


# Speech Recognition and Listening Effort of Meaningful Sentences Using Synthetic Speech

Trends in Hearing  
Volume 26: 1–14  
© The Author(s) 2022  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/23312165221130656  
journals.sagepub.com/home/tia  


Saskia Ibelings<sup>1,2,3</sup> , Thomas Brand<sup>2,3</sup> and Inga Holube<sup>1,3</sup>

## Abstract

Speech-recognition tests are an important component of audiology. However, the development of such tests can be time consuming. The aim of this study was to investigate whether a Text-To-Speech (TTS) system can reduce the cost of development, and whether comparable results can be achieved in terms of speech recognition and listening effort. For this, the everyday sentences of the German Göttingen sentence test were synthesized for both a female and a male speaker using a TTS system. In a preliminary study, this system was rated as good, but worse than the natural reference. Due to the Covid-19 pandemic, the measurements took place online. Each set of speech material was presented at three fixed signal-to-noise ratios. The participants' responses were recorded and analyzed offline. Compared to the natural speech, the adjusted psychometric functions for the synthetic speech, independent of the speaker, resulted in an improvement of the speech-recognition threshold (SRT) by approximately 1.2 dB. The slopes, which were independent of the speaker, were about 15 percentage points per dB. The time periods between the end of the stimulus presentation and the beginning of the verbal response (verbal response time) were comparable for all speakers, suggesting no difference in listening effort. The SRT values obtained in the online measurement for the natural speech were comparable to published data. In summary, the time and effort for the development of speech-recognition tests may be significantly reduced by using a TTS system. This finding provides the opportunity to develop new speech tests with a large amount of speech material.

## Keywords

synthetic speech, speech recognition, speech quality, listening effort, audiology

Received 4 May 2022; Revised 1 September 2022; accepted 16 September 2022

## Introduction

Speech-recognition tests are used not only in the clinical diagnosis of hearing impairment, but also in the evaluation of hearing systems such as hearing aids. In speech-recognition tests, the task is to repeat the recognized words. It is possible to perform tests in quiet at different sound pressure levels (SPL), or in noise at different signal-to-noise ratios (SNR). Measurements in noise yield a psychometric function describing a speech-recognition score as a function of SNR. The SNR of a specific speech-recognition level, often 50%, is called the speech-recognition threshold (SRT). Currently used speech recognition-tests have the common feature that they were recorded using real (natural) speakers. For the recordings themselves, not only are efforts in terms of time and technical knowledge needed, but also professional equipment. The question

arises whether the complex process of developing speech tests can be simplified using synthetic speech.

Matrix tests, for example in German the Oldenburg sentence test (OLSA; Wagener et al., 1999), consist of 50 well-known words, and all sentences have the same grammatical structure. To generate the OLSA, 100 sentences were

<sup>1</sup>Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, Oldenburg, Germany

<sup>2</sup>Medizinische Physik, Universität Oldenburg, Oldenburg, Germany

<sup>3</sup>Cluster of Excellence Hearing4All, Oldenburg, Germany

### Corresponding Author:

Saskia Ibelings, Institute of Hearing Technology and Audiology, Jade University of Applied Sciences, Ofener Str. 16/19, D-26121 Oldenburg, Germany.  
Email: saskia.ibelings@jade-hs.de



recorded so that all possible word transitions were considered. The recordings took place in a sound attenuated booth. The sentences were cut into segments, which include every single word with its specific co-articulation at the end. These words were concatenated into new combinations to generate new sentences. To achieve homogeneous word intelligibility, level adjustments were necessary (Kollmeier et al., 2015).

For the German Göttingen Sentence Test (GÖSA; Kollmeier & Wesselkamp, 1997), which consists of everyday sentences, psychometric functions for each sentence and weighting factors for the individual words of a sentence were measured. Based on the results, level corrections were applied to reduce inhomogeneities. The GÖSA finally consists of ten test lists of 20 sentences each. These include not only declarative sentences, but also exclamations and questions (Kollmeier & Wesselkamp, 1997). The different test lists should be presented only once within a reasonable time period, because participants might remember the sentences and speech-recognition scores might thus increase (Yund & Woods, 2010). However, the relatively small number of ten test lists limits the test's applicability in research and in clinical care, e.g., for regular repetition in cochlear implant validation. Therefore, a sentence test with many more lists would be desirable. However, existing sentence tests with natural speakers cannot easily be extended, because the voice's characteristics are not constant over a larger age range (Schötz, 2007), the manner of speaking (e.g., speech rate; Schlueter et al., 2014) might not be replicated, and technical equipment might differ, which may result in differences in speech-recognition scores.

To simplify the process of speech-test development, it is possible to use Text-To-Speech (TTS) systems. Current, TTS systems are not only less expensive, but can also reduce the effort by saving some optimization steps. Furthermore, there is then no need to hire a speaker or purchase recording equipment. TTS systems also offer the advantage that any amount of material can be post-produced. Nuesse et al. (2019) already showed that a TTS system is suitable for generating the sentences of OLSA using a female speaker (Wagener et al., 2014). A comparison between the natural speaker and the synthesized speaker using the voice "Claudia" (Acapela Group, Mons, Belgium), revealed an SRT difference of only 0.5 dB and comparable slopes (Nuesse et al., 2019). Another study using the German Freiburg monosyllabic test also found that natural and synthetic speech resulted in comparable SRTs and slopes (Schwarz et al., 2022).

Although speech recognition is comparable, listening effort might be influenced by synthetic speech. An explanation of listening effort is given by the "Ease of Language Understanding" model (ELU) by Roennberg et al. (2013). Speech is the input signal to the listener's intermediate memory. The information contained in the speech regarding phonology, syntax, semantics, and prosody is automatically

compared with representations from the listener's long-term memory. If the information matches, recognition of speech is easy. Due to hearing loss or complex environmental noise, however, recognizing speech may be hampered due to a mismatch of information. Synthetic speech may have an impact on this recognition process, in which case additional mental resources in the form of further processing steps and conscious and active processes are needed to recognize the speech. These processes include using context effects in sentence recognition. Listening effort is thus the increased demand on mental resources needed to identify the speech that is not understood (Roennberg et al., 2013).

Both objective and subjective methods have been used to determine listening effort (Klink et al., 2012a, 2012b; McGarrigle et al., 2014; Pichora-Fuller et al., 2016). Subjective methods include questionnaires and scales (Krueger et al., 2017). Brain activity, as an objective measure, is examined using electroencephalography (EEG; Obleser et al., 2012). Other objective methods include physiological and cognitive measures, such as pupil size (Koelewijn et al., 2015), heart activity (Mackersie & Calderon-Moultrie, 2016), or skin conductance (Holube et al., 2016). Simantiraki et al. (2018), as well as Govender and King (2018), used pupillometry to measure listening effort for synthetic speech, and showed that synthetic speech generated with an HMM (Hidden-Markov Model)-based system led to larger pupil dilations than natural speech, indicating an increase in listening effort. It is worth noting that pupil size also showed a relationship with the quality of the TTS systems. Synthetic sentences that were rated higher in quality by the participants resulted in smaller changes in pupil size and reduced listening effort (Govender & King, 2018).

Another method to gauge listening effort is to measure the Verbal Response Time (VRT; Houben et al., 2013; Meister et al., 2018; Pals et al., 2015; Visentin et al., 2021). VRT describes the time delay between the end of the stimulus presentation and the beginning of the response of the participant. According to Pals et al. (2015) the VRT is a good indicator for listening effort, but it is still unclear, whether it directly measures the listening effort itself or another dimension of listening effort. Both the presence of noise and poorer SNR (Houben et al., 2013; Meister et al., 2018; Visentin et al., 2021) led to higher VRT values.

In the last few years, TTS systems have been continuously improved, resulting in the assumption that synthetic speech closely matches natural speech (King, 2014). Older systems often use Unit Selection (King, 2014; Taylor, 2006), in which speech is stored in a library. To find the right segment, the written text is decomposed into phonetic units. Before concatenating the segments, the selected segments are adjusted in duration, intensity, or even frequency. Nuesse et al. (2019) used such a system (Virtual Speaker, Acapela, Mons, Belgium) and found comparable speech-recognition scores for natural and synthetic speech. However,

disadvantages of Unit Selection are both the diminished fluency of the sound and the high storage load (Taylor, 2006). Statistical models based on HMMs (Taylor, 2006) can, however, remedy these problems. The automated training based on many representative speech materials using statistics makes the model more robust. In addition, the use of features like Mel-Frequency Cepstral Coefficients (MFCC) can reduce storage requirements (Taylor, 2006).

According to the Blizzard Challenge (King, 2014), statistical models result in an improved intelligibility compared to systems with Unit Selection. Modern systems often use Deep Neural Networks (DNN; Zen et al., 2013). DNN-based systems are not only rated as sounding more natural, but an objective improvement over HMM systems was also observed, e.g., a lower error rate for voiced and unvoiced utterances (Zen et al., 2013). One example of a DNN-based system is Acapela Cloud (Acapela Group, Solna, Sweden). In HMM- and DNN-based systems, the parameters obtained are used as the input of a parametric synthesizer or vocoder to generate audio signals. A vocoder uses a source-filter model of speech, i.e., a source signal (noise or pulse train) is passed through a filter representing the human vocal tract (Bunnell, 2022; Zen et al., 2013). Although the Wavenet technology (Shen et al., 2018) is also based on neural networks, it models the raw waveform of the audio signals sample by sample instead of using a vocoder.

The overarching aim of the current study was to examine whether TTS systems can be usefully applied to speech-recognition tests using everyday sentences. In the first experiment, the qualities of three different TTS systems were compared in different dimensions. Using the best-rated TTS system, the sentences of the GÖSA were synthesized for the second experiment for a male ( $TTS_{\text{male}}$ ) and a female speaker ( $TTS_{\text{female}}$ ).  $TTS_{\text{male}}$  was used to allow comparisons with the natural male speaker of the original GÖSA.  $TTS_{\text{female}}$  was used because many international speech-recognition tests were recorded using female speakers (Kollmeier et al., 2015). Both  $TTS_{\text{male}}$  and  $TTS_{\text{female}}$  were modified to match as far as possible the speech rate of the original material. The following listening tests were performed by normal-hearing participants. Speech-recognition scores and VRT values were compared for synthetic and natural speech. Overall, the research questions were:

1. Is it possible to reduce the effort required for generating speech test material of everyday sentences by using a TTS system? This would be expected, since the production time of the synthetic speech material is shorter than for natural speech, because optimization steps such as level adjustments may no longer be required due to the more consistent properties of the synthetic speech.
2. Are the VRT values for synthetic speech and natural speech comparable?
3. Are the speech-recognition scores for synthetic speech comparable to those of natural speech? It was expected

that the differences would be in the range of those for different natural speakers, i.e., up to 5 dB (Hochmuth et al., 2015).

## Experiment I: Comparison of TTS Systems

### Methods

**TTS Systems.** To generate speech material using a synthetic voice, a suitable TTS system had to be chosen among the large number of TTS systems found using online research. The selection criteria were the availability of a male and a female speaker, and a subjective sound quality for the German language, as rated by the first author of this contribution. Additionally, the conditions of use were reviewed, especially in terms of their usability for a publicly available speech test and of a guarantee of the stability of the voices over a longer time period, for adding speech material in the future. None of the freely available TTS systems had sufficient sound quality for the German language without further training needs. One of the potential candidates was the commercial Virtual Speaker by the Acapela Group (Mons, Belgium) used in Nuesse et al. (2019). Since 2021, the Acapela Group (Solna, Sweden) offers a Cloud Service with modified speech synthesis, which was also preselected. Although Google Wavenet (Dublin, Ireland) does not guarantee the stability of the voices, it was preselected because of its application for synthesized digits in the digit-in-noise test (Polspoel et al., 2020). Table 1 gives an overview of the three preselected TTS systems, which differ in their signal-generation processes and costs.

**Stimuli.** Based on Nuesse et al. (2019), 12 different sentences were chosen as stimuli. These were taken from the following speech-recognition tests:

- 6 everyday sentences from the GÖSA (Kollmeier & Wesselkamp, 1997)
- 3 sentences without semantic context from the OLSA (Wagener et al., 2014)
- 3 sentences with low semantic context from the Oldenburg Linguistically and Audiologically Controlled Sentences corpus (OLACS, Uslar et al., 2013)

An excerpt from “North wind and sun” (Holube et al., 2010) spoken by a female speaker was also presented. All sentences were generated using the three different TTS systems for both speakers. The sentences generated were not optimized, but calibrated to the same digital root-mean-square (RMS) level.

**Measurement Process.** Systematic subjective ratings of the different systems and speakers were carried out using a MUSHRA test (MULTI Stimulus test with Hidden Reference and Anchor; ITU-R BS.1534-3, 2015). The quality dimensions evaluated were naturalness, prosody (stress and intonation), and speech flow in combination with subjective

**Table 1.** Preselected TTS Systems.

TTS system	Signal generation	Voice (female)	Voice (male)	Costs	Abbreviation
Virtual Speaker (Acapela Group, Mons, Belgium)	Unit Selection	Claudia	Klaus	1.500 € for 5h	Acapela <sub>US</sub>
Acapela Cloud Service (Acapela Group, Solna, Sweden)	Deep Neural Network	Claudia	Klaus	1.500 € for 75 min	Acapela <sub>DNN</sub>
Google (Dublin, Ireland)	Wavenet	de-DE-Wavenet-F	de-DE-Wavenet-B	1 million characters free per month, otherwise 16 \$ per 1 million characters	Google <sub>wav</sub>

intelligibility (Hinterleitner et al., 2013). Due to the Covid-19 pandemic, measurements took place online using webMUSHRA (Schoeffler et al., 2018). Based on ITU-R BS.1534-3 (2015), the original material was presented as a reference and also named as such, so that the reference was known to the participants. In addition to the synthesized speech, a hidden reference and an anchor were also presented to the participants. According to ITU-R BS.1534-3 (2015) the anchor was derived from the reference by filtering using a low-pass filter with a cut-off frequency of 3.5 kHz. The order of the sentences, TTS systems, and speakers were randomized. The participants could switch between the hidden reference, the anchor, and the synthetic speech, and they could listen to them as often as necessary. The participants' task was to compare these stimuli and rate them on a scale, that included values from 0 (very poor) to 100 (excellent). After completing the MUSHRA tests, participants were asked to sort the TTS systems for each speaker by preference (1 = favourite, 3 = last).

**Participants.** The quality of the speech was evaluated by 14 participants from 22 to 60 years of age (median age: 25.0 years, eleven females, three males). The participants were students and employees recruited via the mailing list of Jade University of Applied Sciences in Oldenburg. According to their own assessment, they had normal hearing. The experiment was approved by the ethics committee (Kommission für Forschungsfolgenabschätzung und Ethik) of Carl von Ossietzky University in Oldenburg, Germany (Drs. EK/2021/063).

**Analysis and Statistics.** The evaluation used Matlab 2020a (The MathWorks, Inc., Natick, Massachusetts) and SPSS 27 (IBM Corp., Armonk, New York). According to ITU-R BS.1534 3, two participants were excluded from the analysis. One participant rated more than 15% of the reference sentences worse than 90; the second scored the anchor greater than 90 in more than 15% of cases. Hence in total, data from 12 participants were used for the statistical analysis. Shapiro-Wilk tests revealed that the data of only one condition (naturalness for Acapela<sub>US</sub> with male speaker) deviated

from a normal distribution ( $p = .016$ ). Hence, for each quality dimension, repeated-measures analysis of variance (ANOVA) was performed with the within-subject factors TTS system and speaker (male/female). Post hoc tests were t-tests for paired samples with Bonferroni correction ( $\alpha = .0167$ ).

## Results

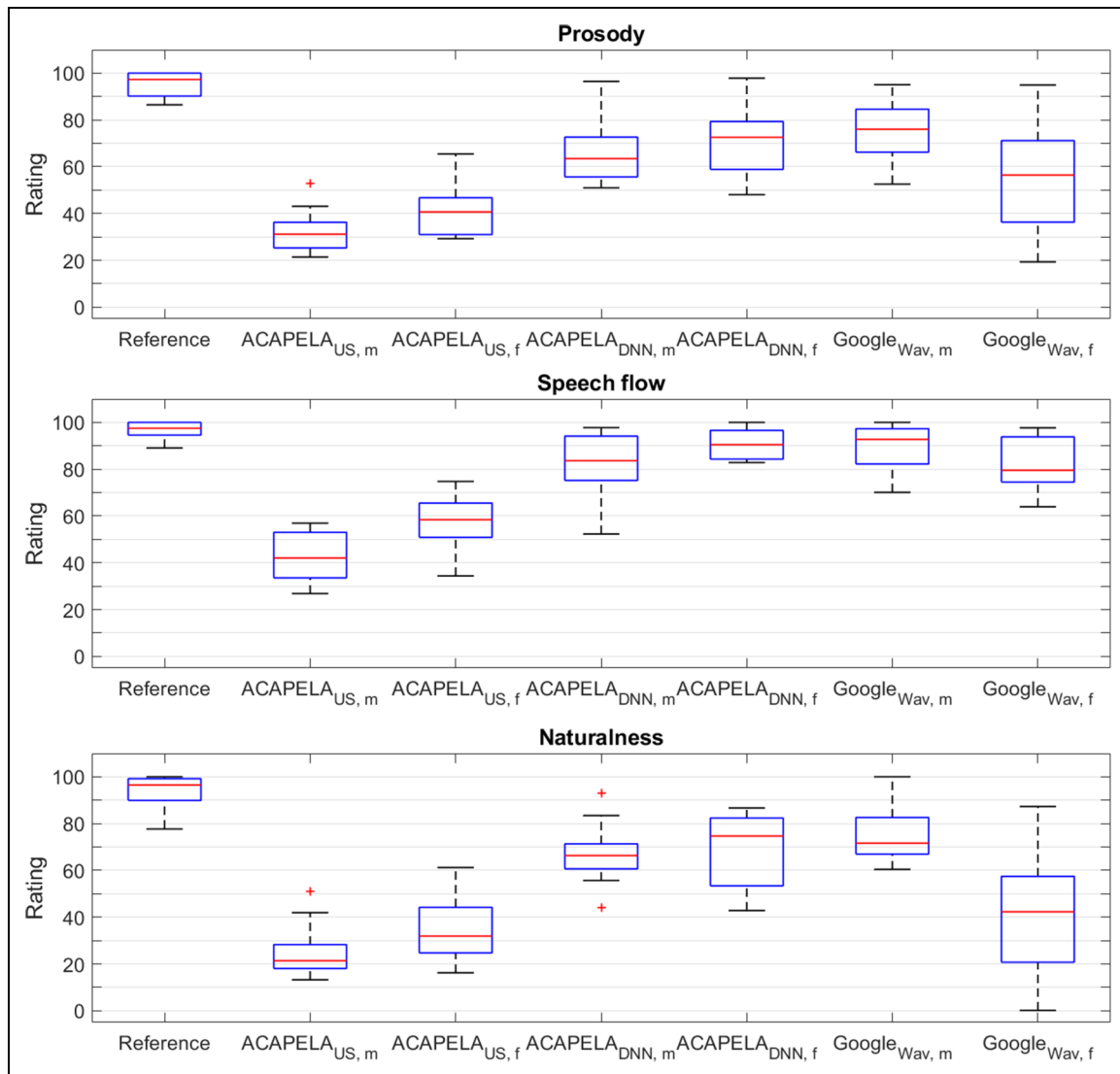
Figure 1 shows the ratings for the various systems in the quality dimensions prosody, speech flow, and naturalness. In addition, the results for the original speech material are shown.

The original material was rated best in all dimensions. Acapela<sub>US</sub> was rated worst in all dimensions, with a trend towards the female speaker being rated better than the male speaker. Acapela<sub>DNN</sub> and Google<sub>wav</sub> produced similar results. However, the participants rated the female Google-speaker worse than the male speaker.

The repeated-measures ANOVA for the quality dimension prosody revealed that the TTS system had a significant effect on prosody ratings [ $F(2, 22) = 115.7, p < .001$ ]. The speaker type (male or female) had no significant effect [ $F(1, 11) = 88.5, p = .529$ ]. The interaction of the two factors was also significant [ $F(2, 22) = 17.1, p < .001$ ]. The post-hoc tests performed with Bonferroni correction showed that Acapela<sub>US</sub> was significantly different from Acapela<sub>DNN</sub> and Google<sub>wav</sub> ( $p < .001$ ), whereas Google<sub>wav</sub> and Acapela<sub>DNN</sub> were not significantly different in prosody ( $p = .242$ ).

Significant effects of TTS system [Greenhouse-Geisser  $\epsilon = 0.661$ ;  $F(1.321, 14.534) = 119.0, p < .001$ ] and speaker [ $F(1, 11) = 10.2, p = .009$ ] on speech-flow ratings were found. Their interaction also proved to be significant [ $F(2, 22) = 18.5, p < .001$ ]. Post-hoc tests indicated that Acapela<sub>US</sub> was significantly different from the other systems ( $p < .001$ ), whereas no significant difference was found between Google<sub>wav</sub> and Acapela<sub>DNN</sub> ( $p > .05$ ).

For naturalness, the ANOVA showed a significant effect of the TTS system on the ratings [Greenhouse-Geisser  $\epsilon = 0.656$ ;  $F(1.312; 14.434) = 105.1, p < .001$ ], but no effect of the speaker type [ $F(1, 11) = 2.97, p = .113$ ]. In addition, there was a significant interaction [ $F(2, 22) = 31.8, p < .001$ ]. The post-hoc tests performed with



**Figure 1.** Ratings using the MUSHA procedure of the quality dimensions prosody (top panel), speech flow (middle panel), and naturalness (bottom panel) for the different TTS systems and speakers (m= male, female=f), evaluated by 12 participants. The TTS systems are abbreviated as in Table 1.

Bonferroni correction showed that the ratings of all TTS systems differed significantly from each other ( $p < .0167$  in each case).

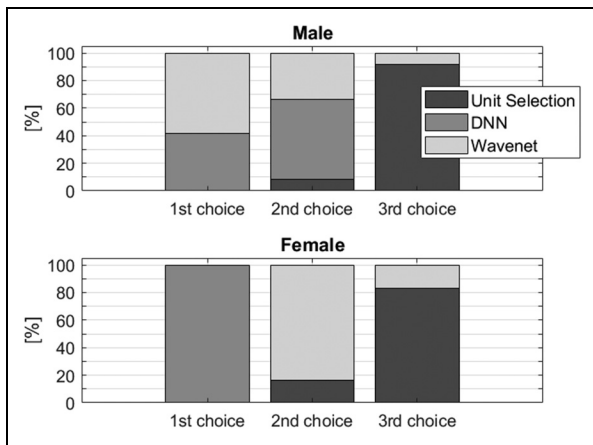
The overall impression given by the participants supports the previously described results (see Figure 2). Among female speakers, ACAPELA<sub>DNN</sub> was the first choice of all participants. For the male speaker, five participants preferred ACAPELA<sub>DNN</sub>, seven Google<sub>Wav</sub>. For most of the participants ACAPELA<sub>US</sub> was the last choice.

### Discussion and Conclusion

Three different TTS systems using different synthesis methods and voices were evaluated. One of the TTS systems, Acapela<sub>US</sub>, outperformed two other systems in

Nuesse et al. (2019). In contrast, in the current study, Acapela<sub>US</sub> was always rated as the worst. That the two other systems outperformed Acapela<sub>US</sub> is presumably related to the advancements in TTS systems and their functionalities over the past few years. This result is in line with Zen et al. (2013), who found that the DNN systems they tested produced better results than other systems. Nuesse et al. (2019) found a similar difference between the subjective ratings for Acapela<sub>US</sub> compared to the original material, as found in the current study for the synthetic speech material.

The two TTS systems Acapela<sub>DNN</sub> and Google<sub>Wav</sub> were often rated as similar, but for naturalness and overall impression, Acapela<sub>DNN</sub> outperformed Google<sub>Wav</sub>. One dimension with similar ratings was *fluency and intelligibility*, based on



**Figure 2.** Overall impression of the three different TTS systems for the male and female speaker. The participants ( $N = 12$ ) had to rank the systems in descending order according to their overall impression.

Hinterleitner et al. (2013). It should be noted that one participant reported difficulties when scoring these two dimensions together. In general, the explanation of the procedures to the participants was limited to the written instructions in the online study. Further explanations, easily and often informally given in lab studies, were not possible. Nevertheless, no influence on the results was detected. The results are consistent in that the natural speech achieved a very high score in all dimensions.

Although Acapala<sub>DNN</sub> was rated as good, the ratings differ from those for natural speech. One possible explanation is that due to the measurement setup the participants knew the reference. Therefore, it was possible to recognize the reference within the presented stimuli and to evaluate it as the most natural. This aspect reveals a problem of using the MUSHRA test in this application. The MUSHRA test is usually used to evaluate intermediate quality differences of audio systems when processing the same source signal. In this study, however, although the same sentence was always used for each condition, these sentences were generated in different ways and differed in their speakers. Therefore, in this study the MUSHRA test only evaluates the different TTS systems against each other, and a comparison to the natural reference does not seem to be appropriate using this setup. However, future developments of TTS systems could possibly lead to natural speech being outperformed by synthetic speech in the assessed dimensions, especially if natural speech is recorded using an untrained speaker.

It remains unclear whether the ratings were influenced by the selected stimuli. It is conceivable that acoustic differences are less obvious in sentences with a high context and that the TTS systems are therefore rated better than in sentences with lower context. Since both meaningful everyday sentences (GÖSA) and low context sentences like OLSA and

OLACS were presented, the ratings could also be analyzed separately for each sentence group. However, the data is limited to 12 participants and only shows deviating trends in the comparisons of ratings for the two sentence groups synthesized with the different TTS systems. Hence, a detailed analysis of the hypothesis should be addressed in a future study. Overall, because Acapala<sub>DNN</sub> yielded ratings (nearly) as good as the original material, it was concluded that Acapala<sub>DNN</sub> is a good choice for synthesizing a new version of the GÖSA.

## Experiment 2: Generating the Göttingen Sentences Using Synthetic Speech

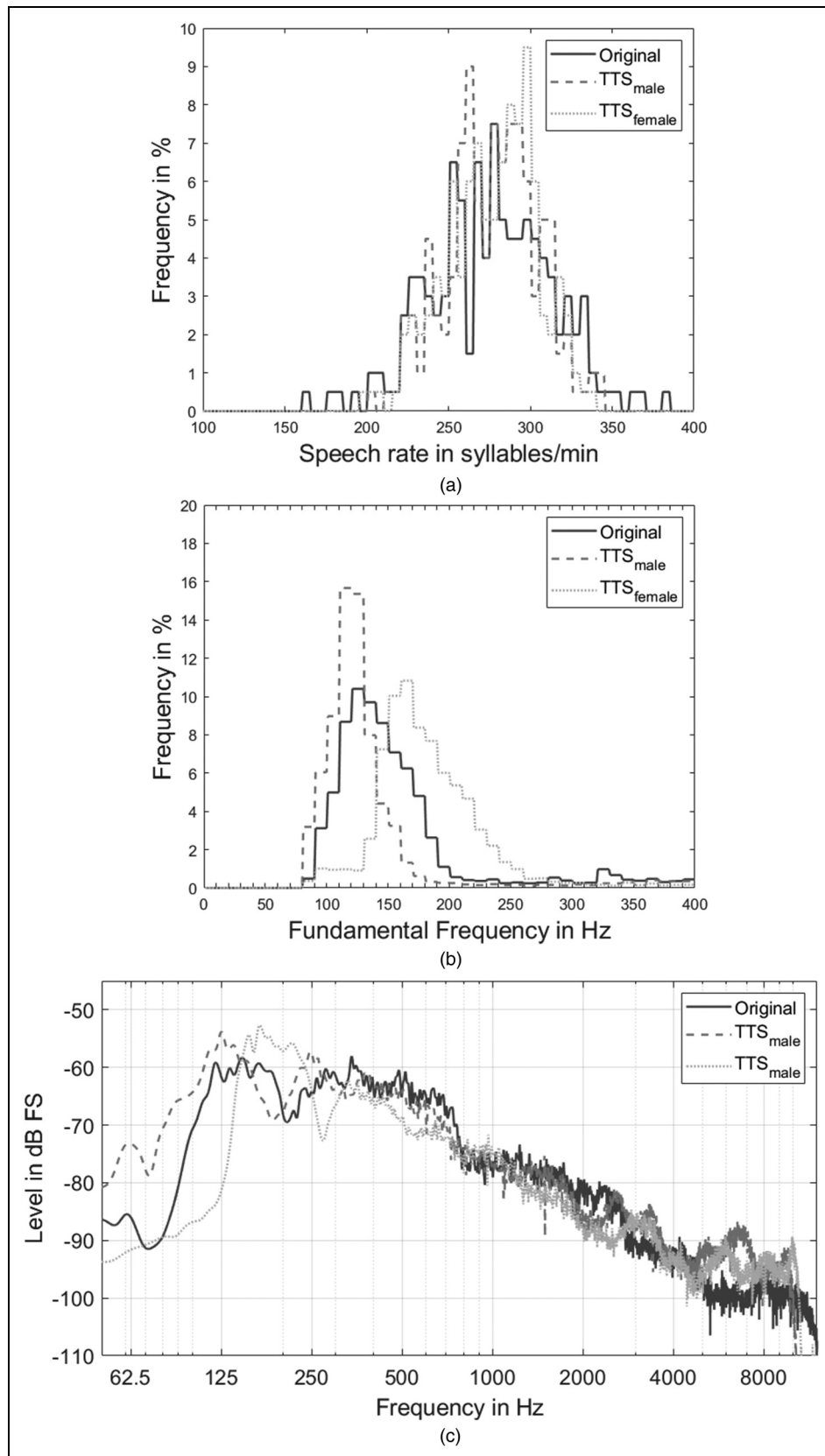
Based on the previous results, the GÖSA sentences were generated using the TTS system Acapala<sub>DNN</sub>. Both speech recognition and VRT were measured for natural and synthetic speech.

### Methods

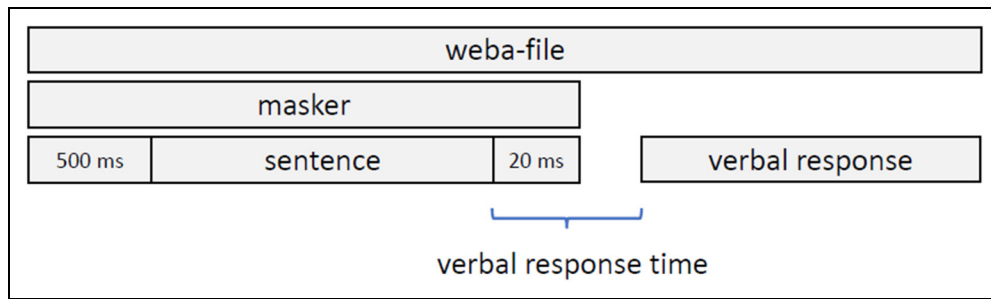
**Procedure for Synthesis.** The speech material of the GÖSA was resynthesized using the Acapela Cloud Service (Acapela Group, Solna, Sweden, <https://www.acapela-cloud.com>, accessed 8/27/2021). Each of the 200 sentences were generated with the German voices Klaus and Claudia. The sentences were entered into the online software and the sampling frequency set to 44.1 kHz.

**Speech Characteristics.** The speaking rate of the sentences was adapted to that of the original recordings. To allow direct comparisons, all audio files were cut directly before the beginning of each sentence and after the end of the sentence. The mean speech rate of the original recordings was  $277 \pm 38$  syllables per minute. The speech rate of the synthesized speech differed from the original recordings, especially in that the sentences using TTS<sub>male</sub> were much faster. Using the overlap-add procedure implemented in Praat (Boersma & Weenink, 2007), the speech rate of the synthetic speech was reduced. Figure 3(a) shows the speech rates of all sentences after adaptation. The adapted synthesized sentences are openly available at Zenodo (Ibelings et al., 2022). The fundamental frequency for TTS<sub>male</sub> was somewhat lower than that of the original speech material (see Figure 3(b)). TTS<sub>female</sub> had the highest fundamental frequency. The long-term average speech spectra of all three variants are shown in Figure 3(c).

**Masker.** To optimize the spectral masking of stationary maskers, the masker should have the same spectral characteristics as the corresponding speech material (Festen & Plomp, 1990). The masker of the natural GÖSA was created from recordings of the same speaker using different speech material and different equipment, resulting in spectral deviations



**Figure 3.** Speech rate (a), fundamental frequency (b), and long-term average spectrum (c) of the original (natural male voice) and the synthetic male and female speakers ( $TTS_{male}$ ,  $TTS_{female}$ ).



**Figure 4.** Schematic illustration of the chronological sequence and the verbal response time. The schematic is only to illustrate when each element began and ended.

(Zinner et al., 2021). Therefore, to facilitate comparisons, each set of speech material was superimposed 30 times (Wagener et al., 2003) to generate a stationary masker. The power density spectra of the resulting maskers, called speech-adjusted noises (SAN, Zinner et al., 2021), differed from the speech materials by less than 0.1 dB in the frequency range from 100 Hz to 12 kHz. Both masker and the sentences were digitally calibrated to the same RMS value.

**Measurement Procedure.** Due to the Covid-19 pandemic, the measurements took place online via Gorilla Experiment Builder ([www.gorilla.sc](http://www.gorilla.sc)). Gorilla allows building experiments using the task builder or scripting. The participants started the experiment in their home environment with their own equipment by opening the Gorilla link. In the first step, the participants were informed about the study. Subsequently, exclusion criteria (age under 18 years, hearing impairment, mother tongue not German, no microphone available, GÖSA known) were clarified. If no exclusion criterion was met, the experiment continued with information, instructions, and the consent of the participants. To ensure that the participant's microphone was functional, a microphone test was conducted by asking the participants to allow access to their microphone and to record one test sentence. If the test recording was audible, a headphone test (Woods et al., 2017), which had been implemented in Gorilla by Milne et al. (2021), was performed. If these technical requirements were met, participants proceeded with the GÖSA.

The measurement consisted of nine test lists, i.e., the three speakers (original, TTS<sub>male</sub>, TTS<sub>female</sub>) were presented at three fixed SNRs (−4, −6, and −8 dB). The SNR values were chosen according to the psychometric function from Zinner et al. (2021) to meet recognition scores of approx. 20, 50, and 80%. The masker started 500 ms before the sentence and ended 20 ms after it. One of the ten lists of the GÖSA, each containing 20 sentences, was randomly selected, but ensured that no test list was measured twice per participant. The order of the sentences within each test list was randomized. After each sentence

presentation, participants were asked to repeat the sentence orally. Guessing was allowed. The responses were recorded using Gorilla's audiorecord zone and saved as a so-called weba file.

**Participants.** Participants were recruited via the bulletin board of Oldenburg University. The link to the study was opened 126 times, including possible second openings by the same person. Hence, individuals may have been counted multiple times. The experiment was started by 67 participants. Six were excluded, because German was not their mother tongue. Of the remaining 61 participants, 24 discontinued the experiment during the instruction and consent process and the check of technical requirements. Thirty-seven participants started the GÖSA, and 25 finished the whole experiment. The age of the 25 participants was between 19 and 40 years (average: 25.6 years, standard deviation: 5.3 years). Seven of the participants were male, 18 female, and all reported normal hearing. They also declared that they did not know the GÖSA and that their mother tongue was German. The experiment was approved by the ethics committee (Kommission für Forschungsfolgenabschätzung und Ethik) of the Carl von Ossietzky University in Oldenburg, Germany (Drs. EK/2021/063). When finishing the experiment, a voucher code of 10 € for a mail-order company was offered.

**Analysis and Statistics.** Gorilla generates a separate table for each questionnaire, information section, and GÖSA test. The verbal responses of the participants were saved as weba files. For three participants, the weba files were incomplete or contained only noise. Therefore, these participants were excluded.

For calculating the VRT, both the sentence offset time and the onset time of the participants' responses are necessary. Figure 4 shows the chronological sequence. To determine the onset time of the response, a speech recognizer of the Fraunhofer Institute for Digital Media Technology IDMT (Branch for Hearing, Speech and Audio Technology, Oldenburg, Germany) was used. The inputs were the weba



files, converted into raw PCM data with a sampling rate of 16 kHz, single channel, and 16-bit signed integer samples. The output files contained start- and end time of the participants' responses per condition (combination of SNR and speaker).

Since it was not possible to measure the sentence offset time with Gorilla, the masker onset was saved instead. Using the knowledge of the sentence length and masker onset, the sentence offset and the VRT were calculated:

$$\begin{aligned} \text{VRT} &= \text{start}_{\text{response}} - \text{end}_{\text{sentence}} \\ &= \text{start}_{\text{response}} - (\text{start}_{\text{masker}} + 500 \text{ ms} + \text{length}_{\text{sentence}}) \end{aligned} \quad (1)$$

According to the Shapiro-Wilk test, the calculated VRT values were not normally distributed. After transformation using the natural logarithm (Baayen & Milin, 2010; Pals et al., 2015), no significant deviation from a normal distribution was detected ( $p > .05$ ).

For measuring the speech-recognition score, all recorded responses were transcribed and word scoring per condition was used. The weighting factors applied in the natural GÖSA (Kollmeier & Wesselkamp, 1997) were used for each speaker. In one of nine conditions (TTS<sub>female</sub> at -4 dB SNR), the speech-recognition scores were not normally distributed (Shapiro-Wilk-Test,  $p = .003$ ). Hence, parametric tests were applied.

For both the logarithmized VRT values and for the speech-recognition scores, individual values deviating from the mean by more than three times the standard deviation were defined as outliers. Thus, one participant was considered as an outlier for speech-recognition scores under two conditions. Therefore, all statistical tests were performed using 21 participants.

Subsequently, based on the speech-recognition scores at the three SNR values, psychometric functions of the form

$$p(L, \text{SRT}_{50}, s_{50}) = \frac{100 \%}{1 + e^{4 \cdot s_{50} \cdot (\text{SRT} - L) / 100}} \quad (2)$$

were fitted for each speaker using the Maximum Likelihood procedure (Brand & Kollmeier, 2002).  $L$  describes the SNR in dB. The slope at the SRT is denoted by  $s_{50}$  and is given in pp/dB.

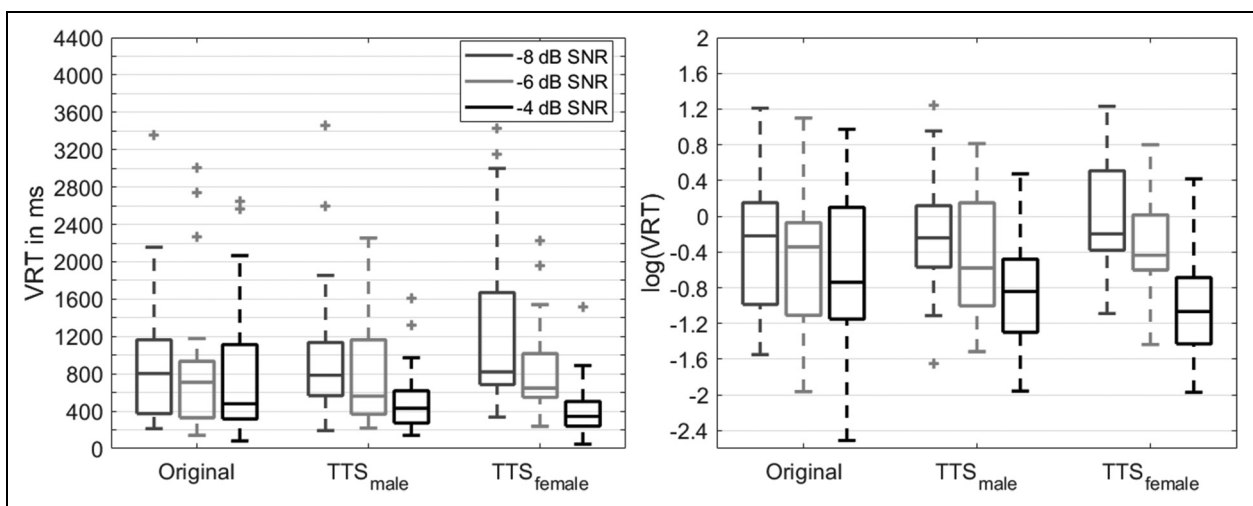
In the results, the VRT and the speech-recognition scores are presented as boxplots. The line in the middle of the box represents the median, the lower and upper bound of the box indicates the 25<sup>th</sup> and 75<sup>th</sup> percentile (box length is the interquartile range); whiskers are drawn from the lowest to the highest value within 1.5 times the interquartile range, + indicates values that are outside 1.5-times the interquartile range.

## Results

### Verbal Response Time

Figure 5(a) shows the VRT in ms and Figure 5(b) the logarithm of the VRT. As expected, poorer SNR led to higher VRTs, indicating an increase in listening effort. The lowest median is about 400 ms for TTS<sub>female</sub> at -4 dB SNR; the TTS<sub>female</sub> at -8 dB SNR led to the highest median (about 820 ms).

An ANOVA for repeated measurements for the logarithmized VRT with the factors SNR and speaker (original, TTS<sub>male</sub> and TTS<sub>female</sub>) confirmed a significant effect of the SNR [ $F(2, 38) = 29.2, p < .001$ ], but no significant effect of the speaker [ $F(2, 38) = 0.363, p = .698$ ]. There was no significant interaction between the factors [ $F(4, 76) = 2.84, p = .205$ ].



**Figure 5.** (a) verbal response time (VRT, delay between end of sentence and verbal response of the participant) for natural (original) and both synthetic speakers (TTS<sub>male</sub> and TTS<sub>female</sub>) at three SNRs (-4, -6, and -8 dB). (b) Log transformed VRTs to avoid deviations from a normal distribution,  $N = 21$ .

For the female speaker, Bonferroni-adjusted post-hoc analysis ( $\alpha=.167$ ) revealed significantly higher VRT values for  $-8$  dB SNR than for either  $-4$  dB SNR or  $-6$  dB SNR ( $p<.001$  each). For the male speaker, significant differences between the VRT values for  $-8$  dB SNR and  $-4$  dB SNR were found ( $p=.001$ ). There were no significant differences between the VRT values for the original speaker ( $p>.0167$ ).

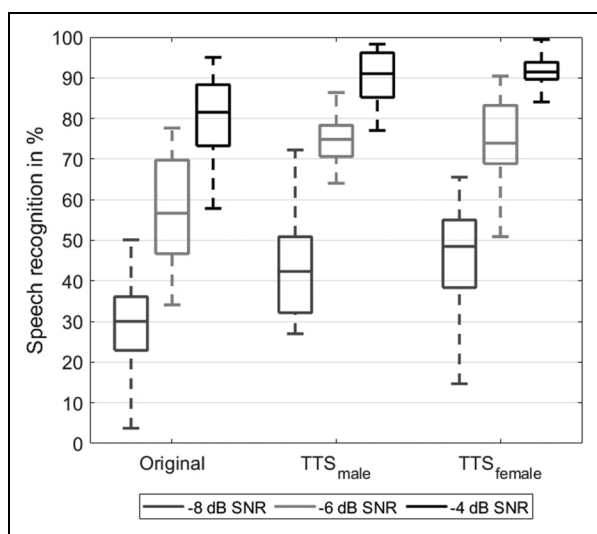
### Speech Recognition

The poorer the SNR, the fewer words were correctly recognized, independently of the speaker (see Figure 6). Furthermore, the synthetic speakers generated higher recognition scores than the natural speaker.

The repeated-measures ANOVA confirmed a significant effect of the SNR on the speech-recognition scores [ $F(2, 40)=519.7, p<.001$ ]. Furthermore, the speakers showed a significant effect [ $F(2, 40)=59.7, p<.001$ ]. There was no significant interaction between SNR and speakers [ $F(4, 80)=1.36, p=.254$ ]. Post-hoc tests with Bonferroni correction ( $\alpha=.0167$ ) revealed significant differences for all SNR values ( $p<.001$ ). The speech-recognition scores for the natural speaker differed significantly from  $TTS_{\text{male}}$  ( $p<.001$ ) and from  $TTS_{\text{female}}$  ( $p<.001$ ). The scores for the synthetic speakers were not statistically different ( $p=.451$ ).

### Psychometric Functions

The psychometric functions for the three speakers based on all speech-recognition scores were fitted using equation (2) and are shown in Figure 7. The SRT for the natural speech is worse ( $-6.5$  dB SNR) than for the synthetic speech (about  $-7.6$  dB SNR). Natural speech resulted in the



**Figure 6.** Speech-recognition scores for the natural (original) and synthetic speakers ( $TTS_{\text{male}}$  and  $TTS_{\text{female}}$ ) at three different SNRs ( $-4$  dB,  $-6$  dB and  $-8$  dB),  $N=21$ .

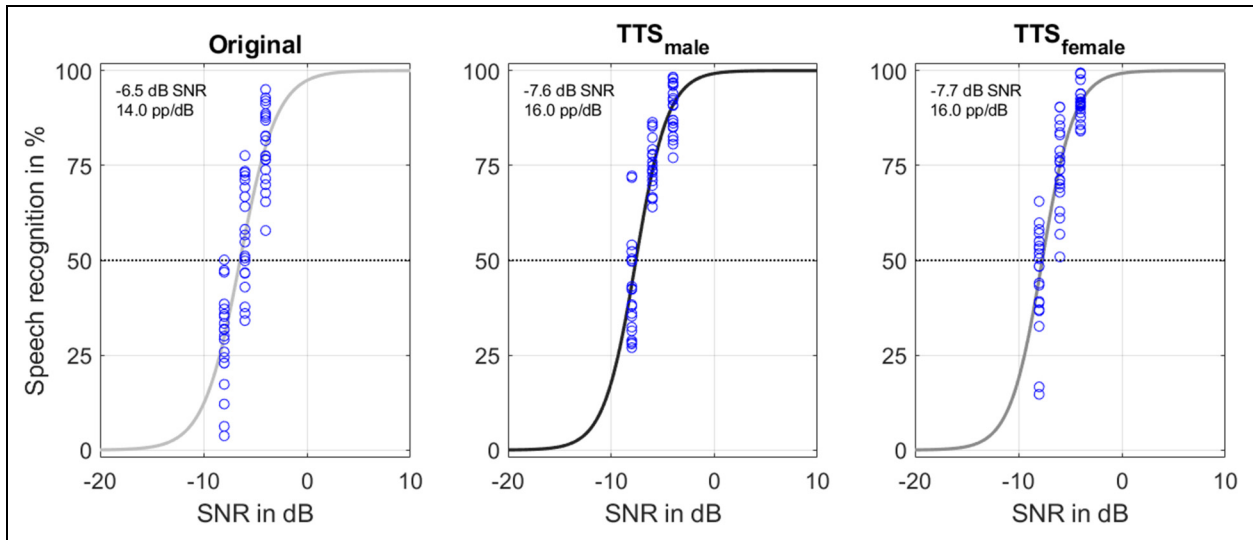
lowest slope at  $14$  pp/dB, the slopes for the synthetic speech being both slightly higher ( $16$  pp/dB). Table 2 gives the measured values in comparison to published data.

## Discussion

### Verbal Response Time

One aim of this study was to compare the listening effort estimated using the VRT for synthetic and natural speech. For both, the results showed that lower SNRs led to higher VRTs of up to a median of about  $800$  ms. This agrees with the results of Meister et al. (2018), who also found an increase in VRT with lower SNR (worse speech-recognition scores). They measured the VRT at two different SNRs corresponding to speech-recognition scores of  $50\%$  and  $80\%$ , and in two different maskers (fluctuating and stationary) for different participant groups (young and old normal-hearing listeners, older hearing-aid users). For all groups, worse SNR led to higher VRTs. A similar study design to that of Meister et al. (2018) was used by Pals et al. (2015), who also found higher VRT values with lower SNR. Quantitatively comparing the results for stationary maskers and comparable speech-recognition scores from those studies to the results of this study reveals that the VRT values of the current study have the same median range, but show a larger spread. A possible explanation is that in the current study, fixed SNRs were used, whereas Meister et al. (2018) and Pals et al. (2015) used fixed SRTs. Furthermore, the age of some participants of the current study is higher than the age of their participants. A greater variance in age may result in a greater variance in VRT (Meister et al., 2018). In addition, the previous studies conducted the experiment in the same booth for all listeners. In contrast, the present study was conducted online at different locations. Therefore, it should be noted that parameters such as the performance of the computer, stability, and speed of the internet connection may have influenced the results (Anwyl-Irvine et al., 2021), hampering comparisons with other studies. According to Anwyl-Irvine et al. (2021), online tests with Gorilla result in a delay of about  $80$  ms (standard deviation:  $8$  ms) for different devices and browsers. Nevertheless, the absolute time delay is not important for the outcome of this study, because all participants performed under all conditions (SNR and speaker); only relative differences between the conditions were analyzed.

It should be noted that there is no agreement on whether VRT directly measures listening effort or whether it is only related to listening effort. Visentin et al. (2021) measured speech recognition using a matrix test at different SNR values with normal-hearing participants. Subjective ratings of listening effort and VRT were measured; in addition, pupillometry was used. The VRT was found to be most sensitive to changes in SNR, which the authors equate with a change in listening effort. At the same time, however, no



**Figure 7.** Speech recognition scores for different SNR and speakers. The circles represent the individual speech-recognition scores of the 21 participants for the natural (Original) and both synthetic speakers ( $TTS_{\text{male}}$ ,  $TTS_{\text{female}}$ ). Based on the scores, psychometric functions were fitted. The numbers in the figure show the  $SRT_{50}$  in dB SNR and the slope in pp/dB.

**Table 2.** Comparison of the Measured SRT Values and Slopes for the Natural Speech Material and the Synthetic Material in the SAN Noise Compared to Published Data. In Each Case, the Values are Based on the Psychometric Functions Fitted to all Data Points.

	SRT in dB SNR	Slope in pp/dB
Original	-6.5	14.0
$TTS_{\text{male}}$	-7.6	16.0
$TTS_{\text{female}}$	-7.7	16.0
Zinner et al. (2021)	-6.2	18.1

correlation with the results of pupillometry could be found, leading to the conclusion that different dimensions of listening effort were captured (Visentin et al., 2021). Pals et al. (2015) call the VRT measurement a “good candidate” for measuring listening effort. Meister et al. (2018) consider VRT a good way to obtain information beyond perceived listening effort. Summarizing these studies, the VRT measurement appears to be an indicator of listening effort, although it is still unclear whether it can be a direct measure of it.

In the current study, no significant difference between the VRT values for the synthetic and the natural speakers was found. Related to the above-mentioned studies, this suggests a similar listening effort for both synthetic and natural speakers. By contrast, Govender and King (2018), who used pupillometry to measure listening effort, observed an increase in pupil size for synthetic speech, i.e., an increased listening effort for synthetic speech compared to natural speech. They observed no clear differences in listening effort between four different TTS systems; in some cases, however, there was a trend toward higher-quality rated systems resulting in lower listening effort. Simantiraki

et al. (2018) also used pupil dilation as an indicator for listening effort. They noted that synthetic speech generated using an HMM based system led to larger pupil dilations than natural speech. None of the systems from these studies were based on neural networks. Hence, it can be assumed that Acapela<sub>DNN</sub> sounds more natural than the systems used by Govender and King (2018) and Simantiraki et al. (2018).

### Speech Recognition

The SRT value for the natural speech measured online was  $-6.5$  dB SNR. Nearly the same value ( $-6.1$  dB SNR) was measured by Kollmeier and Wesselkamp (1997) using a different noise masker (original GÖSA noise). Zinner et al. (2021) found an SRT of  $-6.2$  dB SNR, closely matching the current result of  $-6.5$  dB SNR for the same SAN masker in the free field in a sound-proofed booth with normal-hearing participants. The good agreement despite a lack of control over the equipment used and possible unrecognized hearing losses in the current study indicates that online measurements are an appropriate tool to measure speech recognition using a speech test of everyday sentences.

The speech-recognition scores for the synthetic speech were significantly different from those of natural speech. Synthetic speech led to better SRTs than natural speech by 1.2 dB. One possible reason is that the natural GÖSA appears a little less clearly articulated, and partly mumbled (Müller-Deile, 2009) compared to the synthetic speech. In contrast, Simantiraki et al. (2018) found worse SRTs for synthetic speech than for natural speech. In their study, four different speech types (e.g., synthetic and plain speech) were used. The authors defined plain speech as sentences spoken

in a normal way using a male speaker and their synthetic speech was generated using an HMM-based TTS system. The normal-hearing participants scored 30% fewer correct responses with synthetic speech than with normal speech. As mentioned by Zen et al. (2013), HMM-based systems are rated less natural than DNN-based systems. It can therefore be concluded that the TTS systems have subsequently improved.

The difference in SRT of 1.2 dB between natural and synthetic speech is in the same range as for different natural speakers. Differences in speech recognition between different natural speakers were already observed for the OLSA, which was recorded using a female and a male speaker. In contrast to the current study, where differences between the male and the female synthetic speaker were negligible, the SRT for the natural female speaker was about 2.5 dB better than for the natural male speaker (Wagener et al., 2014). For different German natural speakers, Hochmuth et al. (2015) found differences in SRT of up to 5 dB. The authors explained the differences as related to a different speech rate and a larger vowel space for the female voice. It should be noted that in contrast to the current study, the speech materials in those studies (Hochmuth et al., 2015; Wagener et al., 2014) were not adjusted to the same speech rate. Nevertheless, the speech rate might not have a significant effect on the SRT for the GÖSA. Winkler et al. (2021) showed that the SRT for the GÖSA for normal-hearing participants at a speech rate of 222 syllables per minute was not significantly different from the SRT at 279 syllables per minute.

The fitted slope for the natural speaker was 14 pp/dB, and the slopes for the synthetic speakers were 16 pp/dB. In other studies, the slope was 17 pp/dB to 18 pp/dB (Kollmeier & Wesselkamp, 1997; Zinner et al., 2021). Since the slopes were shallower not only for the synthetic speakers but also for the natural speaker, the effect could possibly be related to the way the measurements were carried out. While measurements are normally conducted in a soundproof booth, in this study the measurements took place in the everyday environment of the participants. Thus, it could not be ensured that the participants were not distracted by other factors (e.g., by using the cell phone or doing other work on the computer screen), reducing attention. Additionally, participants were not tested for hearing impairments, and the age range was rather broad. Differences in participants' performance lead to shallower discrimination functions that are fitted to pooled data (Wagener, 2004). Nevertheless, overall, the observed slopes for the synthetic speakers almost match the literature values for natural speakers, despite the absence of optimization steps typically applied to natural recordings. This indicates that optimization steps might not be necessary when producing speech tests with TTS systems. To facilitate comparisons, the measurements should be repeated or performed under controlled conditions in the laboratory and using more participants.

It is unclear whether the similarity of natural and TTS materials with respect to the results of this study was influenced by the fact that everyday sentences were used. However, it can be assumed that for the purpose of speech audiometry, TTS systems can also be used for other materials. The reason for the assumption is that also for the German matrix test, which consists of sentences without semantic context as well as for the Freiburg monosyllabic test, it was shown that SRTs and slopes of the psychometric functions are similar for natural and synthetic speech (Nuesse et al., 2019; Schwarz et al., 2022). Nevertheless, it should be noted that the results of this study apply only to the GÖSA. This test consists of everyday sentences with three to seven words, and includes questions as well as declarative sentences and exclamations. The grammatical structure is mostly simple (subject-predicate-object) and there are no sub-clauses. Whether the results are also valid for other - or more complex - sentence structures can be examined in future studies.

## Conclusion

Overall, it was shown that the use of a TTS system can simplify the generation of speech material for a speech-recognition test by reducing the time effort required for recording and subsequent optimization. Although the selected TTS system was not rated equally or better than the natural reference, this study confirmed that audiological measurements using synthetic speech are possible. The speech-recognition measurements resulted in about 1.2 dB lower SRTs for the synthetic speakers compared to the natural speaker recording of the original GÖSA. However, the slopes of the psychometric functions were slightly shallower than reported in other studies. Verbal response time, which can be interpreted as indicating listening effort, was comparable for synthetic and natural speech. It should be noted, that the results presented here only apply to everyday sentences. Other tests may lead to different results. However, for further measurements and for generating new speech-recognition tests, the use of a TTS system such as Acapela Cloud is a reasonable choice.

## Acknowledgements

English language services were provided by stels-ol.de.

## Authors' Note

Part of the results were presented as a video at the online meeting on Computational Audiology VCCA 2022, as a poster at the International Hearing Aid Research Conference (IHCON), Lake Tahoe, CA, in 2022, and as a presentation at the annual meeting of the German Audiological Society (DGA) in Erfurt, Germany.


## 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 work was supported by the Graduation program of Jade University of Applied Sciences (Jade2Pro 2.0).

## ORCID iD

Saskia Ibelings  <https://orcid.org/0000-0002-6607-4884>

## References

- Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2021). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*, 53(4), 1407–1425. <https://doi.org/10.3758/s13428-020-01501-5>
- Baayen, H. R., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, 3(2), 12–28. <https://doi.org/10.21500/20112084.807>
- Boersma, P., & Weenink, D. (2007). PRAAT: Doing phonetics by computer (Version 5.3.51).
- 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. <https://doi.org/10.1121/1.1479152>
- Bunnell, H. T. (2022). Speech synthesis: Toward a “Voice” for all. *Acoustics Today*, 18(1), 14–22. <https://doi.org/10.1121/AT.2022.18.1.14>
- Festen, J. M., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, 88(4), 1725–1736. <https://doi.org/10.1121/1.400247>
- Govender, A., & King, S. (2018). Using pupillometry to measure the cognitive load of synthetic speech. Interspeech 2018, 2838–2842. ISCA. <https://doi.org/10.21437/Interspeech.2018-1174>
- Hinterleitner, F., Norrenbrock, C. R., & Moeller, S. (2013, September). Is intelligibility still the main problem? A review of perceptual quality dimensions of synthetic speech. *Gehalten auf der Eighth ISCA Workshop on Speech Synthesis*, Barcelona, Catalonia, Spain.
- Hochmuth, S., Jürgens, T., Brand, T., & Kollmeier, B. (2015). Talker- and language-specific effects on speech intelligibility in noise assessed with bilingual talkers: Which language is more robust against noise and reverberation? *International Journal of Audiology*, 54(sup2), 23–34. <https://doi.org/10.3109/14992027.2015.1088174>
- 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. <https://doi.org/10.3109/14992027.2010.506889>
- Holube, I., Haeder, K., Imbery, C., & Weber, R. (2016). Subjective listening effort and electrodermal activity in listening situations with reverberation and noise. *Trends in Hearing*, 20, 1–15. <https://doi.org/10.1177/2331216516667734>
- Houben, R., van Doorn-Bierman, M., & Dreschler, W. A. (2013). Using response time to speech as a measure for listening effort. *International Journal of Audiology*, 52(11), 753–761. <https://doi.org/10.3109/14992027.2013.832415>
- Ibelings, S., Brand, T., & Holube, I. (2022). *Synthetic Göttingen Sentence Test material created with a text-to-speech system* Zenodo. <https://doi.org/10.5281/ZENODO.6513570>
- ITU-R BS.1534-3. (2015). *Method for the subjective assessment of intermediate quality level of audio systems*. International Telecommunication Union.
- King, S. (2014). Measuring a decade of progress in text-to-speech. *Loquens*, 1(1), e006. <https://doi.org/10.3989/loquens.2014.006>
- Klink, K. B., Meis, M., & Schulte, M. (2012a). Measuring listening effort in the field of audiology—A literature review of methods, part 1. *Zeitschrift für Audiologie - Audiological Acoustics*, 51(2), 60–67.
- Klink, K. B., Meis, M., & Schulte, M. (2012b). Measuring listening effort in the field of audiology—A literature review of methods, part 2. *Zeitschrift für Audiologie - Audiological Acoustics*, 51(3), 95–105.
- Koelewijn, T., de Kluiver, H., Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E. (2015). The pupil response reveals increased listening effort when it is difficult to focus attention. *Hearing Research*, 323, 81–90. <https://doi.org/10.1016/j.heares.2015.02.004>
- Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Usilar, V., 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. <https://doi.org/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. <https://doi.org/10.1121/1.419624>
- Krueger, M., Schulte, M., Brand, T., & Holube, I. (2017). Development of an adaptive scaling method for subjective listening effort. *The Journal of the Acoustical Society of America*, 141(6), 4680–4693. <https://doi.org/10.1121/1.4986938>
- Mackersie, C. L., & Calderon-Moultrie, N. (2016). Autonomic nervous system reactivity during speech repetition tasks: Heart rate variability and skin conductance. *Ear and Hearing*, 37(1), 118S–125S. <https://doi.org/10.1097/AUD.0000000000000305>
- McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group ‘white paper’. *International Journal of Audiology*, 53(7), 433–445. <https://doi.org/10.3109/14992027.2014.890296>
- Meister, H., Raehlmann, S., Lemke, U., & Besser, J. (2018). Verbal response times as a potential indicator of cognitive load during conventional speech audiometry with matrix sentences. *Trends in Hearing*, 22, 1–11. <https://doi.org/10.1177/2331216518793255>
- Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*, 53(4), 1551–1562. <https://doi.org/10.3758/s13428-020-01514-0>
- Müller-Deile, J. (2009). Sprachverständlichkeitsuntersuchungen bei Kochleaimplantatpatienten. *HNO*, 57(6), 580–592. <https://doi.org/10.1007/s00106-009-1930-3>
- Nuesse, T., Wiercinski, B., Brand, T., & Holube, I. (2019). Measuring speech recognition with a matrix test using synthetic speech. *Trends in Hearing*, 23, 1–14. <https://doi.org/10.1177/2331216519862982>
- Obleser, J., Westmann, M., Hellbernd, N., Wilsch, A., & Maess, B. (2012). Adverse listening conditions and memory load drive a

- common alpha oscillatory network. *Journal of Neuroscience*, 32(36), 12376–12383. <https://doi.org/10.1523/JNEUROSCI.4908-11.2012>
- Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear & Hearing*, 37(1), 5S–27S. <https://doi.org/10.1097/AUD.0000000000000312>
- Pals, C., Sarampalis, A., van Rijn, H., & Başkent, D. (2015). Validation of a simple response-time measure of listening effort. *The Journal of the Acoustical Society of America*, 138(3), EL187–EL192. <https://doi.org/10.1121/1.4929614>
- Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L. & Wingfield, A. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear & Hearing*, 37(1), 5S–27S. <https://doi.org/10.1097/AUD.0000000000000312>
- Polspoel, S., Kramer, S. E., Van Dijk, B., & Smits, C. (2020). Aladdin: Automatic LAnGuage-independent Development of the Digits-In-Noise test. In *Virtual Conference on Computational Audiology (VCCA)*, online. <https://computationalaudiology.com/aladdin-automatic-language-independent-development-of-the-digits-in-noise-test/>
- Roenneberg, J., Lunner, T., Zekveld, A., Soerqvist, P., Danielsson, H., Lyxell, B., Dahlström, Ö., Signoret, C., Stenfelt, S., Pichora-Fuller, M. K. & Rudner, M. (2013). The Ease of Language Understanding (ELU) model: Theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience*, 7, 1–17. <https://doi.org/10.3389/fnsys.2013.00031>
- Schlueter, A., Lemke, U., Kollmeier, B., & Holube, I. (2014). Intelligibility of time-compressed speech: The effect of uniform versus non-uniform time-compression algorithms. *The Journal of the Acoustical Society of America*, 135(3), 1541–1555. <https://doi.org/10.1121/1.4863654>
- Schoeffler, M., Bartoschek, S., Stöter, F.-R., Roess, M., Westphal, S., Edler, B., & Herre, J. (2018). WebMUSHRA—A comprehensive framework for web-based listening tests. *Journal of Open Research Software*, 6, 8. <https://doi.org/10.5334/jors.187>
- Schötz, S. (2007). Acoustic analysis of adult speaker age. In C. Müller (Hrsg.), *Speaker classification I* (pp. 88–107). Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-540-74200-5\\_5](https://doi.org/10.1007/978-3-540-74200-5_5)
- Schwarz, T., Frenz, M., Bockelmann, A., & Husstedt, H. (2022). Untersuchung einer synthetischen Stimme für den Freiburger Einsilbertest. *GMS Zeitschrift für Audiologie - Audiological Acoustics*, 4(Doc04), 94–101. <https://doi.org/10.3205/ZAUD000022>
- Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R. A., Agiomvrgiannakis, Y., & Wu, Y. (2018). Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4779–4783).
- Simantiraki, O., Cooke, M., & King, S. (2018). Impact of Different Speech Types on Listening Effort. *Interspeech 2018*, 2267–2271. <https://doi.org/10.21437/Interspeech.2018-1358>
- Taylor, P. (2006). Unifying unit selection and hidden Markov model speech synthesis. *Interspeech 2006*, paper 1456-Wed2A3O.5-0. ISCA. <https://doi.org/10.21437/Interspeech.2006-487>
- Uslar, V. N., Carroll, R., Hanke, M., Hamann, C., Ruigendijk, E., Brand, T., & Kollmeier, B. (2013). Development and evaluation of a linguistically and audiological controlled sentence intelligibility test. *The Journal of the Acoustical Society of America*, 134(4), 3039–3056. <https://doi.org/10.1121/1.4818760>
- Visentin, C., Valzolgher, C., Pellegatti, M., Potente, P., Pavani, F., & Prodi, N. (2021). A comparison of simultaneously-obtained measures of listening effort: Pupil dilation, verbal response time and self-rating. *International Journal of Audiology*, 61(7), 561–573. <https://doi.org/10.1080/14992027.2021.1921290>
- Wagener, K. (2004). *Factors influencing sentence intelligibility in noise*. Oldenburg: Bibliotheks- und Informationssystem der Univ.
- Wagener, K., Hochmuth, S., Ahrlich, M., Zokoll, M. A., & Kollmeier, B. (2014). Der weibliche Oldenburger Satztest. Proceedings of 17 Jahrestagung der Deutschen Gesellschaft für Audiologie, CD-Rom, 4 pp. Oldenburg, Germany.
- Wagener, K., Josvassen, J. L., & Ardenkjær, R. (2003). Design, optimization and evaluation of a danish sentence test in noise: Diseño, optimización y evaluación de la prueba danesa de frases en ruido. *International Journal of Audiology*, 42(1), 10–17. <https://doi.org/10.3109/14992020309056080>
- Wagener, K., Kuehnel, V., & Kollmeier, B. (1999). Entwicklung und Evaluation eines Satztests für die deutsche Sprache I: Design des Oldenburger Satztests. *Zeitschrift für Audiologie*, 38(1), 4–15.
- Winkler, A., Schlüter, A., Gebauer, T., Seifert, J., Tuschen, L., Radeloff, A., & Holube, I. (2021). Einfluss von Sprechtempo und Störgeräusch auf das Sprachverstehen im Göttinger und im HSM-Satztest. *GMS Zeitschrift für Audiologie - Audiological Acoustics*; 3 (Doc02). <https://doi.org/10.3205/ZAUD000014>
- Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, 79(7), 2064–2072. <https://doi.org/10.3758/s13414-017-1361-2>
- Yund, E. W., & Woods, D. L. (2010). Content and procedural learning in repeated sentence tests of speech perception. *Ear & Hearing*, 31(6), 769–778. <https://doi.org/10.1097/AUD.0b013e3181e68e4a>
- Zen, H., Senior, A., & Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 7962–7966. Vancouver, BC, Canada: IEEE. <https://doi.org/10.1109/ICASSP.2013.6639215>
- Zinner, C., Winkler, A., & Holube, I. (2021). Vergleich von fünf sprachtests im sprachsimulierenden Störgeräusch. *GMS Zeitschrift für Audiologie - Audiological Acoustics*, 3(Doc04), 1–12. <https://doi.org/10.3205/zaud000016>