How Face Mask Design Affects Speech Intelligibility

Liam Scott

May 5, 2021

This study analyses the effects of face masks on speech intelligibility to determine if they cause a significant reduction. To achieve this, three methods are used: the Speech Transmission Index (STI), Visual Modified Rhyme Test (V-MRT) (a proposed modification on the Modified Rhyme Test (MRT) to include the effects of visual cues) and a subjective test (ranking the masks, 1 - 5). It concludes that (whilst in one test all of the masks had a significant effect on speech intelligibility) due to discrepancies between the tests results, there is insufficient data to arrive at a single, unchallengeable outcome and therefore further testing is required.

Contents

List of Tables

List of Figures

1. Acknowledgements

I want to take this opportunity to thank my supervisor, Lee Davison for his guidance and support throughout working on this project. My thanks also goes to Sky Hawkins for their advice on this project's direction. I, of course, want to recognise all the participants for taking part in my testing and the Solent University Media Tech group for facilitating my dumb questions and putting up with my barrage of DM's.

I also want to offer my congratulations to Hendrik Erz for creating Zettlr, a free and open-source program which made collecting my thoughts and writing this document an absolute pleasure to do!

Word	Meaning
V-MRT	Visual Modified Rhyme Test
MRT	Modified Rhyme Test
STI	Speech Transmission Index
COVID-19/COVID	The Coronavirus pandemic (SARS-CoV-2)

2. Introduction and Objectives

In a global pandemic, it is vital to limit the spread of infection. In the case of COVID-19, wearing face masks is expected to be an integral step in slowing transmission. The World Health Organisation (WHO) (2020) has put out guidance to wear masks and the UK government has chosen to enforce this (GOV.UK 2021).

Whilst there have been investigations and research on the effect of wearing a face mask on speech intelligibility (SI), they are mainly focused on the medical industry's application. However, as of COVID-19, average citizens are now wearing masks daily and must communicate effectively whilst wearing them. Face masks provide both physical and acoustic hindrances like the covering of the mouth and effecting the ability to lip read and overall being able to effectively communicate with others. This has led to complaints of struggling to talk, hear and generally be understood whilst wearing a mask (Ong 2020; Valerie Fridland 2020).

Although the ability to communicate effectively in a crisis is extremely important, there is no mention of speech intelligibility in both old and new medical face mask standards (BSOL 2009; British Standards Institution 2019). This document also aims to provide actionable SI statistics for the most common mask types used by the public during COVID.

These results will be compared against IEC 60268-20 (International Standards: Sound system equipment–Part 16: Objective rating of speech intelligibility by speech transmission index) in order to determine if the effect of wearing masks on SI is considered significant.

Measurement methods were carried out according to existing standards and purpose-modified standards to meet the requirements for COVID lockdown conditions. Available research and data were analysed to determine gaps in the existing literature and current findings.

3. Background

The following section discusses the background theory related to speech intelligibility (SI). It focuses on the key factors affecting SI and offers explanation on how to factor them into testing. The main existing methods for testing SI are introduced for both predictive and perceptual testing, analysing their strengths and weaknesses.

3.1. Acoustic & Psychoacoustic Factors

Speech intelligibility is the measurement of how comprehensible speech is to a listener. It is affected by a number of factors: speech level, background noise, reverberation, frequency and visual cues (French and Steinberg 1947).

3.1.1. Background Noise and Signal-to-Noise Ratio

Signal to noise ratio (SNR) is the ratio of the desired information (in this case speech) versus the background noise and is a key factor affecting speech intelligibility (IEC 2020).

When speech suffers from a low SNR, it becomes less intelligible due to psychoacoustic masking from background noise.[1] Normal conversation level is around 60 dBA and 'can be understood fairly well in background levels of 45 dBA,' which would correspond to a signal to noise ratio of +15dB (WHO, n.d.). However, Lane and Tranel (1971) noted that a speaker will involuntarily amplify their voice in the presence of loud noise to enhance their intelligibility (Lombard effect).

Multiple studies on the effect of background noise have shown its detrimental effects on concentration and communication (Hodge and Thompson 1990; Banbury and Berry 2005; Soli and Sullivan 1997). For these reasons, during testing background noise should match realistic levels as much as possible and accurately simulate its impact in a real world scenario.

When performing background noise measurements, an A-weighted curve is generally selected for low to mid-level noise levels as it represents the frequency response of the human ear at 40 phons (Robinson and Dadson 1956). It is vital to take measurements over a reasonable period of time and to average the measurement across this time to produce an equivalent constant level, this limits the impact of abnormal, impulsive noises which would otherwise unduly alter the measurement. This is referred to as an LEq (Equation 1).

$$L_{Aeq} = 10log[\frac{(t_1 \times 10^{\frac{L_1}{10}}) + (t_2 \times 10^{\frac{L_2}{10}}) + ... + (t_n \times 10^{\frac{L_n}{10}})}{T}], dBA \quad (1)$$

[1] Psychoacoustic masking is the phenomena of when a sound with greater energy masks a sound with less energy. This can be in the both the frequency or time domain (Moore 2012).

Where t_n is the measured period of time for noise level L_n.

Table 2: Speech level (dB A-weighted) of male and female talkers at 1-meter (ISO 9921; Berger et al. 2003)

Vocal Effort	Male Talker	Female Talker
Low (relaxed)	54	50
Normal	60	55
Raised	66	62
Loud	72	71
Shout	78	82

3.1.2. Frequency

Within speech, consonants provide informational content. This can be demonstrated by the fact that when removing all vowels from a word, it usually remains intelligible, whereas when all consonants are removed, the word is usually unintelligible. Howard and Angus (2017) noted that consonants are normally -20dB quieter than vowels in speech, and operate at a higher frequency than vowels, making them more vulnerable to intelligibility loss by acoustic absorption (Peterson and Barney 1952).

A study by Alexander Goldin, Barbara Weinstein, and Nimrod Shiman (2020) found that (of the three medical masks tested) all masks acted with low pass characteristics' and attenuated frequencies above 2 kHz. This suggests that face masks will have an effect on SI due to it's filtering effect increasing the loss of constants.

3.1.3. Reverberation Time

Room Reverberation produces a similar masking effect, reflecting altered and time delayed reflections of the speech. 'Reverberation tends to attenuate the rapid modulations of speech by filling in the less-intense portions of the waveform' (Assmann and Summerfield 2004). This causes a low-pass filtering effect, attenuating the higher frequency consonants, reducing SI.

However, a study by J. Bradley, Reich, and Norcross (1999) found that 'the SNR aspect is shown to be much more important than room acoustics effects.' This suggests that if testing is carried out in an environment where reverb is not a factor, you can still receive usable results e.g., wearing headphones or in an anechoic chamber.

3.2. Visual Cues

A great deal of speech intelligibility processing occurs subconsciously. Humans perceive audio and visual content as interconnected and so rely on both for SI, this is achieved via a combination of observation, and contextual information (McGurk and MacDonald 1976). A study by Sumby and Pollack (1954) involving speech intelligibility with and without being able to see a video of person speaking found that SI was significantly higher when participants were able to see the speaker compared to an audio only test condition. Furthermore, they indicated that 'as the S/N ratio decreases, the importance of the visual cues to listener-intelligibility increases.' Later testing by Neely (1956) found that (when combined with auditory cues) the presence of visual cues can improve SI results by around 20%.

The Royal National Institute for Deaf People estimate around 12 million people suffer from hearing loss in the UK alone (RNID 2018). Hearing loss usually manifests in the average person as they grow older (Presbycusis) and consists of progressively losing the ability to perceive the higher frequency of the audible spectrum (above 3000 Hz) (Nadol Jr 1993). This results in a low pass filtering effect and can cause the loss of certain higher frequency consonants (such as *s*, *h* or *f)*). These people often report that they can hear, but not understand (Turner and Cummings 1999).

A key tool used by individuals with hearing loss is lip reading (speech reading), however with many masks designs the mouth is partially or even completely obscured, making this task very difficult. Furthermore, multiple studies have determined that the average person with normal hearing also has a high level of speech reading ability and rely on this subconsciously (VanBebber 1954). This demonstrates that to accurately measure SI, visual cues should be factored into testing.

3.3. Existing Methods for Testing SI

There are many established methods to measure speech intelligibility. They fall into two categories: perceptual (participant based) and predictive (measurement based) methods for testing speech intelligibility.

3.3.1. Perceptual Tests

Perceptual tests of speech intelligibility measure the percentage of correctly guessed and understood speech units e.g., words, phrases, sentences. The testing can use synthetic, recorded or live speech (with effects like the Lombard effect considered).

MRT (Modified Rhyme Test) The Modified Rhyme Test (MRT) is a closed-set subjective method for calculating speech intelligibility and is specified in ANSI/ASA S3.2-2009 (ANSI 2009). Subjects are played a carrier sentence "*You will mark ___ now.*" and are made to identify the spoken monosyllabic word from a small list of other similar words whilst artificial background noise is played simultaneously. This allows the simulation of a realistic environment and is commonly used to measure the effects of noise on SI. The MRT is used extensively by the US military to test their communication systems (Pollard and Garrett 2017).

As the MRT only uses words as its speech unit there is no subjective bias[2] to the test. Furthermore, multiple studies have found that subjects show very little evidence of learning during repeated tests, allowing for the testing of multiple masks in one large test (House et al. 1965; Williams and Hecker 1967; Nye and Gaitenby 1973). However, using only monosyllabic words produces results which are less realistic than full sentences or passages.

Connected Speech Test (CST) The Connected Speech Test (CST) is another subjective method for calculating speech intelligibility and is mainly used for evaluating hearing aids. Subjects are first shown the contextual topic of the test phrase and are then played a CST passage (in the presence of babble). They must repeat back what was said, which is recorded and converted to rationalised arcsine units (rau) for further analysis (Cox, Alexander, and Gilmore 1987).

The CST performs better as a predictor of real-world intelligibility than the MRT as it uses full sentences. However, despite being told what the context is before seeing the

[2] "Subject bias, also known as participant bias, is a tendency of participants (subjects) in an experiment to consciously or subconsciously act in a way that they think the experimenter or researcher wants them to act." ("Subject Bias in Psychology: Definition & Examples" 2017)

sentence, subjective bias is still likely to occur, albeit in a more controlled fashion. Furthermore, the CST has not been adopted by any notable technical standards, despite multiple papers using the method for testing the effects of masks on SI (Mendel, Gardino, and Atcherson 2008; Atcherson et al. 2017).

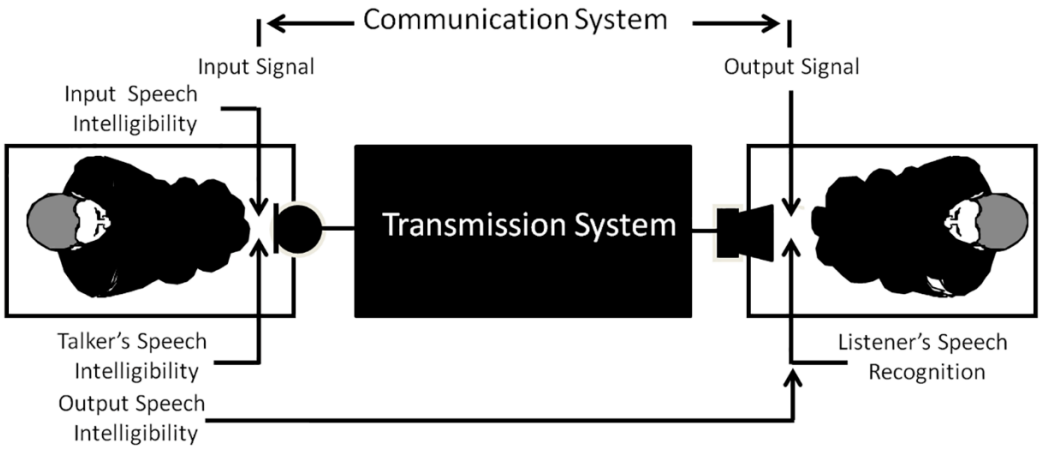

Figure 1: Generic Perceptual Test Setup (Letowski and Scharine 2017)

Visual Testing All discussed methods in this section concentrate on the audio aspect of SI and do not account for the SI effects of visual cues. There are ways to measure the effect of visual cues, but they rely on specialist equipment and training which is not reasonably accessible for the purposes of this study. Furthermore, these methods would vastly limit the sample size and over complicate the testing process. Therefore, this paper proposes a modification on the existing audio-only MRT methodology called the *V-MRT* (subsection 4.3).

3.3.2. Predictive (Technical) Tests

As perceptual ratings and tests are based on the perception of speech by listeners, they are not limited by the characteristics of the test environment. However, they are time-consuming and costly to perform.

On the other hand, predictive tests are fast and inexpensive to perform, and SI can be calculated in real time. 'They do not actually measure speech intelligibility but predict it on the basis of the transmittability of the input signal' (Letowski and Scharine 2017).

There are 3 main standardised predictive speech intelligibility testing predictors detailed in ANSI/ASA S3.5 and IEC 60268-16:

Articulation Index (AI) Articulation Index (AI) is "based on the concept that the overall speech intelligibility of a communication system results from independent intelligibilities calculated from the peak-speech-to-root-mean-square measured in selected frequency bands from 200 to 7000 Hz" (Fletcher and Steinberg 1929; French and Steinberg 1947). There are three common methods for calculating AI: octave band method, third-octave band method and the critical band method.

This produces a resulting value between 0 (poor) and 1 (perfect intelligibility). A score of 0.3 or below is considered bad, 0.3 - 0.5 is poor/satisfactory, 0.5 - 0.7 is good and above 0.7 is excellent.

Despite AI remaining a valid way of testing SI, it is commonly replaced by the SII method.

Speech Intelligibility Index (SII) Speech Intelligibility Index (SII) is generally considered as the updated version of the Articulation Index (AI) and is based on the same underlying theory. However, it provides a more generalised framework;[3] allowing more flexibility when defining the basic input variables and reference point of the measurement (Hornsby 2004). "The concept of the SII is based on measuring the SNR across a number of frequency bands, and weighting each SNR by band-importance functions that are dependent on speech material" (Letowski and Scharine 2017). It is calculated as follows:

$$SII = \sum_{i=1}^{n} I_i A_i \tag{2}$$

Where n refers to the number of frequency bands, I_i refers to the importance of a given frequency band (i) to speech understanding and A_i is band of audibility (ranging from 0 to 1).

It is susceptible to giving artificially low scores if any compression is introduced and distortions from reflections and echoes can invalidate the test (Letowski and Scharine 2017). However, performing tests in an anechoic environment significantly reduces the risk of these events and results in accurate, usable data.

Speech Transmission Index (STI) The Speech Transmission Index (STI) is another popular method for objectively evaluating speech intelligibility. It is widely used in Europe and is specified in IEC 60268-16:2020 (IEC 2020).

Due to the length and complexity of the Full-STI test, several simplified versions of STI

[3] A structure that is used to build something.

have been created. RASTI (Rapid Speech Transmission Index) was developed which only uses the 2 octave bands centred around 500 - 2000 Hz (with a set of 9 modulation frequencies). However, it was known to frequently overestimate SI (when compared to results from the AI and SII tests) and as such is no longer a generally accepted method (IEC 2020).

An alternative simplification utilises a testing procedure using a STIPA (Sound Transmission Index Public Address) signal which reduces the time for taking measurements significantly. This comes at the cost of the frequency range tested, as the STIPA signals usable range is between 125 Hz to 8 kHz. Despite this, a study from J. S. Bradley, Reich, and Norcross (1999) found that there is a "just noticeable difference" between the results of a STIPA test and a full STI test and so can be used.

Although predictive tests are fast and easy to perform, they are limited by their chosen algorithm. All discussed predictive methods, do not account for the effects of nonstationary noise on speech communication and assume either monaural or diotic (the same signal at both ears) speech reception (rather than binaural/dichotic hearing) (Letowski and Scharine 2017).

Common Intelligibility Scale To compare different SI methods results, a common denominator is required. The Common Intelligibility Scale (CIS) was developed by Barnett and Knight (1996) and was designed to do exactly that. The main advantage of the CIS is SI scores are easily converted for different methods. Barnett based the CIS in relation to STI (Equation 3).

$$CIS = 1 + log(STI) \tag{3}$$

However, Barnett and Knight (1996) admit that where the gradients of curves are shallow, the CIS values are not reliable. As a result, despite the CIS allowing an approximation of the values predicted from one testing procedure to the other, it cannot be depended on for any rigorous testing.

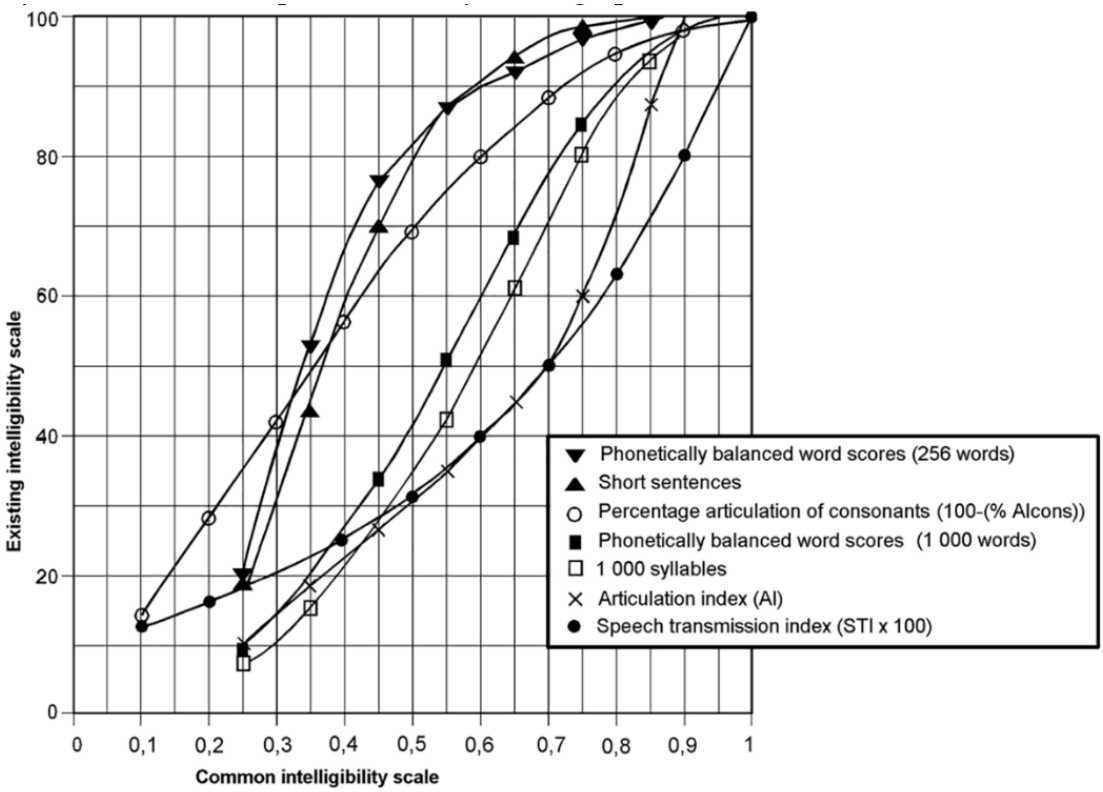

Figure 2: Relationship between various SI expressed on the CIS (Barnett and Knight 1996)

4. Method

This study occurred during a period of restrictions connected to the SARS-CoV-2 pandemic. As such testing was designed to operate in compliance with the Health Protection (Coronavirus, Restrictions) (England) Regulations 2020.

4.1. Masks Tested

When measuring the effect of mask design on speech intelligibility, it is important for an appropriate variety of commonly used masks be tested. Therefore, a range of locally available and used masks were chosen. From this sample, the following masks were selected.

1. Control (no mask)
2. Medical (Type II Disposable surgical mask meeting BS EN 14683:2019 (British Standards Institution 2019))
3. Snood (Buff/tube scarf)
4. Homemade (following guidelines from WHO (2020))
5. Face Shield
6. N95 (meeting FFP2 standards (British Standards Institution 2009))

Figure 3: Face Masks Types

4.2. Participants

Measures were in place to ensure that Participants had the following attributes:

Listener:

- English as first language
- Age: 20 - 30
- Had practised for the V-MRT before undertaking the final test
- Were confident on the day of taking the test
- Had given consent to participate within the study and understood their right of withdrawal.

4.3. Perceptual Method (V-MRT)

4.3.1. Modifications to the MRT

The MRT method was the backbone of the created V-MRT (Visual Modified Rhyme Test) and was adjusted to include visual cues by integrating video of a person (speaker) reading the test content aloud, switching between masks as needed.

A pilot study was conducted to investigate the feasibility of the use of this modified test procedure. This had 5 participants complete an early version of the V-MRT where the same 50 speech units (from the MRT) was used per mask. This study found that (despite findings from House et al. (1965) suggesting otherwise) selecting the same words over again resulted in bias, with some participants answering before the question was fully read out.

Therefore, the existing MRT word list was used but divided to allow for 10 different words per mask resulting in no repetition of selected words between masks, eliminating this issue. This also reduced the time to complete the test, which can reasonably be expected to reduce the dropout rate of the study.

Table 3: Equipment list for V-MRT video production

Equipment	Use
Computer (w. monitor)	To run software used in pre and postproduction. Windows x64 was used.
Webcam (Logitech C920)	To record video of V-MRT. 1280x720p (HD) minimum.[4]

[4]The final video was recorded at 1920 x 1080p, 30 FPS in MOV with PCM audio stream (muted in editing).

Equipment	Use
Measurement Microphone (Superlux ECM999)	For recording audio for the V-MRT and measuring background noise.
Audio Interface (Focusrite Clarett 2PreUSB)	For controlling and amplifying the input signal. Ensuring a balanced interface is used with no pin 1 problems (Fox and Whitlock 2010).
XLR Cable	Connecting microphone with audio interface.
Software	
Webcam Software (Open Broadcast Software Studio)	For recording the webcam's output, ensuring webcam focus or any visual artefacts whilst recording.
Editing Software (Black Magic DaVinci Resolve)	For editing the V-MRT video and normalising audio.
DAW (Audacity)	To monitor audio level when recording and basic trimming in post. The Audacity project was set with 48 kHz sample rate (standard in video production), 32-bit float, mono.
Text Editor (Vim)	Containing the list of correct V-MRT answers to be read out.

V-MRT production

The measurement microphone was positioned 10 cm away from the speaker and was angled towards the speaker's mouth. During the recording of test data, the speaker monitored their voice levels to ensure reasonably consistent levels and no clipping occurred. The test data was performed and recorded 5 times to ensure good takes for all questions.

As the test data was recorded in sub-optimal conditions, a measurement of the background noise was performed to determine any hidden high frequency or loud low/mid frequency content (a continuous noise around 130 Hz was found but was disregarded as the added background noise would mask it). The video was then sent for post processing.

Due to the remote nature of this adapted technique, it was necessary to ensure participants would be listening at an appropriate level, as each participant would have a different audio system, and would likely have a different normal listening volume.

At the beginning of the video a 1 kHz tone at -60 dB compared to the main content was played with instructions (on screen text) advising participants to set their volume so 'they

were just able to hear it.' This was designed to match the 60 phons curve and normalise the tests (Figure 4).

Additional instructions on how to perform the test where added, along with the background noise (paragraph 4.3.2) (-15 dB below test data)[5]. This was then rendered as an MOV file with a PCM audio stream (to avoid the effect of digital compression). This file was uploaded to Tresorit Send[6] as an 'uncompressed file' and distributed to participants.

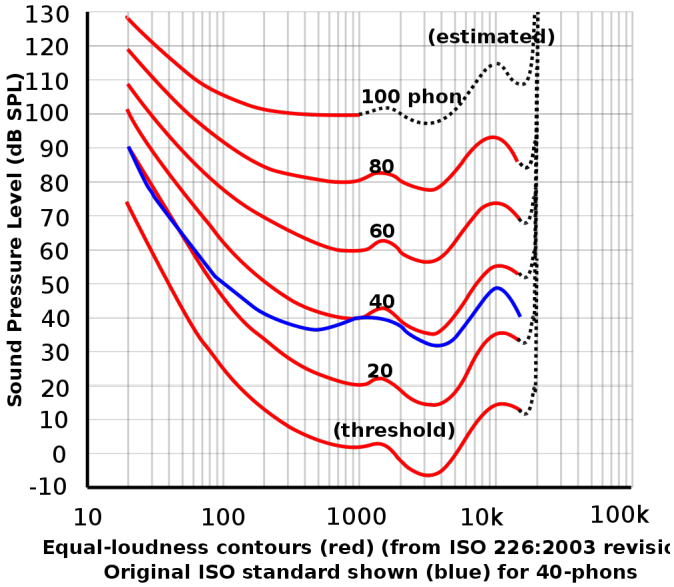

Figure 4: Equal Loudness Contours (60 Phons) (Commons 2017)

4.3.2. Visual Modified Rhyme Test (V-MRT) Methodology

Preliminary Background Noise Measurements

Background noise measurements were conducted using an NTi XL2-TA and measurement microphone (this can be substituted for a measurement microphone, audio interface and computer using Audacity). The LEq of a local pub (at peak time) was measured 5 times to give an average and realistic worst case scenario. This data is then processed in Audacity's filter curve plugin to created a filter curve that would represent its sonic properties (Figure 5).

Pink noise was chosen as it has equal power at each octave across the audible spectrum and results in humans not having a bias to a certain frequency band (Szendro, Vincze, and Szasz 2001). The use of pink noise allows for continuous noise level without any transients that may occur in a location recording, allowing for precise control of the signal

[5]Following guidance from the WHO and data from noise measurements later described (WHO, n.d.)

[6]Tresorit is a cloud storage provider with end-to-end encryption (any cloud provider can be used, so long as minimal/no compression is used) ("End-to-End Encrypted Cloud Storage for Businesses | Tresorit" n.d.). Although sharing files across physical drives is better.

to noise ratio. A subtractive filter curve was applied to the pink noise (also generated with Audacity) and was used as the background noise. This is used in all further testing to achieve similar masking effects to actual location background noise.

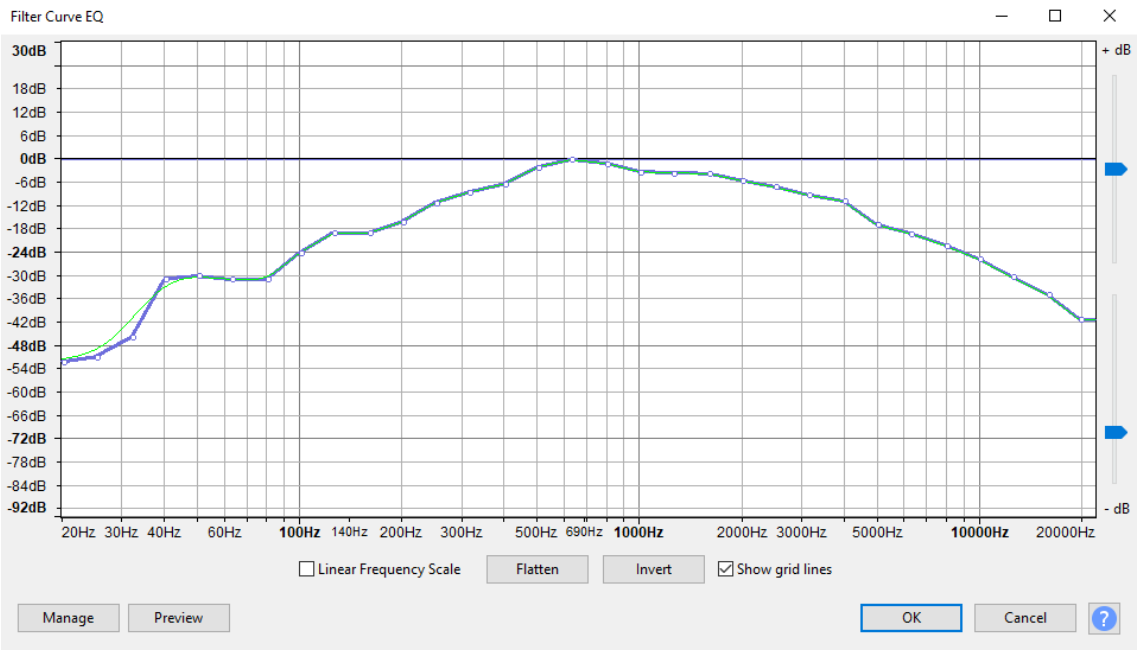

Figure 5: Audacity Filter Curve Plugin w. location filter curve

Table 4: Participant equipment list for V-MRT testing

Equipment	Use
Headphones	To listen back to the V-MRT. Headphones are varied depending on participants choice (preferably meeting ITU-R BS.708-R).
Computer (w. monitor)	To open the online Form and play the V-MRT video.
Software	
Online Form (Google Forms)	For presenting the V-MRT questions and allowing for responses.
Media player (VLC)	For playing back the V-MRT video. VLC was chosen as it is free, open-source and compatible with all major operating systems (Windows, Mac, Linux, OpenBSD).

Method

The form begins by asking for the consent of being part of the test and does not allow the participant to go further without agreeing. Next, they are asked their sex (male or female) and details of any (known) hearing problems. Next their model and brand of headphones used for testing is given (including any specialist audio gear in their system). This is designed to help catch any anomalous results that may occur by using any digital bass enhancements or if the headphones cannot replicate the required frequency range (accurately).

The participant was then asked to watch the V-MRT video in which a person wearing a mask (or not for the control) says the carrier phase 'Mark the word ___ ' and then a word (following the MRT methodology) in the presence of background noise. The participant is then asked to pause the video and select their response out of the 6 choices on the Google form. A practise question is played to help introduce the levels of background noise (SNR) and confirm the method of answering the questions. This is repeated for the duration of the test (50 questions) after each 10 questions changing to a different mask type (or control). The participant then sends off their form for analysis.

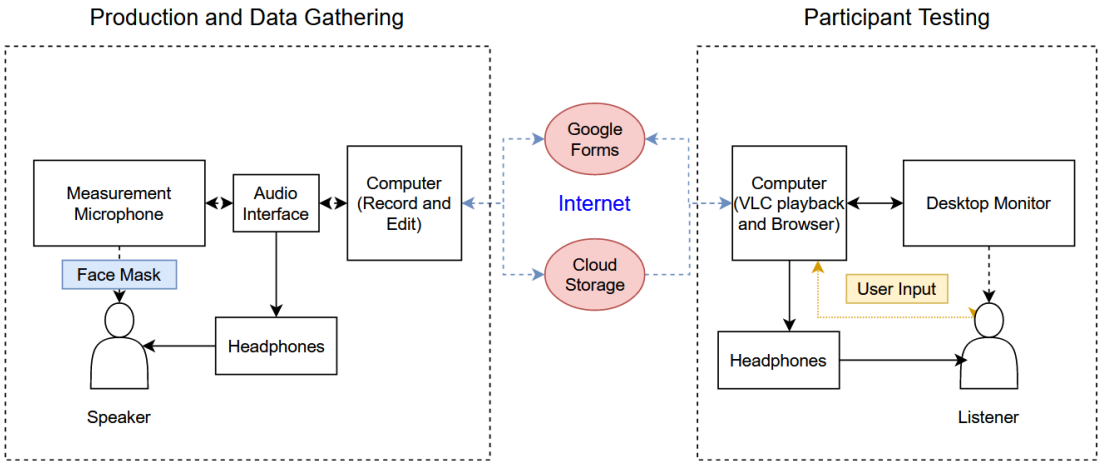

Figure 6: V-MRT Workflow

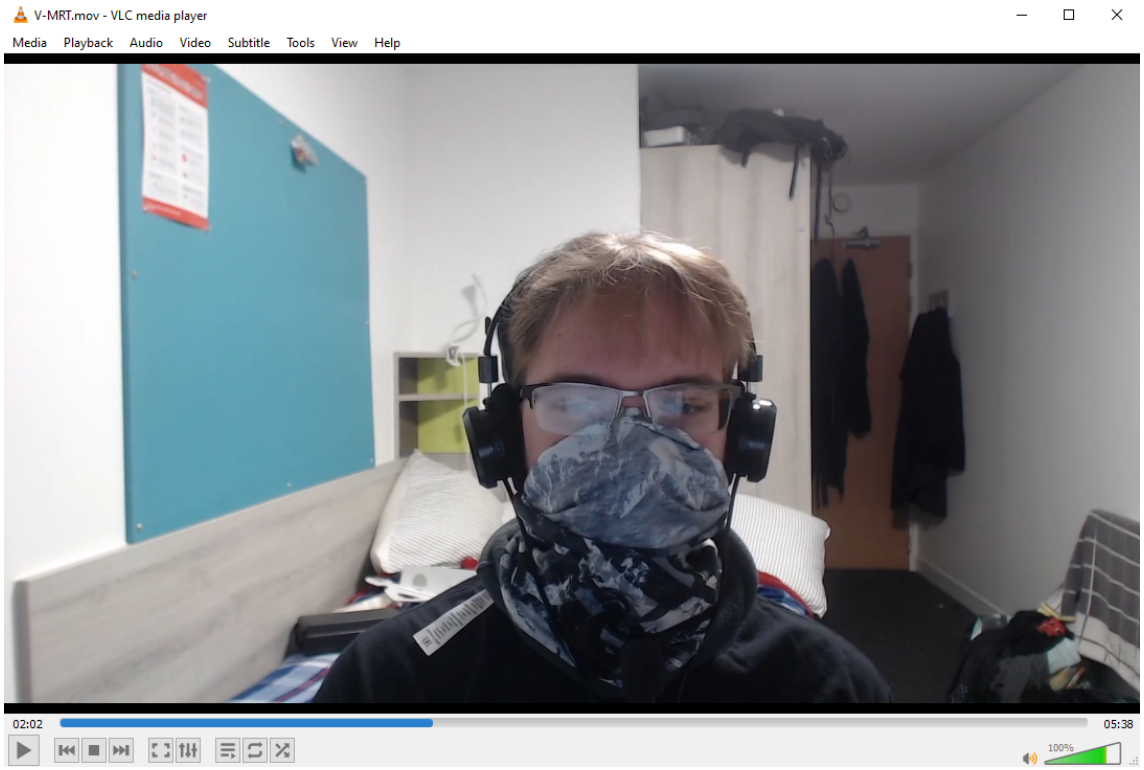

Figure 7: Still from V-MRT video - VLC Player

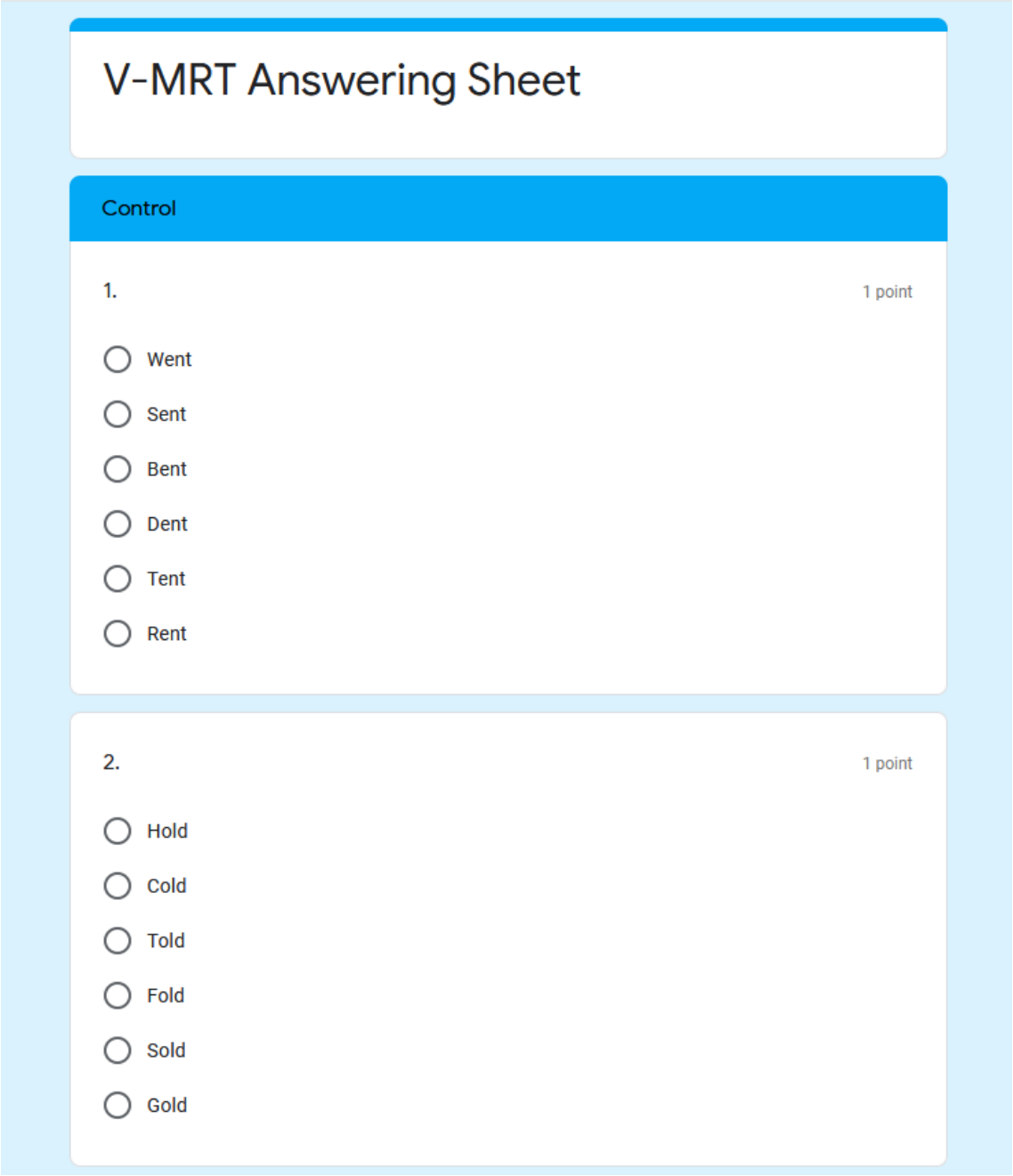

Figure 8: V-MRT Google Form

4.3.3. Subjective Testing

At the end of the V-MRT test form, there is the option to answer additional subjective questions. These fell into two categories:

- The perceived differences between the mask types on their effect on speech intelligibility (ranking each mask),
- The quality of the V-MRT proceedings

An option for additional comments was included in the test in case a participant wanted to leave a more detailed review or other queries.

To edit this text, Microsoft's Azure Cognitive Services Text Analytics service was used ("Text Analytics | Microsoft Azure" n.d.). This allowed for unstructured text to be analysed using a machine learning technique called natural language processing (NLP). This automatically detected and output key identifying features of the text including topics, people, language used (i.e. English) and the text sentiment.[7]

[7] Sentiment analysis (or emotional mining) is the used to identify a subjects emotional state to something (i.e. text) ("Text Analytics | Microsoft Azure" n.d.)

4.4. Predictive Method (STI)

4.4.1. Method

This methodology is based on the standard IEC 60268-16:2020:

Table 5: Equipment list for STI testing

Equipment	Use
Hemi/Anechoic Chamber	Measurement location: To avoid the effect of early reflections, reverberation or external noise.
NTi Minirator MR-PRO (Signal generator)	Used to generate Pink noise and STIPA signal (standard NTi Minirator can be used instead by loading the STIPA signal onto an SD card).
BRÜEL & KJÆR TYPE 4128-C Head and Torso Simulator (HATS)	To simulate a realistic mouth, output test signals and allow easy placement of face masks.
Unity Gain Power Amplifier (DIY in-house build)	To provide unity gain power to the HATS.
NTi XL2-TA (HANDHELD AUDIO AND ACOUSTIC ANALYZER)	Used for SPL and STIPA measurements.
NTi M4260 (Class 2, 1/4” Measurement Microphone)	Working in conjunction with the XL2-TA to measure SPL and STIPA.
Microphone Stand	To elevate, angle and keep steady the measurement microphone
XLR Cable	To connect the the XL2-TA and Measurement microphone
Banana Leads (x2)	To connect the HATS and the unity gain power amplifier
Measuring Tape	To ensure alignment and measure the distance between of microphone and HATS.
OPTIONAL: TYPE 4231 Sound Calibrator	To calibrate the measurement microphone with XL2-TA.

The equipment is setup as seen in Figure 9 in an hemi-anechoic chamber. The measurement microphone was placed 1m away[8] on a microphone stand (to avoid contact noise) and directed at the HATS mouth. The NTi Minirator MR-PRO was then used to generate pink noise and an NTi XL2-TA was used to measure the SPL level. The noise level was adjusted to 60 dB A (representing a human at normal conversation level (Table 2)). Then the NTi Minirator MR-PRO was set to use the STIPA test signal and the XL2-TA to the STIPA measurement setting. Three control STI measurements (without a face mask) were made and recorded. The first mask was then placed over the mouth of the HATS, and measurements where taken and recorded. This was repeated for the other masks; snood, Homemade, disposable surgical mask, face shield and N95.

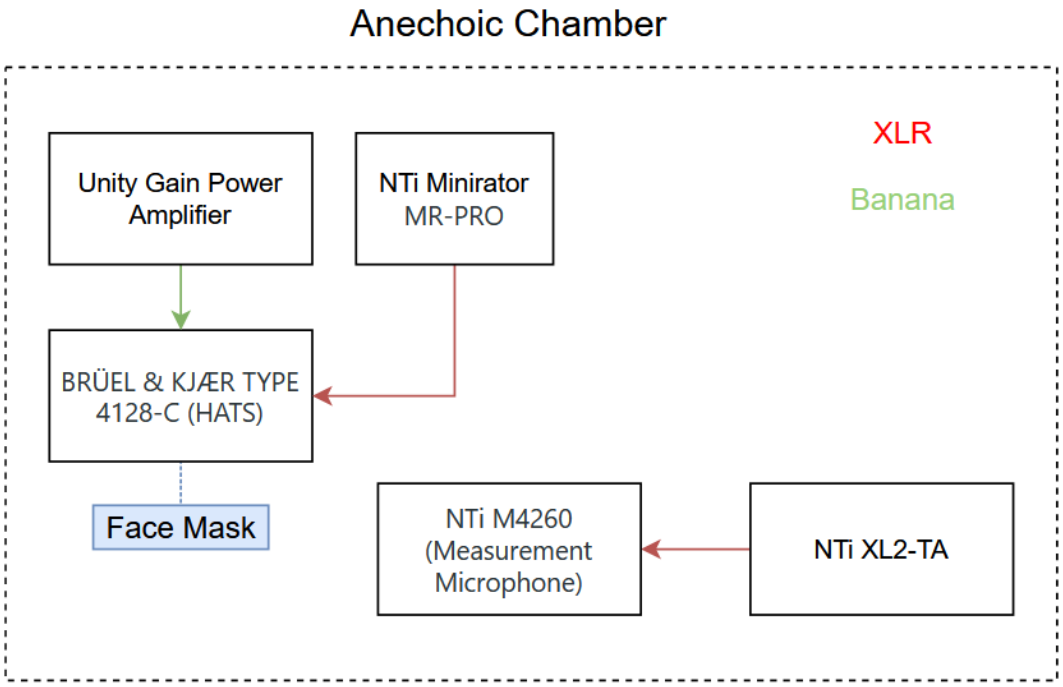

Figure 9: STI Flow Diagram

[8]For these measurements the amplifier was not powerful enough to generate 60 dB A at 1m. Subsequently, the microphone was placed at 0.5m with the SPL adjusted accordingly.

Figure 10: STI equipment setup (HATS with N95 mask)

Figure 11: STI Equipment Set Up (Control)

5. Results Analysis

This chapter presents the results from the V-MRT, STI and other gathered data from the online form. It offers analysis and description of the data and (where relevant) comparisons between the different SI measurements to find correlations.

5.1. Data Processing

The V-MRT data was first processed by identifying and removing anomalous results. To achieve this, three factors where considered:

- Suspiciously high or low scores
- Known low quality headphone brands and models
- Participant identified hearing problems

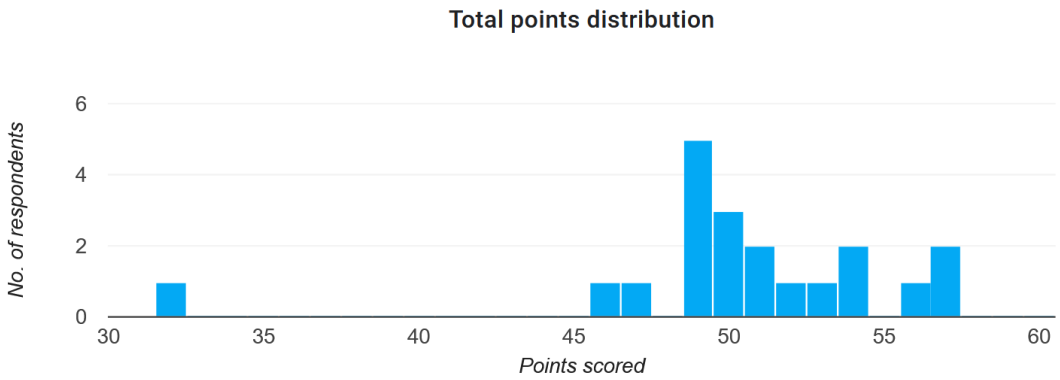

Figure 12: RAW V-MRT Total Points Distribution

Figure 12 shows one outlier result with a score of 32 out of 60. This participant self-reported suffering from a known 'glue ear in the right ear.' Therefore, it was disregarded from all the results (as well as the subjective test results). This was the only outlier removed from the test data.

5.2. Comparing STI and V-MRT

Table 6 shows the relationship between STI and alternative subjective measures/SI ratings (Palmiero et al. 2016; IEC 2020). The V-MRT uses words as it's speech unit, as so it is comparable to the score given in the percentage intelligibility of words (ANSI 2009). These results where compared to quantify the effects of visual cues and subjective bias to that of the results based solely on acoustic and frequency effects.

Table 6: Relationship Between STI, Subjective Intelligibility Measures and Intelligibility Ratings

STI Value	Quality according to IEC 60268-16	Intelligibility of Syllables in %	Intelligibility of Words in %	Intelligibility of Sentences in %
0 - 0.3	bad	0 - 34	0 - 67	0 - 89
0.3 - 0.45	poor	34 - 48	67 - 78	89 - 92
0.45 - 0.6	fair	48 - 67	78 - 87	92 - 95
0.6 - 0.75	good	67 - 90	87 - 94	95 - 96
0.75 - 1	excellent	90 - 96	94 - 96	96 - 100

As the initial data collected by the V-MRT is considered RAW data, it must be adjusted to account for random chance success. Each test had 6 options to select from, meaning guessing purely at random would be expected to achieve a 17% score (ANSI 2009).

$$R_A = R - W/(n-1) \tag{4}$$

Where R_A is the number of correctly guessed items (adjusted), R is the raw number of items guessed correctly, W is the number of incorrectly guessed items, and n is the number of alternative choices per item.

5.2.1. V-MRT Results

A pass/fail criterion of ≥ 87% correctly guessed words or 0.6 STI was established based on recommendations in IEC 60268-16:20. This means that a pass would meet the minimal intelligibility needed to be rated as 'good.'

Due to the small sample size (19 participants) the results validity of the V-MRT were a concern. To counter this, one tailed paired T-tests were performed to test if the data collected was valid.

Table 7 shows all T-test tests achieved over 99% confidence, with 95% being the accepted threshold for confidence ($p \leq 0.05$) .

Therefore, the results can be considered as not random, and it can be concluded there is a significant difference between the control results and all other mask types.

Table 7: V-MRT data

	Control	Face Shield	Home Made	Medical Mask	N95	Snood
RAW	182	165	160	151	157	159
R_A	180	161.25	155	143.75	149.75	153
Avg Mean	95.3%	82.6%	83.7%	84.2%	79.5%	86.8%
Std Dev (SD)	0.061	0.124	0.125	0.083	0.126	0.088
n	19	19	19	19	19	19
p-value		2.62442E-05	0.000927634	6.4834E-05	0.000147282	0.000172673

Figure 14 illustrates the normal distribution of the V-MRT mean scores compared to their respective mask types. Testing found that the snood performed the best with a mean intelligibility of 86.8%, from there the medical mask performed second best (84.2%), followed by the home made mask (83.7%), the face shield (82.6%) and the N95 (79.5%). All of these masks produced a notable reduction compared to the control (unmasked, normal speech).

As shown in Figure 13, none of the masks tested met the pass criteria of 87% intelligibility. The mean average standard deviation for the face masks was 0.1092, suggesting that the final mean scores are widely varied and display significant crossover between different mask types. For the masks with a higher SI score (medical mask and snood) the standard deviation is like the controls (0.061). However, the homemade mask, N95 and face shield (which score lower) have a greater standard deviation (between 0.124 - 0.126).

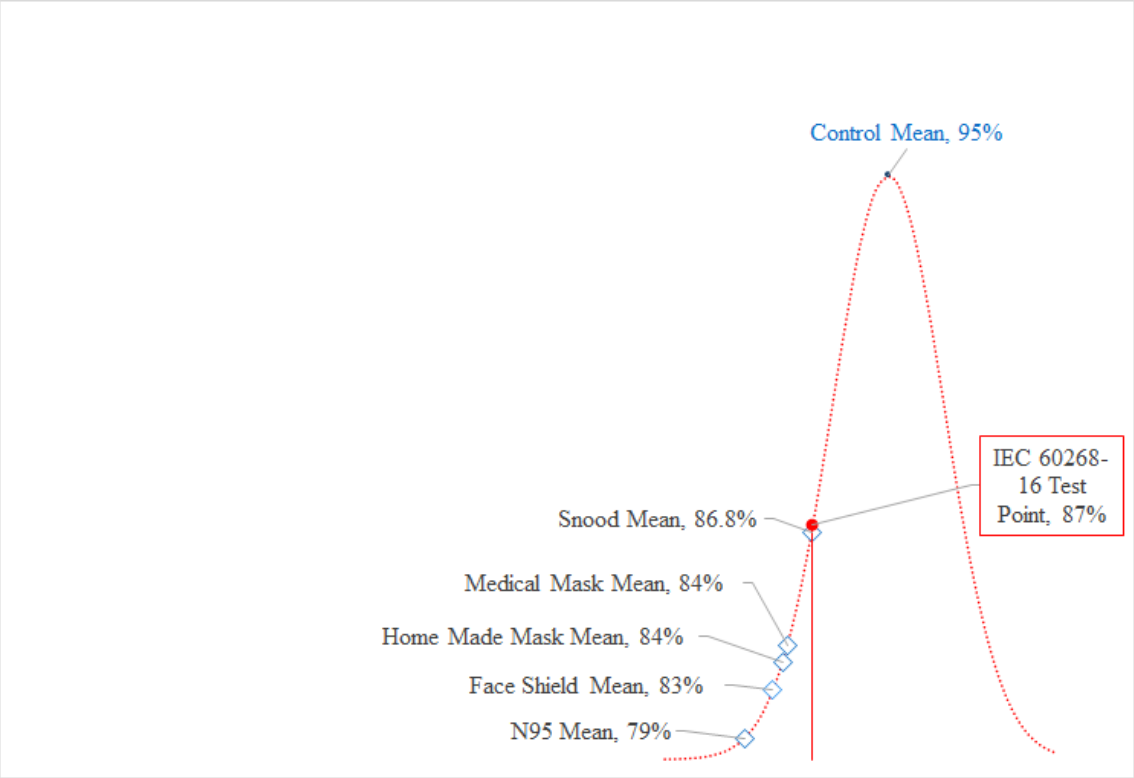

Figure 13: % Accurate Responses by Mask Type Mean of Samples mapped to Control Normal Distribution (V-MRT)

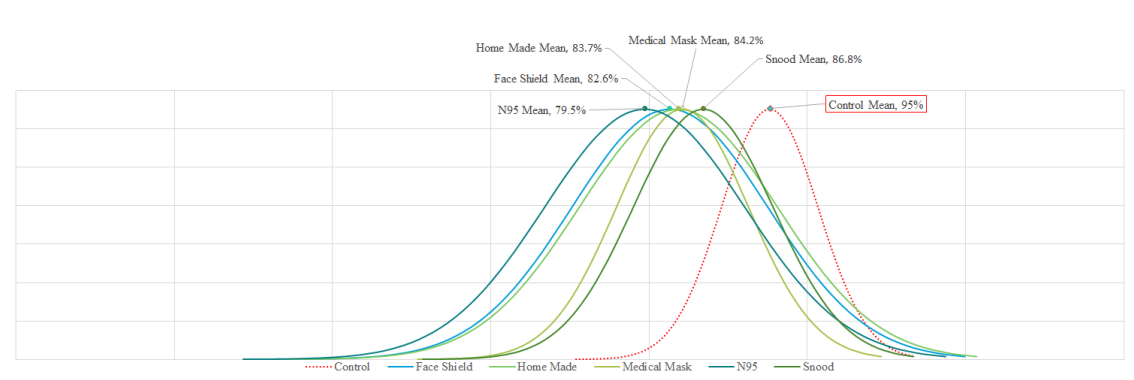

Figure 14: Normal Distribution showing % Accurate Responses by Mask Type

5.2.2. STI Results

The STI acts as a representation of the physical acoustic reduction in intelligibility caused by the masks. This notably does not account for the visual cues incorporated into the V-MRT score. Table 8 presents the data collected from the STI tests.

Table 8: STI results

Mask Type	1	2	3	Mean (STI)	SD
Control	0.94	0.98	0.97	0.963	0.017
Snood	0.94	0.97	0.96	0.957	0.012
Medical	0.97	0.95	0.97	0.963	0.009
N95	0.87	0.84	0.84	0.850	0.014
Face Shield	0.86	0.82	0.79	0.823	0.029
Homemade	0.94	0.92	0.92	0.927	0.009

Although the V-MRT and STI results are very different from each other, performing a Pearson coefficient correlation shows there is still a strong/medium positive correlation between the two tests ($r = 0.532$, $p = 0.0497$).

Table 9, shows there is no noticeable difference between the control and medical mask results (both showing a mean average of 0.963 STI) and only a slight reduction with the snood (0.957 STI). However, the V-MRT scores 95% for the control (0.75 - 1 STI and 'excellent') followed by a large reduction, with the snood mask (86.8%) continuing with the medical mask (84.2%), a 2.6% reduction. This places them in the 'excellent,' 'fair' and 'fair' categories respectively and is equivalent (at a minimum) of a 0.15 STI reduction. The homemade mask scores 0.927 STI (within the 'excellent' category) but receives an 88.7% V-MRT rating (in the middle of the 'fair' category). The N95 and face shield have the greatest effect on SI for both tests with the STI test scoring 0.85 STI and 0.82 STI (still considered 'excellent') and the V-MRT scoring 83% and 79%.

The STI test results have an average standard deviation of 0.01246 (this could be related to the small sample size) and finds both the highest and lowest scoring mask types to have the largest variation.

Table 9: Comparing V-MRT score with STI results and IEC 60268-16 rating

Mask Type	Mean (V-MRT %)	IEC 60268-16 rating (V-MRT)	Mean (STI)	IEC 60268-16 rating (STI)
Control	95.3%	Excellent	0.963	Excellent
Snood	86.8%	Fair	0.957	Excellent
Medical	84.2%	Fair	0.963	Excellent
Homemade	83.7%	Fair	0.927	Excellent

Mask Type	Mean (V-MRT %)	IEC 60268-16 rating (V-MRT)	Mean (STI)	IEC 60268-16 rating (STI)
Face Shield	82.6%	Fair	0.823	Excellent
N95	79.5%	Fair	0.850	Excellent

5.2.3. Subjective Questions Results

Participants were asked to self-assess the perceived intelligibility of each mask through a rank order list of all five masks.

As can be seen in Table 11, the subjective rating on the effect of mask design on SI follows a very similar trend to that of both the V-MRT and STI test results. The medical mask performs the best with a score of 28, this is closely followed by the snood (31). The homemade mask causes a large score increase (+25) scoring 56, this is seen again with the face shield (+25) scoring 81 and finally the N95 performs the worst (89).

The correlates with both the STI and V-MRT results, finding a very similar mask ranking (with the snood following the STI results as the best mask types, but finding the N95 the worst type like the V-MRT). However, Table 11 also shows there is a wide variation between results for all mask types, with an average standard deviation of 0.5306.

Table 10: SI measurements mask type ranking (1 - 5) (1 is best, 5 is worst)

Mask Type	V-MRT Ranking	STI Ranking	Subjective Ranking
Snood	1	2	2
Medical	2	1	1
Homemade	3	3	3
Face Shield	4	5	4
N95	5	4	5

Table 11: Participant perceived mask intelligibility ranking (lower the better)

Mask Design	Total (/285)	SD
Medical	28	0.612
Snood	31	0.597

Mask Design	Total (/285)	SD
Homemade	56	0.404
Face Shield	81	0.562
N95	89	0.478

5.2.4. Test Evaluation (Additional Comments)

Participants were asked to provide feedback on the test, which was then used to prompt further investigation into potential limitations of the test procedure.[9] Six comments were given, with two removed due to irrelevance. This left four comments providing usable feedback, all of which represented participants who performed to a reasonable standard within the test.

The length was well received, and participants found that most but not all participants felt the test was easy to understand. Some commented on the quality of audio, highlighting that some plosive sounds were able to form distortions despite the use of a pop filter. Furthermore, one participant indicated surprise at the level of attenuation provided by the Face Shield test portion. One participant flagged the use of Google Forms as an issue, citing potential privacy concerns relating to the operating model of Google LLC.

[9]The full text from these comments are not reported, as they contained identifying information about the participants.

6. Evaluation of the V-MRT

As a result of the COVID pandemic, resources where sparse. Therefore, the proposed methodology for the V-MRT had limitations that would not be required with the correct access to personnel and equipment. On the other hand, this also gave the opportunity to design a method that had benefits the original MRT did not have. This section covers those limitations and benefits in detail also giving examples from the data collected in this investigation.

6.1. Limitations of the V-MRT

6.1.1. Audio setup variation

Ideally, every participant should have used the same audio setup. However, the COVID-19 lockdown made this very difficult, and a compromise was made to allow participants to use headphones they owned. Although the effect of this was minimised by asking participants for their model and brand of headphones, each test still had its own unique sonic properties colouring the results.

This can be solved by either, increasing the data set (which would result in a better average of results and help minimise the effects of anomalous results) or (after lockdown) testing each of the participants in a controlled environment where the same audio setup can be used for each test.

6.1.2. Speech Patterns

Finding and training speakers and participants was time consuming. As a result of this, in testing a single male speaker was selected (author of this paper). Not only does this add gender bias but also only tests one speech pattern and accent.

When analysing the frequently missed questions (classified as questions that scored below 50%), most of the incorrect answers where very close to the correct answers e.g., ‘kin’ and ‘king.’ This misidentification could have been caused by all manner of external problems but was most likely due to the speaker’s enunciation. The speaker tended to soften the ‘ng’ sound making it sound like an ‘n,’ likewise when comparing a hard ‘c’ or ‘k’ to a ‘t’ (sometimes not even saying the word’s ending).

As the speaker in question is the author of this paper (and subsequent creator of the test), it is likely their accent was not factored in when producing the V-MRT video. Despite this not coming up in feedback, to improve the test and avoid bias in the future, the speaker should be an external party.

Question	Correct responses
12.	4/19
24.	7/19
25.	7/19
37.	4/19
39.	8/19
48.	9/19
56.	9/19

Figure 15: List of frequently missed questions from the V-MRT

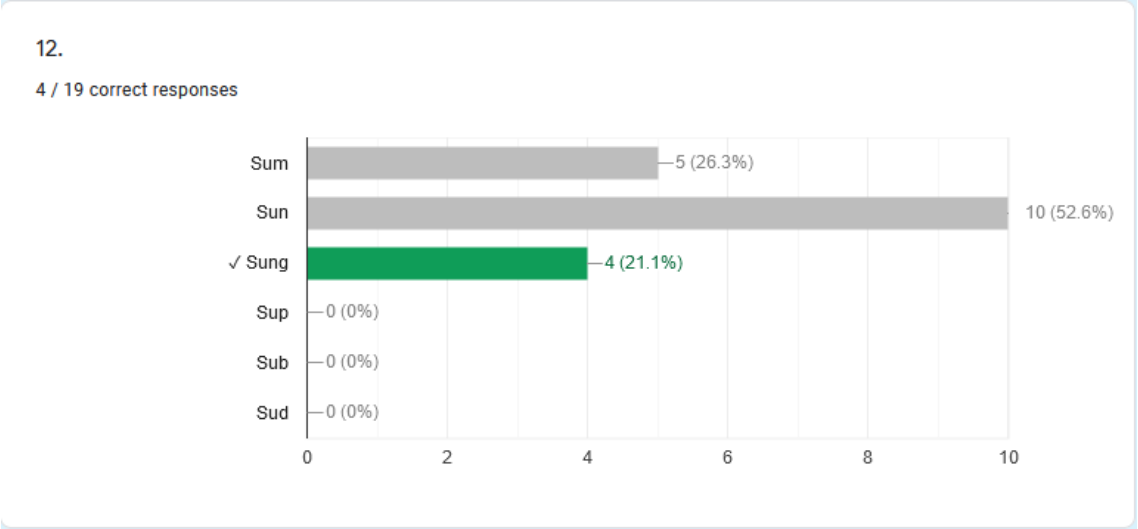

Figure 16: V-MRT Question 12

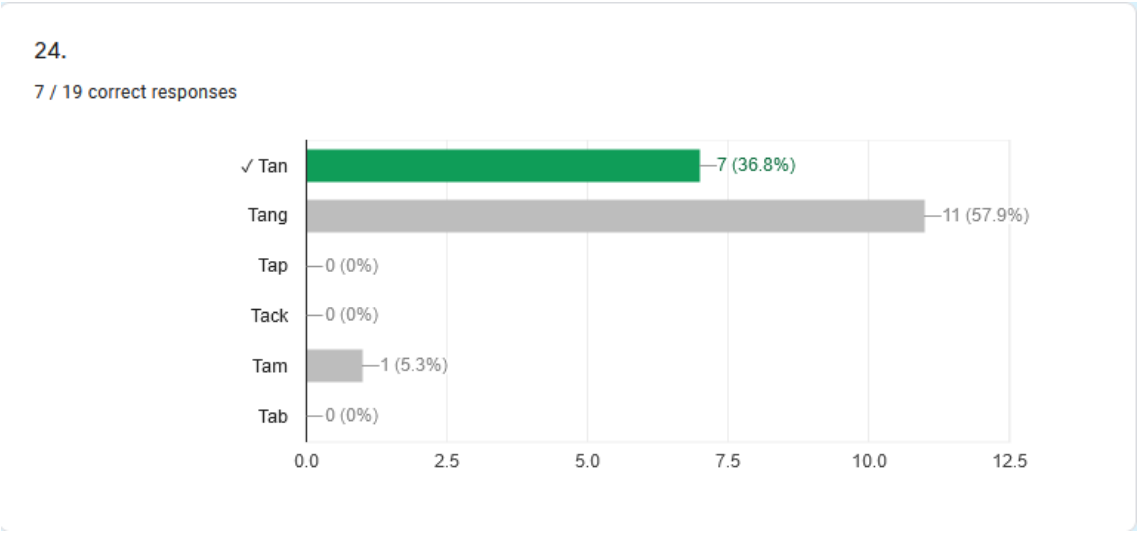

Figure 17: V-MRT Question 24

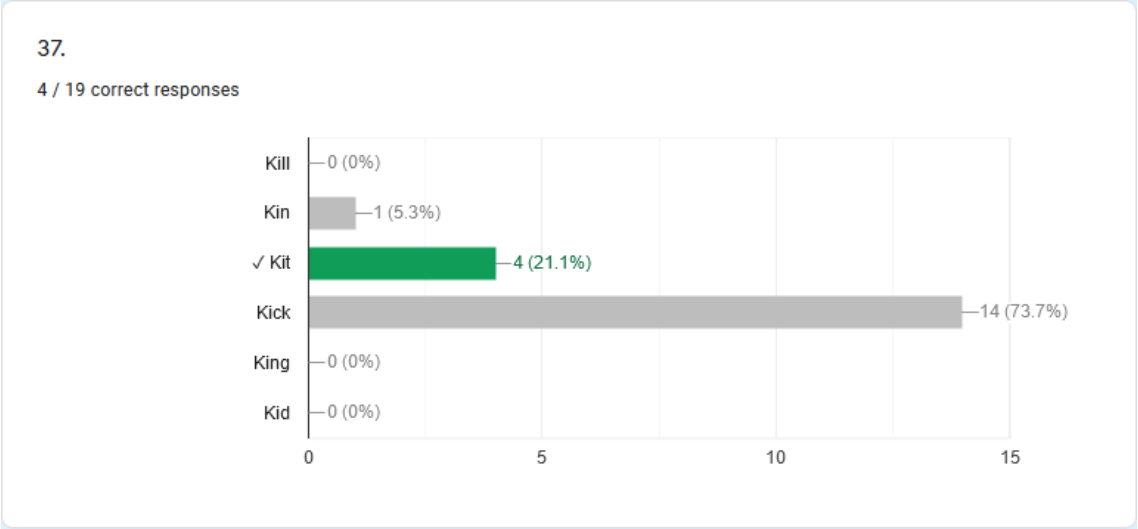

Figure 18: V-MRT Question 37

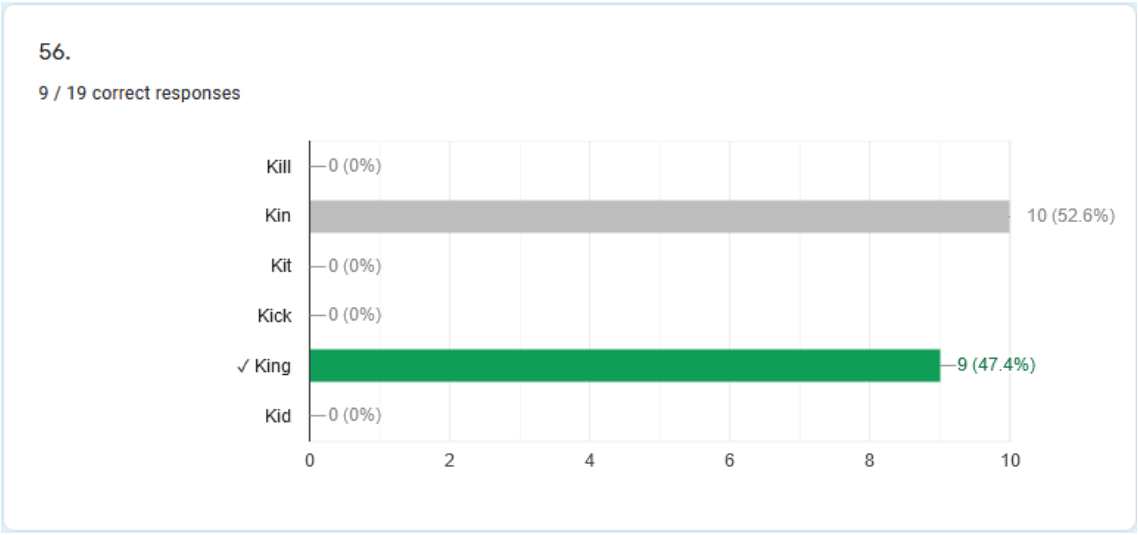

Figure 19: V-MRT Question 56

6.1.3. Sex Bias

Data from the V-MRT shows that the number of males taking the test is disproportional to females taking the test (Figure 20). The MRT solves this by using multiple speakers with a 50/50 split between male and female and recommends the same split for participants (ANSI 2009). In the future, the V-MRT should be adjusted to this ratio which can be easily done with access to more participants (this was attempted but due to lack of participants was unable to be met).

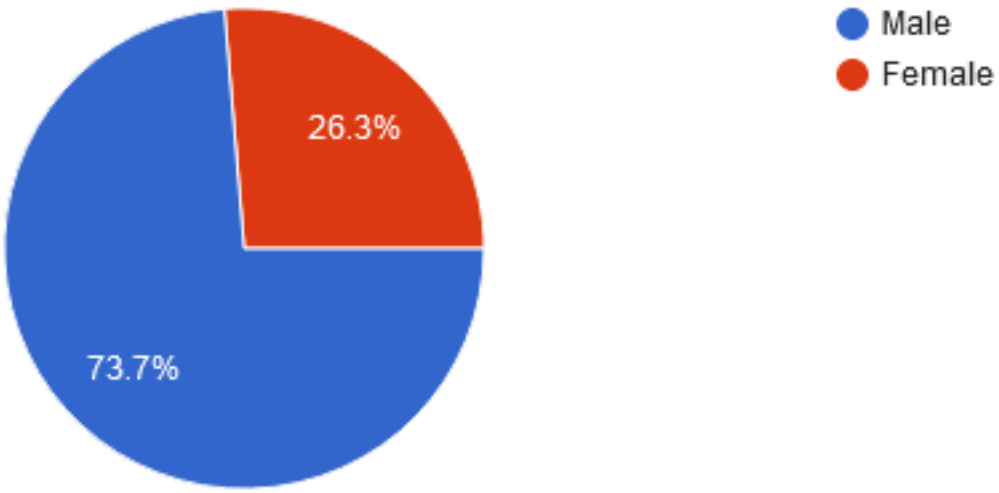

Figure 20: What is your given sex at birth? (V-MRT)

6.1.4. Speech Units

As the V-MRT was limited by time constraints, only 10 speech units where tested (190 answers per mask) compared to the MRT which performs with 50 speech units (950 answers per mask). This causes a significantly larger effect for incorrect answers on the results. Modifying the V-MRT to use 50 speech units (per mask) would likely solve this, however an increased data set would also achieve the same thing (and still allow for time constraints).

6.2. Benefits of the V-MRT

6.2.1. Effect of Visual Cues

One of the advantages of the V-MRT test over alternatives was its ability to factor in (and potentially measure) the effect of visual cues on SI. This allows for certain observations to be made that may have been missed in the STI.

For example, despite the STI results finding the face shield having the worse effect on SI (0.82), both the V-MRT and subjective analysis find it as the second worse even with one additional comment, pointing out 'the face shield was surprising bad.' This is likely due to a large reduction in SPL (potentially as much as 20 dB (Atcherson et al. 2017)). But also suggests a bias inflicted using the face shield.

This could be the effect of visual cues making the mask perform/seem to perform better than the N95 because of the placebo effect when seeing the whole face (especially the lips). The STI results show its worse acoustically, however by not taking into account visual cues it ignores this very human aspect.

However, since starting this report the UK government have deemed face shields not effective enough to stop the corona virus (GOV.UK 2021). This is results in a potentially dangerous bias and should be considered when choosing a face mask.

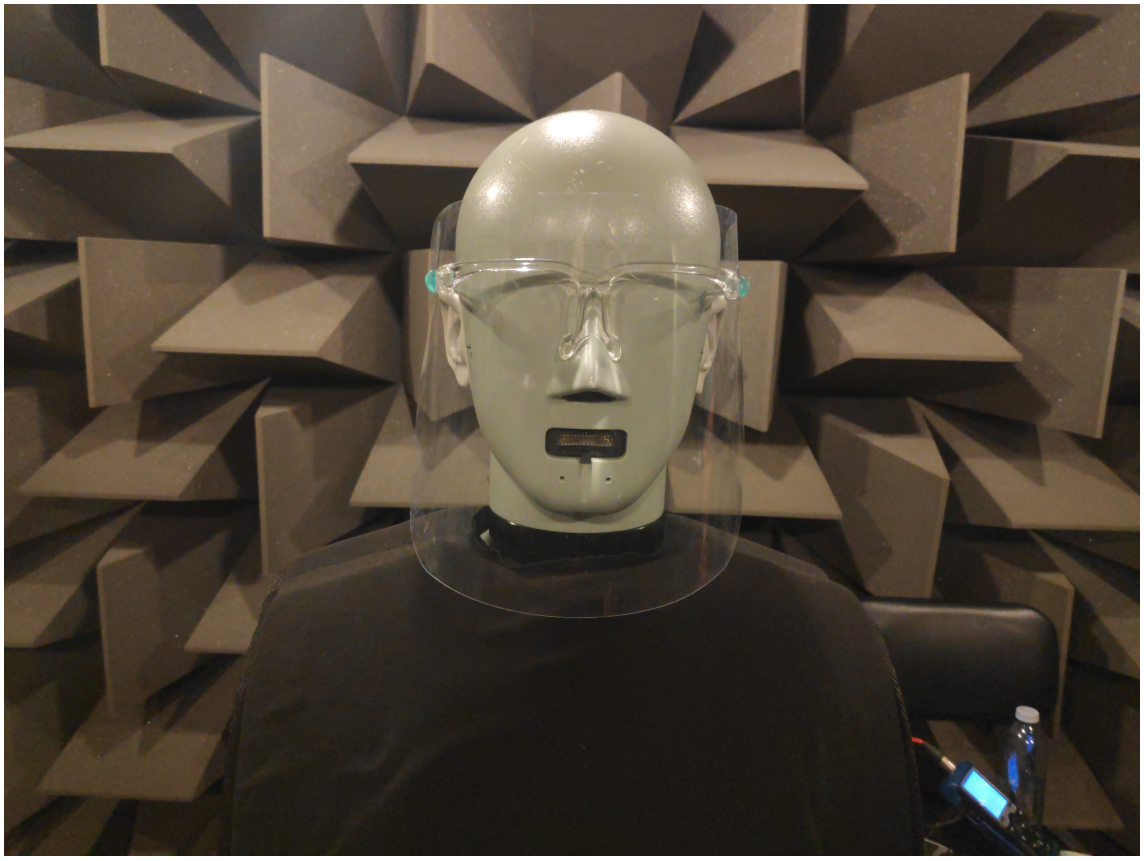

Figure 21: HATS wearing face shield during STI test

6.2.2. Access to participants

During the COVID lockdown the public was strongly advised to go home and stay home as much as possible (GOV.UK 2021). This made getting access to enough participants a challenge. The V-MRT needed to meet the criteria of at least 100 data points to do any meaningful statistical analysis and as such was factored in during the development (Schlaifer and Raiffa 1961).

The V-MRT could be posted online (on social media), allowing for anyone interested (with the correct equipment) to partake in the research. Furthermore, the equipment needed was designed to be easily accessible (needing only a computer and some 'good' headphones) to maximise potential participants. This flexibility makes the V-MRT potentially a very versatile test to perform.

Figure 22: UK government advertisement during COVID-19 lockdown (GOV.UK 2021)

6.2.3. Speed of testing

One the main benefits for performing the V-MRT over an alternative perceptual test (like the MRT or CST) is the time it takes to perform. Taking the test should take an average of 10 minutes (as the video is 6 minutes long and allowing for time to fill out the form) making the test very accessible and have comparable times to predictive methods (like STI or SII). This is also a factor when gathering participants as a quicker test is far easier to fit within others busy lives and is therefore more likely to attract more people.

6.2.4. Background Noise

As previously mentioned, the V-MRT is more realistic than the STI, considering the human element of SI. Background noise is nearly always present in day-to-day life (excluding anechoic environments) and as such, during testing, background noise was played at 50 dB (-15 dB below the speech signal) (Table 2). This will likely have caused a masking effect on the speech and worsened SI. Multiple studies have also shown the negative effects of background noise on concentration and performance doing tasks (J. Bradley, Reich, and Norcross 1999; J. S. Bradley 1986; Ishikawa et al. 2017).

It should also be noted that in real life transient sounds occur. This can lead to temporal masking[10] which would effect SI. However, the results gathered in this study, does not convey this effect due to the use of pink noise as a continuous background noise source (explained further in paragraph 4.3.2).

Although the V-MRT data has been collected from a single location in this paper, it can easily be adapted to work for multiple noise locations (by producing inverse filters for each

[10] A type of auditory masking that occurs when a sound with high energy occurs (usually transient) and masks the sound slightly before and after it (pre and post-masking).

location tested (paragraph 4.3.2)). This opens the possibility of measuring the differences between spaces and could offer further data to produce a more accurate general model for the masks tested.

7. Discussion

The results analysis shows there is a discrepancy between the data collected by the V-MRT and STI test.

Figure 13 shows in the V-MRT test all face masks performing below the 87% standard (equivalent 0.6 STI or the minimal for 'good' speech intelligibility from IEC 60268-16:2020) and therefore concludes face masks do have a significant effect on speech intelligibility.

However, Table 8 shows that (from the STI test data) all face masks perform with an 'excellent' SI and therefore there is not a significant effect on speech intelligibility.

As the STI methodology is an up to date, standardised method (IEC 60268-16:2020). It is more likely that the STI results are more accurate than the V-MRT results.

Furthermore, presuming the issues laid out in section 6 are valid and the V-MRT video itself is to blame, the subjective tests results can also be considered as invalid.

By only asking for the ranking of masks (1 - 5), the subjective test failed provide enough data to make a meaningful comparison between the other tests results. Also, the subjective test also did not ask for feedback on the control which may have revealed important data on the V-MRT tests validity.

Nevertheless, there is still a strong/medium positive correlation between the V-MRT and STI test results, which both also correlated closely with the results of the subjective test results (the wide variation of data from the V-MRT could have caused this data to not align perfectly).

7.1. Participant Error

As result of the COVID-19 pandemic, participants were not able to be supervised during testing and despite being given instructions and practising the test may have used the incorrect methodology, incorrect listening levels or simply inserted random results.

If the V-MRT was performed with speakers instead of headphones or at the incorrect volume, it could have been potentially masked by external noise (from faulty equipment or room noise). This would result in a worse performance overall and SI rating. This would have been easily solved with more participants.

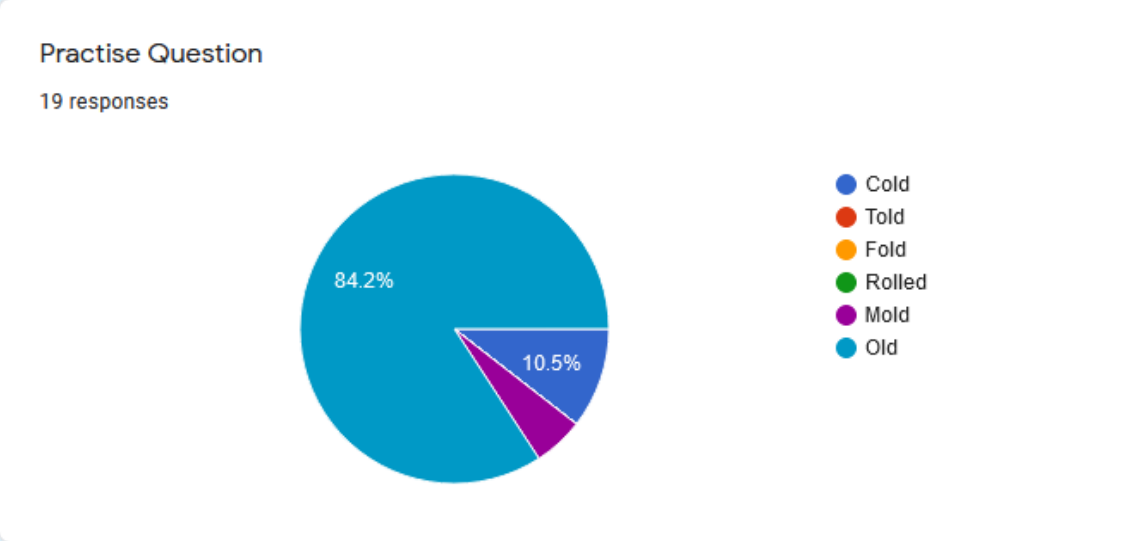

Figure 23: V-MRT Practise Question

8. Conclusion

This study presented measurements and analysis on the effect of commonly on speech intelligibility. It did this by using three testing methods:

The STI method which collected SI data based on the acoustic and frequency effects of each mask performed with the STIPA test signal in an anechoic environment (outlined in IEC 60268-16:2020).

The V-MRT, based on a modification on the standardised MRT (from ASA 3.2-2009) that considered the effect of visual cues and used participants ability to identify and choose the correct word from a group of similar sounding words, whilst background noise was played.

A subjective ranking of each of the masks (1 - 5) based on their perceived effect on SI.

This research took place during the SARS-CoV-2 UK lockdown restrictions which made access to necessary equipment and participants difficult. This was dealt in part by producing modifications on existing measurement standards (V-MRT) that could perform in sub-optimal conditions (using the participants headphones instead of standardised ones (ITU-R BS.708-R), not supervising testing, using Google Forms, recording the V-MRT in an untreated room).

With the current data collected in this study the results are inconclusive if face masks have a significant effect on speech intelligibility. The study's findings could be potentially be effected by the number of participants, participant selection and results in insufficient data to properly generalise these findings to the general population.

Whilst it is likely that the STI results have a higher degree of accuracy than the other methods, it is still not possible to determine if the effects of visual cues produced such

different results. Further research is required to determine if this is the case.

Nevertheless, there is still a strong/medium positive correlation between the V-MRT and STI test results (also corresponding the results of the subjective test results) suggesting that (although off) there is still an order to the face masks effects (medical and snood performing best, face shield and N95 performing worse with the homemade mask performing in the middle).

8.1. Recommendations and Future Work

As this study was intended to gather data useful during the COVID-19 pandemic, compromises where made to meet legal and safety requirements. However, a replication of this work without pandemic conditions should be performed to achieve and determine the answer to the question laid out in this study. It is recommended to perform either an updated and corrected version of the V-MRT (to observe the effect of visual cues) or the MRT to provide accurate perceptual data. The STI results from this study can be used to judge accuracy of these results however, it would be beneficial to perform the STI test again.

Whilst the V-MRT (in its current form) should be used with caution for any serious scientific work (with considerations of its constraints). With some modifications, it has the potential to offer a fast perceptual speech intelligibility test that could be used in situations where access to participants is scarce, making it ideal for use in pandemic like situations (as this study was performed during SARS-CoV-2 this can be validated) or to perform tests with others across internet.

Corrections could include:

- Using a larger number of speech units per mask tested (> 10).
- Using a larger data set (> 20 participants).
- Ensuring the V-MRT speaker is NOT the author of the study being conducted.
- Supervision of the tests.

With events like the SARS-CoV-2 outbreak causing the increased use of masks in general were communicating effectively is critical, future medical mask standards should be required to meet certain speech intelligibility criteria to ensure wearing does not have a significant effect on speech intelligibility (> 0.6 STI, $\geq$ 87% Intelligibility of Words, minimal 'good' standard from IEC 60268-16:2020).

8.2. Project Management

As can be seen in Figure 24, the project varied significantly throughout development. The project management time plan was split into groups: writing, perceptual testing and predictive testing.

The writing group (blue) consisted of writing the dissertation and was given the entire length of the project to complete it. This allowed for maximum flexibility and time for completion setting a goal of submission by 05/05/2021 to allow for external errors (with the final submission data set for 07/05/2021).

The predictive testing group (red) consisted of anything regarding the V-MRT (or MRT) including gathering participants, noise source measurements and the V-MRT testing procedure.

Finally, the predictive testing group (green) was formed of STI (with dynamic range and frequency testing) related factors. Including gathering equipment and performing the test. The STI testing was time sensitive (due to COVID lockdown restrictions) and was completed in a single day.

The original plan was to perform four tests: The STI test, frequency analysis of the mask's effects, the mask's effects on dynamic range and the MRT. However, due to the SARS-CoV-2 lockdowns, bookings and access to facilities to perform these tests where lost (the STI test was performed during a brief relax of the rules). Further, the standard for the MRT (ASA 3.2-2009 was unavailable from the University library and so had to be sent in from another University. This led to a large period of time where (despite having gathered all necessary participants for a minimal test) the MRT (and subsequent V-MRT) could not be performed. In case ASA 3.2-2009 did not arrive on time, multiple papers where the MRT was used in testing were collected, and a model of how to perform an altered version of the MRT was created. Fortunately, the document was delivered on time, which meant the V-MRT could be adapted from the official standard.

Six total supervisor meeting were made with minutes taken (See Appendix C). These consisted of 20 - 30 minute long conversations over video call and involved a question-and-answer format. However, most conversations with the supervisor were made as quick direct messages sent over an instant messaging platform.

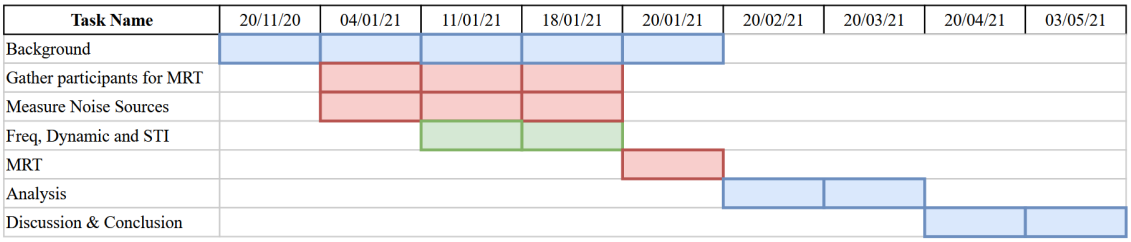

Figure 24: Original Gantt Chart (blue = writing, red = perceptual testing, green = predictive testing)

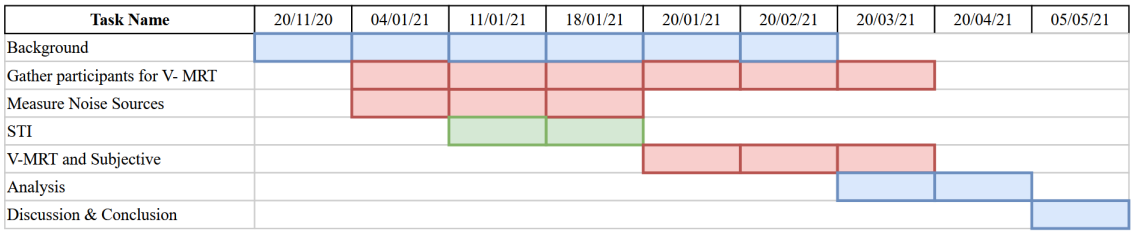

Figure 25: Updated Gantt Chart

8.3. Risk Management

Table 12 outlines the potential risks to the project and how they were mitigated. This study meets all research ethics requirements as outlined by Solent University and has no actionable ethics impact. Existing methods and testing where modified to meet all standards for data collection during COVID-19 (Solent University 2021).

Table 12: Risk Analysis and mitigation steps before and after COVID

Risk Factor	Planned Steps before COVID	Mitigation Steps (During COVID)
Access to equipment and facilities	Originally booked but cancelled due to COVID lockdowns.	Equipment measurements was bought by the researcher. For testing, participants used previously owned personal equipment.
Access to participants	Face to face contact or email to book availability.	Used email and social media to hold video calls and send data. Contingency to upload tests to social media and gather from anonymous third parties.
Access to technical standards	Access via Library resources.	Access to standards via library resources but contingency of combining available research papers to build models for testing. In this case, more time would be given to produce a suitable model.
Participants personal data	No identifiable information given. Participant can ask for removal of their test results.	Same as before COVID. Potential to perform test over alternative forms system to prevent Google LLC privacy concerns.

Risk Factor	Planned Steps before COVID	Mitigation Steps (During COVID)
Loss of data	Data backed up to cloud provider.	Same as before COVID.

9. Bibliography

Alexander Goldin, Barbara Weinstein, and Nimrod Shiman. 2020. "How Do Medical Masks Degrade Speech Reception? - Hearing Review." April 1, 2020. https://www.hearingreview.com/hearing-loss/health-wellness/how-do-medical-masks-degrade-speech-reception.

ANSI. 2009. "ANSI/ASA S3.2-2009 - Method for Measuring the Intellibility of Speech over Communication Systems." Acoustical Society of America.

Assmann, Peter, and Quentin Summerfield. 2004. "The Perception of Speech Under Adverse Conditions." In *Speech Processing in the Auditory System*, 231–308. Springer.

Atcherson, Samuel R., Lisa Lucks Mendel, Wesley J. Baltimore, Chhayakanta Patro, Sungmin Lee, Monique Pousson, and M. Joshua Spann. 2017. "The Effect of Conventional and Transparent Surgical Masks on Speech Understanding in Individuals with and Without Hearing Loss." *Journal of the American Academy of Audiology* 28 (01): 058–67. https://doi.org/10.3766/jaaa.15151.

Banbury, Simon P., and Dianne C. Berry. 2005. "Office Noise and Employee Concentration: Identifying Causes of Disruption and Potential Improvements." *Ergonomics* 48 (1): 25–37.

Barnett, P. W., and R. D. Knight. 1996. "The Common Intelligibility Scale." *PROCEEDINGS-INSTITUTE OF ACOUSTICS* 17: 201–6.

Bradley, John S. 1986. "Speech Intelligibility Studies in Classrooms." *The Journal of the Acoustical Society of America* 80 (3): 846–54.

Bradley, John S., R. Reich, and S. G. Norcross. 1999. "A Just Noticeable Difference in C50 for Speech." *Applied Acoustics* 58 (2): 99–108.

Bradley, JS, R Da Reich, and SG Norcross. 1999. "On the Combined Effects of Signal-to-Noise Ratio and Room Acoustics on Speech Intelligibility." *The Journal of the Acoustical Society of America* 106 (4): 1820–28.

British Standards Institution. 2009. "BS EN 149:2001 + A1:2009 - Respiratory Protective Devices — Filtering Half Masks to Protect Against Particles — Requirements, Testing, Marking." BSI.

British Standards Institution. 2019. "BS EN 14683:2019 - Medical face masks - Requirements and test methods." BSI. https://www.lisungroup.com/wp-content/uploads/2020/05/BS-EN14683-2019-Standard-Free-Download.pdf.

BSOL, BSOL. 2009. "BS EN 60268-5:2003+A1:2009 - Sound System Equipment. Loudspeakers."

Commons, Wikimedia. 2017. *File:Lindos1.svg — Wikimedia Commons, the Free Media Repository*. https://commons.wikimedia.org/w/index.php?title=File:Lindos1.svg&oldid=260111271.

Cox, Robyn M, Genevieve C Alexander, and Christine Gilmore. 1987. "Development of the Connected Speech Test (CST)." *Ear and Hearing* 8 (5): 119S–126S.

"End-to-End Encrypted Cloud Storage for Businesses | Tresorit." n.d. Accessed March 9, 2021. https://tresorit.com/.

Fletcher, Harvey, and JC Steinberg. 1929. "Articulation Testing Methods." *The Bell System Technical Journal* 8 (4): 806–54.

Fox, Jamie, and Bill Whitlock. 2010. "Ground Loops: The Rest of the Story." In *Audio Engineering Society Convention 129*. Audio Engineering Society.

French, Norman R, and John C Steinberg. 1947. "Factors Governing the Intelligibility of Speech Sounds." *The Journal of the Acoustical Society of America* 19 (1): 90–119.

GOV.UK. 2021. "Face Coverings: When to Wear One, Exemptions, and How to Make Your Own." GOV.UK. March 30, 2021. https://www.gov.uk/government/publications/face-coverings-when-to-wear-one-and-how-to-make-your-own/face-coverings-when-to-wear-one-and-how-to-make-your-own.

GOV.UK, GOV.UK. 2021. "New TV Advert Urges Public to Stay at Home to Protect the NHS and Save Lives." GOV.UK. January 10, 2021. https://www.gov.uk/government/news/new-tv-advert-urges-public-to-stay-at-home-to-protect-the-nhs-and-save-lives.

Hodge, B., and J. F. Thompson. 1990. "Noise Pollution in the Operating Theatre." *The Lancet* 335 (8694): 891–94. https://doi.org/https://doi.org/10.1016/0140-6736(90)90486-O.

Hornsby, Benjamin W Y. 2004. "The Speech Intelligibility Index: What Is It and What's It Good For?" 57 (10): 6.

House, Arthur S., Carl E. Williams, Michael HL Hecker, and Karl D. Kryter. 1965. "Articulation-Testing Methods: Consonantal Differentiation with a Closed-Response Set." *The Journal of the Acoustical Society of America* 37 (1): 158–66.

Howard, David M., and Jamie Angus. 2017. *Acoustics and Psychoacoustics*. Taylor & Francis.

IEC, IEC. 2020. "IEC 60268-16:2020 - International Standard: Sound System Equipment-Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index."

Ishikawa, Keiko, Suzanne Boyce, Lisa Kelchner, Maria Golla Powell, Heidi Schieve, Alessandro de Alarcon, and Sid Khosla. 2017. "The Effect of Background Noise on Intelligibility of Dysphonic Speech." *Journal of Speech, Language, and Hearing Research* 60 (7): 1919–29.

Lane, Harlan, and Bernard Tranel. 1971. "The Lombard Sign and the Role of Hearing in Speech." *Journal of Speech and Hearing Research* 14 (4): 677–709. https://doi.org/10.1044/jshr.1404.677.

Letowski, Tomasz R., and Angelique A. Scharine. 2017. "Correlational Analysis of Speech Intelligibility Tests and Metrics for Speech Transmission." US Army Research

Laboratory Aberdeen Proving Ground United States.

McGurk, Harry, and John MacDonald. 1976. "Hearing Lips and Seeing Voices." *Nature* 264 (5588): 746–48.

Mendel, Lisa Lucks, Julie A. Gardino, and Samuel R. Atcherson. 2008. "Speech Understanding Using Surgical Masks: A Problem in Health Care?" *Journal of the American Academy of Audiology* 19 (9): 686–95.

Moore, Brian CJ. 2012. *An Introduction to the Psychology of Hearing*. Brill.

Nadol Jr, Joseph B. 1993. "Hearing Loss." *New England Journal of Medicine* 329 (15): 1092–102.

Neely, Keith K. 1956. "Effect of Visual Factors on the Intelligibility of Speech." *The Journal of the Acoustical Society of America* 28 (6): 1275–77.

Nye, P. W., and J. H. Gaitenby. 1973. "Consonant Intelligibility in Synthetic Speech and in a Natural Speech Control (Modified Rhyme Test Results)." *Haskins Laboratories Status Report on Speech Research, SR* 33: 77–91.

Ong, Sandy. 2020. "How Face Masks Affect Our Communication." BBC Future. June 9, 2020. https://www.bbc.com/future/article/20200609-how-face-masks-affect-our-communication.

Palmiero, Andrew J., Daniel Symons, Judge W. Morgan, and Ronald E. Shaffer. 2016. "Speech Intelligibility Assessment of Protective Facemasks and Air-Purifying Respirators." *Journal of Occupational and Environmental Hygiene* 13 (12): 960–68. https://doi.org/10.1080/15459624.2016.1200723.

Peterson, Gordon E, and Harold L Barney. 1952. "Control Methods Used in a Study of the Vowels," 10.

Pollard, Kimberly A, and Lamar Garrett. 2017. "Speech Intelligibility of Aircrew Mask Communication Configurations in High-Noise Environments," 34.

RNID, RNID. 2018. "Facts and Figures." Action on Hearing Loss. 2018. https://actiononhearingloss.org.uk/about-us/research-and-policy/facts-and-figures/.

Robinson, Derek W, and R So Dadson. 1956. "A Re-Determination of the Equal-Loudness Relations for Pure Tones." *British Journal of Applied Physics* 7 (5): 166.

Schlaifer, Robert, and Howard Raiffa. 1961. *Applied Statistical Decision Theory*.

Solent University. 2021. "Research Integrity." 2021. https://www.solent.ac.uk/research-innovation-enterprise/researcher-support/research-integrity.

Soli, Sigfrid D., and Jean A. Sullivan. 1997. "Factors Affecting Children's Speech Communication in Classrooms." PhD Thesis, Acoustical Society of America.

"Subject Bias in Psychology: Definition & Examples." 2017. Study.com. December 1, 2017. https://study.com/academy/lesson/subject-bias-in-psychology-definition-examples.html.

Sumby, William H., and Irwin Pollack. 1954. “Visual Contribution to Speech Intelligibility in Noise.” *The Journal of the Acoustical Society of America* 26 (2): 212–15.

Szendro, P., Gy Vincze, and A. Szasz. 2001. “Pink-Noise Behaviour of Biosystems.” *European Biophysics Journal* 30 (3): 227–31.

“Text Analytics | Microsoft Azure.” n.d. Accessed April 12, 2021. https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/.

Turner, Christopher W, and Karolyn J Cummings. 1999. “Speech Audibility for Listeners with High-Frequency Hearing Loss.”

Valerie Fridland. 2020. “Why Do Masks Make It so Hard to Understand Each Other?” University of Nevada, Reno. October 21, 2020. https://www.unr.edu/nevada-today/news/2020/atp-masks-hard-understand.

VanBebber, Mary Lillian. 1954. “A STUDY OF FACTORS INFLUENCING IMPROVEMENT IN SPEECH READING ABILITY,” 61.

WHO. 2020. “Advice for the Public on COVID-19 – World Health Organization.” 2020. https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public.

WHO, WHO. n.d. “Comnoise-4.” WHO. https://www.who.int/docstore/peh/noise/Comnoise-4.pdf.

Williams, Carl E., and Michael HL Hecker. 1967. “Selecting an Intelligibility Test for Communication System Evaluation.” *The Journal of the Acoustical Society of America* 42 (5): 1198–98.

A. MRT Test Sheet w. Answers

Control

1. Tent
2. Told
3. Pan
4. Lake
5. Kit
6. Dust
7. Teach
8. Dim
9. Red
10. Pin

Mask 01

12. Duck
13. Sung
14. Seen
15. Got
16. Test
17. Pip
18. Back
19. Pay
20. Dig
21. Page

Mask 02

22. Cave
23. Mop
24. Boil
25. Tan
26. Fig
27. Fame
28. Peel
29. Bark
30. Heat
31. Cud

Mask 03

32. Paw
33. Men
34. Pun
35. Bean
36. Seat
37. Hip
38. Kit
39. Fang
40. Took
41. Mass

Mask 04

42. Ray
43. Safe
44. Kill
45. Sill
46. Gale
47. Wick
48. Peak
49. Buff
50. Sass
51. Bun

Mask 05

52. Peace
53. Sad
54. Run
55. Dip
56. Meat
57. King
58. Rang
59. Man
60. Race
61. Cane

B. Git Log

commit e502fa896c29c9d7c9a0588f358c0d0d74e7cebc (HEAD -> working) Date: Sat May 1 11:53:18 2021 +0100

```
First Draft submitted

- Completed results analysis (chose T-tests for V-MRT)
- Completed suggestions from Sky
- Exported needed image files to folder
```

commit 76898c48ca9ad02c14fc8fd398a9c5d03dade16e Date: Mon Mar 15 17:03:09 2021 +0000

```
Lockdown update

- Ordered files properly
- Developed V-MRT video and ran initial testing
- Completed both Methods for testing SI and V-MRT methodology
```

commit b6b16465e461f89268ed559d14a849d36897ecde Date: Fri Nov 20 10:54:32 2020 +0000

```
Major update

- Conversion from old methods to new
- Gathering resources for MRT testing
    - Ordered ASA 3.2-2009
- Gathered 12 participants (require more)
```

commit 80dc3d13fa5ef933025f8263d13f8767260a9eb4 (origin/working) Date: Wed Oct 14 19:50:24 2020 +0100

```
Added folder structure and added to various docs

- First meeting with supervisor (Direction of project and various BG)
- Continued collecting sources in Zotero (ref sys)
- Setup Gantt Chart using issues and milestones
- Tested pandoc and modified tex file
- Basic Draft layout created with links
```

C. Minutes

C.1. Session 01

C.1.1. Methodologies

- **Reading words**.
  - Differing complexities (monosyllabic, Disyllabic and trisyllabic??
  - random generator to choose words
    * Assign number to words
- **Basic STIPA test**
  - Mask -> Hats (voice module)

C.1.2. Other

- What kind of signal to use a speech analogue
- **program noise signal**
- Filter pink noise
- Audacity
- Export curves (Filter Curve EQ)
- XML files
- Edit in plain text
- Crest factor of speech
- Masking

C.1.3. Frequency effects

- Small change in f might make large compound effect
- Hearing Aids Standards
  - Looking at specific f band
  - only test up to 8kHz
- Crest factor

C.1.4. Dynamic range

Measure background noise?

- Classroom
- Circumstances

- BG noise
  - insertion loss

Biggest risk - Masks make no difference - Adjust direction of project

C.2. TODO

- Look into program noise and crest factor
- Learn how to use the HATS voice module
- Keep reading papers

C.3. Session 02

STI: Feasibility How to use HATS w. mouth piece
Update on progress

C.3.1. Options Analysis

- Go through different ways to achieve same goal
- Not really alternative methods so think broader
  - Look at difference between MRT and ABC-MRT
  - Use HATS over NTi Talkbox (Need to use a mask, anthropomorphic)
  - Different ways to measure STI (Computers, NTi, etc.)

Visuals?

- Mention it but this is my further research (not in the scope for this project)

C.3.2. Literature Review

- **Find gap in project**
- There's a gap in the literature around this subject so my investigation is…
- Gap: Public perspective from public perspective (COVID-19)
- Not necessarily medical masks

C.3.3. POTENTIAL: Comparing SI to stopping viral transmission

- If there is **good data** (official) = Good
  - Even just verbal rating system (Good, Bad, etc.)
  - Good vs Bad vs STIPA, Good vs Bad vs Dynamic Range etc.

- Look at nearer end of project
  - **Get data what is it telling you**

C.3.4. MRT

- Average Noise
  - **Option 1**: Expand on hospital noise references: but also we are gonna have to look into other sources of noise
  - **Option 2**: Go somewhere and measure (Supermarket, streets, hospital (found online))
- Measure average of place - **extract EQ curve from place** and **apply to pink noise** (avoids transients from place (bus goes past), Gets time invariance so is stable)
- **1 minute** - enough time for anything that could happen to happen ;) , more you take the better representation
- Graph each place
- Test with each location
- **⅓ Octave DATA**
- Could then test with all noise sources

C.3.5. COVIDify

- Send MRT to people
  - BG noise with recording and send files to participants
- Issues level matching
- Speakers (EQs)
- **Large sample size would even out**
- "In the future do larger project"
- Can't destroy all sources of error so **acknowledge them** and explain why you can't deal with it

C.3.6. Next Steps

- Get NTi XL2 and take measurements

C.4. Session 03

Testing in Chamber - Friday from 09:00 - 16:00
Preliminary testing

What's SPL for human conversation at 1m? - STIPA standards or research - 60268 objective intelligibility

HATS mouth attachment tested by 31-11-2020

Potential testing over Christmas period (Empty) - With more supervision

Chamber has 2 person limit.

C.5. Session 04

C.5.1. Questions

Modified MRT

- **Can I just use headphones?**
  - Apply frequencies curve of masks to MRT audio and noise with volume from in Spoons
  - Loose verb but SNR is most important

C.5.2. YES! You can modify for headphones

- Just justify

Have people lined up for testing only 10 but 5 male, 5 female - Don't turn data down - Always get it

Made MRT test sheets

What do I need participants to sign to do testing

Swapped from doing stuff online with Github to just offline

C.6. Minutes 05

C.6.1. Issues with frequency data

- User error has led to initial measurements being distorted
- Lockdown has lead to being unable to access the anechoic chamber for new measurements
- Solution
  - Email: Hickling about access and likely date

C.6.2. Alternatives

- Look further into The effect of visuals on speech intelligibility
- By using measurement mic (owned) to be taken as flat frequency response

- Still has the audible effect of wearing a mask + the visual effect
  - Better than initial idea
- Ask participant to wear headphones already has

C.7. Minutes 06

Results analysis

- Work on the narrative of the piece

D. Google Forms CSV conversion Script

‘import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
df = pd.read_csv(‘MRT.csv,’index_col=None,na_values=[‘NA’])
new = df.drop([‘Timestamp’],axis=1)
new = new.dropna(how=‘all,’axis=1)
new = new.loc[:,::2]
dfs = new.replace(to_replace=[“0.00 / 1”],value=0)
dfs = dfs.replace(to_replace=[“1.00 / 1”],value=1)
dfs = dfs.drop([‘Do you consent to being part of this test? [Score],’‘What is your given sex at birth [Score],’‘If you have a known hearing disability, please specify. Otherwise, leave this question blank. [Score],’‘Please enter the model and brand of your headphones used for testing. If you have any specialist audio gear in your system, please also specify. [Score]’],axis=1)
dfs.to_csv(r’cleanedata.csv’)‘

If performing the V-MRT using Google forms, it should be noted that the csv file it generates the ‘score’ results as strings which need to be converted to integers/floats to be worked with. During testing a script was derived to automate this process: