



university of
 groningen

campus fryslân

Assessing the relationship between stimulus duration and Mean Opinion Score for speech synthesis evaluation

Brandi N. Hongell



university of
groningen

campus fryslân

University of Groningen - Campus Fryslân

**Assessing the relationship between stimulus duration and Mean Opinion
Score for speech synthesis evaluation**

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. Matt Coler (Voice Technology, University of Groningen)

Brandi Hongell (S5541727)

June 21, 2024

Acknowledgements

I am greatly appreciative to my supervisor, Matt Coler, for his continuous support not only during the thesis, but throughout the entire duration of the program. He fostered an environment that was perfect for my consistent wish to always learn and do more, and I am forever grateful for the opportunities he has afforded me whilst a student in this program. His drive to see me succeed and prompt responses to numerous e-mails have allowed me to proudly present this thesis. Additionally, I would like to thank Matt for suggesting the primary topic of my thesis.

I would also like to express my appreciation to each of the instructors in the program. Each instructor was patient and helpful as I navigated not only becoming a student again for the first time in 6 years, but also becoming a student in a new country. Thank you to Joshua, Matt, Phat, Shekhar, and Vass for fostering an enjoyable learning environment. The knowledge that I have learned from each of your courses played critical roles in the completion of my thesis.

In addition, I would like to give a very special thank you to my partner, Tim. He is the primary reason that I left my unfulfilling corporate job in America to take a chance on a new career path by joining this program. He has supported me in every way from the application process to the completion of this thesis.

Furthermore, I would like to acknowledge the support from my mother, Sylvia. She has done everything in her power and even beyond, from the day I was born, to help me succeed and accomplish my dreams. Thank you for always fighting for me and cheering for me, and thank you for always believing in me and pushing me to be my best—even when I didn't think it was possible. Additionally, thank you to my brother Ryan, and to my Nanny and Papa, for supporting me in this journey.

Lastly, I would like to express my gratitude to all the individuals who have played a part, no matter how small, including my fellow classmates, in shaping my academic journey and the successful completion of this thesis.

Abstract

Despite the rapid advancements in speech synthesis, the Mean Opinion Score (MOS), established in the 1990s and relatively unchanged since, remains the standard for evaluating speech synthesis. Lack of reassessment of MOS over time has raised many questions about the reliability and robustness of the field's predominant evaluation metric. Therefore, this study critically assesses how non-standardized testing variables may affect MOS, using listening tests to measure how four different durations of synthetic speech clips interact with the MOS ratings of three different synthetic voices. While the results show that duration does not have a statistically significant impact on the MOS of a synthetic voice, therefore producing inconclusive results, there is promise in continued research as shown by the 33.9% reported effect size. This moderately strong effect size suggests the possibility of a meaningful association between duration and MOS. Overall, this study highlights the lack of standardization present in MOS evaluation and the questions the reliability of this evaluation metric. It is therefore suggested to continue MOS research on not only duration, but also other unstandardized variables, as well as the implementation of best practices in MOS testing.

Keywords: speech synthesis, speech synthesis evaluation, mean opinion score, speech duration

Contents

1	Introduction	8
2	Literature Review	11
2.1	Mean Opinion Score	12
2.1.1	Origin and Significance	12
2.1.2	Limitations and Criticisms	13
2.2	Factors Affecting the Perception of Synthetic Speech of Longer Durations	14
2.2.1	Complexity of Utterances	14
2.2.2	Listener Fatigue	15
2.3	Conclusion	16
3	Research Question and Hypothesis	19
4	Methodology	22
4.1	Research Design	22
4.2	Experimental Design	22
4.2.1	Variables	22
4.2.2	Additional Considerations	23
4.3	Pilot Test	23
4.3.1	Participant Selection	23
4.3.2	Stimuli	24
4.3.3	Procedures	24
4.3.4	Summary of Results and Discussion	24
4.3.5	Refinements Based on Pilot Test	25
4.4	Primary Listening Test	25
4.4.1	Participant Selection	25
4.4.2	Stimuli	25
4.4.3	Procedures	26
4.5	Ethical Considerations	26
4.5.1	Voluntary Participation	26
4.5.2	Anonymity	26
4.5.3	Participant Comfort	27
4.5.4	Bias Prevention	27
4.5.5	Use of AI	27
4.5.6	Replicability	27
5	Results	30
5.1	Presentation of Results	30
5.2	Interpretation and Analysis of Results	31
5.2.1	Visualized Data	31
5.2.2	Statistical Analysis	32
5.3	Reflection on the Hypothesis	32
5.4	Limitations of the Study	33

6	Conclusion	36
6.1	Key Findings	36
6.2	Future Work	36
6.2.1	Expanded Duration Study	36
6.2.2	Standardization of MOS	37
6.2.3	Research of Different Variables	37
6.3	Broader Impact on Speech Synthesis Evaluation	38
	References	39
	Appendices	41
A	The North Wind and the Sun	41
B	Questionnaire Workflow	42
C	Pilot Questionnaire and Stimuli	44
D	Primary Listening Test Questionnaire and Stimuli	47

1 Introduction

Recent advancements in neural network-based speech synthesis have improved both the quality and accessibility of synthetic voices. Additionally, the development of end-to-end speech synthesis models like WaveNet and Tacotron have made these high-quality, natural synthetic voices more accessible with the ability to generate speech directly from text without the need for intermediate vocoding steps, as noted by Mehrish, Majumder, Bhardwaj, Mihalcea, and Poria (2023). These advanced TTS systems have become integral in the daily lives of many people, being used in capacities such as virtual assistants like Siri and Alexa, language learning software like Duolingo, and accessibility software such as Voice Dream Reader for individuals with disabilities, like visual impairment.

One aspect of speech synthesis that has seen notable improvement in recent years is the quality and naturalness of the voices generated by these advanced TTS systems. The evaluation of speech quality is an essential practice for every TTS model to ensure that the voices produced are pleasant to listen to, and therefore, desired by listeners. One of the most widely used metrics for speech synthesis evaluation is the Mean Opinion Score (MOS), the output of the Absolute Category Rating (ACR) system. To calculate MOS, listeners are asked to rate the overall quality of speech samples on a 5-point Likert scale, where 1 represents poor and 5 represents excellent. The MOS is then calculated as the arithmetic mean of all the individual ratings provided by the human raters. MOS sees popularity primary due to the simplicity of its implementation and its ability to provide a quantitative measure that can be used in many applications, such as comparing different systems or changes in a system over time.

Despite its presence as the predominant metric in speech synthesis evaluation, MOS has remained relatively unchanged since its inception in the 1990s via listening test protocol presented in the ITU P.800 recommendation. The metric's failure to evolve alongside the speech which it evaluates has led many researchers to criticize MOS and question its reliability as an evaluation tool. Two of the most notable works contributing towards this conversation come from Le Maguer, King, and Harte (2024) and Kirkland et al. (2023). These works assess the MOS methodology and highlight the highly sensitive nature of the metric as well as its inconsistent use overall, due to a lack of standardization of MOS listening tests. While these downfalls are apparent, there is a need to systematically investigate and understand **how non-standardized variables impact MOS scores across voices of all qualities.**

Motivated by this gap, this thesis further examines the limitations inherent in MOS, particularly focusing on **how the duration of speech samples influences its reliability and validity as an evaluation tool.** To achieve this, this study will conduct listening tests measuring listeners' perception of naturalness across four different durations spanning three different synthetic voices. By doing so, this study seeks to provide valuable insights to the developing conversation surrounding MOS, as well as contribute research-backed suggestions supporting the standardization of the MOS evaluation metric.

The results of this study are significant for several reasons. Firstly, this research contributes to the foundation of a trending topic in the field of speech synthesis by addressing an area of MOS listening tests that currently has no directly-related research. These findings will contribute to the ongoing discussion surrounding accurate, reliable, and fair evaluation of synthetic speech through the use of MOS. Secondly, deeper understanding of factors affecting listeners' perception of naturalness can aid in the development of new synthetic speech technologies. Moreover, insights from the study may assist in future research seeking to build industry standards and best practices for speech

synthesis evaluation. Overall, this thesis hopes to aid in kickstarting the advancement and development of the relatively static metric, MOS, which will, in turn, better the field of speech synthesis as a whole.

Now that a brief introduction and motivation for this research has been presented, the structure of the thesis is the following: section 2 provides an extensive literature review that sets the stage for the research question and hypothesis presented in section 3. In section 4, the methodology is covered, including experimental setup and design, as well as the pilot test. Then, section 5 presents the results from the primary experiment alongside a statistical analysis and discussion including implications on the hypothesis and limitations of the study. Lastly, section 6 summarizes the thesis and presents the conclusions drawn, along with recommended future work and perceived impact on the field of speech synthesis.

2 Literature Review

This section is dedicated to providing a comprehensive review of the literature pertaining to the evaluation of synthetic speech, with a specific focus on the Absolute Category Rating (ACR) and its output, the Mean Opinion Score. In addition to this, this review will consider how synthetic speech outputs may be affected by the complexity of their utterances, as well as how humans perceive audio in differing durations. By conducting a thorough and critical analysis of the literature in this field, this review aims to offer valuable insights into the limitations of the MOS evaluation method and highlight potential gaps that may be filled by research, particularly the effect of speech clip duration on MOS.

To those ends, the section is structured as follows. To begin, I will delineate the databases and search terms used during the comprehensive literature search described above and describe the inclusion/exclusion criteria used in selecting the literature. After that, I offer a succinct overview of the key findings and contributions of the selected papers. This will be followed by a formal conclusion in section

To allow for comprehensive coverage of the literature, five primary platforms were employed. Firstly, the International Speech Communication Association (ISCA) Archive was used. The ISCA Archive was chosen as it is the largest open collection of research papers in speech communication. To capture works that may fall outside of this database, searches were also conducted on ScienceDirect, arXiv, and IEEE Xplore. Lastly, Google Scholar was used to capture works like grey literature, which may not be listed within the aforementioned traditional databases.

Due to the differing search engines amongst the databases, the search criteria varied slightly between each chosen database. Below, I have listed each database and the corresponding search criteria to ensure replicability of the review. Searches on ISCA Archive were conducted within the Papers table. Searches on ScienceDirect used the 'Title, abstract or author-specified keywords' field. Searches on arXiv searched within the 'Abstract' of works. Searches on IEEE Xplore searched within 'Author keywords'. All searches were conducted in May 2024.

- **ISCA Archive:** 'duration', 'evaluation', 'listening test', 'mean opinion score', 'MOS'
- **ScienceDirect:** 'mean opinion score' AND 'speech synthesis', 'MOS' AND 'speech synthesis', 'speech synthesis' AND 'evaluation', 'speech synthesis' AND 'duration', 'speech synthesis', 'listener fatigue'
- **arXiv:** 'mean opinion score' AND 'speech synthesis', 'MOS' AND 'speech synthesis', 'speech synthesis' AND 'evaluation', 'speech synthesis' AND 'duration', 'speech synthesis', 'listener fatigue'
- **IEEE Xplore:** 'mean opinion score' AND 'speech synthesis', 'MOS' AND 'speech synthesis', 'speech synthesis' AND 'evaluation', 'speech synthesis' AND 'duration', 'speech synthesis'
- **Google Scholar:** 'mean opinion score mos speech synthesis', 'mos speech synthesis evaluation', 'mos synthetic speech evaluation', 'mean opinion score listening test', 'speech synthesis complexity', 'synthetic speech duration', 'listener fatigue', 'survey fatigue'

After assessment of eligibility, 13 results were ultimately selected to be included in the qualitative synthesis. To be eligible for inclusion, all works must have met the following set of inclusion criteria:

1. Selected works must be written in English.
2. The works must be peer-reviewed journal articles, conference papers, or academic books. Non-peer-reviewed sources, such as blogs, websites, and forum posts were not considered.
3. Works that primarily focus on the evaluation of synthetic speech must utilize MOS. Variants of MOS, such as DMOS and CMOS, were not considered for this review.

By applying these filters and exclusion criteria, I aimed to ensure the inclusion of the most pertinent and up-to-date literature directly related to MOS, factors affecting the perception synthetic speech outputs, and the potential effects of varying clip durations on human listeners' perceptions, aligning with the research objectives and scope of this study.

The literature review is organized into different subsections, based on the general topic of which they are part. Subsection 2.1 discusses the literature regarding the Mean Opinion Score as an evaluation tool, including its origin, significance and limitations. Subsection 2.2 discusses factors that may affect listeners' evaluation of synthetic speech, including particular characteristics of synthetic speech and listening test design.

For simplicity and readability, table 1 provides a full list of references appended with notes, sorted by order of appearance in the following subsections of the literature review.

2.1 Mean Opinion Score

The literature in this section primarily serves two purposes. Firstly, it is important to define MOS and highlight its significance as an evaluation tool of synthetic speech. Secondly, it is important to understand its downfalls and limitations to understand gaps in research that stand to be filled.

2.1.1 Origin and Significance

The Mean Opinion Score is the pre-dominant metric used for the evaluation of the quality and naturalness of synthetic speech generated by text-to-speech (TTS) systems. MOS and the Absolute Category Rating (ACR) protocol for speech synthesis evaluation originated with the publication of the ITU-T P.800 recommendation from the International Telecommunication Union (1996). The recommendation describes methods and procedures for the evaluation of quality of speech codecs and transmission systems through many different types of tests, including listening tests. The recommendation utilizes the ACR protocol, wherein listeners are presented individual speech samples. These listeners then rate the overall quality of each sample on a 5-point Likert scale, ranging from 1 (bad) to 5 (excellent). The ratings from all listeners for each sample are then averaged to obtain the MOS. This recommendation has since been adopted as a framework for speech synthesis evaluation.

MOS sees widespread use as a synthetic speech evaluation tool due to its ability to easily quantify the perceived naturalness and quality of TTS systems. In a survey conducted by Kirkland et al. (2023), approximately two-thirds of speech synthesis papers from INTERSPEECH 2021, INTERSPEECH 2022, and SSW 2021 used MOS for evaluation. Beyond the evaluation of isolated TTS systems, MOS is also commonly used to compare the performance of TTS systems as well as to track the progress of a TTS system over its development life cycle.

2.1.2 Limitations and Criticisms

Despite its stronghold as an evaluation method, MOS does not exist without criticism. A recent paper by Le Maguer et al. (2024) highlights many limitations of MOS. In this work, the authors look to determine the reliability of the Absolute Category Rating protocol and MOS by conducting four experiments, following the 2013 edition of the Blizzard Challenge. These experiments were created to answer questions about the stability of MOS over time on the same TTS system, how scores of lower quality systems influence scores of higher quality systems, how new technologies affect the scores of older TTS systems, and how the MOS of modern TTS systems evolve in isolation. The results of the experiments concluded in the confirmation of the relative nature of MOS in spite of its beginnings as an Absolute Category Rating.

One of the primary limitations highlighted by Le Maguer et al. (2024) is that the ACR protocol **does not impose the use of anchors**. In this context, anchors refer to reference samples or signals included in the set of evaluation stimuli. This lack of anchors leads to listeners choosing their own implicit anchors, further supporting the authors' conclusion of MOS being a relative metric. The mandated inclusion of standardized anchors is recommended for reliable use of MOS. Similarly, the presence of low and high quality stimuli in the evaluation set lead to listeners rating samples in relativity to the range of the samples present.

A previously referenced paper by Kirkland et al. (2023) also offers a critical analysis of MOS. Particularly, this research showcases the **inconsistent use of MOS and the underreporting of listening test details**. The experimentation conducted in the study demonstrates that variations in the design of MOS listening tests significantly impact the MOS results. One common variation in MOS listening tests is what aspect of the stimuli the participants are meant to evaluate. In the previously mentioned survey of papers conducted by this study, over half of the papers measured naturalness while 22.6% of papers measured perceived quality. A staggering 16.5% of surveyed papers did not explicitly state what the MOS listening test was measuring. The study goes on to mention that although the surveyed papers state what they intend to measure, it is not clear whether the participants were aware of this measure due to a lack of reporting of listening test questions.

In addition, the paper from Kirkland et al. (2023) also discusses the underspecification of the MOS scale increments and the scale labels. Most of the surveyed works did not make any specifications in either area. Of those that did specify, nearly half of the studies used full-point increments in their ratings while the other half used half-point increments. In terms of the scale labels, although most did not specify, most of those that did used a scale with labels recommended by the International Telecommunication Union (1996) standard. The standard recommendation for labels is: 1 (Bad), 2 (Poor), 3 (Fair), 4 (Good), and 5 (Excellent). For their experiment, the authors designed 4 MOS listening test which varied in terms of scale increments and what aspect of the speech samples the participants were asked to evaluate (overall quality or perceived naturalness). The findings showed that each of these changes did have meaningful effects on the MOS outcomes. The study concludes by offering concrete reporting suggestions for researchers, conference organizers and reviewers.

The lack of reporting of listening test details is also discussed in a study by Chiang, Huang, and yi Lee (2023) wherein the authors surveyed all papers from INTERSPEECH 2022 that cover speech synthesis, excluding those that did not use MOS evaluation. This survey looked at whether or not the papers reported on a defined set of factors, and the following experiment tested to see if the differing factor affect the results of the MOS evaluation on the quality of speech samples. These factors

included the recruitment platform, language background and geographic location of participants, qualification of the participants, instructions given to participants, and number of participants and speech samples. The experiments found that the variance in these factors were highly influential to the results.

A study by O'Mahony, Oplustil-Gallegos, Lai, and King (2021) also found that the **wording of instructions** has an impact on listeners' ratings of presented stimuli, though this study specifically experimented on synthetic speech in context. Nevertheless, the study found that these variations, as well as prosodic differences, both in context and in isolation, affected the evaluations by the participants. Another study on contextual factors by Cooper and Yamagishi (2023) found that MOS tests are further affected by the overall range of quality of the presented stimuli, a phenomenon known as range-equalizing bias.

The lack of standards in MOS listening tests is further highlighted by studies such as one by Wester, Valentini-Botinhao, and Henter (2015) which reports that over 60% of papers evaluating synthetic speech at INTERSPEECH 2014 concluded results with less than 20 listeners. The papers goes on to show that a stable level of significance when measuring naturalness with MOS can only be reached with more than 30 participants. In addition, a study conducted by Rosenberg and Ramabhadran (2017) highlights how biases, primarily participant bias and utterance bias, can affect the outcomes of MOS. The authors caution against relying on outcomes that do not account for biases.

From the research that has been conducted on MOS and the many variables affecting it, it is apparent that there are many unexplored variables that may also affect the reliability of MOS. For this thesis, I have chosen to focus on the possible effects of speech clip duration on MOS.

2.2 Factors Affecting the Perception of Synthetic Speech of Longer Durations

This section of the review serves to explore possible factors that may affect listeners' perception of synthetic speech, primarily as a consequence of a longer utterance of speech.

2.2.1 Complexity of Utterances

There is some literature discussing how the utterance length of synthetic speech may affect the perception of quality and naturalness. A study by Nusbaum and Pisoni (1985) suggests that the lack of redundant acoustic-phonetic cues in synthetic speech that are present in natural speech may affect listeners' perception of naturalness. Particularly, as utterances get longer, the lack of these cues may become more apparent to listeners. In addition, longer synthesized speech utterances introduce more opportunities for error to occur and accumulate.

A study by Clark, Silen, Kenter, and Leith (2019) states that synthetic speech utterances are traditionally rated in isolation rather than in longer contexts. The authors suggest that providing longer utterances to listeners may affect their evaluations by increasing their cognitive load. This claim is further supported by the Nusbaum and Pisoni (1985) study. The context study specifically evaluates long-form TTS utterances in three contexts: sentences in isolation (traditional), full paragraphs, and context-stimulus pairings. The authors found that the results differed across each context. The researchers also prove that it is difficult and inconclusive to derive the MOS of a paragraph from the MOS of its individual sentences.

A previously referenced study by O'Mahony et al. (2021) mentions the need for prosodic variation in newer TTS systems to keep up with the standards of naturalness. However, longer utterances

need further considerations for these variations. Synthetic voices that are trained on isolated sentences may struggle to create context-appropriate prosody for longer utterances, which may further reduce the perceived naturalness by listeners.

2.2.2 Listener Fatigue

Listening fatigue is a condition that sometimes occurs after prolonged exposure to auditory stimuli. Similarly, survey fatigue occurs when participants in a survey (for the purposes of this thesis, a listening test) become unwilling or unmotivated to provide feedback, which may affect the quality of responses. An article by Sinickas (2007) suggests that survey fatigue can be avoided by following a set of recommended protocol, notably by presenting shorter surveys. Another article by Jeong et al. (2023) supports this protocol, reporting a correlation between survey participation and the amount of time spent on the survey.

While most literature around listening fatigue concerns listeners with hearing impairments, the issue is still important to consider, as synthetic speech should ideally be accessible to all who wish to use it. Further, in many situations, listeners with hearing loss may use synthetic speech as a helpful tool. Results from a journal article by Hornsby (2013) suggest sustained speech-processing demands can lead to mental fatigue in adult persons with hearing loss, yet also notes that additional research is needed.

In a work by Sarem, Marashi, and Siyyari (2019), the authors study the relationship between L2 listening comprehension and listening fatigue among Iranian intermediate EFL learners. This research finds that the listening comprehension of these participants is adversely affected by their listening fatigue levels. This is important to consider as many L2 learners utilize speech synthesis tools to communicate and learn. Therefore, it is important to consider these users in evaluation.

This concludes the literature review section, which provides an extensive overview of the prior research on Mean Opinion Score, speech synthesis and its evaluation, and human factors such as listener and survey fatigue. This review covers the foundational framework of MOS as a synthetic speech evaluation tool and highlights its prominence in past and current speech synthesis research. The majority of discussion lies within the lack of standards of MOS listening tests and the under-reporting of listening test variables. This leads to a high level of potential manipulation of MOS scores. Further, the review covers potential reasons for variable scores that may exist due to the capabilities of synthetic speech production itself as well as human factors such as listening fatigue and survey fatigue.

While previous studies have contributed to our understanding of the unreliability of the Absolute Category Rating protocol and the Mean Opinion Score, there are still gaps in knowledge to address. This thesis, particularly, will focus on the manipulation of one unstandardized variable in MOS listening tests, duration of speech clips. This experiment will attempt to establish a correlation between MOS and duration while keeping all other variables static. The design of the listening test and utterances will also be created with consideration to the literature over fatigue to promote a high number of enthusiastic responses, which will in turn produce meaningful results. The methodology that helps address this knowledge gap will be described in the following section.

2.3 Conclusion

In conclusion, research on synthetic speech evaluation and the human perception of synthetic speech leave many gaps and questions to be answered about whether differing variables impact each other and if so, to what degree. While the table 1 offers a high level overview of these topics, this conclusion serves to draw connections between each of them.

The literature assessing Mean Opinion Score as an evaluation method highlights many limitations and downfalls of the predominant evaluation tool for speech synthesis. Criticism of MOS exists in many forms, such as the lack of detailed reporting as researched by Kirkland et al. (2023) as well as Chiang et al. (2023). MOS also suffers a high level of sensitivity to variations, such as the presence and absence of anchors, as studied by Le Maguer et al. (2024). It is clear from this critical literature that there are many variables possibly affecting MOS that are still yet to be researched, and MOS is in severe need of standards in order to remain relevant and reliable as an evaluation tool for speech synthesis. This thesis seeks to understand how varying speech clip durations may affect the MOS of a synthetic voice as a means to understand further variables affecting MOS.

To support this duration study, it was important to seek out literature that may answer why duration could affect MOS. To answer this, two avenues were explored. The first, complexity of longer synthetic speech utterances, is primarily supported by research in two areas. Firstly, Nusbaum and Pisoni (1985) discusses certain key acoustic cues present in human speech that are not present in synthetic speech. From this, it can be inferred that longer exposure to synthetic speech gives humans more time to notice the absence of these cues. Secondly, Clark et al. (2019) discusses how long-form TTS is generally evaluated in the form of isolated sentences and suggests that longer utterances may lead to increased cognitive load.

In addition to this, I wanted to consider the human experience of listening to longer instances of synthetic speech. This topic was primarily covered by research over survey fatigue, as detailed by Sinickas (2007). This article specifically recommends the use of shorter surveys to increase honest and enthusiastic responses from participants. Longer synthetic speech clips logically create longer listening tests, so this is an important factor to consider. While I was unable to find any specific literature discussing listener fatigue in general, research by Hornsby (2013) and Sarem et al. (2019) highlight the existence of listener fatigue in special groups such as hearing aid users and second language (L2) listeners. Synthetic speech is largely used as an accessibility tool, so it is important to consider these groups of listeners for potential evaluation.

After the review of this literature primarily focused on lengthier audio, it is apparent that there is a lack of research surrounding the duration of synthetic speech and how it affects not only the direct perception of synthetic speech, but also how longer synthetic speech clips may affect the listeners' experience in practice and participation in listening tests for evaluation. The research question that has been logically derived from this review will be presented in the following chapter, as well as the hypothesis.

Table 1: List of references for subsections 2.1-2.3, summarized

Reference	Brief description	Subsection
International Telecommunication Union (1996)	ITU-T Recommendation P.800, foundational framework for MOS	2.1
Kirkland et al. (2023)	A critical analysis of MOS test methodology in TTS evaluation, highlighting the inconsistent use and underreporting of listening test details	2.1
Le Maguer et al. (2024)	A study to determine the reliability of MOS, showing that it is a relative score that is highly sensitive to variations. Calls for a new standard in speech synthesis evaluation.	2.1
Chiang et al. (2023)	Discusses the underreporting of listening test details and how influential these details are to the results of MOS	2.1
O'Mahony et al. (2021)	How MOS is affected by different variables of synthetic speech in context	2.1
Cooper and Yamagishi (2023)	The effect of the range of speech clip quality on the overall MOS of a TTS system	2.1
Wester et al. (2015)	Measuring the effect of the number of participants in a listening test on the stability of the the MOS	2.1
Rosenberg and Ramabhadran (2017)	The effect of different types of biases on the levels of participant and utterance on the MOS	2.1
Nusbaum and Pisoni (1985)	Identifies factors that must be considered when evaluating the human perception of synthetic speech	2.2.1
Clark et al. (2019)	Discussion of the evaluation of long-form TTS synthesis	2.2.1
Sinickas (2007)	Article discussing common causes of survey fatigue and proposed solutions	2.2.2
Hornsby (2013)	Article discussing listening effort and mental fatigue associated with sustained speech processing demands amongst hearing aid users	2.2.2
Sarem et al. (2019)	Listening fatigue in L2 listening comprehension among intermediate Iranian EFL learners	2.2.2

3 Research Question and Hypothesis

In light of the preceding review, it is apparent that there are many unexplored variables in MOS listening tests that may effect its reliability as an evaluation tool. MOS, which has existed since the 1990s, serves as the predominant metric for the evaluation of synthetic speech, and has served as an anchor in the rapid evolution of synthetic speech over the past 30 years. However, MOS itself has failed to evolve alongside the technology it evaluates, and therefore, it is important to assess and criticize the evaluation metric itself before it get left behind by the continued development of speech synthesis. Critiques have begun to surface, such as Le Maguer et al. (2024) noting the lack of standardized anchors and Kirkland et al. (2023) exposing the inconsistent use of MOS overall. This thesis aims contribute to this conversation by investigating an unknown relationship in MOS testing: duration. The insights from Nusbaum and Pisoni (1985) on acoustic cues, Clark et al. (2019) on cognitive load, and Sinickas (2007) on survey fatigue converge to suggest that longer synthetic speech clips could negatively impact MOS. These findings underscore the need to systematically investigate the duration variable in synthetic speech evaluation.

Thus, we arrive at our central inquiry; the research question at the core of this study:

How does the duration of synthetic speech clips affect the Mean Opinion Score across different speech quality levels?

My hypothesis is that the MOS will decrease as the duration of the speech clips increase, particularly when evaluated across clips of varying perceived quality levels. I predict the effect to be measurable and significant. The review in section 2 revealed that MOS is highly sensitive to variations, such as the presence and absence of anchors, showcased by Le Maguer et al. (2024). Additionally, a study by Clark et al. (2019) shows that traditional utterances for synthetic speech evaluation are conducted in isolation and suggest that longer utterances may affect the evaluation negatively, however they do not conduct an experiment on this assumption.

Outside of the pitfalls of MOS, longer synthetic speech clips allow listeners more time to process and analyze the clip which they are evaluating. One factor that may be subconsciously noticed by evaluators is the general lack of redundant acoustic cues in synthetic speech that are present in human speech, as suggested by Nusbaum and Pisoni (1985). Additionally, the lack of prosodic variation in longer utterances, may further reduce the perceived naturalness by listeners, among other factors.

Logically, longer synthetic speech clips will also lead to longer listening tests. It is important to consider how this factor may negatively affect participants, particularly pertaining to the phenomenon of listening fatigue. Sinickas (2007) discusses how longer surveys often lead to participants becoming unwilling and/or unmotivated to provide feedback. This may lead to issues such as dishonest answers, lower ratings, or lack of answers at all. Additionally, longer listening tests may have a negative effect on participants of specialized groups, such as hearing impaired or second-language learners. Studies by Hornsby (2013) and Sarem et al. (2019) suggest this increased cognitive load may negatively impact listeners, therefore potentially affecting evaluation scores. With speech synthesis largely being used as an accessibility or learning tool, it is important to consider these groups when evaluating TTS systems.

If my hypothesis falsified, it would indicate that the duration of synthetic speech clips does not significantly impact the MOS. This would mean listeners would not be particularly influenced by the length of the speech clips they are evaluating. Additionally, a falsification of the hypothesis would suggest that MOS as an evaluation tool may be more robust and reliable than predicted, and

therefore, MOS can be reliably used for evaluating synthetic speech quality regardless of the length of its utterances. Lastly, a falsification of this hypothesis would open new research opportunities towards specific factors that do significantly influence MOS.

In the following chapter, I will discuss the methodology used to test this hypothesis, including the research design, experimental design, pilot test, primary experiment, and ethical considerations made throughout the research.

4 Methodology

In this section, I will outline the methodology used to address the research question and validate the hypothesis. First, in subsection 4.1, I will discuss the research design of this thesis, including the research approach and the justification for the experimental design. Next, subsection 4.2 will focus on the experimental setup, including the design and variables of the experiment. After this, in subsection 4.3, I will detail the pilot test used to determine suitability of the experiment, including refinements made after pilot feedback. Logically following, subsection 4.4 will then discuss the primary experiment of this thesis, the listening test. Finally, in subsection 4.5, I will reflect on the ethical considerations inherent in conducting this research.

4.1 Research Design

This thesis employs a quantitative experimental design to answer the research question presented in section 3, how the duration of synthetic speech clips. This approach has been chosen as it allows for the collection of numerical data necessary to measure MOS, which can further be analyzed to identify patterns and correlations and test the hypothesis, also presented in section 3.

The justification of this experimental design lies in that it provides a controlled environment where in the effect of speech clip duration on MOS can be isolated. This level of control minimizes the influence of other variables, which have been shown to effect MOS, as discussed in the literature review, particularly in subsection 2.1.2. Controlling these variables increases the validity of the study. Further, the quantitative nature of the experimental design allows for the use of statistical analyses to test the hypothesis and generalize findings among a broader context.

4.2 Experimental Design

The experimental setup involved creating speech clips of varying durations. The chosen durations were: single-word utterance, short phrase, single sentence, and multi-sentence. These durations were chosen to cover a range of listening experiences from very short to relatively long speech samples. The longer length of multi-sentence was chosen in contrast to the traditional testing method of sentences in isolation, as referenced by Clark et al. (2019). This literature also motivated the choice of qualitative duration categories rather than quantitative, timed durations. The text selections were derived from one of Aesop's Fables, The North Wind and the Sun. The full text selection can be seen in the appendix, A. This text was chosen due to its prominence as a commonly chosen text for phonetic transcriptions.

4.2.1 Variables

The independent variable in this experiment is the duration of the presented speech clips, which is hypothesized in section 3 to have a measurable effect on the dependent variable, MOS. Several control variables were also considered to ensure the validity of the reported results. Below I have detailed these control variables:

- **Synthetic voice:** Although this study does measure effects across multiple voices, each participant was only presented one voice to evaluate. The voices were selected randomly, yet

distributed evenly. The randomization and distribution algorithms were stock options provided by the Qualtrics software. The exact workflow and logic of the listening tests can be found by opening the associated QSF files in Qualtrics; these files are available in the GitHub repository¹. A visual representation of the workflows can also be found in Appendix B.

- **Text prompts:** Each voice was generated using the same text prompt for each duration.
- **Question and selection wording:** Each question was presented with the same wording, and each set of answers was also the same. The only variance among questions was the associated clip.
- **Testing platform:** All listening tests were conducted using Qualtrics survey software.

4.2.2 Additional Considerations

In addition to these variables, it is important to note that many specific test design choices were considered and should be documented for replicability and transparency of the experiment. Firstly, this experiment explicitly uses MOS to measure the perceived naturalness of the speech clips. This choice was made due to the majority of surveyed papers by Kirkland et al. (2023) reporting to use MOS in this way.

A 5-point Likert scale was used with whole integer increments, and the labels used were those recommended by the International Telecommunication Union (1996) recommendation. The labels can be found in Appendix D. Further, no granularity was provided with the scale. This was a conscious choice to highlight the lack of standards in MOS testing, specifically the issue of lack of anchors, described by Le Maguer et al. (2024). While acknowledging this may affect the scores of individual participants, the research question presented in subsection 3 is interested in the effect of duration across varying levels of naturalness, therefore this variance should not affect the results. To further highlight the lack of standards in MOS testing, participants were not limited to the number of times they could listen to the stimuli, possibly introducing recall bias.

Lastly, demographic data of participants was not considered in the participant selection process, and no demographic data was collected. This was a conscious choice particularly due to my personal wish for synthetic speech to be inclusive and accessible to all that wish to use it. Therefore, as this experiment is a generalized listening test not targeted towards and specific listener groups, demographic data was not deemed necessary.

4.3 Pilot Test

Prior to the experiment, a pilot test was conducted to assess the suitability of the study, as well as to offer an opportunity to refine the research design. It additionally served as method to ensure the reliability of the data collection process with the survey software and identify any potential issues.

4.3.1 Participant Selection

A small sample of 18 participants was selected for the pilot test. Participants were recruited via word of mouth and social media. There were no inclusion or exclusion criteria set for the selection

¹<https://github.com/branaphy/msc-vt-thesis>

of participants.

4.3.2 Stimuli

The stimuli for the pilot test were created from two voices. The first voice was a synthetic voice generated using a pre-trained model of MMS TTS from Pratap et al. (2023). The text prompts used to produce the stimuli can be found in Appendix C. The second voice was a human voice, with the recording being provided with consent by a classmate. The human voice was not altered in any way outside of being trimmed to fit the varying durations.

4.3.3 Procedures

Participants were presented with 4 speech clips, each of a unique duration. The clips were presented in a randomized order to each participant to prevent order effects. Additionally, each participant was only presented with a clip set from one of the two voices. The voice presented to the participant was chosen at random, yet distributed evenly. Additionally, participants received, at random, one of two possible stimuli for the phrase and sentence durations.

Participants were asked to rate the naturalness of each speech sample on a scale of 1 to 5, with 1 representing Bad and 5 representing Excellent. There were no additional labels on the numbers. In addition, participants were required to rate the speech clips one at a time and were unable to return to past questions to limit comparison. The full question format can be seen in Appendix C.

4.3.4 Summary of Results and Discussion

The results for each voice are presented below in summary tables 2 and 3.

Table 2: Summary of results for the MMS voice

Duration	Mean	Std	Min	Max	Count
Word	3.22	1.03	2.00	5.00	9
Phrase	3.11	0.93	2.00	4.00	9
Sentence	2.33	0.76	1.00	4.00	9
Multi-sentence	1.44	0.50	1.00	2.00	9

Table 3: Summary of results for the human voice

Duration	Mean	Std	Min	Max	Count
Word	1.56	0.68	1.00	3.00	9
Phrase	3.22	1.15	2.00	5.00	9
Sentence	4.00	1.15	2.00	5.00	9
Multi-sentence	4.00	1.33	1.00	5.00	9

The results suggest that the hypothesis presented in section 3 may be supported for instances of synthetic speech, but not for human speech. The MOS scores for the synthetic voice trended downwards as the utterances became longer, and the human voice saw the opposite effect. I expect that the primary cause of this stems from the differences between human and synthetic speech highlighted in section 2.2.1.

An additional consideration to make is that the utterances from MMS were created in isolation, while the utterances of human speech were trimmed from one continuous passage. This is the most likely explanation for the increasing MOS scores alongside increasing context.

4.3.5 Refinements Based on Pilot Test

Based on the results and feedback from the pilot test, several refinements were made for the primary listening test:

- **Instructions:** Changes were made to the instruction page to enhance the presentation value and professionalism of the survey. Additionally, an 'Important to Note' section was also added.
- **Stimuli:** The stimuli was adjusted to consist of 3 different synthetic voices for more coverage of voice qualities, including the introduction of female voices and different speech rates. Human voices were removed from the scope of the research, as the results of the human voice did not support the hypothesis. Additionally, while the number of presented stimuli remained at 4, the number of options per duration category was reduced to 1 for consistency and simplicity.
- **Labeling:** Labels were added to each individual answer choice, following the recommendation from International Telecommunication Union (1996).

4.4 Primary Listening Test

Following the pilot listening test, an expanded listening test was performed to further explore the potential relationship between speech clip duration and the evaluation of synthetic voices.

4.4.1 Participant Selection

Participants were recruited via word of mouth and via social media platforms, Facebook, LinkedIn, and Instagram. There were no explicit inclusion or exclusion criteria set for participants. A total of 110 participants were recruited for this research, ensuring the validity benchmark of 30 participants per voice as detailed by Wester et al. (2015) was met.

4.4.2 Stimuli

The speech clips were created across three different synthetic voices to test how duration affects MOS across different speech quality levels. The selected voices were chosen by reviewing the trending pre-trained models on Hugging Face². Two of the chosen voices were created using Parler-TTS Mini: Espresso from Lacombe, Srivastav, and Gandhi (2024). This is a fine-tuned model trained on the Espresso dataset from Nguyen et al. (2023). The original model is a reproduction of work

²https://huggingface.co/models?pipeline_tag=text-to-speech&sort=trending

by Lyth and King (2024). This model offers 4 voices, 2 male and 2 female. Additionally, the model offers variances in emotion and speaking rate. Manipulating each of these variables allows for the produced voices to be distinct from each other, despite being generated from the same pre-trained model. The third chosen voice was created using the SpeechT5 model from Ao et al. (2022). The text prompts used to produce the stimuli can be found in Appendix D.

4.4.3 Procedures

Participants were presented with 4 speech clips, each of a unique duration. The clips were presented in a randomized order to each participant to prevent order effects. Additionally, each participant was only presented with a clip set from one of the three voices. The voice presented to the participant was chosen at random, yet distributed evenly.

Participants were asked to rate the naturalness of each speech sample on a scale of 1 to 5. The answers were labeled with the representation of their perceived naturalness, using the recommendation from International Telecommunication Union (1996). In addition, participants were required to rate the speech clips one at a time and were unable to return to past questions to limit comparison. The full question format can be seen in Appendix D.

4.5 Ethical Considerations

It is of utmost importance to ensure that research is always conducted responsibly and ethically. This is particularly important for this thesis, as the experiment relied on human participation. Prior to experimentation, ethical clearance was granted for this research by the Ethics Committee at the University of Groningen - Campus Fryslân³. In this section, I will address ethical considerations made across all stages of my research.

4.5.1 Voluntary Participation

Firstly, participation in both the pilot test and the primary listening test was entirely voluntary. Participants were not given any sort of compensation as motivation to participate in the studies. Additionally, participants were freely able to withdraw from the study at any time without any consequences.

4.5.2 Anonymity

All responses collected in this research were anonymous by the use of the Anonymize Responses tool in Qualtrics. Additionally, surveys were only distributed via the anonymous link and no personalized invitations were created. There were no requests or opportunities for participants to submit personal or identifying data. Access to individual responses is limited to only the author of the research. The responses have been exported from Qualtrics and exist in a private branch of a repository hosted on OSF⁴ to maintain security.

³<https://www.rug.nl/cf/onderzoek-gscf/ethics-committee/ethics-committee-campus-fryslan>

⁴<https://osf.io/>

4.5.3 Participant Comfort

The listening test was purposefully designed to be as short as possible whilst maintaining validity to mitigate any possible discomfort that may occur due to prolonged listening activity or survey taking, a topic discussed in subsection 2.2.2. In addition to this, participants were not time limited in any way and were able to take breaks or discontinue participation freely.

4.5.4 Bias Prevention

Bias in many areas was considered as well. Firstly, to avoid any biases towards characteristics of the synthetic voices, three distinctly different voices were chosen and distributed evenly and at random amongst participants. Further, the questions within the listening test were presented in a random order to each participant to prevent order bias. Participants were not permitted to revisit past questions in an attempt to limit any bias caused by comparison. While demographic data was not collected in this study, the listening test was distributed across many different public social platforms globally in an attempt to ensure a diverse and representative sample of participants.

4.5.5 Use of AI

In accordance with the AI policy, this subsection will detail the ways in which I engaged with AI throughout the course of the research. All AI assistance was only used in a supplementary fashion, and no text or work done in the research was done solely with AI. All work assisted by AI tools has been critically assessed for accuracy and relevance.

The AI tool, Perplexity⁵, was used in initial idea generation as a means to quickly search for any research that may have already been conducted on the chosen topic. The GPT-4o model from OpenAI⁶ was used to refine the research methodology and to assist in the structure of the literature review and methodology chapters. This model was also used to brainstorm ideas for the presentation of results and statistical analysis methods, however the tasks themselves were carried out manually. Lastly, GPT-4o was used for self-evaluation of the research text, to highlight areas of written improvement based on the grading rubric. Prompts used for the GPT-4o model were refined with the use of the PromptPerfect⁷ tool for efficiency.

4.5.6 Replicability

Although MOS is a subjective measure of quality, all results in this study are reported objectively, ensuring honesty and transparency. In section 5.4, limitations to this study as well as potential sources of error are discussed to further support this. This research has been conducted in a way that allows it to be replicable by any interested parties. All code used to produce the stimuli, the actual stimuli, and the listening test files (including the randomization and distribution workflow) are available via GitHub⁸. All steps and details necessary to reproduce the literature review are listed in section 2. All research was conducted in May 2024.

⁵<https://www.perplexity.ai/>

⁶<https://openai.com/index/hello-gpt-4o/>

⁷<https://promptperfect.jina.ai/>

⁸<https://github.com/branaphy/msc-vt-thesis>

This concludes the methodology section which explains the research, experimental, and ethical design of this research. In the next section, the results of the primary listening test will be presented and discussed via visual presentation and statistical analysis.

5 Results

In this chapter, I will present the results of the primary listening test set up in section 4.4 in section 5.1. Directly following, I will interpret and analyze these results in section 5.2 by referencing data visualizations of the results and performing statistical analysis. After this, I will use section 5.3 to discuss the implications of these results on the hypothesis. Finally, I will discuss the limitations of this study in section 5.4.

For simplicity and clarity, the voices from the Parler-TTS Mini: Espresso model will be referred to by their speaker names, Jerry and Elisabeth, while the voice from the SpeechT5 model will be referred to as SpeechT5.

5.1 Presentation of Results

In this subsection, I will visually present the results in a grouped bar chart in figure 1. I will also present the results for each voice in their own summary tables, tables 4, 5, and 6.

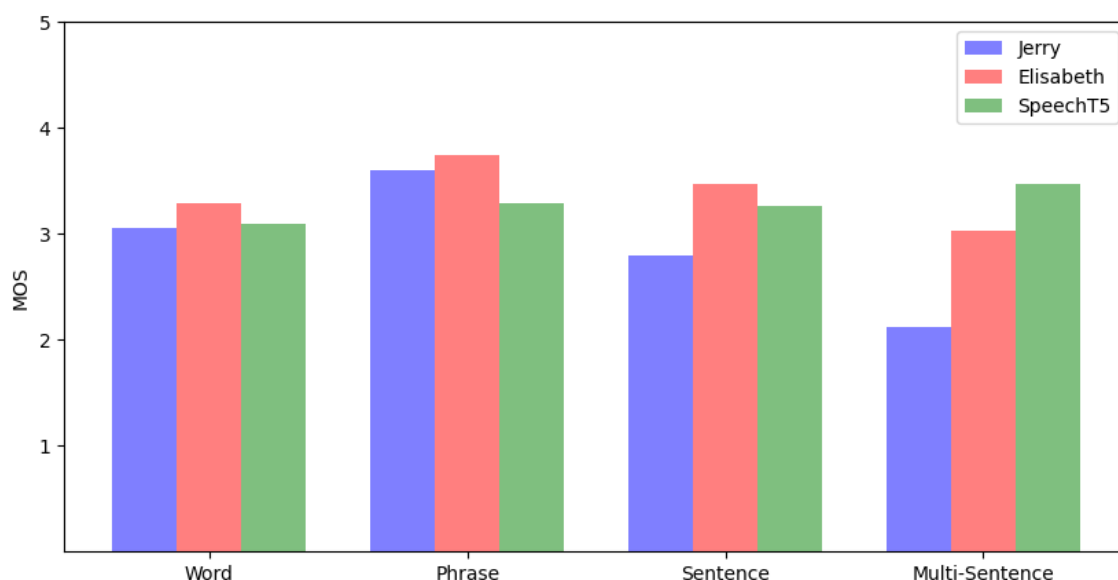


Figure 1: Results for the primary listening test comparing the effect of speech clip duration on the MOS of synthetic speech.

Table 4: Summary of results for the Jerry voice

Duration	Mean	Std	Min	Max	Var	Count
Word	3.06	0.89	1.00	5.00	0.80	35
Phrase	3.60	0.99	1.00	5.00	0.98	35
Sentence	2.79	0.96	1.00	5.00	0.93	34
Multi-sentence	2.12	1.05	1.00	5.00	1.10	34

Table 5: Summary of results for the Elisabeth voice

Duration	Mean	Std	Min	Max	Var	Count
Word	3.29	1.07	1.00	5.00	1.15	34
Phrase	3.74	0.98	2.00	5.00	0.96	34
Sentence	3.47	1.06	1.00	5.00	1.13	34
Multi-sentence	3.03	1.04	1.00	5.00	1.09	34

Table 6: Summary of results for the SpeechT5 voice

Duration	Mean	Std	Min	Max	Var	Count
Word	3.09	1.17	1.00	5.00	1.37	34
Phrase	3.29	1.02	1.00	5.00	1.03	34
Sentence	3.26	1.07	1.00	5.00	1.14	34
Multi-sentence	3.47	1.01	1.00	5.00	1.01	34

5.2 Interpretation and Analysis of Results

In this section, I will briefly discuss the data visualizations presented in subsection 5.1. After this, I will perform a statistical the results using ANOVA. The primary objective of this analysis is to determine the relationship between the duration of a speech clip and its MOS, if any, identifying any notable patterns or anomalies in the data.

5.2.1 Visualized Data

The provided grouped bar chart in figure 1 depicts the Mean Opinion Score (MOS) across different speech clip durations for three synthetic speech voices: Jerry, Elisabeth, and SpeechT5. The durations considered in this study are segmented into four categories: Word, Phrase, Sentence, and Multi-Sentence.

The MOS of voices at the word duration appear to demonstrate a relatively consistent performance, with MOS values clustering around 3. There is a noticeable collective improvement across all voices as the duration increases from word to phrase. The introduction of context and linguistic structure may contribute to this increase. As the duration increases from phrase to sentence, there is a collective decline in MOS. Finally, as the duration further increases to multi-sentence, two of the voices see quite noticeable decreases in MOS while the remaining voice, SpeechT5, stays consistent.

The chart in figure 1 does not reveal and prominent collective patterns, yet it does reveal patterns about individual voices. The Jerry voice sees its best performance at the phrase level, with highly visible declines in MOS as the utterances become longer. The Elisabeth voice also follows this pattern to a lesser degree of severity. The SpeechT5 voice stays consistent throughout all durations.

Comparing these visual observations to the summary tables 4, 5, and 6, it is confirmed that the Jerry and Elisabeth voices experience their lowest MOS at the longest duration (2.12 and 3.03, respectively), while in contrast, SpeechT5 actually sees its highest MOS of 3.47 at the longest duration. There are no other patterns that are shared across all 3 voices.

5.2.2 Statistical Analysis

ANOVA, or Analysis of Variance, is a statistical method used to compare means of three or more samples to see if at least one of the sample means significantly differs from the others. With three different voices and four different durations, ANOVA can be used to compare the means of these groups to find any statistically significant differences. The ANOVA for this study was conducted in Python.

To perform ANOVA on the results, I first combined all results into a single pandas dataframe. Then, I checked assumptions by running a Shapiro-Wilk test for normality and Levene's test for homogeneity of variances. These tests produced non-significant p-values (p greater than 0.05), which suggests that the residuals are normally distributed and indicates that the variances are equal across groups. After this, I performed a one-way ANOVA in Python using the `scipy.stats` module. The ANOVA test results returned an F-statistic of 1.37 and a p-value of 0.32.

The F-statistic measures the ratio of the variance between the groups to the variance within the groups. Thus, a higher F-statistic generally indicates stronger evidence against the null hypothesis. The F-statistic of 1.37 indicates that any observed differences in MOS scores between the different durations are not statistically significant and are relatively small in comparison to the variation of scores within each duration.

The p-value indicates the probability that the observed differences among group means occurred by chance. A p-value less than 0.05 typically indicates statistically significant differences. Since the p-value of this experiment is 0.250, we fail to reject the null hypothesis, meaning there is not enough evidence to conclude that there are statistically significant differences in the MOS across different durations for the presented voices. Thus, while there are some observable visual differences, they are not sufficient to conclude a significant impact of duration on MOS across the different voices analyzed.

Given that the ANOVA did not show statistically significant differences, post-hoc testing was not conducted. However, the effect size was calculated to understand the magnitude of the differences among the durations. The measure used in this study to calculate effect size was Eta squared, which is calculated as the ratio of the variance explained by the treatment (between durations) to the total variance. The Eta squared value calculated was 0.339, which indicates that approximately 33.9% of the variance in MOS can be explained by differences in duration. Despite ANOVA results not showing statistical significance, the moderately strong effect size suggests a meaningful association between duration and MOS.

5.3 Reflection on the Hypothesis

This research aimed to test the hypothesis that the Mean Opinion Scores (MOS) would decrease as the duration of speech clips increased, particularly when these clips were of varying perceived quality levels. However, the results provided by ANOVA did not find statistically significant differences in MOS across different durations (Word, Phrase, Sentence, Multi-sentence) for the tested voices.

The lack of significant findings suggests that the duration of speech clips may not be a predominant factor influencing MOS, contrary to the initial hypothesis.

While research previously done by Kirkland et al. (2023) notes that MOS is highly influenced by variations, it is possible that MOS is more robust to certain variation, such as sample length. Additionally, Le Maguer et al. (2024) found that modern speech synthesis is perceived as more natural than the top-tier systems from 2013. The use of modern advanced TTS systems in the primary experiment may have minimized the perceptual differences that participants could discern over varying durations. Additionally, the relatively short overall duration of the listening test (less than 2 minutes, on average) may have nullified negative effects predicted over longer durations, as suggested by the literature reviewed in section 2.2.2.

It is possible that listeners' evaluations of speech quality may not be as sensitive to the duration of the content as was originally hypothesized. The short design of the test (2 minutes) may have also negated all effects of survey fatigue as researched by Sinickas (2007) and cognitive load as researched by Hornsby (2013) and Sarem et al. (2019) that were considered when forming the hypothesis. However, it is important to note that the Eta squared value does suggest a moderate-level of association between duration and MOS. Additionally, the relationship between speech duration and MOS may not be a strictly linear one. Rather, it is possible that the relationship is much more complex, and therefore, there may be non-linear effects that were not evaluated within the scope of this study.

5.4 Limitations of the Study

There are several factors and limitations within this study that may have contributed to the lack of significant findings. There were several deliberate choices made in the experimental design, detailed in section 4.2, that may have affected the outcomes. For instance, the measurement of "naturalness" versus "quality" as previously discussed by Kirkland et al. (2023) as well as the lack of granularity in the answer labels, also discussed in the same work. The general lack of standardization in MOS listening tests as highlighted by Le Maguer et al. (2024) likely introduced many unknown variables that affected the results.

Additionally, the experiment was made very short by design to reduce fatigue of the participants, but the relatively small sample size (three voices and one utterance per duration category) may have limited the outcomes. Additionally, the generation of two voices from the same model may have limited the results, although they were initially perceived to be distinct from each other. Although the reported values for the two voices from the Parler-TTS Mini: Espresso model were not the same, both voices followed the same trends, albeit with varying degrees of severity.

This study primarily focused on shorter utterances of synthetic speech, ranging from 1 second to 24 seconds. It is still possible that duration may be a factor, but the upper end of duration (spanning minutes) has yet to be explored. This study also categorized the utterances in a qualitative manner, leading to an uneven distribution of lengths across the range of durations. It is possible that a quantitative distribution, such as: 1 second, 5 seconds, 15 seconds, 30 seconds, may have produced different results.

This section covered the results obtained from the primary experiment, a listening test measuring MOS of different voices across varying durations. The results were presented visually via a group bar chart and summary tables, and then these visualizations were interpreted. The results were then

statistically analyzed using ANOVA, deeming the findings statistically insignificant. The results were then discussed in relation to the hypothesis presented in chapter 3. Finally, the limitations of the study were discussed. In the following chapter, I will conclude the study, present suggestions for future work, and discuss the impact and relevance of the study.

6 Conclusion

In summary, this research seeks to answer the question presented in chapter 3: **How does the duration of synthetic speech clips affect the Mean Opinion Score across different speech quality levels?** This specific research ultimately contributes to the overall discussion surrounding the reliability of MOS as an evaluation tool.

Following this introduction, a summary of key findings will be covered in section 6.1. After this, in section 6.2, the implications of these findings on future work will be discussed alongside recommendations for continued research. Finally, I will discuss how the findings of the research impact the field of speech synthesis and its evaluation in section 6.3.

6.1 Key Findings

The primary experiment of this research compared the MOS of three synthetic voices across four different duration categories through a listening test. To test this relationship, listeners were presented with four clips of a singular randomly-chosen voice, each of a different duration. These clips were also presented in a randomly-selected order. All other aspects of the listening test remained the same for each participant.

While each individual voice exhibited visually noticeable patterns in figure 1 over the different duration categories, after statistical analysis via ANOVA, this research was unable to determine statistically significant effects of the duration of a speech clip on the MOS of the voice. Because of this, the hypothesis presented in section 3 can not be supported. This result suggests that listeners' evaluations of speech quality may not be as sensitive to the duration of speech clips on the presented qualitative scale (one word to three sentences). The breadth of quality of synthetic voices may have also not been wide enough, considering the evaluation of only 3 voices across 2 models.

Although the initial hypothesis cannot be supported, this does not completely eliminate the possibility of a relationship between MOS and speech clip duration. Despite the results of the study being statistically insignificant, the effect size of 33.9% calculated in section 5.2.2 suggests some level of meaningful association between the duration of speech clips and MOS, across all voices. Although the overall results of this study are inconclusive and the hypothesis cannot be validated, this metric leaves a level of promise and motivation for future research.

6.2 Future Work

As the results of this particular thesis are inconclusive, it lends itself well to continued future research. I would suggest this research to be continued via three primary avenues: expanded duration study, standardization of the MOS evaluation method, and research of different independent variables in MOS listening tests.

6.2.1 Expanded Duration Study

The scope of this research was limited to an experiment size that could be achieved within an 8-week master's thesis. The primary experiment could likely be expanded in a variety of ways to further investigate the relationship between clip duration and MOS.

Firstly, the experiment could be expanded to encompass a wider range of voice qualities. It could encompass a larger quantity of tested models, as well as models of larger complexity ranges. This study only tested three voices across two pre-trained models. While an attempt was made to capture different genders, qualities, and prosodic variations, a more rigorous voice selection process could be done to ensure a more holistic range of TTS models is represented.

Additionally, the listening test conducted in this thesis was completed in under 2 minutes, on average. An expanded study could offer more stimuli across more duration categories and more precise durations. This study investigated qualitative duration categories, like word and sentence, when in reality, the length of these categories can be highly varied within themselves; this includes short and long words, a lack of standard definition of a phrase, short and long sentences, and a multi-sentence length of two or six sentences. I would suggest future research to measure quantitative durations, measured in seconds.

Lastly, this experiment was well-documented in its experimental design, including the choice of wording and labels. Research discussed in section 2.1.2 shows that each of these factors has an affect on the MOS of a synthetic voice. Additional, yet similar, experiments can be conducted by changing some of these variables to see if any statistically significant relationships between duration and MOS emerge. This would further highlight the need for research in the standardization of MOS.

6.2.2 Standardization of MOS

Although the initial hypothesis presented in section 3 was not supported, this research serves, in the broader sense, to highlight the lack of standardization in MOS.

Within the design of this research and the associated experiment in chapter 4, I was afforded many liberties. In fact, the "standards" that I chose to adhere to from International Telecommunication Union (1996) are merely recommendations. Knowing from the research reviewed in section 2.1, the reliability of MOS generally suffers from the lack of standards. This lack of standards introduces many variables, known to influence MOS, that are left to the choice of researchers. To this end, I suggest and support further research suggested by Le Maguer et al. (2024) and Kirkland et al. (2023) to implement standardization and best practices for speech synthesis evaluation, particularly for the Mean Opinion Score.

6.2.3 Research of Different Variables

Research reviewed in section 2.1 made it clear that there are several factors in MOS listening tests that are highly influential in the reported scores. The literature also raises the issue of underreporting by researchers of variables used in listening tests.

Although this research did not present a statistically significant linear relationship between speech clip duration and the MOS of a voice, it is apparent that there is a knowledge gap in the effect of many other variables on MOS. I would suggest further research focused on independent variables such as background noise, accent, emotional expression, and interaction context to assess their potential impact on the MOS of synthetic voices.

6.3 Broader Impact on Speech Synthesis Evaluation

This research impacts the field of speech synthesis by contributing early research to a currently evolving conversation around standardization and best practices of speech synthesis evaluation. It is apparent from the review conducted in chapter 2 that there is a lack of research done in this area in comparison to other areas of speech synthesis. It is important to continue this conversation and push for the evolution of speech synthesis evaluation to remain current alongside the technology which it assesses.

References

- Ao, J., Wang, R., Zhou, L., Wang, C., Ren, S., Wu, Y., . . . Wei, F. (2022, May). SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 5723–5738).
- Chiang, C.-H., Huang, W.-P., & yi Lee, H. (2023). Why We Should Report the Details in Subjective Evaluation of TTS More Rigorously. In *Proc. interspeech 2023* (pp. 5551–5555). doi: 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*.
- Cooper, E., & Yamagishi, J. (2023). *Investigating range-equalizing bias in mean opinion score ratings of synthesized speech*.
- Hornsby, B. W. (2013). The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear and hearing*, 34(5), 523–534.
- International Telecommunication Union. (1996, August). *Methods for subjective determination of transmission quality* (ITU-T Recommendation No. P.800). Author. Retrieved from <https://www.itu.int/rec/T-REC-P.800-199608-I>
- Jeong, D., Aggarwal, S., Robinson, J., Kumar, N., Spearot, A., & Park, D. S. (2023). Exhaustive or exhausting? evidence on respondent fatigue in long surveys. *Journal of Development Economics*, 161, 102992. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0304387822001341> doi: <https://doi.org/10.1016/j.jdeveco.2022.102992>
- 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. In *Proc. 12th isca speech synthesis workshop (ssw2023)* (pp. 41–47). doi: 10.21437/SSW.2023-7
- Lacombe, Y., Srivastav, V., & Gandhi, S. (2024). *Parler-tts*. <https://github.com/huggingface/parler-tts>. GitHub.
- Le Maguer, S., King, S., & Harte, N. (2024). The limits of the mean opinion score for speech synthesis evaluation. *Computer Speech & Language*, 84, 101577.
- Lyth, D., & King, S. (2024). *Natural language guidance of high-fidelity text-to-speech with synthetic annotations*.
- Mehrish, A., Majumder, N., Bhardwaj, R., Mihalcea, R., & Poria, S. (2023). *A review of deep learning techniques for speech processing*.
- Nguyen, T. A., Hsu, W.-N., d’Avirro, A., Shi, B., Gat, I., Fazel-Zarani, M., . . . others (2023). Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*.
- Nusbaum, H. C., & Pisoni, D. B. (1985). Constraints on the perception of synthetic speech generated by rule. *Behavior Research Methods, Instruments, & Computers*, 17(2), 235–242.
- O’Mahony, J., Oplustil-Gallegos, P., Lai, C., & King, S. (2021). Factors Affecting the Evaluation of Synthetic Speech in Context. In *Proc. 11th isca speech synthesis workshop (ssw 11)* (pp. 148–153). doi: 10.21437/SSW.2021-26
- Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., . . . Auli, M. (2023). Scaling speech technology to 1,000+ languages. *arXiv*.
- Rosenberg, A., & Ramabhadran, B. (2017). Bias and Statistical Significance in Evaluating Speech Synthesis with Mean Opinion Scores. In *Proc. interspeech 2017* (pp. 3976–3980). doi: 10

- .21437/Interspeech.2017-479
- Sarem, S. N., Marashi, H., & Siyyari, M. (2019). The relationship between listening comprehension and listening fatigue among iranian intermediate efl learners.. Retrieved from <https://api.semanticscholar.org/CorpusID:199410298>
- Sinickas, A. (2007). Finding a cure for survey fatigue. *Strategic Communication Management*, 11(2), 11.
- Wester, M., Valentini-Botinhao, C., & Henter, G. E. (2015). Are we using enough listeners? no! — an empirically-supported critique of interspeech 2014 TTS evaluations. In *Proc. interspeech 2015* (pp. 3476–3480). doi: 10.21437/Interspeech.2015-689

Appendices

A The North Wind and the Sun

The full text for the chosen text, The North Wind and the Sun, can be found below.

The North Wind and the Sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak. They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other. Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveler fold his cloak around him; and at last the North Wind gave up the attempt. Then the Sun shined out warmly, and immediately the traveler took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.

B Questionnaire Workflow

In this appendix, the workflow and logic of the both the pilot and primary tests are presented.

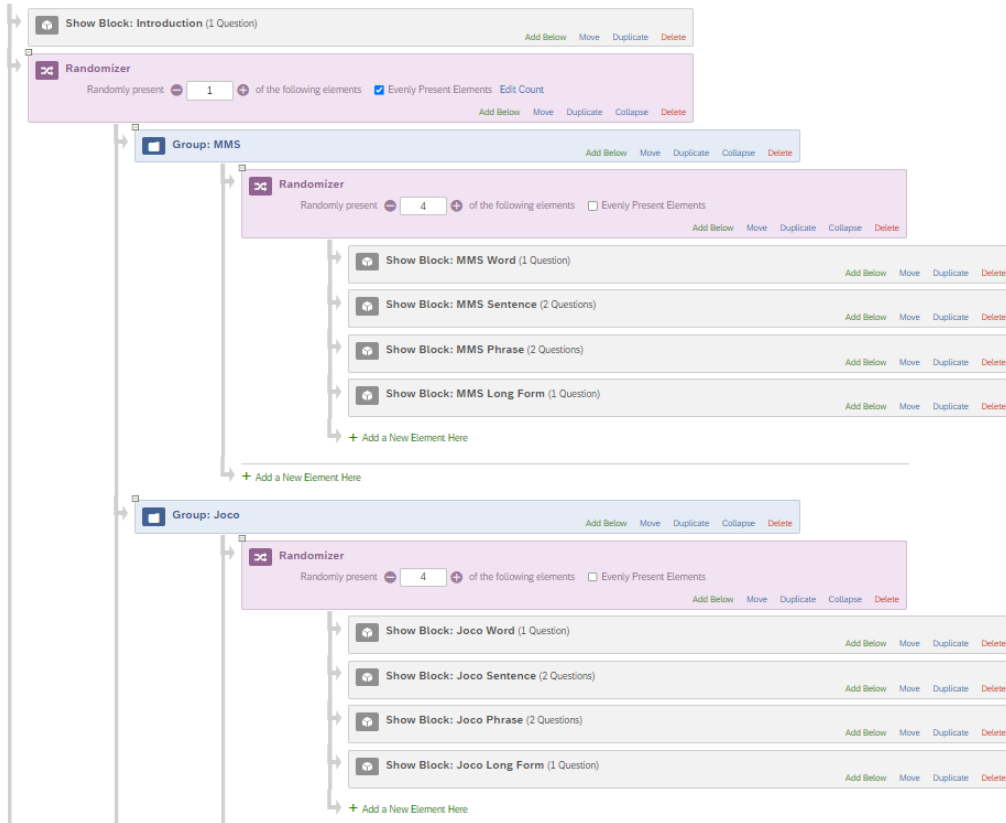


Figure 2: Flow of the pilot test.

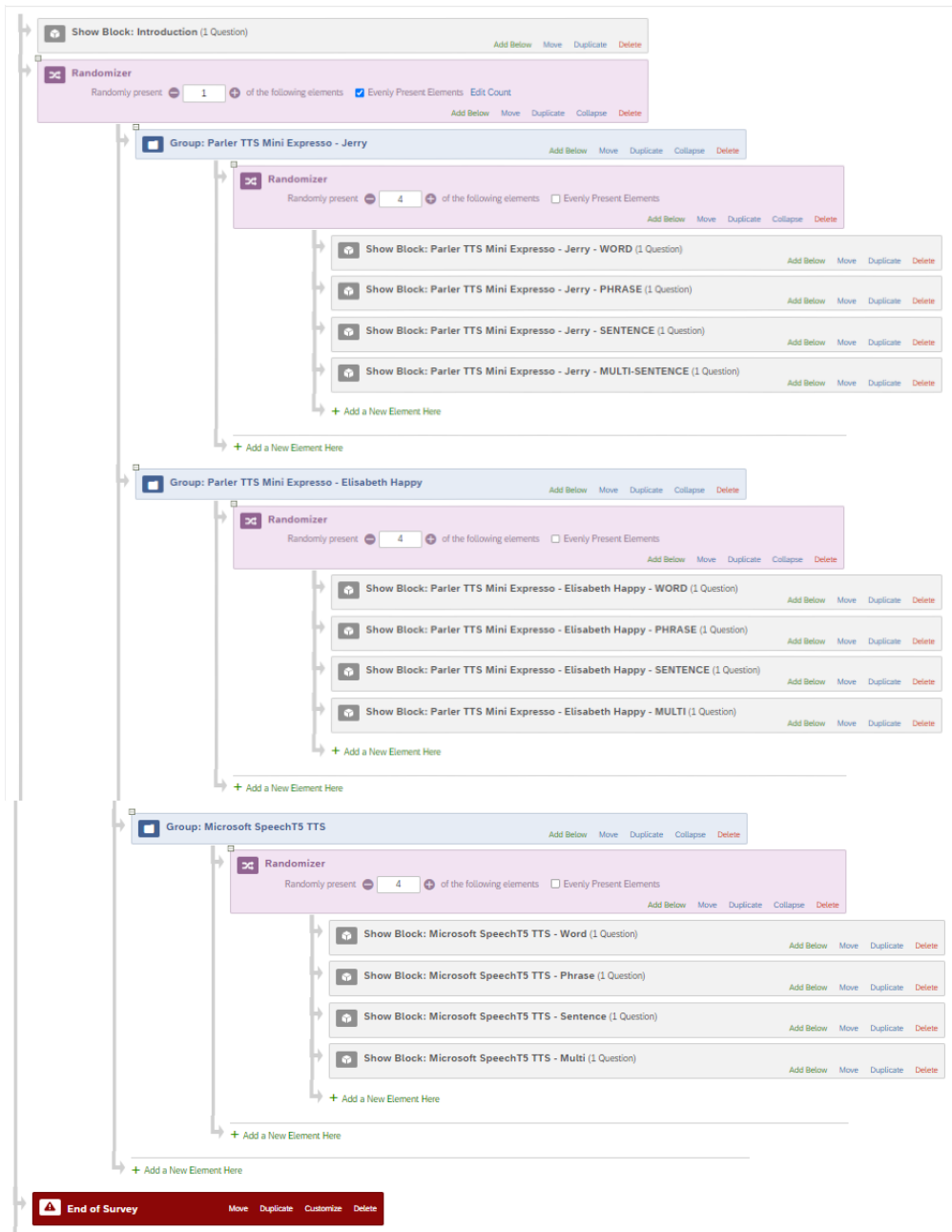


Figure 3: Flow of the primary listening test.

C Pilot Questionnaire and Stimuli

In this appendix, the introduction and question format for the pilot listening test are presented. Additionally, the text prompts for the stimuli are shared.

Table 7: Text prompts used to generate the stimuli for the pilot

Duration	Text Prompt
Word	more
Phrase	immediately the traveler took off his cloak
Phrase	then the Sun shined out warmly
Sentence	The North Wind and the Sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak.
Sentence	They agreed that the one who first succeeded in making the traveler take his cloak off should be considered stronger than the other.
Multi-Sentence	Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveler fold his cloak around him; and at last the North Wind gave up the attempt. Then the Sun shined out warmly, and immediately the traveler took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.



Thank you for participating in this study!

I am interested in learning more about what makes a voice sound "natural." In this survey, you will listen to a series of speech samples and rate their naturalness.

Instructions:

1. You will hear several 4 voice samples--some very short.
2. After listening to each sample, please rate the quality of the voice on a scale from 1 to 5, where 1 means "Bad" and 5 means "Excellent."
3. There are no right or wrong answers. Please provide your honest opinion based on your personal perception of the voice's naturalness.

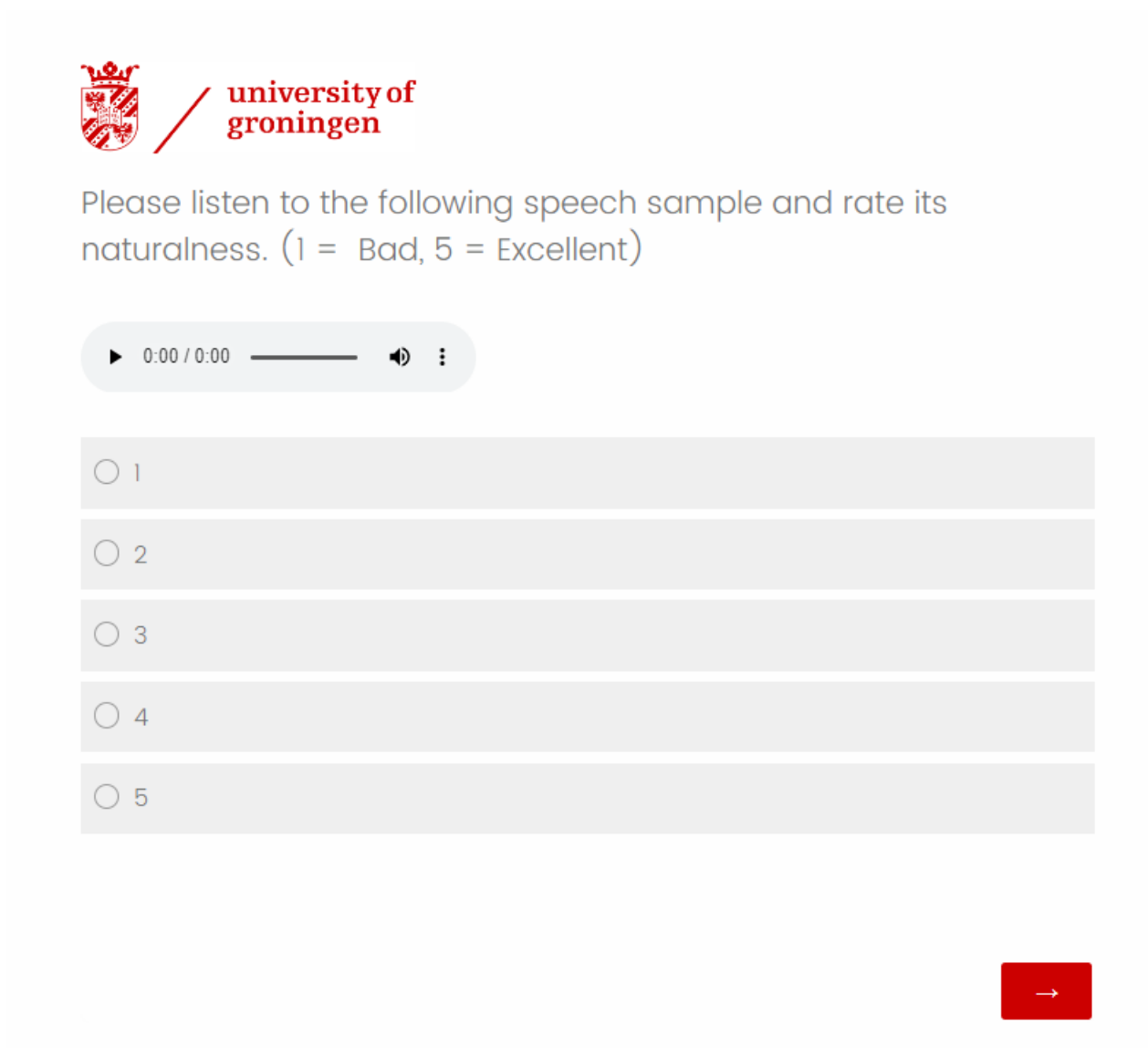
Your responses are anonymous, and all data will be kept confidential. The survey should take approximately 2-3 minutes to complete.

When you are ready, click the arrow to begin.

Thank you for your participation!



Figure 4: Introduction page for the pilot listening test.



The image shows a screenshot of a listening test question format. At the top left is the University of Groningen logo, which consists of a red shield with a white cross and a crown above it, followed by the text "university of groningen" in red. Below the logo is the instruction: "Please listen to the following speech sample and rate its naturalness. (1 = Bad, 5 = Excellent)". Underneath the instruction is a media player interface with a play button, a progress bar showing "0:00 / 0:00", a volume icon, and a settings menu icon. Below the media player are five radio button options labeled "1", "2", "3", "4", and "5", each on a separate light gray background. At the bottom right of the form is a red square button with a white right-pointing arrow.


Figure 5: Question format for the pilot listening test.

D Primary Listening Test Questionnaire and Stimuli

This appendix presents the introduction page and question format for the primary listening test, as well as the text prompts used to produce the stimuli.

Table 8: Text prompts used to generate the stimuli for the primary listening test

Duration	Text Prompt
Word	first
Phrase	they agreed
Sentence	The North Wind and the Sun were disputing which was the stronger, when a traveler came along wrapped in a warm cloak.
Multi-Sentence	Then the North Wind blew as hard as he could, but the more he blew the more closely did the traveler fold his cloak around him; and at last the North Wind gave up the attempt. Then the Sun shined out warmly, and immediately the traveler took off his cloak. And so the North Wind was obliged to confess that the Sun was the stronger of the two.



Thank you for participating in this study! This survey should take approximately 2-3 minutes to complete.

This survey is being conducted as part of my master's thesis to fulfill the requirements of the Master of Science in Voice Technology at the University of Groningen.

I am interested in understanding how human listeners perceive and evaluate speech. In this survey, you will listen to a series of speech samples and rate their naturalness.

Instructions:

- You will be presented with 4 speech samples.
- After listening to each sample, please rate the naturalness of the voice on a scale from 1 to 5, where 1 indicates "Bad" and 5 indicates "Excellent."

Important to Note:

- There are no right or wrong answers.
- Please provide your honest opinion based on your personal perception of the voice's naturalness.
- Your responses are anonymous, and all data will be kept confidential.
- It is recommended to use headphones for this survey.

When you are ready, click the arrow to begin. Thank you for your participation!




Figure 6: Introduction page for the primary listening test.



Please listen to the following speech sample and rate its naturalness.



1 - Bad

2 - Poor

3 - Fair

4 - Good

5 - Excellent



Figure 7: Question format for the primary listening test.