Examining the effects of text support and noise during video meetings on Listening Effort and Comprehension

Fredrik Fernlund

729G40 - Kandidatuppsats

Handledare: Erik Marsja
Examinator: Carine Signoret
Upphovsrätt

Detta dokument hålls tillgängligt på Internet – eller dess framtida ersättare – under 25 år från publiceringsdatum under förutsättning att inga extraordinära omständigheter uppstår.

Tillgång till dokumentet innebär tillstånd för var och en att läsa, ladda ner, skriva ut enstaka kopior för enskilt bruk och att använda det oförändrat för ickekommersiell forskning och för undervisning. Overföring av upphovsrätten vid en senare tidpunkt kan inte upphäva detta tillstånd. All annan användning av dokumentet kräver upphovsmannens medgivande. För att garantera äktheten, säkerheten och tillgängligheten finns lösningar av teknisk och administrativ art.

Upphovsmannens ideella rätt innefattar rätt att bli nämd som upphovsman i den omfattning som god sed kräver vid användning av dokumentet på ovan beskrivna sätt samt skydd mot att dokumentet ändras eller presenteras i sådan form eller i sådant sammanhang som är krävande för upphovsmannens litterära eller konstnärliga anseende eller egenart.

För ytterligare information om Linköping University Electronic Press se förlagets hemsida https://ep.liu.se/.

Copyright

The publishers will keep this document online on the Internet – or its possible replacement – for a period of 25 years starting from the date of publication barring exceptional circumstances.

The online availability of the document implies permanent permission for anyone to read, to download, or to print out single copies for his/hers own use and to use it unchanged for noncommercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission. All other uses of the document are conditional upon the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility.

According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement.

For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its www home page: https://ep.liu.se/.

© Fredrik Fernlund
Abstract

Many companies implemented remote work procedures during the pandemic and for many organizations video meetings have since remained a staple. Remote working has enabled employees to be more flexible with their schedules and technical solutions such as live captioning has been identified as potentially enabling deaf/hard-of-hearing employees during meetings. However with new procedures comes concern about how we potentially can be affected by the changes. Some earlier research has shown that speech intelligibility can be improved by the inclusion of text support, but they also raised the possibility that it could have unwanted adverse effects on cognitive abilities (Zhong, Noud et al., 2022). This study was conducted with this focus, studying the effects of text support on specifically listening effort and comprehension during normal as well as adverse conditions (featuring added noise).

To investigate the effects of text support a 2 (Noise, No Noise) x 2 (Text Support, No Text Support) design was used. The participants were shown 16 short videos simulating video meetings and after each video were asked to rate their perceived listening effort as well as a comprehension question about the contents of the discussion. Each of the four conditions were equally represented but the order of the specific video files and conditions that applied were randomised for each participant to mitigate undue effects.

The results of the study indicate that the presence of captions decrease effort and raise comprehension in both normal and adverse conditions. Noise was found to strongly effect the listening effort required by participants but no significant effect was found upon comprehension. Some concerns regarding the ecological validity were identified during the course of the study such as only studying energetic noise and unrealistic presentation of captions. However the results are nonetheless believed to be generalizable in most regards and showcase that captions can have a positive influence during video meetings.
Contents

1 Introduction 5
2 Purpose 6
3 Theory 7
   3.1 Listening Effort ........................................... 7
   3.2 Comprehension ............................................. 7
   3.3 Noise ....................................................... 8
   3.4 Listening Effort and Noise ................................ 8
   3.5 Comprehension and Noise .................................. 8
   3.6 Text Support ............................................... 9
   3.7 Borg CR10 .................................................... 9
   3.8 LIX ......................................................... 9
   3.9 PsychoPy .................................................... 10
4 Method 11
   4.1 Participants ................................................. 11
   4.2 Procedure .................................................. 11
   4.3 Ethics ....................................................... 12
   4.4 Material ..................................................... 12
   4.5 Pre-study results .......................................... 13
   4.6 Video Material ............................................. 13
   4.7 Statistics .................................................. 14
5 Results 15
   5.1 Listening Effort ........................................... 16
   5.2 Comprehension ............................................. 18
6 Discussion 21
   6.1 Consequences ............................................. 22
   6.2 Method Criticism .......................................... 22
      6.2.1 Listening Effort Measurement ........................... 22
6.2.2 Comprehension Measurement ........................................... 23
6.2.3 PsychoPy Criticism ...................................................... 23
6.2.4 Text Support Criticism ................................................ 23
6.2.5 Noise Criticism .......................................................... 24
6.2.6 LIX Criticism ............................................................ 26
6.3 Limitations and Future Research ......................................... 26

7 Conclusions ......................................................................... 27

8 References ........................................................................ 28
1 Introduction

In the wake of the pandemic remote working and teleconferencing has been an emerging and rapidly growing phenomena (Robinson, n.d.). This has enabled many to switch fields and achieve better working conditions and work-life balance (Tsipursky, n.d.). However at the same time several difficulties and problems connected to this new kind of working lifestyle have become a part of public discourse. Examples such as a coworker’s cat walking in front of the camera or something embarrassing accidentally being left in the shot of a web camera has led to many viral moments spread on the internet. Technical difficulties such as lag, disconnections and noise are also aspects that almost everyone who has used teleconferencing has experienced. The possibilities are great for using video meeting applications, such as Zoom or Microsoft Teams, for purposes such as allowing a greater reach of higher education to those who otherwise might not be able to attend classes. However it is important to study what problems can typically arise in these kinds of contexts and how they affect the participants in these activities.

While some problems such as latency are more or less impossible to fully address there are potentially opportunities in adding tools to help lessen negative effects. One such tool is a feature that for example Microsoft Teams have introduced in their software, automatic text support (Microsoft Support, n.d.). Using voice-recognition technology, this tool is able to create subtitles in real time for other meeting participants. This technology can sometimes be unreliable and can create a mismatch for what was said by a participant (Normand, 2022). However this technology has tremendous potential in enabling deaf/hard-of-hearing meeting participants, clarifying what a speaker in a noisy environment actually is saying and similar. If higher accuracy is achieved. This likely means that this technology will be developed further in future and will likely improve overtime.

Some other questions arise when adding text support to video conferencing. Namely the potential impact on effort. A meta-study by Zhong, Noud et al. (2022) showed that the addition of text support during telecommunication improved speech intelligibility for all participants. This alone is great and could, for many, be seen as enough reason to fully justify the addition of text support in video meeting applications. However in their study they also claim that not enough is known about what effects the addition of text support actually has on cognitive aspects such as listening effort or working memory (WM). While this may seem trivial at first, a heightened demand on the effort required during meetings might make the listener less receptive to the information that is being communicated. This could lessen the effectiveness of the meeting as a whole or have other negative effects that would be detrimental or undesirable. As such it would be worthwhile to get a better understanding of how text support during video meetings affect the participants using a lens based on their potential effects upon the cognitive aspects of the participant.
2 Purpose

The effects of text support on effort and comprehension in video meeting contexts has not currently been examined widely. Therefore the purpose of this study is to contribute to a greater understanding regarding the effects of text support upon cognitive aspects during live video meetings both in normal but also adverse conditions (added noise). The following research questions are thusly the primary aim of this report:

Research question 1: What is the effect of text support in video meetings on listening effort and comprehension?

Research question 2: What is the effect of noise in video meetings on listening effort and comprehension?

Research question 3: What effect does text support have on listening effort and comprehension in adverse conditions (e.g. added noise)?

Terminology

The terms Captions, Subtitles, Text Supplementation and Text Support are used interchangeably throughout the study.
3 Theory

This section details the different frameworks and theories necessary for answering the research questions outlined earlier in the paper.

3.1 Listening Effort

One of the main subjects of interest in this study is listening effort and how it is affected by text support and noise. According to McGarrigle et al. (2014) Listening Effort can be defined as “mental exertion required to attend to, and understand, an auditory message”. However in a response by Römberg et al. (2014) they also stress the importance of understanding mental exertion within it’s cognitive constraints. Exertion is needed when speech can not rapidly or automatically be decoded by the recipient. Sometimes this decoding can be interrupted due to noise or other external or internal phenomena which leads to attempts from the recipient to repair the message by using the fragments that were received to infer what the sentence probably intended to communicate. This increased effort in turn puts a higher demand on an individual’s WM.

While there are no fully agreed upon ways in the audiology community on how to measure Listening Effort, self report measures are generally seen as acceptable tools for measuring listening effort. However in an attempt to standardize research studying these phenomena the framework for understanding effortful listening (FUEL) was developed during a workshop at Eriksholm (Pichora-Fuller et al., 2016). This study used a simpler self report measure by asking participants to rate listening effort on the BORG CR10 exertion scale (described in 3.8), this allowed us to show a greater number of stimuli and as such have more power for our results.

3.2 Comprehension

There is no singular agreed upon way to measure comprehension in listening studies. However commonly a questionnaire about the contents of speech can be used to easily see if participants were able to effectively understand or hear what was said, or by asking participants to verbally repeat phrases that they were listening to (Margolis et al., 2021). For this study participants were first exposed to the stimuli and then answered a yes/no questionnaire featuring a question related to the contents of what was said. This choice was made since it was easier to implement in the software used for the experiments. It differs somewhat from the more commonly used method for studying speech recognition, where participants are often tasked to verbally reproduce an utterance. A study by Dewyer et al. (2017) showed that results from a multiple choice questionnaire they developed for speech recognition, and measurements done using traditional audiologists’ speech recognition tests were highly correlated. As such, even tough there are distinct differences between the methods they have both shown to be valid and reliable ways to measure comprehension.
3.3 Noise

Noise can be categorized as either energetic or informational noise (Brungart, 2001; Cooke et al., 2008; Mattys et al., 2009). Energetic noise can generally be seen as what we typically associate with noisy environments. Machines, people talking unintelligibly in the background or general static noise can all be classified as energetic noise. Informational noise would in contrast be speech that is more easily understood for the participant. The two kinds of noise affect cognitive resources somewhat differently. Energetic noise is most commonly used in a way where it drowns out or masks the task-related item (Mattys et al. 2012). Informational noise can also be used this way but it can also be a distractor directing the attention of participants away from their task. Both kinds of noise put an increased strain on an individuals’ WM. As such it is expected that both dependent variables will be impaired in the presence of noise.

During the current study the noise used to simulate adverse conditions was of the energetic variant featuring modulated speech that was entirely unintelligible. It is a speech-shaped noise (SSN) and was developed in a study by Hällgren et. al (2006) studying speech recognition in noise for Swedish speech, as such it was deemed an appropriate choice for the study. This due to the fact that the videos featured Swedish speech and the participants recruited for the study were native Swedish speakers.

3.4 Listening Effort and Noise

There have been some studies showcasing that the presence of noise leads to a increase of rated listening effort (Rudner et al., 2012). Likewise Sarampalis et al. (2009) showed that a reduction in noise leads to a reduction in cognitive effort. They showed this using a simultaneous separate task, which is a different measure than what will be used for this study. In a study by Larsby et al. (2005) they noted similarities between objective measures (dual task) and subjective measures (Borg CR10) but also distinct differences depending on the hearing - status of the participants. Specifically they noted that hearing impaired participants rated higher degrees of perceived effort in comparison to those that reported normal hearing when compared to their performance in the objective measure.

3.5 Comprehension and Noise

The effect noise can have on comprehension have been studied somewhat extensively, for example the Hearing In Noise Test (HINT) has been widely used for many studies (Nilsson et al., 1994). Generally results from these studies have shown that noise consistently impairs the ability of listeners to recognize speech (Friesen et al., 2001). This is due to the masking effects energetic noise can have where it partially or fully conceals or distorts utterances in a way where it is more difficult for the listener to decode the meaning of the phrase (Mattys et al., 2012).
3.6 Text Support

Subtitling has been a staple in much of media from the early 1900s (Ivarsson, 2004). The technology has mainly been used for translation but a large body of work also exists within closed-captioning which aims to make media more accessible for those with Hearing Impairments or complete deafness (Mahoney, 2022). In an effort to emulate these accessibility features companies such as Microsoft have begun to include captioning features in their video meeting software (Microsoft Teams, n.d.).

The effects of real-time captioning of content have not been widely studied but preliminary efforts generally show that it can be of value in improving speech intelligibility (Zhong, Noud et al., 2022; Zhong, Ricketts, et al., 2022). However, questions remain in regards to the importance of the accuracy of the captions as well as how it can potentially affect other cognitive resources. When listening to speech in adverse, noisy conditions more strain is put on cognitive resources such as WM (Rudner et al., 2011). It has not been widely studied, but can be considered likely, that similar strain is put upon WM when needing to decode captioning at the same time as listening to the content.

3.7 Borg CR10

To measure listening effort participants fill out a self-report questionnaire featuring an exertion scale known as CR10. The Borg CR10 scale has been used in similar contexts as the current study. For example Hua (2014) and Stenbäck et al. (2021), used it to measure listening effort in different speech recognition tasks. The scale was originally developed by Borg (1982) for use in measuring physical exertion levels in medical contexts. The scale is a self-report scale where the participant is asked to rate their exertion on a scale ranging from 0 (no exertion) to 10 (extreme exertion). To avoid the ceiling effect there is also an absolute maximum that is outside of the range that participants can report. The scale has shown to be reliable between tests and the exertion measured in the scale has correlated with heart-rate measurements in medical studies (Borg, 2013).

3.8 LIX

LIX is a measure developed by Swedish pedagogic researcher Carl-Hugo Björnsson in the 1960s. The measure is a simple way to measure complexity of a text by looking at sentence length and frequency of long words. The measure produces slightly different levels of complexity in different languages but for comparing texts using the same language, the scores are often seen as reliable and able to at least somewhat predict the approximate complexity of a sentence compared to others (Björnsson, 1983). The measure outputs a score from 0 to 100 where high scores translate to high complexity. For this study the measure was used to compare transcripts against each other to ensure relative similar complexity. Generally the scores have been divided into the following sub-groups (See Table 1; Lix Räknare, n.d.).
### Table 1: Shows the different categories that different LIX scores are described as.

<table>
<thead>
<tr>
<th>LIX</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 30</td>
<td>Very Easy to Read, Children’s Books etc.</td>
</tr>
<tr>
<td>30 - 40</td>
<td>Easy to Read, Fiction and Magazines</td>
</tr>
<tr>
<td>40 - 50</td>
<td>Medium difficulty of reading, Newspapers etc.</td>
</tr>
<tr>
<td>50 - 60</td>
<td>Hard to Read, Government texts etc.</td>
</tr>
<tr>
<td>&gt; 60</td>
<td>Very Hard to Read, Bureaucratic Speech</td>
</tr>
</tbody>
</table>

#### 3.9 PsychoPy

PsychoPy is an open-source tool developed to enable anyone with internet-access to be able to conduct experiments with millisecond accuracy (Peirce, 2007). The tool is very flexible and is well fitted for studies examining the effects of visual or auditory stimuli. It is possible to integrate eye-tracking but no such external tools were necessary for the experiments conducted in this study. The routines used for the experiments were custom-made to enable full randomization of how the conditions were paired with videos. All data was then saved in CSV-files (comma-separated values) which were then imported into statistical software.
4 Method

To be able to fully understand the impact text support can have on effort and comprehension the study used a 2x2 design where text support was studied both in normal and adverse conditions. By examining the listening effort and comprehension of participants after being exposed to content with and without text support in normal and adverse conditions, a reliable significant effect should be measurable if it exists. The participants watched several videos featuring two other meeting members who conducted a fairly normal meeting conversation (see Appendix 1 for the full transcripts in Swedish of each video. The videos underlined in red were not used for the experiment). After each video the participant answered a short survey featuring questions which measure their listening effort as well as a comprehension question regarding the contents of the video shown.

The study featured two independent factors each having two different levels. The factors were text support either being present or absent as well as noise either being present or absent. The videos that the participants were presented feature one of four conditions. Either the video has no added noise and no text support (condition 1, control condition), added noise but no text support (condition 2), no added noise but text support present (condition 3) or both added noise and text support (condition 4). Each participant was presented 16 videos featuring four videos from each condition, to avoid bias the order of videos and conditions was randomized for each participant. To measure the effect of text support and noise repeated measures ANOVA was used.

4.1 Participants

The participants of the study were primarily recruited from students attending Linköping University by sending out forms where participants could submit interest in participation, recruitment using word of mouth also took place. No compensation was awarded to participants. The total number of participants were 22 (female = 8). The mean age was 24.2 (SD = 3.27), the lowest age was 20 and the highest was 36. All participants self-reported as having normal hearing in the manner of not needing any form of hearing aids during their everyday procedures or video meetings.

4.2 Procedure

For the experiment all participants were seated in similar quiet and calm environments. All tests were done using the same setup (Laptop: Lenovo Legion Y540-15IRH, Display: 1920x1080 144Hz, Headphones: Audio Technica M50x) and the sound volume was the same for all participants. The participants were told not to remove the headphones during the video and that they could ask eventual questions to the test-leader in the pauses inbetween the videos. No other specific directions were given in regards for how participants should act or how they should be seated. This made it so that the participants could listen in a manner they found comfortable and hopefully more reflective of how they would be seated when attending a real video meeting.
Before the stimuli were presented a short text was shown preparing the participant to lessen effects of boredom and lack of vigilance. After this text a video plays for 35 seconds either with or without subtitles and added noise. The order of the conditions that were applied to each video was randomly generated for each participant and ensured that each condition would be equally represented. After each video finished playing the participant was asked to rate how effortful they found their listening experience on the BORG CR10 scale, as well as one question pertaining to the contents of the video to estimate how well they comprehended what was said. Each video had it’s own associated comprehension question (See Appendix 1.) which was developed and tested through the pre-study process. After answering these two items the participants were able to continue to the next video at their own leisure. After all 16 videos had been finished the participant was thanked for their participation and the program closed.

4.3 Ethics

Before the start of the experiment the participants were given a small introduction to what kind of material they would be presented and a consent form was also administered which specified how the data collected would be used. All data connected to the participant was anonymised after the experiment had concluded and it was not possible to connect specific participants to specific data points published in the finished article. Participants were also given all necessary contact information to remove their data points from the study if they wished to do so before or after the experiment had concluded. Due to these considerations the study is deemed to be compliant with The European Code of Conduct for Research Integrity (ALLEA, 2017) and their core principles of reliability, honesty, respect and accountability.

4.4 Material

The transcripts used were developed in collaboration with another researcher studying different stimuli in similar conditions. The dialogues were modeled after what the researchers themselves estimated to be probable and likely interactions to take place in a video meeting and the contents varied in subject from directly relevant to work conducted in the workplace, to social discussions that are in some way workplace related. Every interaction sequence was of roughly the same length (Number of sentences in sequence Mean = 6.5, $SD = 0.882$) and exhibited the same approximate complexity (LIX Score Mean = 22.6 $SD = 3.93$). For each sequence the researchers also created a yes/no question pertaining to the contents of the manuscript. The question was related to information that was in the middle of the script and some care was taken to not make the questions unclear or more difficult than necessary.
After 28 sequences had been developed which all fit within this span a pre-study was launched to select the 16 best sequences for use in the following main study. The pre-study was conducted by using a survey. The participants were asked for GDPR-consent as well as basic demographic information (age and gender). The participants were also asked how often they attend video-meetings on a 7-point likert-scale ranging from “never” to “almost daily”. For each dialogue sequence the participant was asked to read through it and then rate it on a 7-point scale for how likely it would be to occur in a workplace video meeting on a scale ranging from “not very likely” to “very likely”. For each sequence the participant also answered the associated yes/no comprehension question which would be used for the comprehension task in the main study. As such potential problems with the questions could be identified prior to the main study.

4.5 Pre-study results

The pre-study featured 26 participants with slightly more women answering (female, N = 15), the participants had a mean age of 38.4 (SD = 14.1) where the minimum age was 24 and the highest age was 62 and thus encompassing a large age range of the workforce. The participants in general rated themselves as attending video-meetings frequently with a mean self reported rating of 5.19 (SD=1.86) out of a possible 7 on a likert scale ranging from 1 corresponding to “never” to a 7 corresponding with “almost daily or daily”.

Of the 28 sequences featured, 24 scored a 4 (out of possible 7) or higher on how likely they would be to occur in a video meeting in a workplace and thus were deemed context appropriate for the study. The average correct answer for the comprehension questions was found to be 83.1%. To select the 16 best options out of this a filtering method was employed, the criteria was that the transcript needed to score above 70 percent on the question as well as having an average meeting likelihood score of over 4.3. After applying these criteria 16 videos were left. The value used for meeting likelihood was seen as the most important and as such a stricter bound was used for this criteria. 4.3 was seen as a value high enough above the middle (4) to signal that the conversation was quite likely to occur in a video meeting context. The limit for question accuracy was deliberately chosen to be as high as possible while still leaving 16 viable options that scored highly on meeting likelihood. The 16 sequences chosen for the study featured an average number of sentences of 6.5 (SD = 0.816) and the average LIX-Score was 21.8 (SD 4.26). All dialogues can be found in Appendix 1, the sequences underlined in red were eliminated during the filtering process.

4.6 Video Material

The 16 sequences chosen were then used as scripts for video recordings that would be used as the primary stimuli in the study. The videos feature two participants, one male and one female, having a video meeting in the software Zoom (n.d.). To limit the potential effects that differences in pronunciation or inflection can have on speech intelligibility
the participants were asked to speak in a neutral and similar manner throughout all videos. Likewise the actors were seated in rooms with controlled lighting and no visual distractions in the foreground or background. Since the videos featured one male and one female speaker both using different recording setups, some minor audio post-processing was done to minimize the effects of differences between the two speakers. Amplitude normalization was done using the limiter function (of the soft limit type) in the audio software Audacity (2023), equalizing the speaking-volume of the two speakers.

### 4.7 Statistics

The data collected during the study was transformed into a wide format CSV file using a script developed by Erik Marsja for the statistical software R (n.d.), and then analyzed using the repeated measures ANOVA function in Jamovi (n.d.). Repeated measures ANOVA was appropriate to assess significance and effect size of each independent variable for each dependent variable. A pairwise Bonferroni corrected comparison was used to further see how each separate condition affected the results by comparing their effect on listening effort and comprehension. Normality of data was tested by plotting residuals. All descriptive statistics presented throughout the study were collated by using Jamovi. The threshold for significance for this study was established as $p < 0.05$. 
5 Results

Due to the randomisation that took place, some videos were more commonly shown in a specific condition by random chance. For example we can see that video number 28 has a mean value of 0.318 in subtitles (see Table 2), meaning that it more commonly was shown without text support than with. With a large enough sample size these numbers would not be as varied. There is a possibility that it could affect the results in a minor way due to one video possibly being more difficult than another during adverse conditions or the presence of text support being more helpful for a specific video. However there are no indications that this affected the results of the study in any meaningful way.

While the playback of the videos always took the same allotted time there was some variance between participants in how much time it took to answer the questions pertaining to the video. Overall the mean time for the entire experiment (excluding information and consent collection before the experiment-program’s start) was 900 seconds or 15 minutes ($SD = 77.2$ seconds).

Table 2: Conditions which each video was shown in. Closer to 0 means it was more commonly shown with the condition absent and closer to 1 means the condition was more commonly present.

<table>
<thead>
<tr>
<th>Video</th>
<th>Subtitles</th>
<th>Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.409</td>
<td>0.545</td>
</tr>
<tr>
<td>3</td>
<td>0.500</td>
<td>0.500</td>
</tr>
<tr>
<td>5</td>
<td>0.591</td>
<td>0.591</td>
</tr>
<tr>
<td>7</td>
<td>0.409</td>
<td>0.455</td>
</tr>
<tr>
<td>8</td>
<td>0.455</td>
<td>0.409</td>
</tr>
<tr>
<td>10</td>
<td>0.455</td>
<td>0.409</td>
</tr>
<tr>
<td>16</td>
<td>0.636</td>
<td>0.409</td>
</tr>
<tr>
<td>18</td>
<td>0.773</td>
<td>0.500</td>
</tr>
<tr>
<td>19</td>
<td>0.591</td>
<td>0.591</td>
</tr>
<tr>
<td>20</td>
<td>0.364</td>
<td>0.545</td>
</tr>
<tr>
<td>21</td>
<td>0.500</td>
<td>0.591</td>
</tr>
<tr>
<td>22</td>
<td>0.545</td>
<td>0.318</td>
</tr>
<tr>
<td>23</td>
<td>0.409</td>
<td>0.500</td>
</tr>
<tr>
<td>25</td>
<td>0.500</td>
<td>0.455</td>
</tr>
<tr>
<td>26</td>
<td>0.545</td>
<td>0.636</td>
</tr>
<tr>
<td>28</td>
<td>0.318</td>
<td>0.545</td>
</tr>
</tbody>
</table>
5.1 Listening Effort

Table 3: Descriptive statistics of perceived listening effort separated by condition.

<table>
<thead>
<tr>
<th>No Text, No Noise</th>
<th>No Text, Noise</th>
<th>Text, No Noise</th>
<th>Text, Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean 1.90</td>
<td>5.85</td>
<td>1.43</td>
<td>4.96</td>
</tr>
<tr>
<td>SD 0.987</td>
<td>1.87</td>
<td>1.000</td>
<td>1.91</td>
</tr>
</tbody>
</table>

Table 3 shows mean values of listening effort separated by condition. We see that the conditions which feature noise have higher effort levels (Mean = 5.85, SD = 1.87 & Mean = 4.96, SD = 1.91) as compared to the condition were no noise was present (Mean = 1.90, SD = 0.987 & Mean = 1.43, SD = 1.000). Looking at Figure 1, we can see that the data is normally distributed for the listening effort measurement.

Figure 1: Residual plots for listening effort measurement
A repeated measures $2 \times 2$ ANOVA was conducted to measure the effects of the independent variables on listening effort. Significant effects were found for Subtitles ($p \leq 0.001$) Noise ($p \leq 0.001$) as well as the interaction effect between the two ($p = 0.015$). Using $\eta^2_G$ (Bakeman, 2005) we measured the size of the effect on effort and found that subtitles had a small to medium effect ($\eta^2_G = 0.050$), noise had a very large effect ($\eta^2_G = 0.616$) and the interaction between the two featured a very small effect size ($\eta^2_G = 0.005$).

**Table 4: Repeated Measures ANOVA showing the effect of noise and subtitles on listening effort.**

<table>
<thead>
<tr>
<th></th>
<th>$p$</th>
<th>$\eta^2_G$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subtitles</td>
<td>$\leq 0.001$</td>
<td>0.050</td>
</tr>
<tr>
<td>Noise</td>
<td>$\leq 0.001$</td>
<td>0.616</td>
</tr>
<tr>
<td>Subtitles * Noise</td>
<td>0.015</td>
<td>0.005</td>
</tr>
</tbody>
</table>

After the pairwise comparison the effects measured are shown to be significant and persistent in each condition (See Table 5). When noise is present the effort is higher as compared to when it is absent. In contrast the presence of subtitles makes the effort decrease when compared within the same noise condition. Comparing NoText - Quiet and NoText - SSN (Mean Difference = -3.945, $p < 0.001$), with Text - Quiet and Text - SSN (Mean Difference = -3.533, $p < 0.001$), we see that the mitigating effect of text support is larger when noise is present. Which is consistent with the interaction effect measured. We can also see this in Figure 2 which shows that the two lines are not perfectly parallel, that is to say the mitigating effect of subtitles is affected by the noise condition.

**Table 5: Post Hoc Comparison of the effect of noise and subtitles on listening effort.**

<table>
<thead>
<tr>
<th>Condition 1</th>
<th>Condition 2</th>
<th>Mean Difference</th>
<th>$P_{bonferroni}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Text &amp; Quiet</td>
<td>No Text &amp; SSN</td>
<td>-3.945</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>No Text &amp; Quiet</td>
<td>Text &amp; Quiet</td>
<td>0.472</td>
<td>0.006</td>
</tr>
<tr>
<td>No Text &amp; Quiet</td>
<td>Text &amp; SSN</td>
<td>-3.061</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>No Text &amp; SSN</td>
<td>Text &amp; Quiet</td>
<td>4.481</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>No Text &amp; SSN</td>
<td>Text &amp; SSN</td>
<td>0.885</td>
<td>$&lt; 0.001$</td>
</tr>
<tr>
<td>Text &amp; Quiet</td>
<td>Text &amp; SSN</td>
<td>-3.533</td>
<td>$&lt; 0.001$</td>
</tr>
</tbody>
</table>
Figure 2: Repeated measures ANOVA showcasing effect of Subtitles and Noise on Listening Effort

5.2 Comprehension

Table 6: Descriptive statistics of comprehension separated by conditions.

<table>
<thead>
<tr>
<th></th>
<th>No Text, No Noise</th>
<th>No Text, Noise</th>
<th>Text, No Noise</th>
<th>Text, Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.818</td>
<td>0.830</td>
<td>0.898</td>
<td>0.920</td>
</tr>
<tr>
<td>$SD$</td>
<td>0.207</td>
<td>0.142</td>
<td>0.148</td>
<td>0.142</td>
</tr>
</tbody>
</table>

In table 6 we can see that mean values for comprehension was quite similar regardless of conditions. Condition with subtitles present feature higher mean values ($Mean= 0.898$, $SD = 0.148$ & $Mean= 0.920$, $SD = 0.142$) in comparison to the conditions when subtitles are absent ($Mean = 0.818$, $SD = 0.207$ & $Mean = 0.830$, $SD = 0.142$). Looking further on the descriptives for each condition we can see that data is not normally distributed (see Figure 3.)
Using a $2 \times 2$ repeated measures ANOVA on comprehension a significant effect was found for subtitles ($p = 0.012$). However this was not the case for noise, nor was an interaction effect present. Looking at $\eta^2_G$, subtitles were shown to have a small to medium effect size on comprehension ($\eta^2_G = 0.068$, see Table 6).

**Table 7: Repeated Measures ANOVA showing the effect of noise and subtitles on comprehension.**

<table>
<thead>
<tr>
<th></th>
<th>$p$</th>
<th>$\eta^2_G$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subtitles</td>
<td>0.012</td>
<td>0.068</td>
</tr>
<tr>
<td>Noise</td>
<td>0.576</td>
<td>0.003</td>
</tr>
<tr>
<td>Subtitles * Noise</td>
<td>0.892</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Through a pairwise Bonferroni corrected comparison (Table 8 & Figure 4) we see that the effect of text is positive, meaning participants in general have a higher comprehension accuracy when subtitles are present as compared to when they are absent.
Table 8: Post Hoc Comparison for the effect of subtitles on comprehension.

<table>
<thead>
<tr>
<th>Condition 1</th>
<th>Condition 2</th>
<th>Mean Difference</th>
<th>$P_{\text{bonferroni}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Text</td>
<td>Text</td>
<td>-0.0852</td>
<td>0.012</td>
</tr>
</tbody>
</table>

Figure 4: Repeated Measures ANOVA showing effect of Subtitles on Comprehension
6 Discussion

The current study set out to assess the effects of noise and text support on comprehension and listening effort using three different research questions. Viewing the results in the context of these questions we can distill the effects of these phenomena. To answer research question 1 we can interpret from the results that subtitles decreased listening effort of participants as well as being a potential positive factor in comprehension related tasks. For research question 2 the results revealed that noise had a major influence on how effortful listening was but did not seem to indicate a worsening performance in the case of the comprehension task. Research question 3 can be answered by looking at the interaction effects of the tests. While no significance was found for the interaction effect of noise and subtitles on comprehension, a significant but small effect was measured for the interplay of the factors upon listening effort (Table 5 & Figure 1). That is to say that the mitigating positive effect that subtitles had on effort was greater when noise was added as compared to when it was absent. Overall the results show a consistent and positive impact of captioning regardless of conditions.

These results are consistent with what has been shown in earlier studies in regards to the positive aspects text support can have upon legibility (Zhong, Noud et al., 2022; Zhong, Ricketts, et al., 2022). Regarding the questions that was posed in regards to the potential negative effects of text support on listening effort, the results of the current study indicate that no increase of listening effort is found with the presence of text support. On the contrary, the results suggest that text support decreases the self reported listening effort both in normal and adverse conditions.

No significant effect of noise on comprehension was found in the study. This was somewhat unexpected since it would be expected based on the assumptions done by the author based on results from earlier research (Friesen et al. 2021, which also used speech-shaped noise). The cause for this is somewhat uncertain but could be due to the task not being difficult enough or perhaps the sample size for the study was just not large enough to be able to show the effect. The fact that data was not normally distributed for the comprehension measurement could have had an effect, however this is not likely something that affected the analysis majorly due to the nature of Repeated Measures ANOVA (Blanca et al., 2023). It is possible that the methodology used is so different from HINT that it is not possible to expect similar effects to be measurable. Another possibility is that the two forced-alternative-choice method (yes/no) used for this study puts less strain on WM than the reproduction or multiple choice questionnaire measures often used in speech recognition studies (Margolis et al., 2021). This could lead to a case where the increase in WM-load due to presence of noise did not lead to a significant difference in comprehension scores in the different conditions.

The age of the participants in the study generally skewed younger, this might lead to some concern regarding how generalizable the results are outside of this specific segment of the population. Generally members within this age group are believed to be more used to video-conferencing and as such might have techniques to more easily make up for technical problems such as noise. However, as shown by the results from this study a measurable
effect of noise upon listening effort exists even for this group and as such it can reasonably be expected that adverse conditions will have similar effects, or maybe even worse, for members outside of this age-range and that captioning likely will have a similar mitigating effect. As noted in Larsby et al. (2005) older participants made more errors in the tasks during adverse conditions impacting comprehension negatively. While they did not find a significant effect of age upon perceived effort other studies have been able to show the impact age can have on listening effort (Degeest et al., 2015). Nonetheless the findings of this present study indicate that comprehension was improved with text support and as such it is believed that it will have a larger positive impact for older individuals due to the difficulties Larsby et al. (2005) mentioned in their study.

6.1 Consequences

Some questions may arise in regards to the generalizability of the results from this study. However the results indicate that the presence of subtitles, at least when they are fully correct, leads to a decrease in listening effort and increase in comprehension. Thus their inclusion in software used for video conferencing could possibly lead to a better working environment for not only deaf or hearing impaired individuals, but really everyone. The results of this study do not indicate any negative effects due to presence of subtitles regardless of conditions, as such their inclusion seems to be beneficial in software for video meetings.

6.2 Method Criticism

While the study ideally would have used a fully simulated active meeting for each participant where they can fully interact with others (e.g. the DiaPix task; Baker & Hazan, 2011. See 6.3 for a discussion), this was not feasible for this project. Instead videos featuring a video-meeting were used where the participant was treated as a passive participant. While this does not fully recreate the conditions that the participant would normally conduct video-meetings in, it can be considered a fair and realistic presentation which would be adequate for studying the phenomena that is of interest to this study. This choice also maintains the same conditions between different participants thus enforcing the internal validity of the study.

6.2.1 Listening Effort Measurement

In FUEL they raise the importance of including some sort of measurement for when a participant will quit in their listening (Pichora-Fuller et al., 2016). This was not done throughout this study and as such the methodology was not fully FUEL compliant. This likely could have been an issue if the amount or length of the videos had been higher or if the volume of the noise was heightened. However no indication was seen during the conducted tests that participants quit in their listening. As such the results are believed to be reliable and well representative of how the listening effort of the participants was affected by the different conditions.
6.2.2 Comprehension Measurement

The method used for measuring comprehension in the current study (yes/no question) is quite unusual and differs from much earlier research which has had a greater focus on speech recognition and/or reproduction after exposure to stimuli (Margolis 2021; Dewyer 2017). It is highly likely that participants could have been able to use context clues found within other speech to be able to answer some of the questions. For example the comprehension related question for video 10 (See Appendix 1.) where the words in surrounding speech do not communicate a sense of discontent which would be likely if the answer to the question had been yes. That means that even if the participant missed out the actual reason for the employees departure in the script they would still be able to correctly answer the comprehension related question. While this is somewhat true to how we would comprehend speech in real video meetings it is a factor which could impact the results of the study majorly. In future it would be advisable to ensure that participants would not be able to answer comprehension question based only on context clues from surrounding speech. That is to say regard should be taken to the design of sequences so that the likelihood of either yes or no being correct would be similar if the target phrase was removed from the script.

6.2.3 PsychoPy Criticism

While PsychoPy is a useful tool which effectively and easily allows researchers to construct experiments, some minor deficiencies were found with the software during the course of the study. One such deficiency was differing delays between audio and video, most often ranging from asynchronity in the range of 0-20 ms (mean = 0.0105s, $SD = 0.0073s$). This should not affect the results of this study in any meaningful way since this range lies comfortably within the Temporal Binding Window for speech (461ms) which means that it should not meaningfully nor negatively impact effort (Stevenson & Wallace, 2013). As such no adjustments were made to correct for this throughout the rest of the study.

6.2.4 Text Support Criticism

The captions used in the study were 100% accurate to the contents of the video. This differs from the expected accuracy of automatic-captioning tools used for most video conferencing tools, for example Microsoft Teams captions generally lies within an accuracy range of 85%-90% (Normand, 2022). While this choice limited the studies’ ecological validity it prevented bias in how subtitling errors could differ between different videos. One further concern is that the subtitles are shown for the entire phrase all at once instead of being generated alongside the utterance of the video participants. This was a limitation of the software used for creating the subtitles and would optimally be improved upon in future studies studying live-captioning to be more similar to real world uses.
6.2.5 Noise Criticism

Analyzing SSN using a spectrogram from Academo.org (n.d.), we can see that the noise features signals in a span from 0 Hz to roughly 13 KHz (Figure 3). However the main intensity of the noise is in the range of 80 Hz to 1.4 KHz, the same span as most peoples natural speaking range. The noise is mostly consistent with minor fluctuations in different frequency bands. Due to the fact that SSN is slightly more intense in the sub-352 Hz band it would be expected that it would affect the understanding of the contents of the male speaker (Right side Figure 4) more than it would affect the female speaker (Left side Figure 4). However several participants reported verbally that they found it more difficult to hear the female speaker. This could be due to a variety of factors. While the sound volume throughout all videos was normalized there were still differences in the recording equipment used for the different speakers. As can be seen in the spectrogram picture the male speakers speech is far more noisy while the female speakers speech is more staccated and as such it might be more susceptible to be drowned out by the noise. It is also possible that the lower frequencies of the male speaker’s voice were easier for some participants to hear through the noise due to the noise not being as intense below 80 Hz. The present study tried to control for this by varying which speaker communicated the comprehension related item in each recording. However for future studies examining noise in similar contexts it would be recommended to ensure the same recording conditions and equipment for all participants to hopefully counteract technical issues that could have an undue effect on results.

While speech-shaped noise has been widely used in earlier research some regard should be taken to how the effects upon cognitive aspects can differ from other more natural types of distractors. For example ”babbel” noises or other informational maskers could quite likely have different outcomes on listening effort and comprehension since the strategies to counteract those kinds of distractors differ from the listening strategy used for the type of energetic noise which was used in the current study. For example Brungart (2001) showed that the ability to recognize what was said was easier with energetic masking when compared to informational masking. While there are strengths to SSN considerations should be taken to how generalizable the results of the current study are in regards to other kinds of noise.
Figure 5: Spectrogram of the SSN - Noise used in the study main intensity 80 Hz - 1.4 KHz

Figure 6: Spectrogram of an excerpt from one of the audio transcripts. Female speaker (Left) main intensity 100 Hz - 1.5 KHz, Male speaker (Right) main intensity 60 Hz - 1.5 KHz
6.2.6 LIX Criticism

Using lix as a generalized measure of readability has been criticized due to its limitations in not separating between commonality of different words nor the phonetic characteristics of words. The measure also takes no consideration of differences that can occur due to conjugation of verbs or compound words. These oftentimes register as long words, and therefore increase the LIX score (less readable) even though they might in fact not be as difficult as shorter, more uncommon words. While this is fair criticism of the measure’s efficacy at its stated goal, the use of the measure for this specific study was limited to ensure relative similar complexity between short transcripts. While this is not the purpose that LIX has been designed for, it was seen as a sufficient way to compare the texts specifically in the context of how it was used for this study.

6.3 Limitations and Future Research

Some regard should be taken to how applicable the results of this study are due to limitations such as number of participants, ecological validity of subtitle technology used, the design of the comprehension questions, static noise vs more natural sounding distractors etc. Each of these were discussed in detail above and future studies would do well to take into regards how these potentially could affect outcomes. While for example the choice of noise and subtitles limit the ecological validity of the study, it is not believed to have impacted the results majorly. It would be optimal to in future study if informational noise has the same impact and how the mitigating effect of text support is affected by it as well as if it impacts comprehension differently from the energetic noise used for this study. Many of the results of the study were significant, however a larger sample size might have been able to lead to more insightful analysis and an overall broader scope of the study. In future it could also be desirable to study the effect that listening effort could potentially have on comprehension, as it would seem reasonable that a higher demand on the effort required during listening might lead to a lessened ability to understand and decode what is being said.

Ideally an interactive environment would be used in future studies. It might be possible to adapt a version of the DiaPix task (Baker & Hazan, 2011). In this adaption two participants could collaborate over video meeting in conditions with and without automatically generated text support in order to solve the task. This would solve some of the issues identified throughout the course of the present study. However new problems would likely arise, mainly related to reliability. The current methods for generating text support in real time have issues with accuracy (Normand, 2022) especially in cases where participants have accented speech. However if possible to overcome such issues it would be useful to study how the interactive nature that video meetings have in naturalistic settings would put a different strain on the cognitive aspects of participants. It would be optimal to develop a framework which could be used to study cognitive aspects (such as WM, listening effort, comprehension etc.) during conversations both in person as well as when communicating over telecommunications.
7 Conclusions

The findings of the study paint a compelling picture regarding the positive impact text support can have on the listening effort of participants. Some minor questions still remain regarding how well text support helps comprehension although the findings are positive in this regard as well. In future it would be beneficial to further study these phenomena in more naturalistic settings and with a larger number of participants.
8 References

ALLEA (2017), The European Code of Conduct for Research Integrity Revised Edition. ALLEA - All European Academies


One platform to connect. (n.d.). Zoom. Retrieved May 9, 2023, from https://zoom.us/


