

Brand, T., & Kollmeier, B. (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. The Journal of the Acoustical Society of America, 111(6), 2801–2810. doi:10.1121/1.1479152

Chang, Y.-Y. (2011). Evaluation of TTS systems in intelligibility and comprehension tasks. In Proceedings of the 23rd Conference on Computational Linguistics and Speech Processing (pp. 64–78). Taipei, Taiwan: The Association for Computational Linguistics and Chinese Language Processing.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Govender, A., & King, S. (2018a). Measuring the cognitive load of synthetic speech using a dual task paradigm. In Proceedings of the 19th Annual Conference of the International Speech Communication Association (ISCA) (pp. 2843–2847). Baixas, France: International Speech Communication Association. https://doi.org/10.21437/Interspeech.2018-1199

Govender, A., & King, S. (2018b). Using pupillometry to measure the cognitive load of synthetic speech. In Proceedings of
the 19th Annual Conference of the International Speech Communication Association (ISCA) (pp. 2838–2842). Baixas, France: International Speech Communication Association. doi:10.21437/Interspeech.2018-1174

Greene, B. G., Logan, J. S., & Pisoni, D. B. (1986). Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. *Behaviour Research Methods, Instruments, & Computers, 18*(2), 100–107. doi:10.3758/BF03201008

Hagerman, B. (1982). Sentences for testing speech intelligibility in noise. *Scandinavian Audiology, 11*(2), 79–87. doi:10.3109/01050398209076203

Hazan, V., & Shi, B. (1993). Individual variability in the perception of synthetic speech. *Third European Conference on Speech Communication and Technology, 3*, 1849–1852.

Hinterleitner, F., Norrenbrock, C. R., & Möller, S. (2013, August 31–September 2). Is intelligibility still the main problem? A review of perceptual quality dimensions of synthetic speech. In *Eighth ISCA Workshop on Speech Synthesis*, Barcelona, Catalonia, Spain.

Hochmuth, S., Brand, T., Zokoll, M. A., Castro, F. Z., Wardenga, N., & Kollmeier, B. (2012). A Spanish matrix sentence test for assessing speech reception thresholds in noise. *International Journal of Audiology, 51*(7), 536–544. doi:10.3109/14992027.2012.670731

Holube, I., Fredelake, S., Vlaming, M., & Kollmeier, B. (2010). Development and analysis of an international speech test signal (ISTS). *International Journal of Audiology, 49*(12), 891–903. doi:10.3109/14992027.2010.506899

Hülsmeyer, D., Buhl, M., Wardenga, N., Warzybok, A., & Schädl, M. R. (2017, August 19). Do models hear the noise? Predicting the outcome of the German matrix sentence test for subjects with normal and impaired hearing. In *Proceedings of International Conference on Challenges in Hearing Assistive Technology (CHAT) 2017*, Stockholm, Sweden.

Isaac, K. B. (2015). The intelligibility of synthetic speech in noise and reverberation (PhD thesis). The University of Edinburgh, Edinburgh. Retrieved from http://hdl.handle.net/1842/15870

ISO 8253-3:2012. Acoustics – Audiometric test methods – Part 3: Speech audiometry. German version EN ISO 8253-3:2012. Berlin, Germany: Beuth Verlag.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America, 61*(5), 1337–1351. doi:10.1121/1.381436

King, S. (2014). Measuring a decade of progress in text-to-speech. *Loquens, 1*(1), e006. doi:10.3989/loquens.2014.006

Knohlauh, K., & Maloney, L. T. (2012). Modeling psycho-physical data in R Use R! New York, NY: Springer. Retrieved from http://www.loc.gov/catdir/enhancements/fy1306/2012943991-b.html

Kollmeier, B. (1990). Messmethodik, Modellierung und Verbesserung der Verständlichkeit von Sprache (Habilitation). Georg August Universität Göttingen, Göttingen.

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V. N., Brand, T., & Wagener, K. C. (2015). The multilingual matrix test: Principles, applications, and comparison across languages: A review. *International Journal of Audiology, 54*(sup2), 3–16. doi:10.3109/14992027.2015.1020971

Kollmeier, B., & Wesselkamp, M. (1997). Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *The Journal of the Acoustical Society of America, 102*(4), 2412–2421. doi:10.1121/1.419624

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America, 95*(2), 1085–1099. doi:10.1121/1.408469

Puglisi, G. E., Warzybok, A., Hochmuth, S., Visentin, C., Astolfi, A., Prodi, N., & Kollmeier, B. (2015). An Italian matrix sentence test for the evaluation of speech intelligibility in noise. *International Journal of Audiology, 54*(sup2), 44–50. doi:10.3109/14992027.2015.1061709

Schädl, M. R., Warzybok, A., Hochmuth, S., & Kollmeier, B. (2015). Matrix sentence intelligibility prediction using an automatic speech recognition system. *International Journal of Audiology, 54*(sup2), 100–107. doi:10.3109/14992027.2015.1061708

Schuetzer, A., Lemke, U., Kollmeier, B., & Holube, I. (2016). Normal and Time-Compressed Speech: How Does Learning Affect Speech Recognition Thresholds in Noise? *Trends in Hearing, 20*, 1–13. doi:10.1177/2331216516669898

Studebaker, G. A. (1985). “A rationalized” arcsine transform. *Journal of Speech, Language, and Hearing Research, 28*(3), 455–462. doi:10.1044/jshr.2803.455

Tang, Y., Cooke, M., & Valentini-Botinhao, C. (2016). Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech. *Computer Speech & Language, 35*, 73–92. doi:10.1016/j.csl.2015.06.002

Uslar, V. N., Carroll, R., Hanke, M., Hamann, C., Ruijendijk, E., Brand, T., & Kollmeier, B. (2013). Development and evaluation of a linguistically and audiologically controlled sentence intelligibility test. *The Journal of the Acoustical Society of America, 134*(4), 3039–3056. doi:10.1121/1.4818760

Wagener, K. C., Brand, T., & Kollmeier, B. (1999). Entwicklung und Evaluation eines Satztests für die deutsche Sprache II: Optimierung des Oldenburger Satztests [Development and evaluation of a sentence test for the German language II: Optimization of the Oldenburg sentence test]. *Zeitschrift für Audiologie/Audiological Acoustics, 38*(2), 44–56.

Wagener, K. C., Brand, T., & Kollmeier, B. (1999b). Entwicklung und Evaluation eines Satztests für die deutsche Sprache III: Evaluation des Oldenburger Satztests [Development and evaluation of a sentence test for the German language III: Evaluation of the Oldenburg sentence test]. *Zeitschrift für Audiologie/Audiological Acoustics, 38*(3), 86–95.

Wagener, K. C., Hochmuth, S., Ahrlisch, M., Zokoll, M. A., & Kollmeier, B. (2014). Der weibliche Oldenburger Satztest
Wagener, K. C., Kühnel, V., & Kollmeier, B. (1999). Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests [Development and evaluation of a sentence test for the German language I: Design of the Oldenburg sentence test]. Zeitschrift für Audiologie/Audiological Acoustics, 38(1), 4–15.

Warzybok, A., Zokoll, M. A., Wardenga, N., Ozimek, E., Boboshko, M., & Kollmeier, B. (2015). Development of the Russian matrix sentence test. International Journal of Audiology, 54(sup2), 35–43. doi:10.3109/14992027.2015.1020969

Zokoll, M. A., Fidan, D., Türkyılmaz, D., Hochmuth, S., Ergenç, I., Sennaroğlu, G., & Kollmeier, B. (2015). Development and evaluation of the Turkish matrix sentence test. International Journal of Audiology, 54(sup2), 51–61. doi:10.3109/14992027.2015.1074735.