Revisiting Three Text-to-Speech Synthesis Experiments with a Web-Based Audience Response System

Christina Tånnander[1,2], Jens Edlund[1], Joakim Gustafson[1]
[1]KTH Royal Institute of Technology
[2]Swedish Agency for Accessible Media
christina.tannander@mtm.se, edlund@speech.kth.se, jocke@speech.kth.se

Abstract

In order to investigate the strengths and weaknesses of Audience Response System (ARS) in text-to-speech synthesis (TTS) evaluations, we revisit three previously published TTS studies and perform an ARS-based evaluation on the stimuli used in each study. The experiments are performed with a participant pool of 39 respondents, using a web-based tool that emulates an ARS experiment. The results of the first experiment confirms that ARS is highly useful for evaluating long and continuous stimuli, particularly if we wish for a diagnostic result rather than a single overall metric, while the second and third experiments highlight weaknesses in ARS with unsuitable materials as well as the importance of framing and instruction when conducting ARS-based evaluation.

Keywords: TTS evaluation, audience response system, evaluation methodology

1. Introduction

Evaluation is singled out as a key weakness in Text-to-speech synthesis (TTS) research and development by leading figures in the field, and the need for TTS evaluation methods that are better tailored to specific evaluation objectives is widely recognised (King, 2014; Wagner et al., 2019). The evaluation of long and information rich texts (sometimes referred to as "long-form TTS") is one of the specific areas where current methods have been found wanting (Clark et al., 2019; Cambre et al., 2020). Furthermore, as modern TTS approaches a quality that is in some respects indistinguishable from human speech (Shen et al., 2018; Shirali-Shahreza & Penn, 2018; Pandey et al., 2023), evaluation must adapt to be better-suited for this new generation of voices, as well as for specific applications such as conversational speech and the presentation of more complex linguistic materials. For the last decade or so, Audience Response Systems (ARS) have occasionally been used for speech analysis and evaluation, and despite convincing initial results, its suitability for different tasks is not well understood. Our purpose, then, is to explore the appropriateness of ARS-based evaluation for different evaluation goals and to learn best practices to facilitate interpretation and avoid pitfalls. We do so by revisiting three sets of TTS stimuli that have been used in three previous studies, and letting new respondents evaluate these in an on-line version of ARS where they click a button whenever they perceive something they do not like. Our results show that ARS is well suited for the evaluation of long and information rich texts. It is particularly useful if diagnostics are required, that is if the evaluation should inform us of what goes wrong and where, rather than provide a single generic number corresponding to some general sense of quality. With other tasks, we see that ARS is less suitable, and/or requires considerable consideration when it comes to framing and instruction.

2. Background & related work

2.1 Text-to-speech evaluation

The de facto standard evaluation for TTS is the mean opinion scores (MOS), which was originally developed as a standard for evaluating audio transmission. The rather deliberate process prescribed by the standard ITU-T (1996) and its numerous amendments is intended to ensure a measure of reliability, but is rarely if ever adhered to these days. This is likely of little consequence, as it would not ensure reliability for the TTS evaluation task in any case: this task is too different from the original purpose of the standard.

In current MOS evaluation, respondents are typically asked to rate the "naturalness" of a single sentence on a scale from 1 to 5. MOS has been criticised since decades for being poorly founded for the TTS task, for suffering from bias, and for lacking general reliability (Hasegawa-Johnson & Alwan, 2003; Polkosky & Lewis, 2003; Vazquez-Alvarez & Huckvale, 2002; Zieliński & Rumsey, 2008) and its lack of diagnostic value is well-known (Pols, 1998).

More recently, the number of papers addressing MOS problems and other TTS evaluation-related topics have grown. In 2023, both Interspeech and SSW (the Speech Synthesis Workshop) had sessions dedicated to TTS evaluation. Some examples of bias discussions in these papers are the respondents' effect on the variance of MOS scores (Chiang et al., 2023; Finkelstein et al., 2023), the impact of the rating range and the instructions given to the respondents (Chiang et al., 2023; Cooper & Yamagishi, 2023; Kirkland et al., 2023), and the task or context in which the voice is used (Lameris et al., 2023). Several papers compared MOS to other TTS evaluation methods, such as elimination test (Kayyar et al., 2023) and A/B test (Camp et al., 2023).

LREC-COLING 2024, pages 14111–14121
20-25 May, 2024.

2.2 Audience Response Systems

Audience Response Systems (ARS) most likely have their origin in the film and television industry, where respondents have been asked to click a button or set a dial continuously while watching a production, with the main advantage that it is useful to spot flaws and/or opportunities (Perebinossoff et al., 2005). In the speech field, the first use we are aware of occurred as a live evaluation at a special session for song synthesis at Interspeech 2007.

In 1995, ITU adopted BT.500-7, in which they describe the Single-Stimulus Continuous Quality Evaluation (SSCQE) method. The method was an outcome of the EU-funded RACE MOSAIC project. Designed for television quality assessment, the method has remained largely unused for speech and TTS evaluation. According to Alpert & Evain (1997) it was initially intended as a three-stage process, but only the first stage is included in BT.500-7, and this stage is despite the lack of acknowledgements in essence a reformulation of ARS. Alpert & Evain (1997) notes that experiments were conducted that verified that the method gave meaningful results for continuous stimuli of durations as long as 30 and 60 minutes.

The use of ARS-style evaluations for speech technology was proposed in Edlund et al. (2008), and a tool dubbed *the Yuck button* with similar functionality was introduced in Poppe et al. (2011) and used frequently in subsequent work. A series of publications about ARS as a TTS evaluation method were published between 2012 and 2015. The first was a proof-of-concept in which three respondent groups were given different instructions about when to click: when they heard (1) something unintelligible, (2) something irritating or (3) something not entirely correct. The number of clicks from group (1) was about 1/10 of group (2) and (3), which had similar numbers of clicks. The click distributions confirmed that there was consensus within and across the three groups (Edlund et al., 2012; Tånnander, 2012). Next, Edlund et al. (2013) examines respondents' reaction times, showing that click latency for simple beep sounds average at least 500 ms. More complex tasks can be assumed to have longer latencies.

Finally, Edlund et al. (2015) shows that peaks resulting from ARS can be reliably related to events in the speech signal. The material from this study is among those revisited here, and the study is described in more detail in section 3.4. A related use of ARS was explored by Strömbergsson et al. (2020, 2021) who validated the ARS method for acceptability and intelligibility of speech produced by children with speech sound disorders.

3. Method

We revisit three published TTS evaluation studies, each with different original research questions and goals. While some elements - notably the stimuli under investigation - are the same as in the initial studies, we emphasise that the new experiments are not replication studies. Instead we are looking to investigate ARS-based evaluation on different evaluation tasks and materials. Some experiment design elements were kept consistent across all three new studies. These are described in section 3.1. On the other hand, several factors in each experiment were deliberately changed. These individual differences between the original and new studies are summarised in the tables leading into experiment specific method sections.

3.1 Experiment design (general)

3.1.1 Stimuli

Human non-TTS recordings, unit selection TTS (**US**) and neural TTS (**NTTS**) were all based on the same Swedish, female voice. The **US** voice was built in 2011 using an in-house TTS framework from the Swedish Agency for Accessible media, *Filibuster* (Sjölander et al., 2008). The **NTTS** was trained on the same data set, using the Tacotron2 framework (Wang et al., 2017) in conjunction with the WaveGlow vocoder (Prenger et al., 2019). Due to limitations of the experiment platform, all audio files were converted to mp3.

3.1.2 Respondents

A participant pool of 39 individuals employed by the Swedish Agency for Accessible Media (MTM) was recruited in order to ensure a level of consistency between experiments. In each experiment, 32 respondents from the pool participated, resulting in seven respondents being unique to either test session 1 or test sessions 2 or 3.

3.1.3 Experiment platform

The web-based experiment platform *Lyss*, which was developed by STTS AB to provide remote functionality similar to that of an Audience Response System, was used for all experiments.

3.2 Process (general)

3.2.1 Instructions

Respondents were directed to use headphones. They were instructed to use their mouse or touchpad to click within a designated area whenever they heard something they did not like. The seminal instruction was (translated from Swedish): "*Listen to the audio file and click the click area whenever you hear something that you dislike*". To ensure consistency in test conditions across all three experiments, this instruction was kept identical, although the introduction to sessions could contain additional information.

3.2.2 Experiment sessions

Experiment 1 was conducted in a separate test session, whereas experiments 2 and 3 were combined into a single session, with an exception for two respondents. These two respondents completed experiments 2 and 3 in separate sessions due to practical concerns. Each session started with a 30-second practice file.

3.3 Analysis (general)

The test results analysed the same way across all three experiments, with the exception of correlations between the outcomes of the original and present experiments, which are detailed in the analysis sections of each individual experiment.

3.3.1 Number of clicks

For each experiment, we conducted an analysis of the total number of clicks per user. By the nature of the task, we expect the click distribution to be more Poisson than normal distributed and assess overall gender and three age groups (25-39, 40-49, and 50-65) differences using non-parametric tests (Mann-Whitney U and Kruskal-Wallis, respectively). Additionally, for each audio stimulus, we calculated the minimum, maximum, median, and quartiles of the number of clicks.

3.3.2 Click distribution

Kernel density estimation was performed on the clicks from each individual respondent. Similar to Edlund et al. (2015), we used a Gaussian kernel with a width of 250 ms, consistent with the approximate duration of a syllable. No correction for estimated reaction times were undertaken, as we wanted to present the data with as little interpretation as possible. The results from (Edlund et al. (2013) show, however, that the event causing a peak is found somewhere in the second preceding the peak. The area under each resulting curve was normalised to 1, creating in effect individual probability density functions for each respondent. These functions were then sampled at a frame rate of 10 ms. The set of samples at each frame was used in two ways. Firstly, the samples were summed, multiplied by the total number of clicks, and divided by the number of respondents. The area under any segment of the resulting curve is a weighted average number of clicks per respondent for that segment. The weight is such that each respondent has the same influence over the curve regardless of the number of clicks the respondent produced, as there is no reason to assume that someone who clicks more is more accurate or more correct than someone who clicks less. The weighting comes with a potential problem, in that a single click from a respondent in a group where all other respondents click frequently can result in a very high peak. Rather than thresholding on the number of clicks, we constructed a complementary plot based on the median value of the set of samples in each frame. This effectively removes the peaks that are purely the result of one or very few respondents. We use the median plot as a filter to select meaningful peaks.

3.4 Experiment 1

A 2015 study investigated the relationship between KDE click peaks on the one hand and events occurring in the unit selection synthesis process (objectively) or in the speech signal (subjectively) on the other (Edlund et al., 2015). An association between the highest peaks and the internal operations of the TTS system was found, particularly at concatenations, where different audio segments were pieced together to create the target message. Additionally, a professional TTS developer associated most of click peaks with perceptual anomalies in the synthesised speech signal. The study primarily served as a methodology paper investigating the use of ARS for TTS evaluation. Table 1 summarises the differences between the original and the present study.

	Original (2015)	New
Evaluation method	ARS	ARS
Purpose	ARS methodology	ARS methodology
Respondents/ stimulus	20 students	16 employees
Stimuli	1 audio file	2 audio files
Duration	3 minutes	6 minutes
Test tool	Xbox in room	Web-based (Lyss)

Table 1: Summary of listening test setup for exp-1: 2015 experiment and new ARS experiment.

3.4.1 Purpose of revisitation

The new study adds a new **NTTS** voice to the **US** voice used in the original study in a bid to explore whether ARS proves equally meaningful and appropriate for assessing both **US** and **NTTS** voices. We expect the number of clicks to be fewer in **NTTS** compared to **US**, and that the click peaks can be linked to events in the speech signal, even though these might differ in nature and exhibit less apparent relationships compared to the peaks in **US**.

3.4.2 Stimuli

The same **US** audio file from the original 2015 experiment was used (**TEXT-1**). This is a text about student support in school and was now also synthesised with **NTTS**. A fabricated error, functioning as a sanity check, was inserted in the original audio, where some words and a pause were deleted in the 55th second. We replicated the same error in the **NTTS** version, placing it in the 59th second to account for slight variations in audio length, as shown in Table 2.

	Duration	Error location
US-TEXT-1	2:55	0:55
NTTS-TEXT-1	3:08	0:59
US-TEXT-2	2:50	1:43
NTTS-TEXT-2	3:03	1:49

Table 2: Length and fabricated error locations of the two texts in experiment 1, with **US** and **NTTS**.

In addition, we included a second text, a Wikipedia entry about Carl Linnaeus (**TEXT-2**). Again, we fabricated an error, this time by duplicating the syllable “Ca” in the proper name “Carl” at 1:43 minutes into the **US** audio and 1:49 into the **NTTS** audio. Each respondent listened to either **US-TEXT-1** and **NTTS-TEXT-2** or **NTTS-TEXT-1** and **US-TEXT-2**. The order of the stimuli was systematically altered.

3.4.3 Respondents

In the 2015 study, 20 students in computer science participated. In the current experiment, 32 individuals from the participant pool described in 3.1.2 took part (16 female, 15 male and 1 other).

3.4.4 Instructions

The instructions were similar in the old and new study. Respondents were instructed to listen to the audio files and asked to click a button whenever they heard something they did not like.

3.4.5 Process

In contrast to the 2015 study, where three separate groups of respondents were gathered in the same room and provided with Xbox controllers as clicking devices, the test environment in the current experiment was web-based, as described in 3.1.3.

An important difference from the 2015 approach is that, at that time, the respondents received verbal instructions within a group setting, decreasing the potential for misunderstandings. With a web-based experiment, we cannot ensure that respondents carefully read and comprehend the provided instructions in the same manner.

3.4.6 Analysis

The analyses shared across all three experiments (see 3.3) were used to compare the number of clicks for **US** and **NTTS**, and to assess whether events in the speech signal likely to underlie the peaks were of different character for the two TTS types.

3.5 Experiment 2

A 2021 study examined different pre- and post-processing techniques for slowing down synthetic speech (Tånnander & Edlund, 2021). Respondents judged how well the different versions matched their expectations of slow speech using five alternatives ranging from “very poorly” to “very well”, which was converted into a five-grade scale used to rank the methods.

Respondents rated actual human slow reading, **SLOW-1,** the highest (see Table 4), followed by a variant where the pause locations and durations from this human slow reading were transposed to the TTS rendition (**SLOW-7**), and on third place the TTS version with pauses inserted between each word (**SLOW-6**). A version with pause durations of several seconds inserted at commas and full stops was used as a sanity check (**SLOW-4**). This method also received the lowest ranking by the respondents. Table 3 summarises the differences between the original and the present study.

	Original (2022)	New
Evaluation method	Rating/ranking	ARS
Purpose	Most appropriate reading	ARS methodology
Respondents/ stimulus	64 from Prolific	32 employees
Stimuli	2 audio files * 8 variations	1 audio file * 8 variations
Duration	8 minutes	4 minutes
Test tool	Web-based (SBTal)	Web-based (Lyss)

Table 3: Summary of listening test setup for exp-2: original experiment from 2022 and ARS experiment.

3.5.1 Purpose of revisitation

The primary goal is to assess the appropriateness of ARS, this time for selecting the most suitable slow version of a text of approximately 30 seconds. The ARS method is unlikely to perform optimally with these stimuli, as the slowing-down techniques used are highly diverse, with the majority applied uniformly throughout the entire audio file, rather than being applied at specific points. Notwithstanding, it is interesting to see what is captured by ARS.

3.5.2 Stimuli

The original stimuli comprised two texts related to covid-19, each with an approximate 30-second duration when narrated at a very slow pace by the Swedish, female speaker used in all experiments. These human recordings serve as our reference stimuli. The same texts were also re-recorded by the same speaker at a normal speaking rate and synthesized using the **NTTS** synthetic voice. To standardise the duration of the latter two audio files, they were time-stretched to match the 30-second reference stimulus. The TTS version was further adjusted, primarily through the strategic insertion of pauses at specified points, as outlined in Table 4, ensuring a consistent 30-second duration for each stimulus. For our current experiment, we used only the first of these texts and systematically varied the order of the eight versions.

File	Description	Orig rank
SLOW-1	Human, slow reading	1
SLOW-2	Human, stretched reading	6
SLOW-3	TTS, stretched	4
SLOW-4	TTS, pauses at minor and major delimiters	8
SLOW-5	TTS, pauses at phrase boundaries	7
SLOW-6	TTS, pauses between words	3
SLOW-7	TTS, pauses mimicking human slow reading	4
SLOW-8	TTS, as (7) + stretched	2

Table 4. Description of the eight slow variants in the original experiment from 2021.

All stimuli and complete KDE analyses of the clicks are publicly available[1].

3.5.3 Respondents

In the 2021 study, 64 respondents recruited from Prolific evaluated the eight variants of each of the two texts. In our current study, 32 respondents from the participant pool took part.

3.5.4 Instructions

In the original experiment, the evaluation question was meticulously framed: *"Imagine that you have requested a short text to be read for you very slowly. How well does this reading match your expectations?"* (Tånnander & Edlund, 2021). In the current experiment, the evaluation question in section 3.2.1 was used together with brief information indicating that the respondents would be listening to slow speech.

3.5.5 Process

Both the 2021 and current experiment were conducted using a web-based tool.

3.5.6 Analysis

In addition to the general analysis (3.3), a rank of the different slow variants was created by ordering the estimated click medians, to provide data that can be related to the original experiment.

3.6 Experiment 3

A 2022 study compared two **NTTS** voice builds on the same training data. **VOICE-1** was trained in a standard manner and **VOICE-2** in a manner that provides a measure of external control over prominence. A preference test showed a strong preference for **VOICE-2** in sentences where prominence of verb particles and numerals was crucial for sentence comprehension (Tånnander et al., 2022).

Table 5 summarises the differences between the original and the present study.

	Original (2021)	New
Evaluation method	A/B test	ARS
Purpose	Best voice	ARS methodology
Respondents/ stimulus	20 employees	16 employees
Stimuli	40 audio files, 20 sentence pairs	20 sentences concatenated
Duration	3 minutes	1,5 minutes
Test tool	Web-based (SBTal)	Web-based (Lyss)

Table 5: Summary of listening test setup for exp-3: original experiment from 2021 and ARS experiment

3.6.1 Purpose

Once again, the purpose is to evaluate the appropriateness of the ARS method, this time when it comes to comparing single sentences. In this case, the expectations are again somewhat low: one of the key benefits of ARS is that it allows the continuous, unbroken evaluation of lengthy continuous stimuli, which is not a requirement for single sentence evaluation. Nevertheless, the diagnostic properties of ARS may add something that other methods miss.

3.6.2 Stimuli

In the 2022 study, 10 sentences containing particle verbs were selected, and synthesised without prominence control (**VOICE-1**) and with prominence control (**VOICE-2**). **VOICE-1** was expected to render certain verb particles and numerals with a conspicuous lack of prominence (this is a typical problem for Swedish TTS) and **VOICE-2** with at least some degree of prominence. 20 sentence pairs were synthesised with **VOICE-1** and **VOICE-2** (Tånnander et al., 2022).

Since the ARS evaluation method is not expected to work well on short stimuli (e.g. individual sentences), the 20 sentences were concatenated into one single audio file. To display that the sentences were unrelated to one another and avoid irritation due to the incoherence of the content as a whole, a two-second silence was inserted after each sentence. The overall duration of the audio files was around 1:20 minutes. To prevent respondents from recognising the sentences, each respondent listened to just one rendition of each sentence per test set. The sets were balanced to include 5 sentences where verb particles were assigned prominence and 5 without, as well as 5 sentences where the numerals were assigned prominence and 5 were not. Consequently, two sets of 20 sentences in randomised order were created, **PREF-1** and **PREF-2**.

3.6.3 Respondents

20 respondents took part in the original preference test where they selected the best reading of the two versions of the same sentence.

3.6.4 Instructions

No extra information was given about the character of these audio files.

3.6.5 Process

Both the 2021 and current experiment were conducted using a web-based tool.

3.6.6 Analysis

An extra initial step to disentangle the concatenated sentences was undertaken before the general analysis, which was done on the separate **VOICE-1** and **VOICE-2** sentences.

[1] https://github.com/christinatannander/ARS-LREC-COLING-2024

4. Results

4.1 General descriptive statistics

Table 6 shows the descriptives of gender and age categories over all three experiments. The Mann-Whitney U test indicated that females clicked significantly more than males, z = 3.47, p < .001. The Kruskal-Wallis H test indicated that there is a non-significant difference in the number of clicks between the different age groups, $\chi2(2) = 2.52$, p = .284, with a mean rank score of 188.28 for 25-59, 176.05 for 40-49, 165.63 for 50-65.

	Median	Min	Max	1Q	3Q
Total clicks/ experiment	3	0	91	1.00	3.00
Male	2	0	61	0.00	7.00
Female	4.5	0	91	1.75	10.0
25-39	3	0	91	1.50	7.00
40-49	5.5	0	47	2.00	11.2
50-65	5	0	62	1.00	10.0

Table 6: Median, minimum, maximum 1st and 3rd percentiles of gender and age categories.

4.2 Experiment 1

4.2.1 Number of clicks

Table 7 presents the descriptive statistics of clicks in experiment 1.

	Median	Min	Max	1Q	3Q
Total clicks/user	30.5	3	98	17.7	50.5
US	24	2	91	15.0	38.2
NTTS	7	1	27	2.75	11.2
TEXT-1	12.5	1	91	4.00	27.0
TEXT-2	12.5	1	62	6.00	24.0
US-TEXT-1	25	3	91	15.2	57.5
US-TEXT-2	24	2	62	16.5	42.7
NTTS-TEXT-1	6	1	27	2.50	13.2
NTTS-TEXT-2	7	1	19	5.20	13.5

Table 7 : Descriptive statistics of experiment 1. US = unit selection TTS, NTTS = neural TTS.

We observe that all respondents clicked at least once per file. The total number of clicks per listener ranges from 3 and 98. A Mann-Whitney U test showed a significant difference between the number of clicks for **US** and **NTTS** ($z = 4.80$ $p < .001$), where the listeners clicked almost 3 times more often when listening to the **US** reading. No significant differences were observed when comparing the number of clicks between the two texts or whether an audio file was presented first or last in the listening test.

4.2.2 Click distribution

Click distributions for **US** and **NTTS** behaved similarly in **TEXT-1** and **TEXT-2**, with more and taller peaks appearing in the **US** files. Figure 1 shows the distributions of average and median estimated clicks per respondent (CPR; right Y axis) for the **US** voice and **TEXT-1** and Figure 2 shows the **NTTS** voice and **TEXT-2**.

Peaks with median CPR over 0 were analysed and connected to an event in the speech signal, as described in section 3.3.2. For **US**, 48% of these peaks were related to audible concatenations, 36% to odd prosody, and 15% to the fabricated errors. The only three **NTTS** peaks that remain in the median CPR were associated with the fabricated error (2 peaks) and prosody (1 peak).

All peaks with a median over 0.01 in **US-TEXT-1** were analysed, as well as a few cases where a high average peak was not accompanied by a median peak. Peak (2) in Figure 1 shows that the sanity check works well: there is a high peak after the fabricated error in the 55th second. Most of the remaining peaks are related to audible concatenations in the **US** TTS voice, but there are also a couple of peaks connected to prosody problems (1, 3 and 4). The 5th and 6th peaks average on around 0,03, but are not accompanied by a median peak, making them less reliable. Nonetheless, these peaks cold be associated to odd prosody (5) and audible concatenations (6).

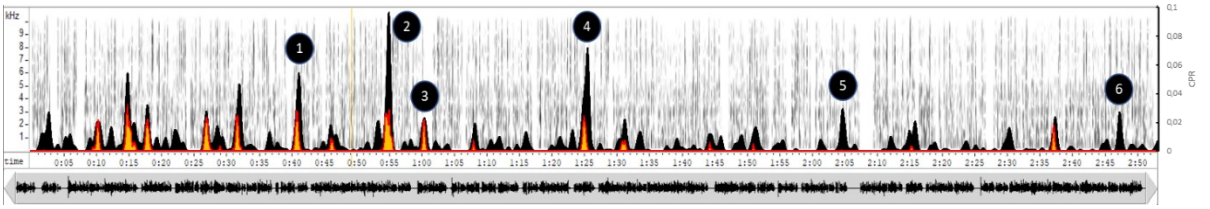

Figure 1: Average (black) and median (orange/bright) clicks per respondent (CPR; right Y axis) of **US-TEXT-1**

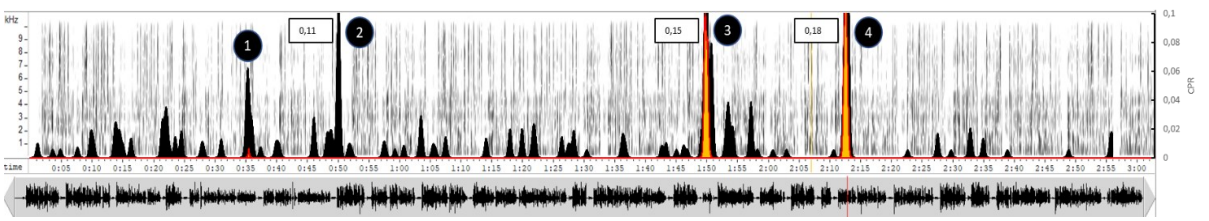

Figure 2: Average (black) and median (orange/bright) clicks per respondent (CPR; right Y axis) of **NTTS-TEXT-2**. Note that three peaks are truncated.

In **NTTS-TEXT-2** (Figure 2), there are three median peaks above 0: the fabricated error (3), which is in fact followed by two connected peaks, and an unprominent verb particle in the sentence *"Linné tyckte inte* **om** *honom"* (En. *Linnaeus did not like him*.) (4). Peak 1 seems to be caused by some archaic language in a quote in the text, and peak 2, with a high average but a median of 0 to a few syllables with creaky voice.

4.3 Experiment 2

4.3.1 Number of clicks

Table 8 presents the descriptive statistics of clicks in experiment 2.

	Median	2021 rank	ARS rank
SLOW-1	0	**1**	**1**
SLOW-2	3	6	6
SLOW-3	2	4	3
SLOW-4	2.5	8	2
SLOW-5	4	7	4
SLOW-6	3	3	8
SLOW-7	3.5	4	6
SLOW-8	3	2	5

Table 8. Click medians, 2021 and ARS ranks of the eight slow variants as described in Table 4.

4.3.2 Click distribution

Peak locations were analysed to provide explanations for the respondents' reactions. In some but not all instances, these peaks could be associated with events in the speech signal. The task was notably more challenging than identifying peaks in the Experiment 1. There are no medians above 0 in four of the slow variants (**SLOW-1**, **SLOW-2**, **SLOW-3** and **SLOW-8**). Instead, a few of the tallest average peaks in these files were analysed. In (1), a tall peak was associated with odd vowel quality. There are no clearly distinct average peaks in **SLOW-2** or **SLOW-3**, and in **SLOW-8**, the most distinct peak seems to be related to creaky voice.

SLOW-4 (Figure 3) had very long pauses inserted at minor and major delimiters, which triggered one tall peak in the first speech section, probably due to staccato-like speech. The longest pause caused an almost even click curve, while the second longest pause triggered a strong reaction from a few respondents.

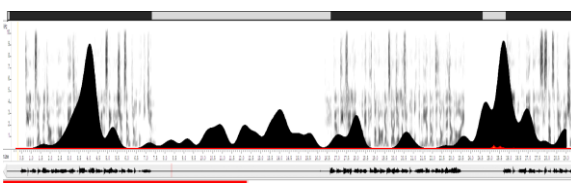

Figure 3: Average clicks per respondent in **SLOW-4**. The top bar shows speech (black) and silence (grey).

Figure 4 shows the clicks per respondent in **SLOW-5**, where pauses were inserted at phrase boundaries. Here, the average peaks (black) seem to be related to the end of phrases and the subsequent pause. Two median peaks (orange) occur in the beginning of the file, also in connection to phrase end and pause.

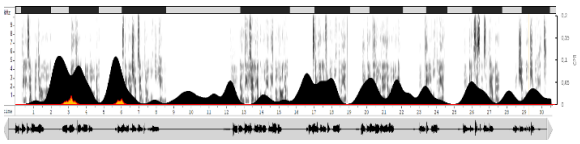

Figure 4. Average (black) and median (orange) clicks per respondent in **SLOW-5**. The top bar shows speech (black) and silence (grey).

SLOW-6, with pauses between each word has considerably more clicks in the beginning of the file, though only one with a median above 0. The same goes for **SLOW-7**, where the TTS was pausing at the same locations as the human slow reading.

4.4 Experiment 3

4.4.1 Number of clicks

Table 9 presents the descriptive statistics of clicks in experiment 3. There was no significant difference in number of clicks between **VOICE-1** and **VOICE-2**.

	Median	Min	Max	1Q	3Q
Total clicks/user	7	0	22	4.75	10.0
PREF-1	6,5	0	14	3.75	9.00
PREF-2	8	0	22	5.00	10.2
VOICE-1	7	0	16	3.75	8.50
VOICE-2	3,5	0	14	2.00	8.00

Table 9. Click statistics of experiment 3.

To compare the current results with the results from the original experiment, the percentage of number of clicks for **VOICE-1** in each utterance pair was used as the relative preference for **VOICE -2**, as shown in Figure 5. Pearson's correlation indicated a non-significant medium positive relationship between the original preference test and the click test ($r(18) = .333$, $p = .151$).

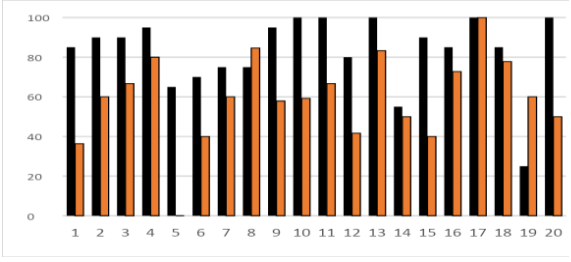

Figure 5: For each sentence pair: percentage of preference for **VOICE-2** from the original test (black) and percentage of clicks for **VOICE -1** (orange).

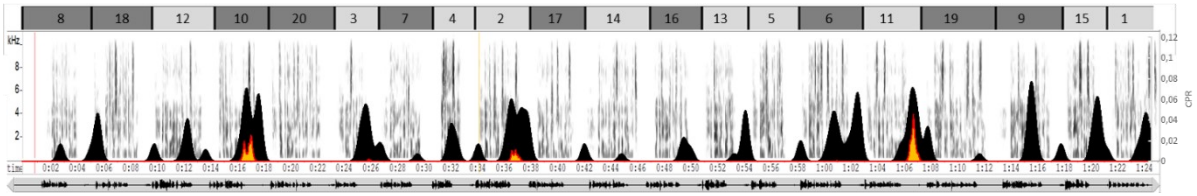

Figure 6: Average (black) and median (orange/bright) clicks per respondent of **PREF-1**. The y axis to the right shows the number of clicks per respondent. The top bar represents the sentence IDs, with bright grey bars for **VOICE-1** and dark grey bars for **VOICE-2.**

4.4.2 Click distribution

The stimuli in this experiment contained 20 unrelated sentences each, separated by pauses of around 2 seconds. We see a tendency that the respondents wait to click until the sentence is finished, resulting in most click peaks located within these pauses.

In total, there were 6 median peaks above 0 in **PREF-1** and **PREF-2**, whereof 2 belonged to **VOICE-2** (with prominence control) and 4 to **VOICE-1**. The **VOICE-2** peaks are suspected to be caused by an odd /a/ in one case, and too much prominence of the verb in the particle verb (this is usually destressed). The click distribution of **PREF-1** is shown in Figure 6.

5. Discussion

We first note that in contrast with most studies, we found a significant gender difference in the data as a whole for willingness to click, where female respondents click more than males. The difference is not significant in the individual experiments. As the respondent groups were not balanced for non-gender demographic characteristics (other than their general place of employment), we simply note that the female and male groups may have shared other common factors than gender and forego further analysis.

Experiment 1 is the only of the three experiments where there was a strong reason to think that ARS-based evaluation would be genuinely meaningful. The experiment evaluated (relatively) long, continuous speech which would be difficult to assess using a method that looks for a comparison (since the first stimuli risks being forgotten as a respondent listens to the second) or a single score (since it is hard to say what part of the stimuli triggered the score). Here' we also expected to see the benefit of the diagnostic properties of ARS-based evaluation facilitating clear indications of actual problems in the stimuli. These expectations were borne out. We find that even the number-of-clicks analysis seemingly provides useful information, but more specifically, we see that the peak positions meaningfully indicate problems in the TTS. We also consider the method validated, as again the manufactured error is found and given a high rank, and the peaks largely coincide with the peaks of the previous experiment.

The main task in experiment 2 is to compare different methods attempting to achieve the same goal: slowed down speech. It is unclear, here, that diagnostics are required as we are looking to a property of the whole utterance, and unless we look to evaluate more lengthy stimuli, it is likely better to use a score, as in the original experiment. We note that the human reading is singled out as the "best" by both methods, but the rest of the ranks are quite disparate (see Table 8), and there is no reason to trust the ARS rank more than the score-based rank. Looking to the details, we see the largest difference in **SLOW-4**, which ranks last based on original scores but second based on click median. This particular method to slow down speech, where pauses at commas and full stops were extended in the extreme to reach the intended file length 30 seconds, was used as a sanity check in the original experiment. The ARS respondents did not click frequently at these absurdly long pauses but might have done so given a slightly different instruction. **SLOW-6**, where short pauses were inserted between each word, shows the second largest difference, ranked third in original experiment and last with ARS. ARS respondents clicked fairly regularly at the pauses in the beginning of the stimuli, causing a relatively large number of clicks but only one peak in the median click count. In general, there are click peaks that point to interesting occurrences in the stimuli, like peaks in the human stretched audio that seem to point to artifacts due to the stretching, but finding such occurrences was not the aim of the study.

In Experiment 3, the expectation in the original experiment, that the prominence-controlled TTS would outperform the one without prominence control on the experiment utterances, was borne out. The ARS-based method adds little value to the original experiment, however. We see that the 2 seconds pauses seem to make the respondents wait for the pause until they click, which makes the test almost identical to a scoring test, and completely eradicates its diagnostic strength. Here, too, a more specific instruction may help, but generally speaking, we find no real reason to abandon the original experiment method.

It is worth noting, however, that one of the problems found by ARS in experiment 1 point to exactly the type of problem that the systems in experiment 3 aims to resolve.

Some of the results are consistent with the sizeable effects of framing demonstrated in (Edlund et al., 2012): the lack of repeated clicks during absurdly long silences as well as the "wait for the pause" clicking behaviour observed with the pause-separated concatenated stimuli could likely be changed by a simple instruction.

6. Conclusion and future work

We conclude that ARS-based evaluation has highly desirable characteristics when the material is long and continuous and diagnostics are of interest; that it is not a particularly appropriate method for evaluating differences that affect the overall audio quality; and that it is not an appropriate method for evaluating single sentences. In the first case, it adds diagnostics and the ability to judge lengthy material in one continuous flow, similar to lengthy materials are usually consumed. In the two latter, it mimics or worse fails to repeat the strengths of other methods, without adding much. We also note that ARS-based evaluation may still be considerably more versatile than what we have shown, but this would require the instructions to be much more tailored to the task.

With respect to ARS and to most if not all other evaluation methods, the framing of the experiment is important and quite overlooked. It is likely that click rates and perhaps also positions are affected not only by what question is asked, but also by how the listeners perceive their role: are they representing themselves in a private capacity, themselves professionally, or some other group that they believe the experiment is designed to help? An informal post experiment survey with our participants showed that the vast majority thought their task was to represent themselves, but this could vary with different respondents in different settings. The most common question used in ARS experiments seem to be similar to the "yuck button" by in Poppe et al. (2011) – simply press when you dislike something, but we believe it would be valuable to systematically experiment with different types of questions and task.

For future work, we are chiefly interested in moving away from the web-based format and back to simultaneous testing in groups, where the respondents can be seen as one complex instrument, and where we are certain that they have operated under the same conditions. This is the environment the method was developed for, and where it is most likely to excel.

7. Acknowledgements

This work is funded in part by the Vinnova funded project Deep learning based speech synthesis for reading aloud of lengthy and information rich texts in Swedish (2018-02427), and by the Swedish Research Council project Connected (VR-2019-05003). The results will be made more widely accessible through the Swedish Research Council funded national infrastructure Språkbanken Tal (2017-00626).

8. References

Alpert, T., & Evain, J. P. (1997). Subjective quality evaluation: The SSCQE and DSCQE methodologies. EBU Technical Review, 12–20.

Cambre, J., Colnago, J., Maddock, J., Tsai, J., & Kaye, J. (2020). Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content. Procs. of CHI'20, 1–13. https://doi.org/10.1145/3313831.3376789

Camp, J., Kenter, T., Finkelstein, L., & Clark, R. (2023). MOS vs. AB: Evaluating text-to-speech systems reliably using clustered standard errors. Procs. of Interspeech 2023, 1090–1094. https://doi.org/10.21437/Interspeech.2023-2014

Chiang, C.-H., Huang, W.-P., & Lee, H. (2023). Why we should report the details in subjective evaluation of TTS more rigorously. INTERSPEECH 2023, 5551–5555. https://doi.org/10.21437/Interspeech.2023-416

Clark, R., Silen, H., Kenter, T., & Leith, R. (2019). Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs. Procs. of SSW10, 99–104. https://doi.org/10.21437/SSW.2019-18

Cooper, E., & Yamagishi, J. (2023). Investigating range-equalizing bias in mean opinion score ratings of synthesized speech. Procs. of Interspeech 2023, 1104–1108. https://doi.org/10.21437/Interspeech.2023-1076

Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. (2008). Towards human-like spoken dialogue systems. Speech Communication, 50(8), 630–645. https://doi.org/10.1016/j.specom.2008.04.002

Edlund, J., Hjalmarsson, A., & Tånnander, C. (2012). Unconventional methods in perception experiments. In Nordic Prosody: Proceedings of the XIth Conference, Tartu 2012. Peter Lang. https://doi.org/10.3726/978-3-653-03047-1

Edlund, J., Moubayed, S. A., Tånnander, C., & Gustafson, J. (2013). Temporal precision and reliability of audience response system based annotation. Proc. of Multimodal Corpora 2013, 6.

Edlund, J., Tånnander, C., & Gustafson, J. (2015). Audience response system-based assessment for analysis-by-synthesis. Proceedings of ICPhS.

Finkelstein, L., Camp, J., & Clark, R. (2023). Importance of human factors in text-to-speech evaluations. Procs. of SSW12, 27–33. https://doi.org/10.21437/SSW.2023-5

Hasegawa-Johnson, M., & Alwan, A. (2003). Speech coding fundamentals and applications. In Handbook of Telecommunications. John Wiley & Sons. http://www.isle.illinois.edu/speech_web_lg/pubs/2002/hasegawa-johnson02handbook.pdf

ITU-T. (1996). Methods for subjective determination of transmission quality. In International Telecommunication Union (Recommendation T-REC-P.800; Series P: Telephone Transmission Quality, Vol. 800, p. 22). ITU-T - Telecommunication standardization sector of ITU.

Kayyar, K., Dittmar, C., Pia, N., & Habets, E. (2023). Subjective evaluation of text-to-speech models: Comparing absolute category rating and ranking by elimination tests. Procs. of SSW12, 191–196. https://doi.org/10.21437/SSW.2023-30

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

Kirkland, A., Mehta, S., Lameris, H., Henter, G. E., Szekely, E., & Gustafson, J. (2023). Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation. Procs. of SSW12, 41–47. https://doi.org/10.21437/SSW.2023-7

Lameris, H., Kirkland, A., Gustafson, J., & Szekely, E. (2023). Situating speech synthesis: Investigating contextual factors in the evaluation of conversational TTS. Procs. of SSW12, 69–74. https://doi.org/10.21437/SSW.2023-11

Pandey, A., Edlund, J., Le Maguer, S., & Harte, N. (2023). Listener sensitivity to deviating obstruents in WaveNet. Proc of Interspeech 2023, 1080–1084. https://doi.org/10.21437/Interspeech.2023-1843

Perebinossoff, Philippe., Gross, B., Gross, L. S., & Vane, E. T. (2005). Programming for TV, radio, and the Internet: Strategy, development, and evaluation. Focal Press. https://books.google.se/books?id=dOo56mkYLU4C&dq=%22Audience+Studies+Institute%22

Polkosky, M. D., & Lewis, J. R. (2003). Expanding the MOS: Development and psychometric evaluation of the MOS-R and MOS-X. International Journal of Speech Technology, 6(2), 161–182. https://doi.org/10.1023/A:1022390615396

Pols, L. C. W. (1998). Speech synthesis evaluation. In J. Mariani (Ed.), Survey of the state of the art in human language technology (pp. 429–430). Cambridge University Press.

Poppe, R., Truong, K., & Heylen, D. (2011). Backchannels: Quantity, type and timing matters. International Workshop on Intelligent. http://link.springer.com/chapter/10.1007/978-3-642-23974-8_25

Prenger, R., Valle, R., & Catanzaro, B. (2019). Waveglow: A flow-based generative network for speech synthesis. Proc. of ICASSP 2019, 3617–3621. https://doi.org/10.1109/ICASSP.2019.8683143

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. Proc. of ICASSP 2018, 4779–4783. https://doi.org/10.1109/ICASSP.2018.8461368

Shirali-Shahreza, S., & Penn, G. (2018). MOS naturalness and the quest for human-like speech. In Procs. of SLT 2018, 346–352. https://doi.org/10.1109/SLT.2018.8639599

Sjölander, K., Sönnebo, L., & Tånnander, C. (2008). Recent advancements in the Filibuster text-to-speech system. Proc. of SLTC 2008. Proceedings of the Swedish Language Technology Conference (SLTC)., Stockholm.

Strömbergsson, S., Edlund, J., McAllister, A., & Lagerberg, T. (2021). Understanding acceptability of disordered speech through Audience Response Systems-based evaluation. Speech Communication, 131, 13–22. https://doi.org/10.1016/j.specom.2021.05.005

Strömbergsson, S., Holm, K., Edlund, J., Lagerberg, T., & McAllister, A. (2020). Audience Response System-based evaluation of intelligibility of children's connected speech – validity, reliability and listener differences. Journal of Communication Disorders, 87, 106037. https://doi.org/10.1016/j.jcomdis.2020.106037

Tånnander, C. (2012). An audience response system-based approach to speech synthesis evaluation. Procs. of SLTC 2012, 75–76.

Tånnander, C., & Edlund, J. (2021). Methods of slowing down speech. Proc. of SSW11, 43–47. ISCA-Archive. https://doi.org/10.21437/SSW.2021-8

Tånnander, C., House, D., & Edlund, J. (2022). Syllable duration as a proxy to latent prosodic features. Procs. of Speech Prosody 2022, 220–224. https://doi.org/10.21437/SpeechProsody.2022-45

Vazquez-Alvarez, Y., & Huckvale, M. (2002). The reliability of the ITU-P.85 standard for the evaluation of text-to-speech systems. Procs. of ICSLP 2002, 329–332. https://www.isca-speech.org/archive/icslp_2002/i02_0329.html

Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Eje, G., Maguer, S. L., Malisz, Z., Székely, É., Tånnander, C., & Voße, J. (2019). Speech synthesis evaluation—State-

of-the-art assessment and suggestion for a novel research program. Procs. of SSW10, September, 105–110. https://doi.org/10.21437/SSW.2019-19

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. Procs. of Interspeech 2017, 4006–4010. https://doi.org/10.21437/Interspeech.2017-1452

Zieliński, S., & Rumsey, F. (2008). On some biases encountered in modern audio quality listening tests: A review. Journal of the Audio Engineering Society, 56(6), 427–451.