UNIVERSITY OF SOUTHAMPTON

Automated Prediction of Speech Intelligibility in Noise for Hearing Aids

by

Robyn Hunt

Faculty of Engineering and Physical Sciences

March 2022

Declaration of Authorship

I, Robyn Hunt, declare that this thesis titled, 'Automated Prediction of Speech Intelligibility in Noise for Hearing Aids' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date: 27/03/2022

UNIVERSITY OF SOUTHAMPTON

Abstract

Human Sciences Group
Faculty of Engineering and Physical Sciences

Doctor of Philosophy

by Robyn Hunt

According to a study by Action on Hearing Loss (2017a), 80% of people with hearing loss have difficulty understanding speech in the presence of background noise. Currently, we rely on behavioural speech-in-noise tests to determine and compare the efficacy of different hearing aids for improvement of speech intelligibility. If a sufficiently reliable automated prediction of intelligibility could be made available, the cost and complexity of testing new hearing aids could be reduced, and may allow bodies such as the NHS to compare device performance more efficiently.

This thesis aims to evaluate several existing speech intelligibility prediction metrics by comparing their outputs against results from behavioural speech-in-noise tests. Behavioural speech-in-noise test scores from 21 normal hearing participants and speech intelligibility predictions from automated metrics were obtained for IEEE sentences (Institute of Electrical and Electronics Engineers, 1969) in stationary, speech-shaped background noise at signal-to-noise ratios from -8 to +3 dB, as processed by three different hearing aid models (currently prescribed by the NHS) with and without noise reduction settings enabled in addition to a control condition with no amplification and a low-cost amplifying device.

All automated prediction metrics tested showed a broad increase in intelligibility with increasing signal-to-noise ratio. However, only one of the three automated metrics tested, the Hearing Aid Speech Perception Index (HASPI) (Kates and Arehart, 2014), was able to detect statistically significant differences between conditions which mirrored those seen in behavioural speech-in-noise test results. HASPI did, however, struggle to accurately predict the behavioural speech-in-noise scores for some specific hearing aid conditions and signal-to-noise ratios. Further investigations attempted to identify the main causes of HASPIs shortcomings including analysis of feature importance and implementation of a range of mapping and machine learning methods, the effects of differing stimulus types and the robustness of HASPIs component features in combination with features from alternative existing automated metrics.

This thesis concludes that currently available automated metrics for speech intelligibility prediction are not fully capable of detecting differences between devices and settings, particularly between efficient noise-reduction programs and a low-cost amplifier. Whilst these metrics form an excellent basis for speech intelligibility prediction, further work is needed to develop existing metrics for use in comparing and tuning hearing aids and settings.

Acknowledgements

I would like to thank my project supervisors, Dr Steven Bell and Professor David Simpson, for their limitless support and guidance over the duration of the project.

Furthermore, I extend my gratitude to Dr Matthew Blyth, Dr Mark Fletcher, Samuel Perry and Dr Alastair Manders for their input and contributions.

It is also important to acknowledge Professors James Kates, Jesper Jensen and Cees Taal for providing the MATLAB code for their speech intelligibility algorithms, on which my research is based.

Finally, and most importantly, I would like to thank the colleagues, friends and family who gave up their precious time to participate in my research study, and the representatives from Oticon, Siemens, GN ReSound and Phonak who kindly gifted me their latest products and software, without whom I would be unable to complete this work.

Contents

List of Figures

List of Tables

Abbreviations

AI	**A**rticulation **I**ndex
ARHL	**A**ge-**R**elated **H**earing **L**oss
BAHA	**B**one **A**nchored **H**earing **A**id
BKB	**B**amford-**K**owal-**B**ench (sentence list)
BKB-SIN	**B**amford-**K**owal-**B**ench **S**peech **I**n **N**oise (test)
BTE	**B**ehind **T**he **E**ar (Hearing Aid)
CSII	**C**oherence **S**peech **I**ntelligibility **I**ndex
CUNY	**C**ity **U**niversity of **N**ew **Y**ork (Sentence Test)
ESTOI	**E**xtended **S**hort-**T**ime **O**bjective **I**ntelligibility
FAAF	**F**our **A**lternative **A**uditory **F**eature (test)
HA	**H**earing **A**id
HASPI	**H**earing **A**id **S**peech **P**erception **I**ndex
HI	**H**earing **I**mpaired/Impairment
HINT	**H**earing **I**n **N**oise **T**est
IEEE	**I**nstitute of **E**lectrical and **E**lectronics **E**ngineers
IBM	**I**deal **B**inary **M**asking
IHC	**I**nner **H**air **C**ell
IHRSL	**I**nstitute of **H**earing **R**eserch **S**entence **L**ist
ITE	**I**n **T**he **E**ar (Hearing Aid)
JND	**J**ust **N**oticeable **D**ifference
KEMAR	**K**nowles' **E**lectronics **M**anikin for **A**coustic **R**esearch
NCM	**N**ormalised **C**ovariance **M**etric
NH	**N**ormal **H**earing
NHS	**N**ational **H**ealth **S**ervice
OHC	**O**uter **H**air **C**ell

ORCA	**O**ffice of **R**esearch in **C**linical **A**mplification
PTA	**P**ure **T**one **A**udiometry
QuickSIN	**Quick S**peech **I**n **N**oise (Test)
RITE	**R**eceiver **I**n **T**he **E**ar (Hearing Aid)
RMS	**R**oot **M**ean **S**quare
SDR	**S**ignal-to-**D**istortion **R**atio
SI	**S**peech **I**ntelligibility
SII	**S**peech **I**ntelligibility **I**ndex
SIIB	**S**peech **I**ntelligibility **I**n **B**its
SIN	**S**peech **I**n **N**oise (test, subjective)
SNHL	**S**ensori**N**eural **H**earing **L**oss
SNR	**S**ignal-to-**N**oise **R**atio
SPL	**S**ound **P**ressure **L**evel
SRT	**S**peech **R**eception **T**hreshold
SSN	**S**peech-**S**haped **N**oise
STI	**S**peech **T**ransmission **I**ndex
STOI	**S**hort-**T**ime **O**bjective **I**ntelligibility
SRMR	**S**tandardized **R**oot **M**ean Square **R**esidual
SVMR	**S**upport **V**ector **M**achine **R**egression
TFS	**T**emporal **F**ine **S**tructure
VAD	**V**oice **A**ctivity **D**etector
WIN	**W**ords **I**n **N**oise (Test)

Dedicated to my beautiful sister, Rhiannon, and my beloved Grandad, who have been my inspiration and my motivation and are the reason I conducted this study.

Chapter 1

Introduction

1.1 General Introduction

According to Action on Hearing Loss (2017a), four in every five people with hearing loss have difficulty understanding speech in the presence of background noise. Noise cancellation methods are present in a number of hearing aids, but these are insufficient to meet users requirements for speech understanding; users rate speech listening in noise as the most valued yet least satisfactory attribute of hearing aids (HAs) (Bridges et al., 2012; Meister et al., 2002). Behavioural speech-in-noise tests, in which the listener is asked to repeat back or otherwise indicate understanding of speech in a noisy signal, are regarded as the 'gold standard' of assessing the performance of hearing aids in background noise; however, automated models for prediction of speech-in-noise performance are becoming increasingly desirable, due to the lower cost and variability and increased timeliness associated with automated methods compared to behavioural trials. Several methods for predicting speech-in-noise performance have been discussed in recent literature, most notably work by Kates and Arehart (2014), but none to date have been evaluated using real hearing aid outputs.

This thesis will therefore aim to address the following key objective:

- To evaluate the accuracy of currently available automated speech intelligibility prediction metrics for predicting the outcome of behavioural speech-in-noise tests in response to noisy, hearing-aid processed speech.

Following this introductory section, the reader is presented with a literature summary of the relevant background to the issues addressed in this work. This begins an overview of hearing aid use in the United Kingdom and basic/common functions included in these devices, in addition to a short discussion of problems users of these devices face, focusing in particular on speech intelligibility. Next, current methods of assessing speech intelligibility are detailed, initially focusing on behavioural tests before progressing onto a description of automated tests, with a spotlight on the methods to be implemented in the research.

The third chapter covers the methodology and experimental procedure which underpins the key research questions to be answered, including the main aims and technical details of the set-up.

The fourth and fifth chapters present and discuss the two main components of the study: firstly, the behavioural speech-in-noise tests which provide a gold standard for the comparison in the second part of the study, which focuses on the predictions of speech intelligibility estimated by automated metrics.

In Chapter 6, a summary of the key shortcomings and limitations to be addressed with automated metrics, supported by evidence from the preceding chapter, is revisited. A discussion of further investigative techniques and the additional results and insights given by these is then provided, including modification of automated metrics using machine learning and remapping, analysis of variations due to a range of input changes (such as noise type and speech type) and an investigation into robustness of the metrics to small signal changes.

The final chapter summarises the key outcomes of the research project as a whole, highlights key outcomes of the project and gives recommendations for further work in the field.

1.2 Original Contributions and Publications

1.2.1 Summary of Original Contributions to Scientific Knowledge

The key contributions of this thesis include:

1. An assessment of speech-in-noise performance of currently available NHS hearing aids with and without noise reduction algorithms using double-blind behavioural speech-in-noise tests.

2. A robust appraisal of currently available automated speech intelligibility metrics for assessment of hearing-aid processed, noisy speech using outputs directly

recorded from real hearing aids, using the aforementioned behavioural speech-in-noise tests as a 'gold standard' for evaluation. The work in this thesis has highlighted and discussed key areas where metrics do not perform as expected using real hearing aid outputs, on which current metrics have not been previously tested.

3. A preliminary analysis of the suitability of current features used in automated speech intelligibility prediction metrics and recommendations for potential improvements, including:

   - Exploration of the stability of different features in response to background noise;
   - Analysis of variations in predictions in response to noise type, sentence corpus and gender of speaker;
   - Exploration of machine learning techniques to improve mapping of objective speech intelligibility correlates to behavioural speech-in-noise scores.

1.2.2 Journal Papers

R. Hunt, S. Bell and D. Simpson. Using HASPI For Automated Comparison of Hearing Aid Speech Intelligibility. In preparation for *Ear and Hearing*.

1.2.3 Conference Abstracts

R. Hunt, S. Bell, and D. Simpson, "Predicting the Impact of Hearing Aid Processing on Speech Intelligibility". Poster Presentation at the Basic Auditory Science conference held by the British Society of Audiology at University College, London, England in September 2019.

R. Hunt, S. Bell, and D. Simpson, "Predicting the Impact of Hearing Aid Processing on Speech Intelligibility". Poster Presentation at the Acoustical Society of America Fall Conference at the Hotel Del Coronado, San Diego, California, U.S. in December 2019.

Chapter 2

Background

2.1 Hearing Aid Use in the United Kingdom

Currently, an estimated 11 million people in the UK (one in every six) are affected by hearing loss; this figure is predicted to rise by 4.6 million by 2035 (Hearing Link, 2018). About twelve thousand people in the UK are fitted with cochlear implants and around 2 million use hearing aids (HAs), but it is estimated that a further 4.7 million could benefit from HA use (Hearing Link, 2018).

Although the National Health Service (NHS) devotes roughly an annual £450 million to addressing hearing loss, the additional indirect cost to the UK government due to associated health, social and economic effects of untreated hearing loss is estimated to be at least £30 billion every year (The Ear Foundation, 2014).

2.1.1 Types of Hearing Assistance

Several types of HA and other types of assistive devices are available for people with hearing loss. Behind-the-Ear (BTE) HAs are most common type of HA fitted by the NHS, and are suitable for the widest range of hearing losses (Action on Hearing Loss, 2018a). Receiver-in-the-Ear (RITE) (similar to BTE HAs with microphones placed at the end of the tube within the ear canal) and In-the-Ear (ITE) HAs are also available through the NHS but are much less common, and so won't be discussed in detail here (Action on Hearing Loss, 2017a). Examples of these types of HAs are shown in Figure 2.1.

Open fits (small, flexible domes which fit in the entrance to the ear canal), similar to the 'mini-BTE' shown in Figure 2.1, are often preferable to closed moulds, made with

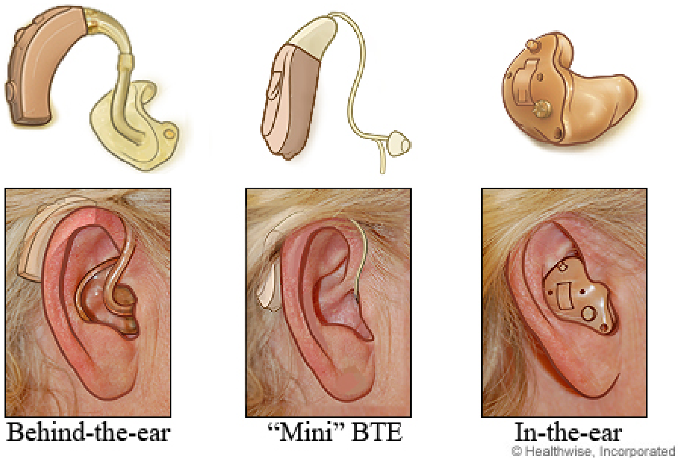


FIGURE 2.1: Illustrations of BTE closed mould, BTE open mould/dome (similar to RITE) and ITE HAs, taken from Healthwise Inc. (2017).

rigid moulded plastic, in terms of user comfort. Total blockage of the outer portion of the ear canal leads to an uncomfortable low frequency boosting effect (the 'occlusion effect'), which is especially prominent for users with a milder hearing loss at low frequencies compared to higher frequencies (Winkler et al., 2016), such as that seen in age-related hearing loss (ARHL) (Action on Hearing Loss, 2017b). Open fits are also almost instant to fit compared to a personalised closed mould of the ear and are considered more aesthetically pleasing (Winkler et al., 2016). However, open moulds also provide an acoustic path for the amplified output of the HA to escape and for unprocessed (direct) sound to enter the ear canal. The amplified sound is then picked up by the HA microphone and amplified again, resulting in an unstable feedback loop, and benefits of noise cancellation by directional microphones are reduced as the direct sound mixes with the processed (and therefore slightly delayed) sound (Valente, 2002) (see Section 2.1.2.2 for more details on directional noise cancellation). Feedback problems are more likely to occur when the gain of the HA is increased (i.e. in those with a higher severity of hearing loss) with open fits compared to closed moulds and as such, closed moulds are often more suitable than open fits for users with moderate to severe hearing loss, where higher gain is necessary.

HAs are not suitable for all types of hearing loss. Individuals with severe and profound sensorineural hearing loss (SNHL), caused by damage or deformity in the cochlea, can be fitted with cochlear implants if HAs do not provide sufficient benefit (Action on Hearing Loss, 2016). Individuals with conductive hearing losses, such as irremovable blockages or damage to the middle ear bones, are often fitted with bone anchored hearing aids (BAHAs), which transmit sound in the form of vibrations directly to the inner ear through the skull (Action on Hearing Loss, 2018a). Since these types of hearing assistance are not assessed as part of the current research, they will not be discussed further in this thesis.

2.1.2 Hearing Aid Functions

BTE HAs detect sounds using microphones located behind the outer ear. These sounds are then processed and amplified by the HA before being reproduced using a small loudspeaker at the top of the HA unit. The sound then travels along a thin tube and is delivered into the ear canal through an ear mould or open-fit insert.

2.1.2.1 Amplification

The gain (amplification level) of the HA at several frequencies is initially fitted using a standard prescription formula, which is based on the individual's hearing thresholds determined using Pure Tone Audiometry (PTA). This amplification level is then limited depending on the input level of the sound, such that very loud sounds are not amplified to an uncomfortable or dangerous level. The gain ratio between the incoming and output sound is reduced at high sound levels. This process is termed 'dynamic range compression' and is shown in Figure 2.2.

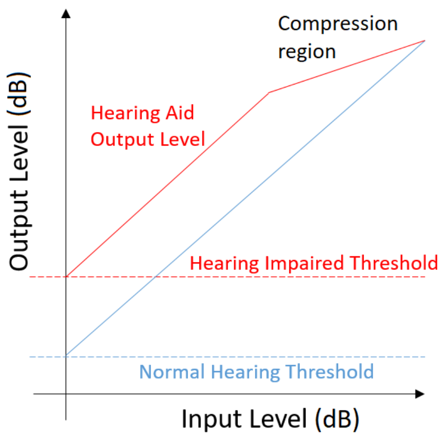


FIGURE 2.2: A simplified representation of dynamic range compression in a HA.

2.1.2.2 Directional Noise Cancellation

Most BTE HAs use a microphone pair or two-ported microphone to attenuate noise arriving from any direction other than the way the listener is facing. The speech is usually assumed to be arriving from in front of the listener. In order to cancel noise arriving from a particular direction, the difference between the signal arriving at the front microphone or port and a delayed version of the signal arriving at the second microphone or port is calculated. The delay can be manipulated to produce the desired directivity of the system (Kates, 2008). A simplified diagram of this type of set-up is shown in Figure 2.3 and an example of directional directivity patterns can be seen in Figure 2.4.

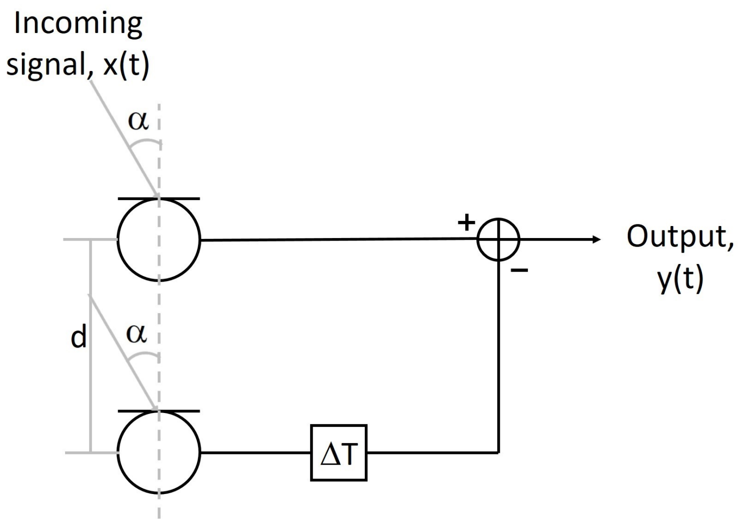


FIGURE 2.3: A schematic diagram of the signal processing involved in a directional microphone set-up within a hearing aid, based on an example from Kates (2008). ΔT, measured in seconds, represents an artificially introduced time delay applied to the signal arriving at the lower microphone. d is the physical distance in meters between the microphones/microphone ports.

The signal arriving at the rear microphone is equal to the signal arriving at the front microphone, $x(t)$, with a delay equal to the difference in the time taken for the sound to travel to each of the two microphones, $\frac{d\cos(\alpha)}{c}$, where $c \approx 343m/s$ is the speed of sound in air. A further delay, ΔT, is then incorporated into the system, the value of which is carefully chosen in order to achieve the desired directivity pattern (i.e. the amount of attenuation applied to the incoming signal, as a function of angle α). The output, $y(t)$, and transfer function, $\frac{Y(f)}{X(f)}$, of the system can be calculated using Equations 2.1 and 2.2 below.

$$y(t) = x(t) - x(t - \frac{d\cos(\alpha)}{c} - \Delta T) \tag{2.1}$$

$$\frac{Y(f)}{X(f)} = 1 - e^{-j2\pi f(\frac{d\cos(\alpha)}{c} + \Delta T)} \tag{2.2}$$

The exact directivity pattern of the set-up is determined by the separation of the microphones, the artificial delay chosen by the designer and the frequency of the arriving sound. For this kind of set-up, directional properties of the HA system can be maintained over a wide frequency range, provided the separation of the two microphones or ports is greater than half a wavelength of the sounds of interest; at frequencies above this point, spatial aliasing occurs (Dillon, 2001). For further reading on spatial aliasing in directional microphone setups the reader is directed to other available literature such as Kates (2008).

An example of the effects of directional microphone set-ups on the signal-to-noise ratio (SNR) can be seen in Figure 2.4 (Aubreville and Petrausch, 2015). Note that the directionality in Figure 2.4 is expressed in terms of the interferer-to-target ratio (ITR), the inverse of the SNR, and as such, the lower the ITR, the higher the SNR. It can be seen that the ratio of noise to the wanted signal is much lower in the rearward direction compared to the forward direction, indicating the effective removal of noise from behind the listener and implying an increased speech intelligibility (SI) compared to an omni-directional setting. The asymmetry seen in about the left/right axis is predominantly due to the head-shadowing effect, which is especially prominent at higher frequencies.

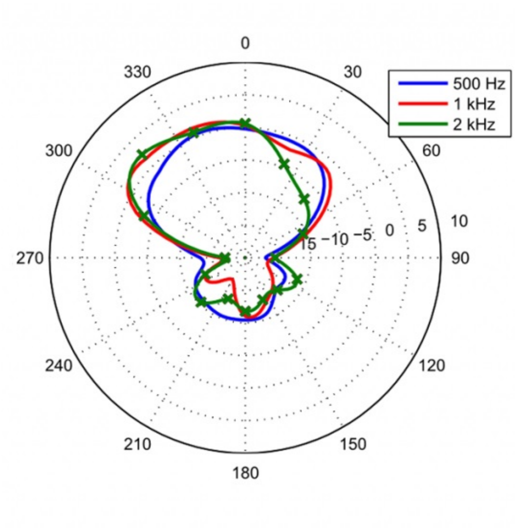


FIGURE 2.4: The directional benefit expressed as the interferer-to-target ratio (the inverse of SNR) for a standard directional setting of a HA, taken from Aubreville and Petrausch (2015). The radius corresponds to interferer-to-target ratio, plotted in dB. The angle of arrival is plotted along the circumference, with zero degrees corresponding to the front of the listener ($\alpha = 0$).

Directional beam-forming to cancel noise from different directions to that of speech is widely considered a more effective way of improving SI than use of single microphone noise reduction algorithms, but is not always possible, as outlined in the following section (Hu and Loizou, 2007; Wouters et al., 2009).

2.1.2.3 Single-Channel Noise Reduction

Directional microphones are not effective for noise cancellation in cases where the speech and noise are collocated (arriving from the same direction). In these situations, it is possible to reduce the noise in the HA output using only one input consisting of a speech signal contaminated by noise. Methods to do this include spectral subtraction and masking algorithms. Spectral subtraction methods work by estimating the relative power of the speech and noise in the signal using speech detection methods, where masking algorithms apply a thresholding technique to either enhance or suppress segments of the signal. Both methods aim to subtract/reduce components of the signal that are dominated by noise, and do so usually in the time-frequency domain (i.e., by subtracting noise from the signal in a time varying manner and from different frequency bands). Neither of these methods are effective for increasing SI unless the noise and the speech are located in separate frequency bands, nor are they appropriate for removal of fast-changing background noise such as competitive talkers, due to difficulties in accurately estimating the noise power contained within the noisy signal (Kates, 2017). They have been reported, however, to improve speech quality and listening comfort in some situations (Bentler, 2006; Dillon, 2001; Mueller et al., 2006). It is likely that listening comfort and speech quality rather than speech intelligibility encourage 49% of HA users to utilise a background noise program involving implementation of single-channel noise reduction algorithms (Action on Hearing Loss, 2017a).

In order to apply spectral subtraction and masking methods, it is necessary to estimate the noise content of the signal. There are several approaches used for separating a noisy speech signal into speech and noise components. One of the most common is to use a Voice Activity Detector (VAD). VAD algorithms are often used to detect whether or not speech is present in a section of the signal. Those parts of the signal where the VAD does not detect speech are used to estimate the noise spectrum (Ramirez et al., 2007). VADs usually detect speech based on some form of feature extraction, for example, typical spectral or temporal fluctuations associated with speech which are likely to be different to the fluctuations of the background noise (Graf et al., 2015). VADs are often effective at high SNRs with a stationary noise source but as the noise approaches a level at which it begins to mask the speech signal or if the background noise is non-stationary, the estimate of the speech and noise components of the signal as predicted by the VAD lose accuracy (Shrawankar and Thakare, 2010).

Spectral subtraction methods involve estimating the magnitude of the noise in each time-frequency cell compared to the magnitude of the combined speech and noise signal. This estimate of the SNR in the signal is then used to adjust the gain of the noise in that time-frequency cell and the weighted noise estimated is subtracted from the noisy signal.

The exact relationship between the estimated SNR of the cell and the applied attenuation depends on the particular spectral subtraction algorithm being used. Some examples of spectral subtraction/multiplication include adaptive Wiener filters, magnitude spectral subtraction and power spectral subtraction (Vaseghi, 2001).

A commonly used algorithm for speech enhancement is the binary mask. A noise estimate for each time-frequency cell is compared to the total signal in order to estimate the SNR, in a similar way to that of spectral subtraction. If the SNR is greater than a predetermined threshold, the gain is set to 1. Otherwise, the gain is set to 0 and as such that cell is suppressed. The time-frequency cells are then recompiled to form the output signal from which the noisy time-frequency cells have been removed, resulting, theoretically, in a less noisy signal. The mask is termed 'ideal' if it estimated from both clean and noisy data, but in practice, only noisy data is available and as such the mask is subject to estimation error (Li and Loizou, 2008).

Auditory masked transformation uses similar methods, but takes into account psychoacoustic effects of masking. As such, the signal attenuation in each time-frequency cell depends on the perceived rather than actual SNR, influenced by the noise level in frequency bands below the frequency of the cell being considered (Kates, 2017).

Due to commercial sensitivities and therefore a lack of published information, it has not been possible to acquire details on the exact methods employed in currently available HAs.

2.1.3 Reported Problems with Hearing Aids

A recent survey conducted by Action on Hearing Loss (2017a) found that 63% of people with hearing loss have at least moderate difficulty hearing in the presence of background noise. HA users rate performance in noise, particularly for speech listening, as the most valued yet least satisfactory attribute of their HAs (Bridges et al., 2012; Meister et al., 2002). After 'using the telephone', 'bars and pubs' are reported as the environment in which listening is the most difficult (Action on Hearing Loss, 2017a).

Many devices and functions are readily available for modern telephones which work in conjunction with or independently from the users' HAs (Action on Hearing Loss, 2018b). Remote microphones are also available to help with speech listening in noisy social situations, but these can be indiscreet and impractical in group situations. As such, following group conversations in noisy environments, such as bars, pubs and restaurants is mainly dependent on the noise cancellation abilities of the users' HA.

The effects of reverberation can be particularly problematic for users of HAs; background noise in the form of reverberation can be extremely difficult for digital noise reduction programs to reduce due to its similarity to the direct speech and in some cases, such algorithms have been shown to degrade SI in reverberation rather than improving it (Reinhart et al., 2020). Due to the complex nature of reverberation in the context of SI, effects of reverberation will not be examined in this thesis.

Although background noise management programs are widely available in a range of NHS HAs, less than half of HA wearers utilise these features, and those who do only report small benefits of doing so (Action on Hearing Loss, 2017a). It is unclear whether the small reported benefit is because noise reduction programs do not perform adequately in noisy situations or because programs are incorrectly used. It is also possible that many users are not aware that background noise cancellation is being implemented in their HAs, since many modern HAs are now capable of switching between programs automatically.

2.2 Behavioural Speech-in-Noise Testing

The most common way to diagnose or quantify a hearing loss is using PTA. This involves presenting tones to the participant at various different sound levels to determine the quietest sounds they can hear at a number of frequencies. HAs are fitted in the NHS using a standard prescription formula, which is based on the thresholds determined using PTA. The exact formula used varies depending on the manufacturer's recommendation but the most common currently in use is NAL-NL2, details of which can be found in Keidser et al. (2011). In addition to PTA, speech discrimination is routinely tested for cochlear implant fitting and assessment, but many audiology professionals do not regularly use speech tests in clinical assessments for HA fitting (British Society of Audiology, 2019). However, correlation between thresholds determined by PTA and difficulty with speech, especially in background noise, is insufficient for assessing difficulty in speech understanding (British Society of Audiology, 2019); it is possible to have a mild hearing loss but still struggle to understand speech in noisy situations, or a more severe hearing loss but cope well with speech in background noise (Chien et al., 2012).

The most common method for assessing speech in noise performance of an individual is to perform a speech in noise (SIN) test. Perhaps the earliest form of rigorous SI assessment was for use in evaluation of early telephones at the Bell Telephone Laboratories in 1910. Varying types of speech test following similar principles have been routinely used, both clinically and for research purposes, since the 1950s (Markides, 1997). Other areas where assessment of SI has been proved beneficial include room acoustics, personal speech communication systems such as telephones and VOIP (Voice Over Internet Protocols), public address systems and entertainment systems.

The aim of SIN tests is usually to determine the SNR at which words or sentences can be understood 50% of the time, known as the Speech Reception Threshold (SRT), or perhaps deduce the psychometric function (the percentage of correctly recalled words or sentences as a function of SNR) (Wright, 1997). A low SRT indicates that SIN performance is good at adverse SNRs, and so overall SIN understanding is good.

Speech intelligibility is defined by the Acoustical Society of America as "that property which allows units of speech to be identified", and can be measured objectively and quantitatively (Acoustical Society of America, 2022). Speech quality, however, is much more difficult to define and is not equivalent to SI; the widely-quoted American National Standards Institute definition of quality is given as "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar" (American National Standards Institute, 1960). This thesis will aim to examine methods relating to objectively defined speech intelligibility rather than the much more subjective aspect of speech quality.

2.2.1 Commonly Used Speech-in-Noise Test Procedures

SIN testing guidance given by the British Society of Audiology (2019) currently recommends the use of sentence-in-noise rather than isolated word or phoneme tests, specifically Hearing in Noise Test (HINT), Quick Speech in Noise test (QuickSIN), Bamford-Kowal-Bench (BKB)-SIN and City University of New York Sentence Test (CUNY), for "statistically meaningful and real-life" assessment of SIN capabilities (Boothroyd et al., 1985; Killion et al., 2006; Nilsson et al., 1994; Niquette et al., 2003). As such, focus will be given primarily to the first three of these tests in the following sections. The CUNY SIN test will not be examined in detail since its application for HA assessment is limited; speech discrimination is assessed at a single (high) SNR and so provides minimal information on SIN performance for listeners with a mild or moderate hearing impairment (British Society of Audiology, 2019).

Matrix tests, which can be administered automatically without the need for an assessor to record the participants' answers, will also be discussed in detail as an alternative to the tests recommended by the British Society of Audiology (2019) (HörTech gGmbH, 2018). A brief overview of two common examples of word-in-noise tests, the Four Alternative Auditory Feature test (FAAF) and the Words in Noise test (WIN), will also be given for comparison (Foster and Haggard, 1987; Wilson, 2003).

Some key features of a range of common SIN tests are given in Table 2.1.

Test	*Noise Type*	*Stimulus type*	*Corpus*	*Forced choice? (Closed set?)*
Bamford-Kowal-Bench Speech in Noise test (BKB-SIN)	Four-talker babble	Sentences	336 BKB sentences	No
HINT	Speech-Shaped Noise (SSN)	Sentences	250 adapted BKB sentences	No
QuickSIN	Four-talker babble	Sentences	720 Harvard (IEEE) sentences	No
Matrix Test (American English)	SSN	Sentences	5 by 10 word matrix (approximately 100,000 possible sentences)	Yes
CUNY	Babble, +10 dB condition only	Sentences	24 lists of 12 sentences	No
Matrix Test (American English)	SSN	Sentences	5 by 10 word matrix (approximately 100,000 possible sentences)	Yes
WIN	Multi-talker babble	Words	70 words from the Northwestern University Auditory Test No. 6 corpus	No
FAAF	SSN	Words	20 matrices of 4 (2 by 2) FAAF minimal pairs of words	Yes
Office of Research in Clinical Amplification (ORCA) Nonsense Syllables test	SSN	Nonsense Words (to assess syllable identification)	2 lists of 32 items (one male, one female)	No

Table 2.1: Details of a variety of SIN tests, including stimulus and noise type (Boothroyd et al., 1985; Etymōtic Research Inc., 2018a,b; Foster and Haggard, 1987; HörTech gGmbH, 2018; Kuk et al., 2010; Vermiglio, 2008; Wilson, 2003).

2.2.2 Stimuli and Speech Corpora

There are several possible options for the stimulus used in a SIN test, including isolated phonemes and syllables, words, sentences (both sensical and nonsensical) and continuous speech.

The shortest stimuli used in speech testing are isolated syllables. Use of syllables hold no true meaning to the listener and as such, inter-subject variation due to vocabulary and/or familiarity with a particular language is eliminated. However, since isolated syllables are not often encountered in everyday life, the stimuli are not ecologically valid and can be difficult for listeners to identify, reproduce or describe to the tester (Markides, 1997).

Some types of SIN test, including the ORCA Nonsense Syllables test, the WIN and FAAF (which presents the listener with four choices of word: the correct spoken word and three alternatives which differ from the spoken word only by one phoneme), present individual words to the listener (Foster and Haggard, 1987; Kuk et al., 2010; Wilson, 2003). Monosyllabic word tests have several advantages and disadvantages. Responding to a word stimulus is easier than identifying sounds in phoneme or syllable tests, but the results depend on the listener's vocabulary. Identification of words in isolation is a more ecologically valid test than presentation of isolated sounds, but is less representative of everyday situations than presenting complete sentences. Word tests, particularly closed-set (forced-choice) tests such as FAAF and the ORCA test can be useful in identifying phonemes of difficulty in a particular listener (Foster and Haggard, 1987; Kuk et al., 2010).

Sentence material is the most ecologically valid stimulus and gives the most clear indication of an individual's ability to follow conversation in a noisy environment. However, there are several key issues with presentation of sentences for speech-in-noise performance, including use of contextual cues, vocabulary, memory effects, learning effects and use of semantic cues. Sentence tests will be discussed in detail in the following paragraphs.

The BKB-SIN test, for example, uses a corpus (speech stimulus set) of twenty-one sets of sentences, with each set containing sixteen simple, British-English, anechoically recorded sentences suitable for school-aged children (Bench et al., 1979; Etymōtic Research Inc., 2018a; Niquette et al., 2003). Using sentences like the BKB sentences has several advantages. The British accented recordings mean they are suitable for British participants. They contain easy vocabulary and are simply grammatically structured so are unlikely to display inter-subject variation related to cognitive ability or memory (Bench et al., 1979). One study by Wilson et al. (2007) showed that the variability in

SRT for both normal hearing (NH) and hearing impaired (HI) groups when presenting two sentence sets is less for BKB-SIN than for HINT or QuickSIN. However, the presence of rich contextual and semantic content can be both advantageous and detrimental; these properties are very common in normal listening conditions and it is known that SI improves when the words can be predicted using context and semantics (Wilson et al., 2007), but very few (if any) automated metrics for prediction of SI are equipped to deal with speech recognition improvements based on the quantity of contextual cues available. The small size of the corpus also presents problems; if the same sentence is repeated to a participant several times, learning effects are likely to skew SI results since the participant may remember parts of the sentence they would not be able to recognise otherwise. However, this effect can be mitigated if the repeated sentences are presented a time apart, for example, in a different testing session (Wilson et al., 2003).

Matrix tests, on the other hand, typically use sentences of a set five-word format (e.g. name, verb, number, adjective, noun, for example, "Thomas sold five small chairs") created from a matrix of fifty words (ten in each of the five word categories) (HörTech gGmbH, 2018; Kollmeier et al., 2015). Since any combination of these words in the allocated format makes grammatical and syntactic sense, a huge database of up to one hundred thousand sentences is available and, as such, learning effects associated with matrix tests are minimised. However, recording each word in isolation and combining them later to form a sentence is problematic; such a method can result in unnatural speech patterns and artefacts in the recording that may be detrimental to SI (Kollmeier et al., 2015; Naylor, 2018). As such, words must be synthesized using a vocal synthesizer, resulting in a less realistic and natural listening condition compared to BKB sentences. The matrix test also has a closed-set format; that is, the participant knows that there are ten specific (and presented) options for each word. This introduces uncertainty in terms of guessing, or educated guessing (based on particular audible phonemes), of which words were present in the sentence without being able to understand the whole word. This kind of speech recognition improvement is also not accounted for by currently available automated metrics. Using the same material in an open-set format is possible, but introduces significant learning effects compared to other open-set tests such as QuickSIN. The Spanish matrix test, for example, has been shown to display a deterioration of 1 dB SNR in the SRT when conducted with an open-set format, with a training improvement of 1-2 dB between the first and sixth SRT measurements (Hochmuth et al., 2012).

Harvard (otherwise known as IEEE) sentences, used in QuickSIN, allow a compromise to be struck between the Matrix and BKB sentence corpora (Etymōtic Research Inc., 2018b). They form a large corpus (seventy-two sets of ten sentences) with an open-set format (no forced choice or options for the participant to choose from), but are designed to contain a low level of semantic and contextual cues, meaning that learning, predictive and cognitive effects are minimised simultaneously (Etymōtic Research Inc., 2018b). The sentences have also been phonetically balanced and equalised such that the variability in intelligibility between sentences is minimised (Etymōtic Research Inc., 2018b). The sentences have been previously recorded anechoically using a British speaker, and as such are natural sounding and appropriate for British study participants.

2.2.3 Noise Types

Different SIN tests use a variety of different types of noise at a number of different SNRs. The majority of testing methods used in clinical situations, for example, BKB-SIN, HINT, Matrix tests and QuickSIN, use either babble noise or SSN (British Society of Audiology, 2019; Etymōtic Research Inc., 2018a,b; HörTech gGmbH, 2018; Nilsson et al., 1994).

Babble noise, a background condition created by combining several streams of (usually unintelligible) speech or speech-like fluctuating voices, is a realistic type of background which simulates conditions likely to be encountered in everyday life (for example, background noise in a restaurant) and is highly representative of the types of situation where speech understanding is most difficult. In contrast, SSN is a stationary noise with the same long-term spectral balance as speech i.e. the time-fluctuating components of speech are removed but the stationary characteristics are preserved. SSN, due to its time-invariant nature, is typically easier for HA noise cancellation algorithms to categorise correctly as noise (and therefore remove from the total signal) than babble noise (Hu and Loizou, 2007). This means that the improvement in noise reduction performance due to a HA or particular program may be overestimated compared to real listening situations when using SSN. Babble noise represents a more ecologically valid background noise than SSN. However, the time-variant nature of babble noise may introduce additional inter-subject variation in SIN performance, particularly for HI participants, due to varying ability to use 'glimpses' (segments of a noisy signal in which the instantaneous SNR is high compared to the long-term SNR) to help identify words or phonemes (Best et al., 2017).

2.2.4 A Note on SNR Calculation

There are three ways to define the sound level of the components, and therefore the SNR, of a signal composed of both speech and noise. The first method is to determine the SNR of a combined signal by calculating the ratio of Root Mean Square (RMS) levels of the speech and noise respectively. A second technique, more commonly applied in SIN tests, is to use the vu-meter method. The third and final method, detailed in

International Telecommunication Union (2011), defines the active speech level as "measured by integrating a quantity proportional to instantaneous power over the aggregate of time during which the speech in question is present (called the active time), and then expressing the quotient, proportional to total energy divided by active time, in decibels relative to the appropriate reference". The vu-meter method assumes the speech level to be equal to the average of two or three peaks in a sentence, or, more generally, the average of 'frequent peaks' in any given segment of recorded material (Killion, 2009). Since SSN is stationary, the level calculated by the RMS and the vu-meter methods are equal. However, for non-stationary signals, like speech, using the vu-meter method results an increase in calculated level of approximately 5 dB compared to that of the RMS method, and therefore triggers an increase in SNR of 5 dB when combining speech with SSN. Although this scaling difference should not affect the relative SNR or SRT calculations for a particular set-up (for example, SNR-loss), it is important to note which method for calculating SNR has been used in order to compare data from other SIN tests or data sets, and apply an approximate 5 dB correction factor if necessary, for example, for comparing the absolute SRTs from the Matrix test SRTs with those from QuickSIN SRTs (Etymōtic Research Inc., 2018b; Hochmuth et al., 2012).

2.3 Automated Assessment of Speech Intelligibility

SIN tests, although effective for assessing SIN capability of individuals, are time-consuming and expensive to conduct on a large scale, for example, for assessment of SIN improvement associated with HAs or settings for an average population of HA users. As a result, it would be beneficial to identify an automated measure which can predict the SIN performance of a particular filtering process, such as the combination of processes used within a HA, without the involvement of a large number of trial participants or potentially biased trained listening experts. Several automated metrics aiming to predict speech and audio quality for a variety of listening situations have been developed, however, none has so far been identified and used to compare the SIN performance of a selection of real (as opposed to simulated) HAs and settings.

2.3.1 Basic Concepts of Speech Intelligibility Prediction

The most basic way to predict degradation of a speech signal is to compare the degraded signal to a clean reference. These kinds of metrics are called 'intrusive'. A basic outline of the process is shown in Figure 2.5.

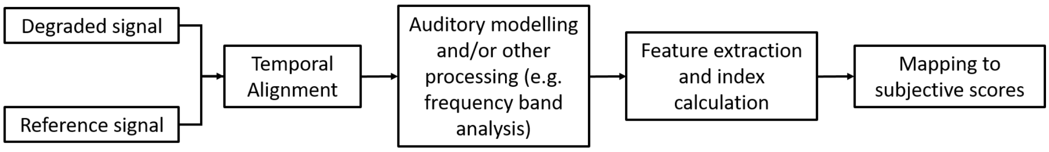


FIGURE 2.5: A flowchart showing the process of intrusive signal comparison, adapted from Kates (2013a).

Most current intrusive metrics (requiring a clean input signal for comparison) for predicting SI involve frequency band analysis of either a form of coherence or envelope correlation between a clean input signal and a degraded (processed/altered) output signal. Many also include some kind of auditory model. These features will be further discussed in the following sections. Non-intrusive metrics, which take only the degraded signal as an input, are used occasionally, predominantly in situations where obtaining a clean reference signal is not possible, for example, when using recordings taken in an uncontrolled (or 'real-life') environment where speech and background noise cannot be easily isolated. They commonly use noise detection techniques to reconstruct an estimate of the clean signal before applying similar processing to that of intrusive metrics. Because of this, non-intrusive metrics tend to be poorer at predicting SI or quality of degraded samples and so will not be discussed in detail.

The majority of intrusive SI metrics are based on one or both of the following principles:

1. Modifications to the Temporal Fine Structure (TFS), or the rapid oscillations in time close to the centre frequency of each auditory filter band (i.e. changes in the spectral content, largely relating to the change in phase between the noisy and reference signals) (American National Standards Institute, 1997; Moon and Hong, 2014),
2. Changes to the envelope of the signal (Steeneken and Houtgast, 1980).

Each of these two core concepts can be implemented in a variety of ways, described in the following section.

2.3.2 An Overview of Previous Methods

Table 2.2 gives a summary of the key features of a selection of automated SIN metrics for reference when reading the concurrent sections.

Test	*Key concepts*	*Designed for use on what kinds of distortion?*

Articulation Index (AI)	Changes to TFS	Distortion in telephone signals, stationary noise
Speech Transmission Index (STI)	Changes to speech envelope	Additive stationary noise, reverberation
Speech Intelligibility Index (SII)	Changes to TFS	Distortion in telephone signals, stationary noise
Coherence Speech Intelligibility Index (CSII)	Changes to TFS, auditory modelling	Additive noise (SSN), peak-clipping, center-clipping
(E)Short-Time Objective Intelligibility (STOI)	Changes to speech envelope	Ideal Binary Masking (IBM) processing, additive noise (SSN, babble and other environmental types of background noise, plus examples with additional temporal modulation for testing of Extended Short-Time Objective Intelligibility (ESTOI))
Normalised Covariance Metric (NCM)	Changes to speech envelope	Noise suppression, non-linear distortion
Hearing Aid Speech Perception Index (HASPI)	Changes to TFS, changes to speech envelope, auditory modelling	Additive noise (SSN and babble noise), peak- and centre-clipping, downsampling, frequency compression, IBM processing, vocoding.

Table 2.2: A table of key features of a range of intrusive speech intelligibility metrics (American National Standards Institute, 1997; French and Steinberg, 1947; Goldsworthy and Greenberg, 2004; Jensen and Taal, 2016; Kates and Arehart, 2005, 2014; Steeneken and Houtgast, 1980; Taal et al., 2010).

The AI is perhaps one of the most commonly used measures of SI in background noise (French and Steinberg, 1947). The AI sets a framework for several other metrics, such as the SII, which in turn forms the basis for CSII and HASPI, all of which calculate the drop in SI due to changes in the spectral content of the signal (the TFS) (American National Standards Institute, 1997; Kates and Arehart, 2005, 2014). These measures follow the same basic principle; the spectrum of the speech sample is divided into frequency bands, some measure of the SNR in each band is computed and a weighted average, based on the importance of each band in relation to SI, is taken (Ma et al., 2009). The CSII extends this procedure by additionally splitting the speech sample into high, moderate and low intensity levels and replacing the SNR with a signal-to-distortion ratio based on the coherence between the clean and distorted samples (Kates and Arehart, 2005).

A second commonly used metric for SI prediction is the STI (Steeneken and Houtgast, 1980). The STI focuses on the changes to the depth of modulation in the signal envelope as a measure of SI. Several metrics such as STOI and HASPI also use modulation depth changes in the signal envelope in order to estimate SI (Kates and Arehart, 2014; Taal et al., 2010). However, rather than using artificial sinusoids as probe signals as per the original STI procedure, newer metrics calculate the cross-correlation between frequency-banded envelopes in order to estimate SNR. This approach provides some compensation for changes to the signal caused by non-linear processing which are incorrectly interpreted by the original approach used by STI (Ma et al., 2009).

2.3.3 Key Metrics for Hearing Aid Speech Intelligibility

2.3.3.1 STOI

The intrusive metric STOI, developed by Taal et al. (2010), was primarily designed to address problems with previous SI metrics for application to signals processed with noise reduction or speech separation algorithms. As the name suggests, STOI is applied to short time segments of the sample in question (256 milliseconds in length), and employs a discrete fourier transform based approach to analyse the relative intelligibility compared to the reference signal.

STOI has been designed for signals sampled at 10kHz; as such, high frequency components above 5kHz (assumed to be outside the relevant frequency band for SI) are not accounted for. The procedure for STOI processing is simple; the signal is split into 256-sample frames using a 50% overlapping window. Each frame is zero-padded with an additional 256 samples before being transferred into the frequency domain and split into fifteen third-octave bands. The processed time-frequency segments are normalized such that the energy within the segment is equal to the clean speech energy within a local sample of 60 segments centred on the segment of interest. The processed segments are also lower-bound clipped, based on the signal-to-distortion ratio of the same local sample. Each of the modified segments, Y', are then given by Equation 2.3, where X denotes the clean signal segment, Y denotes the processed signal segment and n denotes the index of a segment relative to the current segment Y (i.e. Y(0)). α is a normalisation term, and the factor $10^{(-15/20)}$ denotes a lower bound clipping of 15 dB signal-to-distortion ratio.

$$Y' = max[min(\alpha Y, X + 10^{(-15/20)}X), X - 10^{(-15/20)}X]; \tag{2.3}$$
$$\alpha = \sqrt{\frac{\Sigma_n X(n)^2}{\Sigma_n Y(n)^2}},$$
$$n \in \{-29, -28, ...29, 30\}$$

The final output of the metric is given by the mean of the linear correlation coefficients between each of the equivalent clean (X) and modified (Y') time-frequency segments. For further details, please see Taal et al. (2010).

STOI has been shown to give SI predictions which correlate highly with behavioural SIN test scores for additive noise with no additional processing (including SSN, cafeteria noise and interior car noise), ideal binary mask and target binary mask processing (both of which can be thought of as speech separation/signal channel noise reduction procedures).

2.3.3.2 CSII

The Coherence Speech Intelligibility Index (CSII) is the first of two methods discussed here that were developed by Kates and Arehart (2005) for intrusive SI prediction for use on hearing aid processed speech. CSII is based heavily on the SII (American National Standards Institute, 1997) and aims to:

1. Generalise the SII in order to improve estimates of degradation in SI due to hearing loss, additive noise and bandwidth narrowing conditions;

2. Extend SII to accommodate for peak and centre clipping distortions.

There are two primary differences between the SII and the CSII. The first is that, in CSII, the signal-to-distortion ratio, based on the magnitude-squared coherence between the reference and degraded signals, replaces the traditional SNR estimate used in SII. The second is an additional step used in CSII which separates the signal into three intensity levels, computing the coherence between the reference and degraded signals separately for each of these levels.

The standard SNR calculation used in SII assumes perfectly separated noise (N) and speech (P) power spectra, each computed in the frequency domain, and is calculated for each frequency band as shown in Equation 2.4, where k denotes the index of the FFT bin.

$$SNR = \frac{\sum_k W(k)P(k)}{\sum_k W(k)N(k)} \tag{2.4}$$

In the case of CSII, the speech and noise spectra are estimated using the magnitude-squared coherence as follows in Equation 2.5, where γ denotes the autospectral density of the combined (total) distorted signal.

$$\hat{P}(k) = |\gamma(k)|^2 S_{yy}(k) \tag{2.5}$$
$$\hat{N}(k) = [1 - |\gamma(k)|^2] S_{yy}(k)$$

The estimates of speech and noise power from 2.5 are then substituted into Equation 2.4 to give the signal-to-distortion ratio.

During the intensity separation step, the small time-frequency segments of the signal are split into high, mid and low intensity levels relative to the overall RMS level of the signal. There is an approximately equal number of segments in each level. High-level segments are classified as those with an RMS level greater than the overall RMS level for the sentence and consist principally of vowel sounds. Mid-level segments, with an RMS level equal to or down to 10 dB below the sentence RMS level, mostly consist of transitions between vowels and consonants. Consonants and inter-word pauses make up the majority of the low-level segments, which encompass segments with an RMS level of between 10 and 30 dB below the overall level. The original implementation of CSII applies a weighting of zero to the high-level coherence (Kates and Arehart, 2005), but later implementations (such as those seen in (Kates and Arehart, 2014)) attach a zero weighting instead to the low-level coherence. In all cases, however, the mid-level coherence is weighted most heavily, implying a high importance of this amplitude region to overall SI, in line with existing literature (Plomp, 1988; Yoo et al., 2004).

2.3.3.3 HASPI

The Hearing Aid Speech Perception Index (HASPI) is the product of a large work by Kates and Arehart (2014), drawing together concepts described in a range of previous publications, including the CSII described in the preceding section (Kates and Arehart, 2005). HASPI extends the CSII in order to improve SI predictions for further common HA processing algorithms, such as frequency compression and noise suppression. This is implemented by combining the existing auditory coherence-based estimates with a measure for speech envelope degradation, in the formal of cepstral correlation, similar to the base principles used in the STI and STOI (Kates and Arehart, 2005; Steeneken and Houtgast, 1980; Taal et al., 2010).

HASPI incorporates a complex auditory model, detailed fully in Kates (2013a), which, in addition to modelling the auditory periphery of normal hearing individuals, allows for adaptations to simulate varying degrees of hearing loss. This model is summarised in Figure 2.6.

The output of the auditory model provides an envelope sample of each of the reference and processed signals, separated into 32 auditory filter bands between 80 Hz and 8 kHz, in addition to basilar membrane motion estimates in the same frequency bands. Two measures of SI are then calculated using these outputs.

The first of these measures is a TFS method heavily based on the CSII. The basilar membrane model output is further split into short time segments, and the intensity and normalised cross-correlation between the reference and processed signals for each time-frequency segment is calculated; this is equivalent (over a narrow frequency band) to the coherence calculated in CSII. Frequency cells over each time segment are combined to produce a broadband intensity measure for the signal, and silent intervals are then removed. Once silent segments have been discarded, a histogram of intensities is constructed and the signal is split into low, mid and high-level intensity portions using the upper, middle and lower thirds of the histogram. The cross-correlation values for the segments within each intensity portion are then averaged to produce the three-level auditory coherence measures similar to those seen in CSII.

The second measure is an envelope fidelity measure, calculated by cepstral correlation. The envelope output of the model is first split into short time-segments. Each of these segments is fit with a set of half-cosine basis functions, which relate closely to the principle components of short-time speech spectra (Zahorian and Rothenberg, 1981). Again, silent segments are removed, the short-time segments are recombined and the cross-correlation is calculated between the cepstral constructions for each segment of the reference and processed signals. Further details on the fitting of basis functions can be found in Kates and Arehart (2014). The final cepstrum correlation value is given by an average of the correlation across all basis functions.

2.3.4 Mapping Speech Intelligibility Correlates to SIN Scores

For an open-set SIN test (see Section 2.2.1) assuming few errors at high SNRs, the shape of the psychometric function for the SIN test is commonly modelled according to Equation 2.6, where α, β and γ correspond to the SRT, slope of the pyschometric

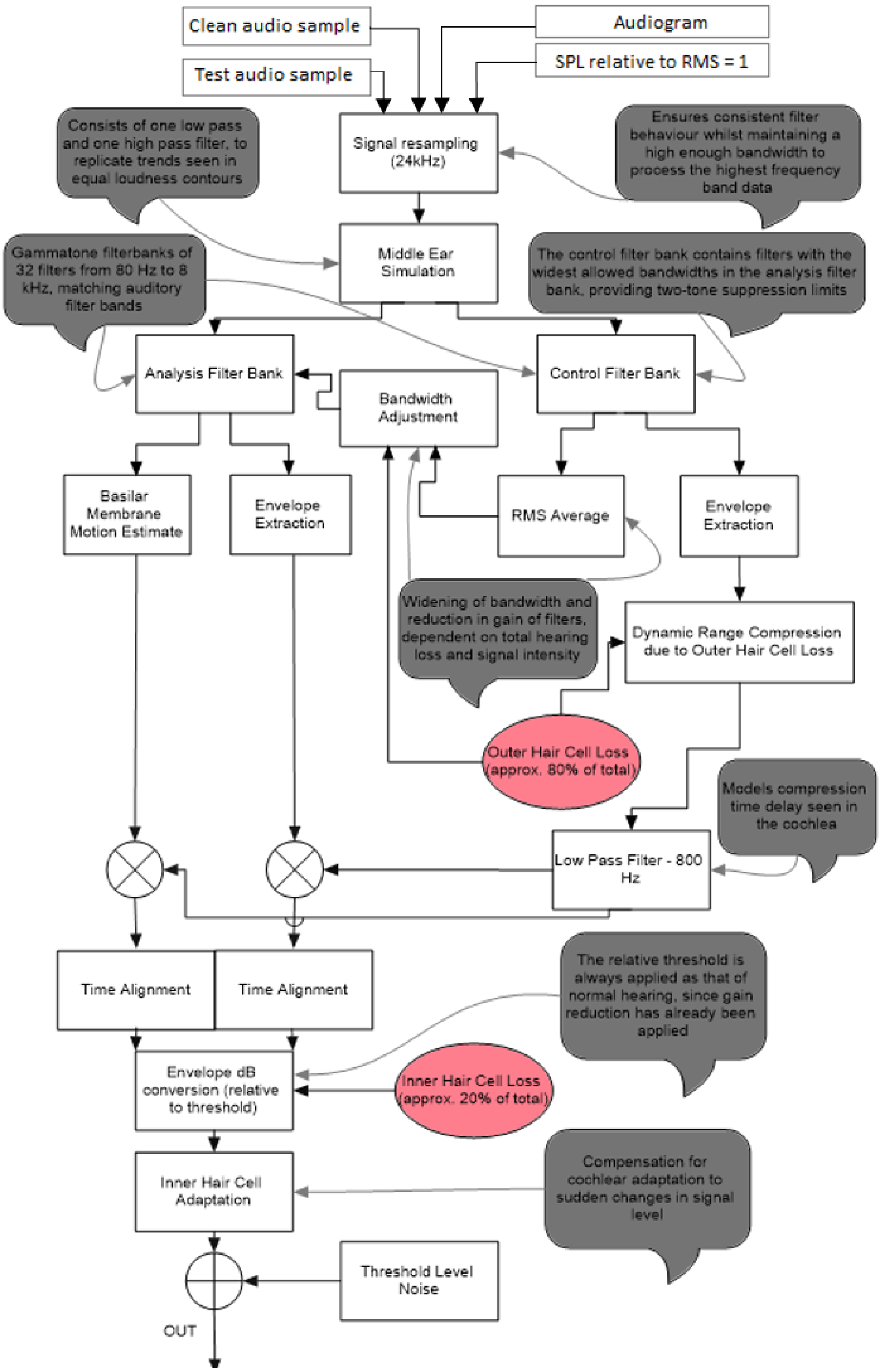


FIGURE 2.6: A flowchart illustrating the auditory modelling applied in HASPI, adapted from Kates and Arehart (2014). Annotations on functions of given steps are given in grey speech bubbles, predominantly taken from text in Kates (2013a).

function at the SRT and presentation SNR respectively (Brand and Kollmeier, 2002; Doire et al., 2017).

$$\%Correct = \frac{100}{1 + e^{\beta(\alpha - \gamma))}} \tag{2.6}$$

More generally, Equation 2.7 can be written as:

$$\%Correct = \frac{100}{1 + e^{f(a,b,c...)}} \tag{2.7}$$

where $f(a, b, c...)$ denotes some function of parameters output by a specific automated metric. Many SI metrics, including STOI, ESTOI, CSII and HASPI, require application of a mapping function in the form shown in Equation 2.7 in order to translate one or a combination of linearly calculated indices into a prediction of SIN score for a given sentence or speech sample. Parameters and weightings are usually fitted to the dataset at hand, but some metrics, for example HASPI, include predetermined mapping functions which have been fitted to a multitude of varying speech processing conditions so they can be applied to a wide range of datasets (Kates and Arehart, 2014).

2.3.5 Verification of Automated Test Measures

Although a wide range of SI metrics are available, the only existing metric which accounts for individual hearing loss is HASPI (Kates and Arehart, 2014). HASPI has been used regularly in recent years for assessment of HA algorithms and processing and has been shown to predict SI well for a variety of listening conditions (Lai and Zheng, 2019; Moshgelani et al., 2019; Van Kuyk et al., 2018).

Very few automated methods have been designed which enable SI to be predicted for HI users, particularly for HA assessment. Kates and Arehart (2014) trained the weighting indices used by HASPI on a large number of conditions. However, none of these conditions are created using a physical HA and instead, each of the HA processing effects (including additive speech-shaped/babble noise and distortion, frequency compression, several noise suppression algorithms and noise vocoding) is modelled and assessed separately (Kates, 2017; Kates and Arehart, 2014). Since HAs in practice use a combination of processing algorithms simultaneously, it is important that the effectiveness of SI predictors is assessed on speech samples processed by several realistic, concurrently occurring HA algorithms. However, research assessing the efficacy of automated metrics in predicting SI under these complex conditions is limited.

In order to accurately predict speech in noise performance, it is necessary for a good metric to predict the behavioural SI within a confidence interval of ±15%, the approximate difference between adequate speech understanding (>75% correct) and difficulty in following conversation (<60% correct) (Alfakhri, 2012)).

Given the literature discussed in this chapter, the main aim of this thesis is to assess the metrics currently available for SI prediction using real hearing aid outputs and their potential use in distinguishing between and ranking available hearing aids based on their associated output SI. Since all current metrics for assessment of SI are based on the normal hearing periphery (the HASPI extension for inclusion of hearing impairment notwithstanding), predictions will be assessed using normal hearing listeners in the first instance, before assessment of hearing impaired listeners is considered.

Chapter 3

Experiment Methodology

3.1 Aims

The main aims of the experiment were to:

- Assess the impact of hearing aid (HA) processing on the outcome of a behavioural speech in noise (SIN) test
- Determine whether speech intelligibility (SI) is significantly different between different HAs and settings for the given stimulus conditions.
- Evaluate the accuracy of various automated SI prediction metrics to predict the outcome of the SIN test for varying listening conditions.

In order to address these aims and determine which, if any, currently available automated metrics are suitable for use when ranking HA in terms of their SI, the study has been conducted in two main parts: behavioural speech-in-noise tests with participants, followed by automated speech-in-noise prediction. The same recordings of real HA outputs are presented in both parts of the study so that results from the behavioural and automated speech-in-noise tests are directly comparable.

3.2 Outline of Experimental Procedure

In order to compare the behavioural SI to automated predictions, each subject must be presented with the same sound samples as are input into automated prediction metrics. As such, recordings of the output of several HAs were taken using Knowles' Electronics

Manikin for Acoustic Research (KEMAR), a mannequin built using average anthropomorphic measurements which allows recordings to be taken at the position of the ear drum (G.R.A.S. Sound and Vibration, 2013). These recordings were presented to subjects through insert earphones, in the form of a speech-in-noise test. The same recordings were also used as the input to each of the automated metrics. The speech-in-noise test results can then be compared to the automated predictions, with error due to differences between the signals used for behavioural and automated assessments minimised.

This approach is similar to those detailed by Kates and Arehart (2014) and Taal et al. (2010), both of which present identical signals for analysis in both behavioural and automated tests. Further work by Kates and Harvey Jr. (2018) additionally tests the Hearing Aid Speech Perception Index (HASPI) algorithm using recorded HA outputs but does no comparison work to behavioural tests with the same stimuli. Another study by Falk et al. (2015) used real HA outputs in both behavioural and automated speech quality tests but did not assess automated speech intelligibility algorithms in the same way. The approach taken in this experiment differs from all other known work to date in that signals are taken from real HAs and these identical signals are used to predict SI in both behavioural and automated tests for direct comparison.

3.3 Selection of Stimulus Materials

Since HASPI, the only automated metric for SI prediction which accounts for hearing loss, gives a prediction of "percentage sentences correct", it is necessary to use sentence-length materials to perform the speech test. These sentences should also have an open-set format to avoid learning effects, and should contain a low level of semantic and contextual cues to reduce predictive and cognitive aspects, as these effects cannot be accounted for by automated metrics. Learning and memory effects can also be reduced if the sentence corpus is large, such that each participant is never presented with the same sentence twice. Short sentences are preferable to reduce inter-subject effects relating to memory capacity. Anechoic recordings of the sentences using a British speaker should be available such that they sound natural to British study participants and can be easily used in the HA recording process.

The Harvard (or IEEE) sentences, used in QuickSIN (Etymōtic Research Inc., 2018b), were chosen since they meet all of the above described criteria (see Section 2.2.1 for further details). The sentences have also been phonetically balanced and equalised such that the variability in intelligibility between sentences is minimised. The Bamford-Kowal-Bench (BKB) sentences were also considered since the vocabulary used is more common in everyday British English than some of the vocabulary in the Harvard sentences, but the high level of semantic content and contextual cues as well as the limited number of available sentences made the BKB corpus less suitable.

Two main options are available for background noise; Speech-Shaped Noise (SSN) and four-, six- or multi-talker babble noise. Babble noise is a more ecologically valid choice than SSN since it more closely represents the type of background noise experienced in everyday life by HA users. However, the time-variant nature of babble noise may introduce unwanted inter-subject variation in SIN performance, particularly for hearing impaired (HI) participants, due to varying ability to use 'glimpses' (segments of a noisy signal in which the instantaneous signal-to-noise ratio (SNR) is high compared to the long-term SNR) to help identify words or phonemes. As such, SSN has been chosen as the background noise for use in this experiment, but further experiments (see Section 6.2.1) explore the effects of noise types more relevant to everyday listening, such as four- and six-talker babble. For more details, please refer to Section 2.2.1.

3.4 Stimulus Generation

A Mackie H824Rmk2 studio monitor (LOUD Technologies Inc., Washington, USA, 2007) was positioned approximately 1 m from the ear position of KEMAR (GRAS Sound and Vibration, Golte, Denmark, 1972), at approximately the same height (1.25 m), as shown in Figure 3.1. An RME Fireface UCX USB Audio Interface (Audio AG, Haimhausen, Germany, 2008) was used to drive the loudspeaker and record the microphone output from the left ear of KEMAR. A calibration sample of SSN was used to calibrate the recording level from KEMAR. A B&K sound level meter was used to confirm that the calibration signal arriving from the loudspeaker at the left ear of KEMAR had an overall sound pressure level (SPL) of 69 dBA. The calibration signal level was higher than the average of the sentences; calibrating this signal to 69 dBA corresponds to an average clean sentence level of 61 dBA. The output of a 94 dB calibrator and a sample of "silence" (background noise) in the room was also recorded in order to establish the level of the noise floor in measurements. The calibrator output was used to verify the measured overall SPL and to ensure that the noise floor in the recordings is sufficiently low that it will not impact the SI (approximately 27 dBA).

A diagram of the set-up, including pre- and post-processing steps, is shown in Figure 3.1. Hearing loss is modelled as an input to HASPI in the form of a 6-frequency audiogram, which is applied to the noisy, processed sound samples only and not the reference signal.

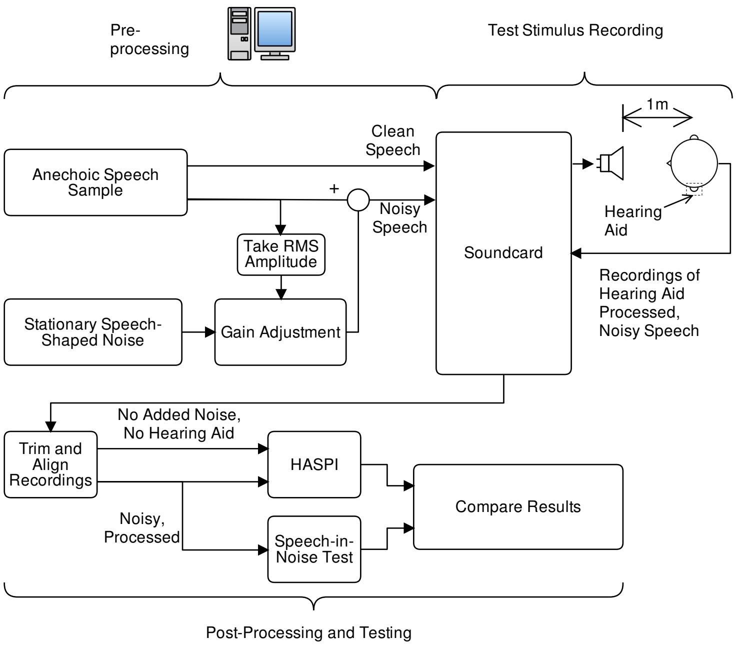


FIGURE 3.1: Experimental set-up showing KEMAR and a loudspeaker, as well as pre- and post-processing steps.

Photographs of the set-up corresponding to the recording stage in Figure 3.1 can be seen in Figure 3.2.

The entire corpus of Institute of Electrical and Electronics Engineers (IEEE) sentences (seventy-two lists with ten sentences per list) was recorded through the above setup at a sampling rate of 22050 Hz. Noisy speech samples at SNRs of -8, -3, 0, and +3 dB, as well as the clean sentences, were recorded for each of the HA conditions detailed below. The SNR of each sentence was created such that the sentence level remained constant and the level of the additive SSN was increased to give the required ratio of Root Mean Square (RMS) amplitude, to a maximum output level (of combined speech, noise and HA processing) of 82 dBA.

The recording conditions were as follows (further details of the exact set-up for each condition are given in 3.5). Full technical specifications for each HA can be found in Appendix A:

- No HA (control condition)
- Signia Contrast S+ Behind-the-Ear (BTE) HA (Erlangen, Germany. First available on the National Health Service (NHS) in 2018)

FIGURE 3.2: Photographs of the recording set-up with KEMAR.

  - In ‘first fit’ configuration
  - With noise reduction algorithms switched on

- Oticon Spirit Synergy BTE HA (Copenhagen, Denmark. First available on the NHS in 2018)
  - In ‘first fit’ configuration
  - With noise reduction algorithms switched on

- Danalogic Ambio 77 BTE HA (Ballerup, Denmark. First available on the NHS in 2018)
  - In ‘first fit’ configuration
  - With noise reduction algorithms switched on

- FK-162 low-cost amplifying aid (volume control only) by GlobalCareMarket/FoKe (£25.85/pair)

The noise reduction algorithms were programmed as recommended by the manufacturer; as such, it is not reasonable to assume condition equality or to therefore directly compare HAs performance.

3.5 Hearing Aid Set-Up

The most recently released NHS HA models (as of August 2018) were obtained from four providers, three of which have been used for this experiment (see Section 3.4). All of the programmable HAs were initially fitted unilaterally to a standard, mild, moderately sloping hearing loss as shown in Figure 3.3 (Bisgaard et al., 2010). Limits of ±5 dB are shown in cyan; these limits indicate the range of audiograms which would be of particular interest to a secondary part of the study involving hearing impaired participants, since they closely match the specified audiogram to which the HAs are programmed (in line with the maximum 5 dB variation between retests detailed in the recommendations for Pure Tone Audiometry (PTA) in clinics (British Society of Audiology, 2011)). These limits represent a range of mild hearing losses (British Society of Audiology, 2011). The fictional test patient was a 63-year-old female with no previous experience with HAs. All HAs were coupled using open-fit domes, as is common for the type of hearing loss shown in Figure 3.3 (see Section 2.1 for more details on HA use). Two programmes were used for each HA (described in detail in sections 3.5.1 to 3.5.3) and volume control was disabled.

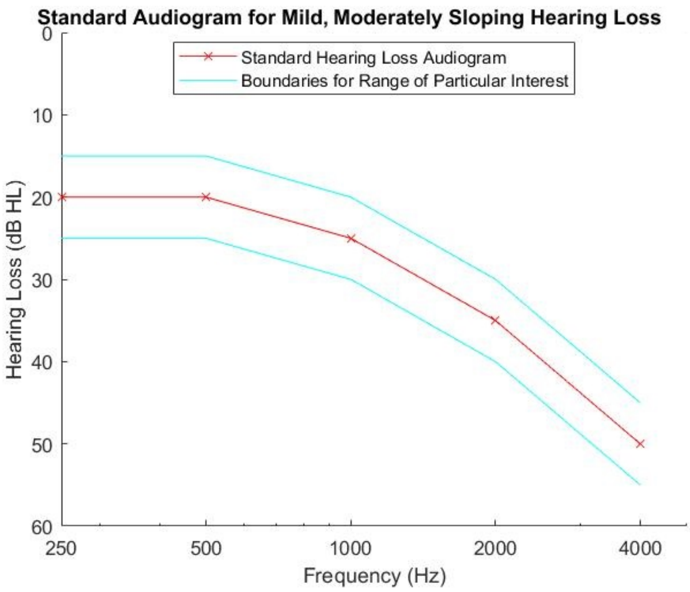


FIGURE 3.3: The average audiogram for patients with mild, sloping hearing loss including limits for the hearing loss of participants of interest, modified from Bisgaard et al. (2010).

3.5.1 Danalogic Ambio 77

The Ambio 77 HA, shown in Figure 3.4, was fitted according to the NAL-NL2 fitting formula.

FIGURE 3.4: Photographs of the Danalogic Ambio 77 device used.

The 'All-Around' program was set to maintain an omni-directional pattern with all noise reduction settings switched off. The second setting had a fixed directionality with a high directional mix (i.e. very little omnidirectional information is maintained). The Noise Tracker II noise reduction algorithm was enabled and set to 'strong'. Both settings were set to a 'syllabic' time constant, corresponding to a fast compression speed.

3.5.2 Signia Contrast S+

The Contrast S+, shown in Figure 3.5, was fitted using Signia's own 'Primax' fitting formula. This formula is closely based on NAL-NL2 but accommodates for new settings available on the device, such as SpeechFocus.

The Universal program was set to use 'TruEar' directionality and had directional speech enhancement and feedback cancellation settings switched on but speech and noise management, SoundSmoothing and eWindScreen turned off.

The Noisy Environments program used a directional microphone pattern and has broadband speech and noise management switched on. Directional speech enhancement remained on and SoundSmoothing and eWindScreen remained off.

FIGURE 3.5: Photographs of the Signia Contrast S+ device used.

3.5.3 Oticon Spirit Synergy

The Spirit Synergy HA, shown in Figure 3.6, was fitted according to the NAL-NL2 fitting formula.

FIGURE 3.6: Photographs of the Oticon Spirit Synergy device used.

The 'first fit' program used the 'surround' or optimal omnidirectional setting with noise management settings switched off. The noise reduction program had a full directional pattern implemented and noise management setting switched on. The 'listening support' level was increased significantly when the noise reduction program is activated. Both programs are set to 'gentle' and as such operate with a minimum compression speed.

3.5.4 FoKe FK-162

The amplifying aid, the FoKe FK-162, purchased online, is shown below in Figure 3.7. The only variable setting on this aid is the volume control, which was set to a low level of amplification (setting number 1 of 5). Two dome sizes were provided with the amplifier; the smaller dome was used for testing.

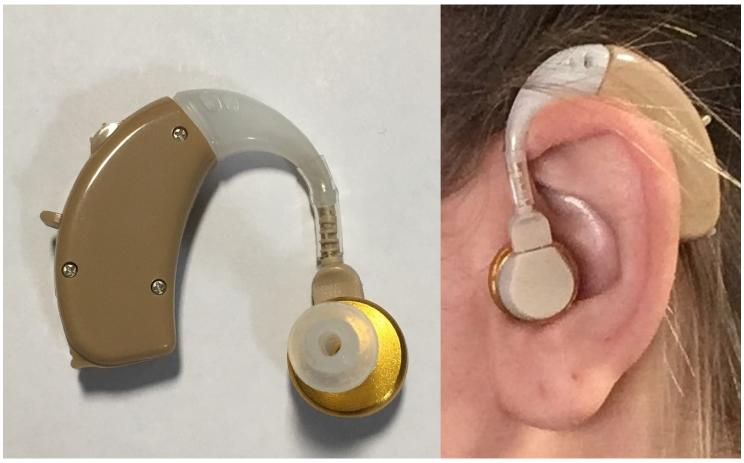


FIGURE 3.7: Photographs of the FoKe FK-162 amplifying device used.

Chapter 4

Speech-In-Noise Performance for Different Hearing Aids According to Behavioural Tests

4.1 Procedure for Speech-in-Noise Tests

Speech-in-Noise tests are used to give a behavioural indication of speech intelligibility. Ethics approval was obtained (University of Southampton ERGO Number 41012, see Appendix B for details) and written informed consent was provided by all participants. Twenty-one normal hearing participants (nine male, eleven female and one undisclosed, aged between 21 and 56 with a mean age 30.0 years) with hearing thresholds less than 15 dBHL participated in this experiment. Each participant was presented with 320 sentences in a random order (forty random sentences for each of the eight hearing aid (HA) conditions - three different hearing aids with two different fits plus a control condition and a low-cost amplifier condition - ten sentences each at signal-to-noise ratios (SNRs) of -8, -3, 0 and +3 dB). Sentences were presented monoaurally to the better ear (i.e., closest to an overall 0 dBHL, four right and seventeen left) through ER2 insert earphones (Etymotic, Illinois, USA, 1984), which are designed to accurately reproduce signals recorded on the KEMAR manikin. The sentences were calibrated to the same level as that at which they were recorded. No adjustment to gain was made according to hearing thresholds. Calibration of sentence presentation level was done using an occluded ear simulating coupler to connect the tube of the inserts to the sound level meter, and adjusting the calibration noise to a level of 69 dBA. The sentence level was then scaled accordingly, with a maximum combined sound pressure level (SPL) of 82 dBA.

Sentences were scored by both keywords correct and sentences correct. If all five keywords in a sentence were recalled correctly, the sentence was marked as correctly recalled, in line with the procedure used by Kates and Arehart (2014).

4.2 Results

Figure 4.1 shows the speech intelligibility (SI) associated with various HA conditions from the experimental procedure described in Section 4.1. From this point, the following abbreviations will be used: HAs 1-3 each represent one of the three National Health Service (NHS) models used, with the suffix 'NR' indicating the noise reduction setting is enabled.

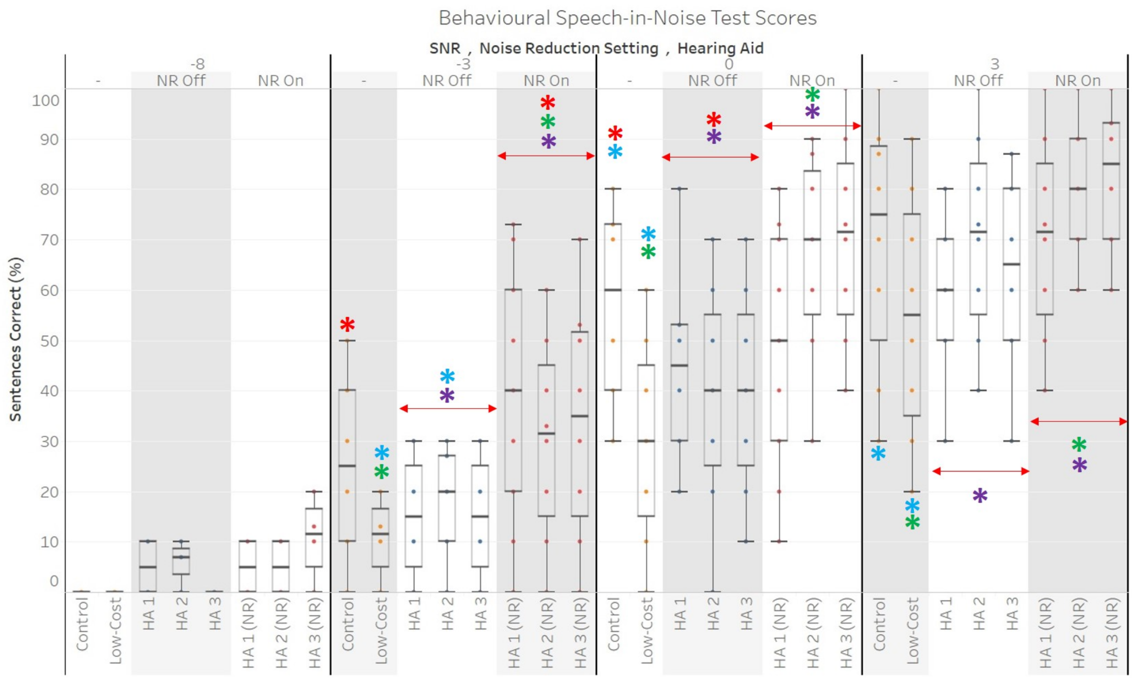


FIGURE 4.1: Results of the behavioural SIN test, as described in Section 4.1. Asterisks of the same colour in each SNR indicate pairs of conditions or groups (indicated by arrows over shaded bands) which are significantly different from each other ($p < 0.05$, equivalent to a Bonferroni-corrected threshold of $p < 0.002$) according to Wilcoxon Signed Ranks tests for paired data; for example, SI associated with the low-cost amplifier is significantly lower than that associated with the group of hearing aid conditions with noise reduction algorithms enabled at 0 dB SNR, as shown by the green asterisk. Each box-and-whisker plot covers the median, minimum, maximum and quartiles across the scores for the 21 participants, with individual participant's scores shown by the red dots.

Kolmogorov-Smirnov and Shapiro-Wilk normality tests reveal that, in the majority of processing conditions, behavioural data is significantly skewed ($p < 0.05$), therefore non-parametric tests have been used to statistically evaluate the behavioural test results. Full statistical details for all comparisons are given in Appendix C.

The effects of gender, presentation side and age were evaluated using SIN scores in the control condition. Correlation between age and score was insignificant at all SNRs ($p > 0.09$). A simulation using 100,000 random permutations reveals that effects of gender are not significant at any SNR ($p > 0.46$).

No significant differences were found at any SNR between any of the HAs with noise reduction switched on ($p < 0.0005$, equivalent to a threshold of $p < 0.05$ with Bonferroni correction for repeated tests, using Wilcoxon Signed Ranks tests for paired data). Similarly, none of the NHS HAs with noise reduction switched off showed any significant differences. It is important to note that the noise reduction algorithms varied in several aspects, such as strength/severity of processing and attack time, and did not necessarily have identical or equivalent settings. This allowed for collection of speech samples with similar but not identical processing applied, to introduce some slight variance for assessing automated metrics, covered in detail in Section 5.

Following comparisons of the NHS HAs with noise reduction switched on and off, further comparisons were made with similar conditions grouped. The control condition, low-cost amplifier condition, the mean score across the three conditions with noise reduction and the mean score across the three conditions without noise reduction were compared. Significant differences ($p < 0.05$ or $p < 0.002$ with Bonferroni correction) were found between several pairs of groups, as shown in Figure 4.1. The key point to note from these comparisons in Figure 4.1 is that noise reduction significantly improves performance compared to conditions with no noise reduction processing and the low-cost amplifier at all SNRs above -3 dB, and also compared to the control condition at -3 dB SNR.

At SNRs of -3, 0 and 3 dB, the average SIN performance across participants is best for the control and three noise reduction conditions, followed by the three NHS conditions with no noise reduction, with the low-cost amplifier giving the worst average performance. At least half of participants achieved the lowest SIN score with the low-cost amplifier at every SNR.

4.3 Discussion

4.3.1 Summary of Key Results

No significant differences in SI were found between different NHS HA manufacturers for the sample group of normal hearing listeners. Significant improvements were found when HAs had noise reduction switched on compared to off (including the low-cost amplifying device), and significant degradation to SI was seen in the low-cost device compared to the control condition.

The lack of significant differences between NHS manufacturers would indicate that, for the sample group of normal-hearing listeners tested, the standard set of HA processing algorithms for each model has a similar effect on the SI of Institute of Electrical and Electronics Engineers (IEEE) sentences in matched speech-shaped noise. At -3 dB SNR, the low-cost amplifier significantly degraded SI performance compared to NHS models and caused a significant loss of SI relative to the unaided condition at positive SNRs. It is possible that, with a larger sample of participants and/or more sentences presented per condition, larger, statistically significant differences may have been seen between conditions; however, it is clear from the current sample that any effects are likely to be small and therefore clinically insignificant (see Section 2.3.5 for details on clinically significant differences).

Although standard HA processing does not improve (and, 0 dB SNR, significantly degrades) SI for normal hearing listeners, single-channel noise reduction algorithms appear to significantly improve SI compared to standard HA processing alone, and at low SNRs can even improve SI compared to the unaided condition. However, differing implementations of single-channel noise reduction across the HA models perform equally well on average across the sample participants.

4.3.2 Interpretation of Results

It was expected that HA processing, especially in the absence of effective noise reduction, should result in a degradation of SI for normal hearing listeners, since basic processing introduces distortion to the signal. This degradation is evident in the results from the low-cost amplifier, but is less severe with the NHS models, which outperform or perform the same as the low-cost model in all cases.

Improvements to SI with single-channel noise reduction are unexpected. The improvement to SI compared to the control condition at lower SNRs (in addition to the improvements compared to standard HA processing) is particularly noteworthy; although it has been shown that ideal masks (which have been generated using separate noise and speech signals) can improve SI, single-channel noise reduction methods which rely solely on the mixed (speech and noise) signal have not generally been shown to improve SI for listeners with normal hearing, particularly where the noise and speech have similar frequency spectra (Brons et al., 2012; Hilkhuysen, 2012; Hu and Loizou, 2007). This is likely due to the nature of the noise introduced in this study; stationary noise is widely known to be more effectively cancelled than other types of fluctuating or more 'realistic' everyday background noises.

The study results discussed here must be interpreted with caution; since the main aim of the study is to provide a 'gold standard' to enable direct evaluation of automated SI metrics, several limitations exist which make direct comparison of HA models and conditions unrealistic. Although the lack of differences between NHS models, improvements compared to non-programmable models and the efficacy of noise reduction is initially encouraging in terms of HA evaluation, these results may not necessarily reflect those for everyday HA users with hearing impairment, especially for those with more severe hearing losses which require more severe signal alteration/amplification than the HA programming tested here provides. The study assesses only a very limited subset of testing conditions: only one sentence corpus has been used, spoken by one individual in a single type of stationary background noise. Different stimulus types, gender and individual vocal characteristics of the speaker, and, in particular, non-stationary noise or noise with differing spectral content, may result in SI variations which differ from those of the subset tested here (see Sections 2.2.2 and 2.2.3 for additional details on the impacts of these aspects on SI). To further this, the SIN test assesses the impact of processing on SI, but does not take into account listening effort, listening comfort, sound quality or naturalness of the resulting audio, all of which are important factors to consider for users of HAs (Bridges et al., 2012; Meister et al., 2002).

Chapter 5

A Comparison of Automated Metrics for Prediction of Speech Intelligibility in Noise

5.1 Automated Metric Application

5.1.1 Selection of Automated Metrics

Three different algorithms were used generate a prediction of speech intelligibility (SI): Hearing Aid Speech Perception Index (HASPI), Short-Time Objective Intelligibility (STOI) and Coherence Speech Intelligibility Index (CSII), as used in Kates and Arehart (2014), Taal et al. (2010) and Kates and Arehart (2005) respectively. MATLAB scripts for HASPI and STOI were obtained directly from the developers. MATLAB scripts for CSII were obtained from the appendix of Loizou (2007). All three metrics are intrusive; these compare the noisy signal to a reference signal with no added noise (a recording of the appropriate sentence with no noise added and no hearing aid fitted to Knowles' Electronics Manikin for Acoustic Research (KEMAR)).

HASPI was chosen since it is the only existing SI metric which can account for changes due to hearing loss, and is widely used to assess hearing aid (HA) processing conditions in relevant literature. HASPI uses audiometric thresholds as an input to account for these changes. Predictions for SI for hearing impaired (HI) subjects matching the audiogram shown in Figure 3.3 as well as participants with normal hearing (NH) were made using HASPI.

STOI is also commonly used to assess HA algorithms, particularly for speech enhancement. The extended version, Extended Short-Time Objective Intelligibility (ESTOI), which has been developed to improve performance in fluctuating background noise, was not used here since an appropriate mapping function for mapping the ESTOI SI correlate to behavioural speech in noise (SIN) scores could not be obtained (see Section 2.3.4 for further background) and no fluctuating maskers have been used in this experiment.

CSII, a precursor to and component of HASPI, has also been included. This is because it has been shown to perform well in a variety of conditions for predicting SIN performance, and uses different methods to determine SI to STOI, allowing a comparison of different approaches to automated prediction of SI to be analysed.

5.1.2 Verification of Implementation

To ensure implementations of the chosen metrics were correct and results obtained from them could be compared to results seen in other publications, test samples were run through each metric.

Full data and results for previous studies using STOI were obtained from Healy and Yoho (2013), so metric outputs could be directly compared with identical input stimuli, verifying that the implementations used in this study and in Healy and Yoho (2013) produce the same results. CSII implementations from Loizou (2007) were checked in the same way using recordings from the same source and comparing outcomes to those seen in Ma et al. (2009). Files for verification of HASPI are given, with expected outcome values, in the user guide provided with the HASPI code obtained directly from the author, and were used directly to check the outputs of the given implementation (Kates, 2013b).

5.1.3 Signal Preprocessing

Recordings of speech with and without background noise and/or HA processing were obtained as detailed in Section 3.2. First, all speech samples were scaled in amplitude using a calibration recording of known sound pressure level (SPL), such that an Root Mean Square (RMS) amplitude of 1 is equal to 65 dBA.

A high-pass Chebyshev Type II filter of order 16 with a cut-off frequency of 45Hz was used to remove low frequency fluctuations below the threshold of hearing.

Each noisy/processed sample was then aligned in time with its corresponding reference sample (with no added noise and no hearing aid processing) using a broadband cross-correlation method.

A recording of 'silence' in the recording space (with no stimulus or intentional noise present) was used to determine the noise floor level in the recording space. The reference speech samples were trimmed of leading and lagging zeroes by removing all samples before the threshold level is reached for the first time and after the threshold level is reached for the final time. The corresponding noisy/processed sample is then trimmed to the same time frame as the reference samples.

Finally, the processed samples are used as inputs to the selected automated SI prediction algorithms (HASPI, CSII and STOI). The outputs of these algorithms are SI correlates which require further interpretation in order to give a direct prediction representative of a behavioural SIN score.

5.1.4 Output Mapping

STOI, HASPI and CSII all require a mapping function in the form of Equation 5.1 (as seen in Section 2.3.4), where $a_{1,2...}$ and b are constants and $d_{1,2...}$ are numbers corresponding to the function outputs related to SI (e.g. cepstral correlation, auditory coherence and other intermediate intelligibility estimates, the values of which depend on various aspects of speech (see Section 2.3.1).

$$\%Correct = \frac{100}{1 + e^{(b+a_1 d_1 + a_2 d_2 + ...)}} \tag{5.1}$$

In the case of HASPI, this mapping function is inbuilt, and so these metrics generate a direct prediction of SI in terms of percentage correct. For HASPI, $b = 9.047$. d_1 is a measure of cepstral correlation and d_2 relates to high-level auditory coherence; $a_1 = 14.817$ and $a_2 = 4.616$ (Kates and Arehart, 2014). In the cases of CSII, STOI and ESTOI, this mapping must be performed in addition to the core estimation procedure (i.e. the parameters a and b must be estimated from the data). In the case of CSII, d_1 and d_2 correspond to coherence in mid and high intensity regions, with $b = -2.623$, $a_1 = 9.259$ and $a_2 = 0.470$, as seen in Kates and Arehart (2014). For STOI, the constants $a = -13.45$ and $b = 9.36$ have been used, as per Coelho and Nascimento (2016). For ESTOI, no suitable mapping parameters have been found in the available literature, so further investigation is necessary before this metric can be used to analyse the given data.

5.2 Results

5.2.1 Automated-Only Predictions

Figure 5.1 shows the SI, as predicted by the HASPI automated metric, associated with various HA conditions from the experimental procedure described in Section 4.1.

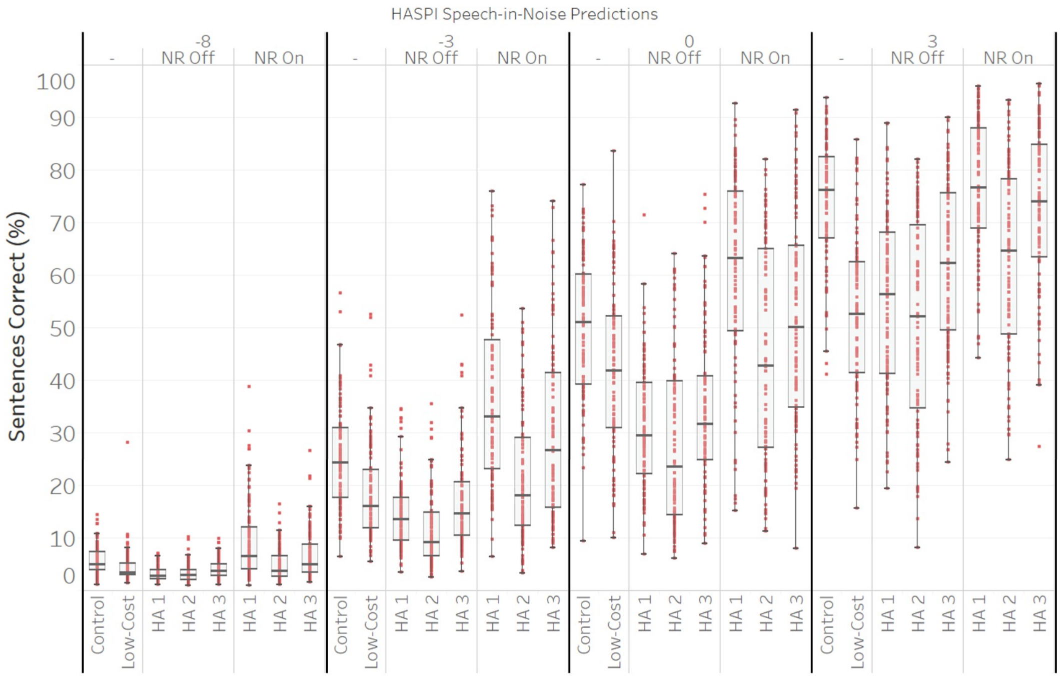


FIGURE 5.1: Box plots of HASPI predictions of speech intelligibility for the same processing conditions as those described in Section 4.1. Individual measurement points consisting of the average score for ten sentences, corresponding to those used in behavioural tests, are given as red dots.

.

The following analysis has been done using the HASPI scores generated from the same samples as those presented to participants; the HASPI score for each of the ten sentences presented to each participant at each signal-to-noise ratio (SNR) and in each listening condition constitutes one data point, resulting in twenty-four sets (four SNRs for eight conditions) of 210 (ten sentences for each of the twenty-one participants) HASPI values. The reader is reminded that the sentences presented to each participant for each listening condition were randomised and no individual participant listened to the same sentence more than once.

Kolmogorov-Smirnov and Shapiro-Wilk normality tests show that, as in the case of the behavioural data, the distribution of HASPI predictions is significantly skewed ($p < 0.05$), therefore non-parametric tests have been used to statistically evaluate the behavioural test results. Full statistical details for all comparisons are given in Appendix C.

Since the HASPI scores are generated independently of listener information, Mann-Whitney tests for non-parametric, unrelated samples have been used to compare scores for each listening condition. In approximately 80% of set comparisons, a significant difference in predicted intelligibility was detected ($p < 0.0005$, equivalent to $p < 0.05$ when Bonferroni correction is applied). A table illustrating this can be found in Appendix C.

Comparisons were also made with similar conditions grouped, as in Section 4.2. The control condition, low-cost amplifier condition, a condition for all samples with noise reduction switched on and another for all the samples with noise reduction switched off were compared. Significant differences ($p < 0.002$, equivalent to $p < 0.05$ with Bonferroni correction) were found between all pairs of groups at all SNRs except between the noise reduction condition and the control condition, and between the low-cost amplifier and the condition with no noise reduction at +3 dB SNR only.

Figures 5.2 and 5.3 below show the SI, as predicted by the STOI and CSII automated metrics respectively, associated with various HA conditions from the experimental procedure described in Section 4.1.

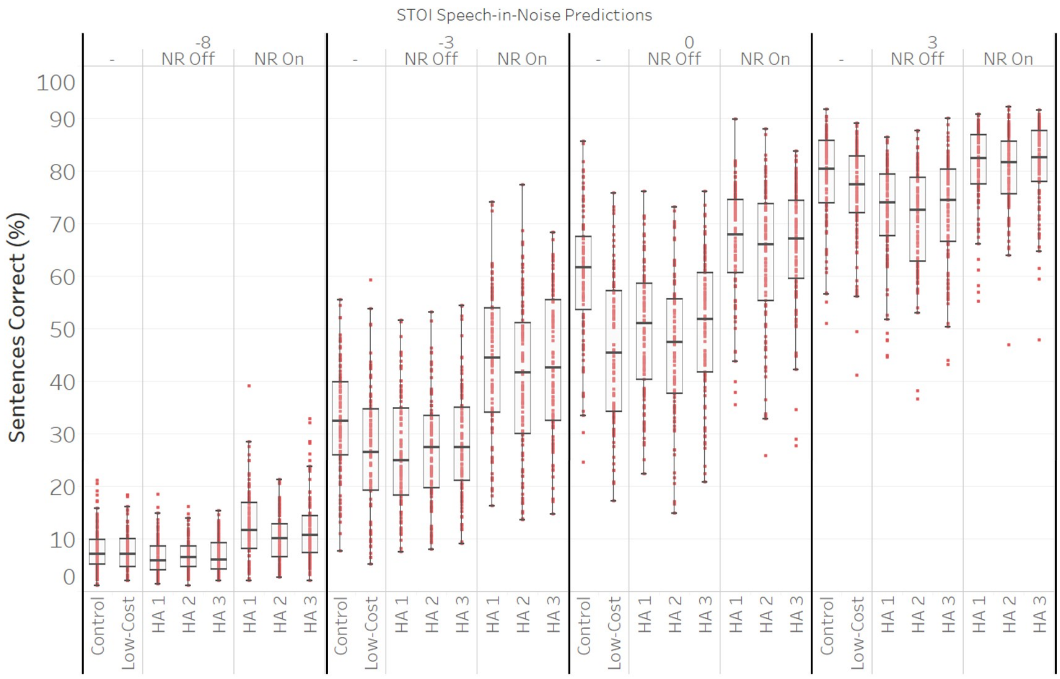


FIGURE 5.2: Box plots of STOI predictions of speech intelligibility for the same processing conditions as those described in Section 4.1. Individual measurement points consisting of the average score for ten sentences, corresponding to those used in behavioural tests, are given as red dots.

.

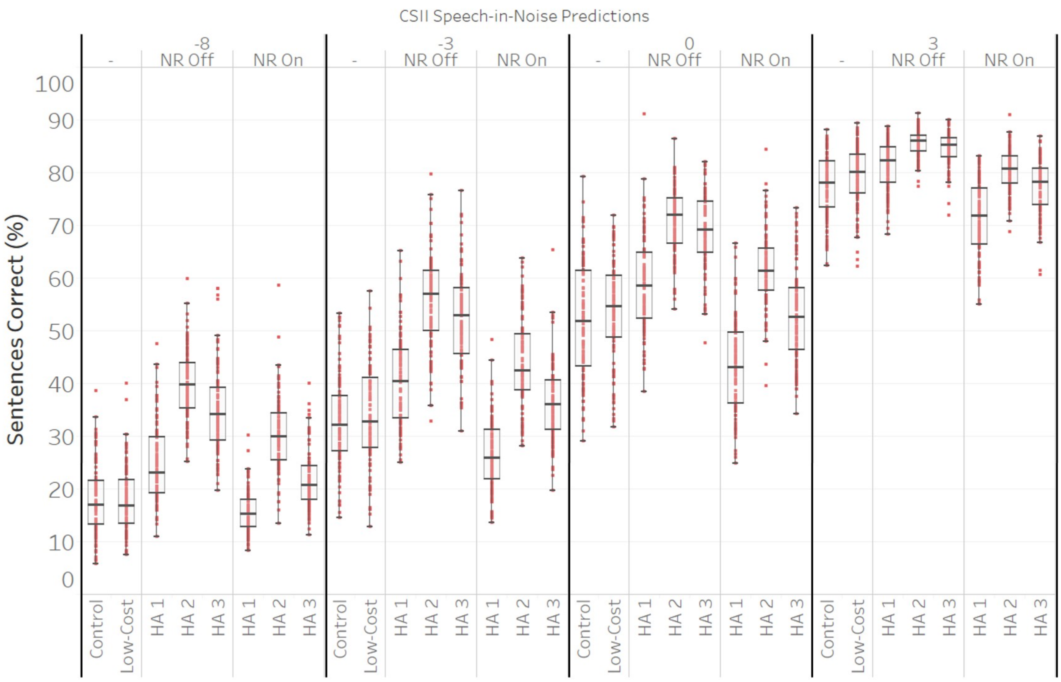


FIGURE 5.3: Box plots of CSII predictions of speech intelligibility for the same processing conditions as those described in Section 4.1. Individual measurement points consisting of the average score for ten sentences, corresponding to those used in behavioural tests, are given as red dots.

.

Using the same analysis as for HASPI data above, no differences between conditions or groups of conditions within each SNR are statistically significant using either the STOI or CSII metric.

5.2.2 Comparison to Behavioural Results

Figures 5.4, 5.5, 5.6 and 5.7 combine the data from Figure 4.1 with that from Figures 5.1, 5.2 and 5.3 respectively in order to directly assess the accuracy of HASPI, CSII and STOI predictions against behavioural SIN scores. In order to create pairwise matches between the automated and behavioural scores, the mean has been taken over HASPI, CSII or STOI values for the ten sentences presented to each participant in each listening condition, such that one mean HASPI, CSII or STOI value (and one behavioural score) corresponds to one listening condition per participant.

Correlation between each metric and the behavioural SIN scores is good overall, with STOI achieving the smallest R.M.S. error of the three and HASPI also achieving a low R.M.S. error. STOI also exhibits trend lines in discrete SNRs which are closest to the expected $y = x$ line. CSII, however, has a much higher R.M.S. error than the other metrics and shows negative correlation with the behavioural SIN scores when analysed in discrete SNRs.

Seven in eight of the listening conditions are associated with a significantly higher predicted SIN score from HASPI compared to the behavioural SIN scores at an SNR of -8 dB. Similarly, in the cases of CSII and STOI, all of the predictions at -8 dB SNR are significantly higher than the behavioural results.

At every SNR, HASPI and CSII significantly overpredict the SI associated with the low-cost amplifier. Similar trends can be seen in the STOI predictions; however, these differences are not significant at higher SNRs.

Differences can also be seen between the behavioural and HASPI scores for the first and second HAs with noise reduction switched on at higher SNRs. Although the difference is not statistically significant in all cases, a general trend can still be seen across SNR; HASPI appears to overestimate the SI associated with the first HA's noise reduction program but underestimates any SI improvement due the second HA's noise reduction. Conversely, the STOI and CSII predictions appear to be more accurate in the cases where noise reduction is switched on compared to those where it is switched off, overpredicting the associated SI, particularly at lower SNRs. However, as is seen in HASPI, STOI also overpredicts SI associated with the first noise reduction condition at 0 dB SNR.

The subjective order of conditions, from best intelligibility to worst, at each tested SNR, are shown in the Figure 5.8. The order of the conditions as predicted by HASPI, STOI and CSII, relative to the subjective data, is also shown. Positive values in red indicate that the objective metric predicts that the condition in question should rank higher than subjective scores suggest, whereas a negative blue value indicates that the objective prediction for that condition ranks lower than subjective scores would suggest. Darker colours show a larger difference in the ranking.

Broadly, the order of the conditions predicted by STOI and HASPI (from best to worst predicted intelligibility, as seen in Figure 5.8) is similar to that shown by the SIN test results in Figure 4.1; noise reduction algorithms are predicted to perform better than first fit conditions at all SNRs. However, the relative improvement due to the cheap amplifier is generally overestimated by STOI and, to a lesser extent, HASPI also. The order predicted by CSII matches less well than the other metrics, with noise reduction algorithms predicted to reduce intelligibility compared to the first fit conditions in all cases.

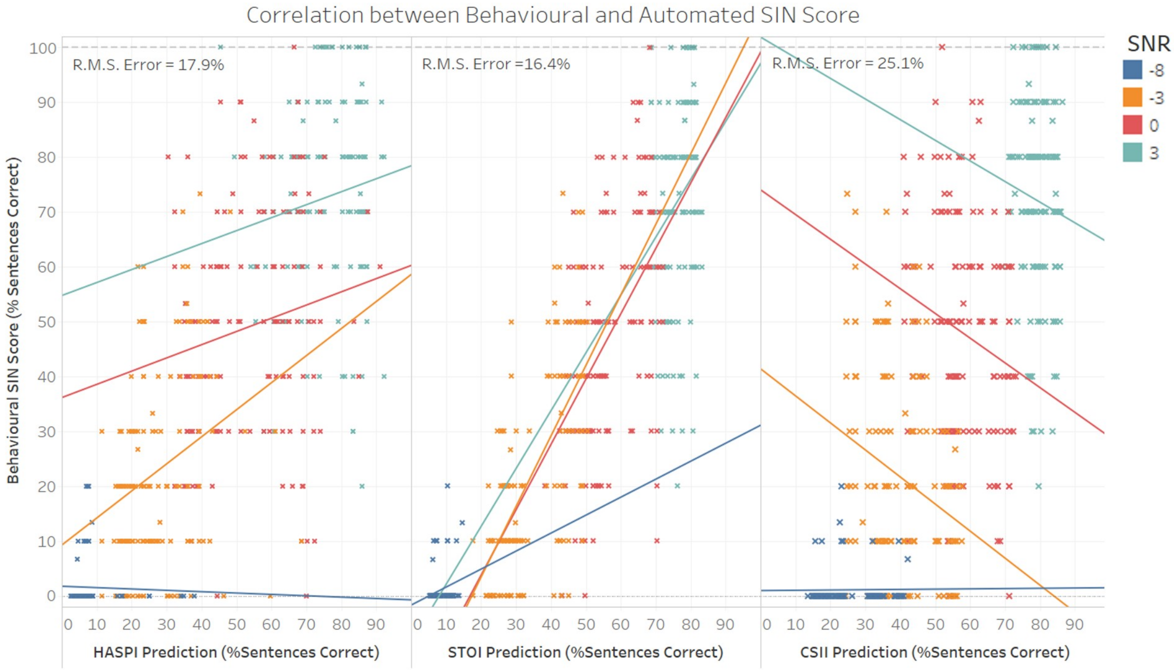


FIGURE 5.4: A comparison of behavioural SIN test results and varying automated predictions for speech intelligibility. The data in this figure is the same as that from Figures 4.1, 5.1, 5.2 and 5.3. Best-fit linear trend lines are given for each SNR, along with the overall R.M.S. error of prediction metric. The overall correlation coefficients for each metric are $\rho = 0.83$, 0.89 and 0.71 from left to right.

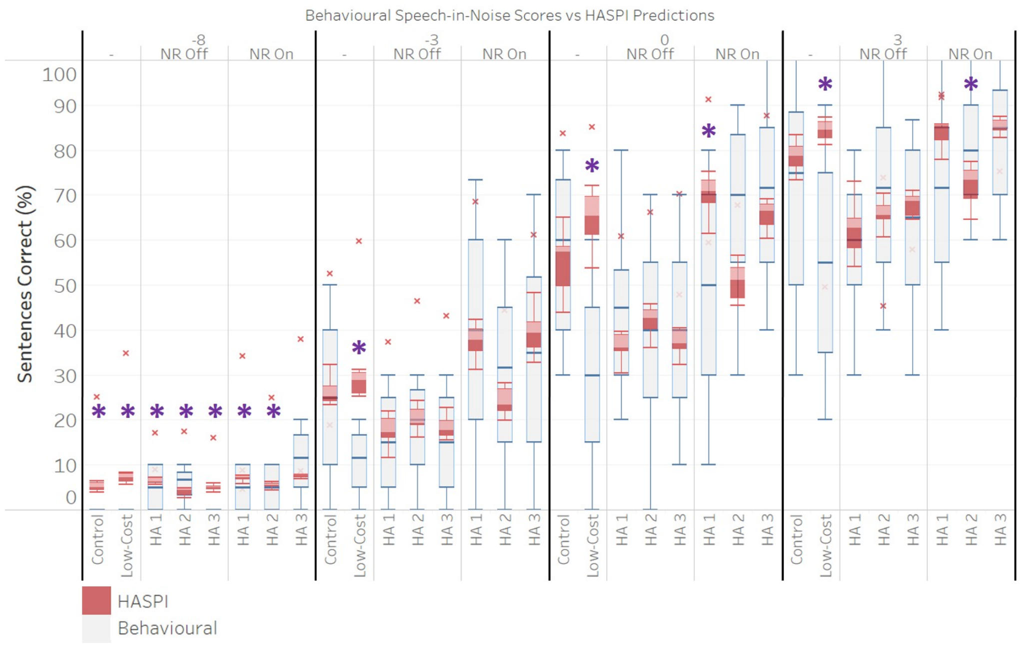


FIGURE 5.5: A comparison of behavioural SIN test results and HASPI predictions for speech intelligibility. The data in this figure is the same as that from Figures 4.1 and 5.1. Asterisks indicate conditions in which the HASPI predictions are significantly different from the behavioural scores ($p < 0.05$ or $p < 0.002$ with Bonferroni correction applied) according to Wilcoxon Signed Ranks tests for paired data.

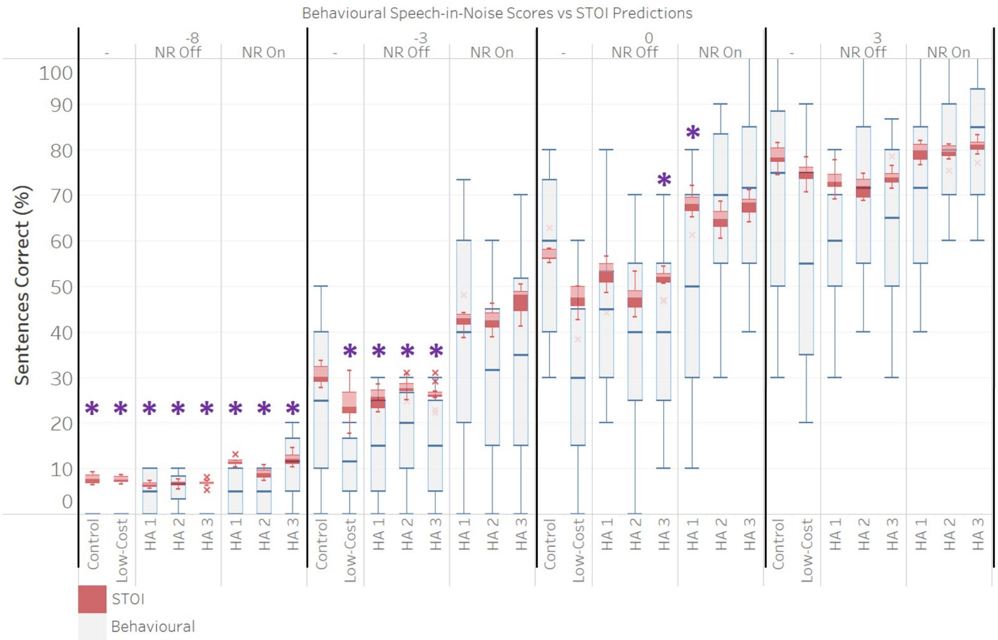


FIGURE 5.6: A comparison of behavioural SIN test results and STOI predictions for speech intelligibility. The data in this figure is the same as that from Figures 4.1 and 5.2. Asterisks indicate conditions in which the STOI predictions are significantly different from the behavioural scores ($p < 0.05$ or $p < 0.002$ with Bonferroni correction applied) according to Wilcoxon Signed Ranks tests for paired data.

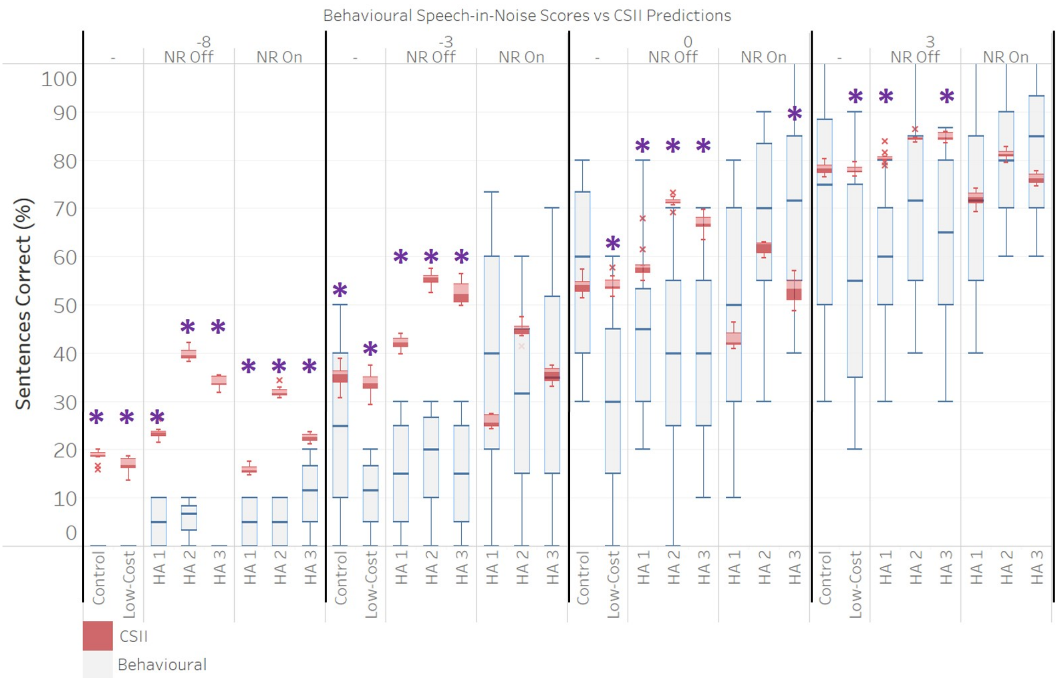


FIGURE 5.7: A comparison of behavioural SIN test results and CSII predictions for speech intelligibility. The data in this figure is the same as that from Figures 4.1 and 5.3. Asterisks indicate conditions in which the CSII predictions are significantly different from the behavioural scores ($p < 0.05$ or $p < 0.002$ with Bonferroni correction applied) according to Wilcoxon Signed Ranks tests for paired data.

Blue/-ve Predicted lower than subjective indicates
Red/+ve Predicted higher than subjective indicates

-8	Subjective	HASPI	STOI	CSII
HA1, NR	2	1	1	-6
HA2, NR	5	2	2	2
HA3, NR	1	-1	-1	-4
NoHA	8	4	4	2
HA1, FF	4	-4	-4	0
HA2, FF	3	-4	-4	2
HA3, FF	8	2	2	6
Cheap	8	3	3	1

-3	Subjective	HASPI	STOI	CSII
HA1, NR	1	0	0	-7
HA2, NR	2	-2	-1	-1
HA3, NR	3	1	1	-2
NoHA	4	1	0	-3
HA1, FF	6	-1	-1	2
HA2, FF	7	-1	-1	6
HA3, FF	5	-1	0	3
Cheap	8	3	2	2

0	Subjective	HASPI	STOI	CSII
HA1, NR	4	3	3	-4
HA2, NR	2	-3	-1	-1
HA3, NR	1	-1	-1	-6
NoHA	3	-1	-1	-3
HA1, FF	6	-1	0	2
HA2, FF	5	-3	-2	4
HA3, FF	7	1	2	5
Cheap	8	5	0	3

3	Subjective	HASPI	STOI	CSII
HA1, NR	4	3	1	-4
HA2, NR	1	-3	-1	-5
HA3, NR	3	1	2	-4
NoHA	2	-1	-2	-2
HA1, FF	6	-1	-1	3
HA2, FF	5	-3	-3	4
HA3, FF	7	2	1	5
Cheap	8	2	3	3

FIGURE 5.8: Tables of order of hearing aid conditions as determined by behavioural SIN scores, and the difference in ranking compared to that from behavioural tests as predicted by HASPI, STOI and CSII. A separate table is provided for each SNR. The conditions with no noise reduction (the 'first fit' configuration) are denoted using 'FF', and the conditions with noise reduction enabled are labelled 'NR'.

5.2.3 Predictions for Hearing Impaired Listeners

HASPI is the only current metric with which it is possible to make predictions for SI for HI listeners. These predictions, based on the audiogram shown in Figure 3.3 are shown in Figure 5.9. The HI predictions have been overlaid with the equivalent data using NH HASPI predictions for comparison.

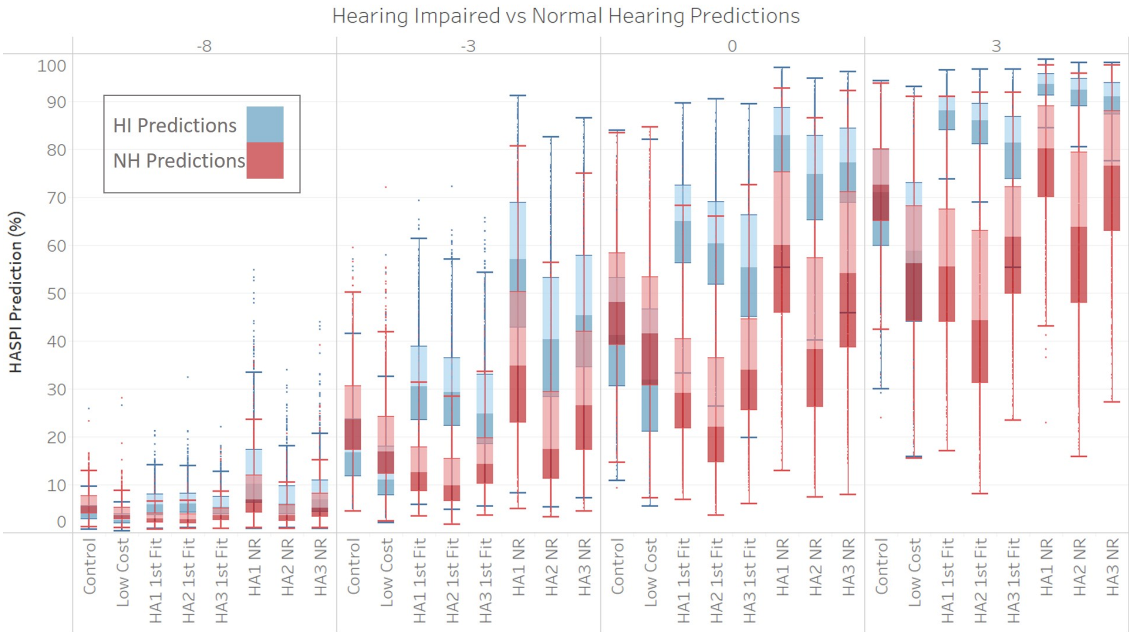


FIGURE 5.9: HASPI SI predictions for subjects with a hearing impairment shown in Figure 3.3, overlaid with predictions for the same conditions for normal hearing subjects.

In all cases, the HASPI prediction SI associated with the National Health Service (NHS) HAs is improved compared to the control condition for hearing impaired listeners, with further improvements seen when noise reduction is enabled, as expected. The SI associated with the control and low-cost amplifying conditions are similar or higher for NH subjects than for listeners with hearing impairment, also in line with expected results.

However, in all cases involving NHS HA models, the SI capabilites of HI listeners is predicted by HASPI to be higher than those predicted for NH listeners; this is very unlikely to be the case and is not a result to be expected.

5.3 Discussion

In order for an automated metric to be considered satisfactory, the predictions should:

- Have a low R.M.S. error compared to and have a high correlation with behavioural results.
- Show statistically significant differences between conditions where differences can be seen in behavioural results (see Figure 4.1).

- Rank data in a similar order.

Of the three metrics tested, both STOI and HASPI display promising initial results in terms of both correlation (overall and in discrete SNRs) and R.M.S. error with behavioural SIN scores, as seen in Figure 5.4. However, CSII seems unlikely to yield promising results due to the higher R.M.S. error, low overall correlation and negative correlation at discrete SNRs.

The only metric to identify statistically significant differences which exist in the behavioural data is HASPI - neither STOI or CSII detect any significant differences between any conditions or groups at all, as described in Section 5.2.1. However, these differences occured so frequently they can be difficult to interpret. It is not possible to decipher whether any significant differences detected by HASPI which are not apparent in the behavioural data are incorrect or simply not detectable in the small sample taken as part of the behavioural experiment. In addition, quantisation error, which occurs in the behavioural data due to the small number of sentences presented per condition, could minimise the possibility of detecting differences at the level of those seen in the HASPI predictions.

When looking at the ranking tables in Figure 5.8, it is easy to see at a glance that the order of conditions from most to least intelligible is reflected poorly at all SNRs by CSII. In contrast, the behavioural ranking is matched well in predictions from both HASPI and STOI with statistically similar conditions (i.e., all first fit configurations, all noise reduction conditions) being ranked similarly. This is particularly notable at 0 dB SNR, with seven of eight conditions being accurately ranked to within two places for HASPI (all conditions in the case of STOI). However, both metrics overpredicted the intelligibility ranking associated with the low-cost amplifier by at least two places at 0 dB and at higher SNRs, this trend continues; HASPI overpredicts the speech intelligibility associated with the low-cost amplifier by up to five places, predicting it as the third best condition for SI at 0 dB SNR compared to rankings from behavioural tests which suggest it may be the worst condition for SI.

In all cases involving NHS HA models, the SI capabilites of HI listeners is predicted to be higher than those of NH listeners. Since this is very unlikely to be the case in any situation, these results indicate that the adaptions made to HASPI to account for hearing impairment require substantial re-evaluation, at least for the mild impairment considered here.

In summary, both STOI and HASPI appear to show considerable merit in the prediction of SI for noisy, HA processed speech. However, the SI in particular conditions (primarily that of the low-cost amplifying device) are much less well predicted by these metrics than would be adequate for the metrics to be considered for use in ranking and comparing hearing aid models or programs. The primary aim of the following chapter is therefore to dig a little deeper into the possible reasons for poor prediction of particular conditions and explore some possible adjustments to the methods which may be able to improve the accuracy of predictions in the more problematic conditions.

Chapter 6

Development of and Improvements to HASPI as an Automated Metric for Speech Intelligibility Prediction

6.1 Speech Intelligibility Correlate Mapping

The main aim of the this chapter is to provide the tools such that a metric can be developed for which high overall correlation and low RMS error with behavioural speech in noise (SIN) results is maintained whilst predictions of speech intelligibility (SI) for noise reduction conditions and the low-cost amplifying condition are improved.

6.1.1 Retraining Logistic Function

The most obvious way to improve the correlation and reduce the error between the behavioural and predicted SIN scores for a particular dataset is to adjust the coefficients used in mapping the intelligibility correlates to the SIN scores (Equation 5.1 seen in Section 5.1.4). This is because changes to the participant speech-in-noise testing procedure, recording process, speech and noise content and types of processing may result in a slightly different relationship between the automated output parameters and the behavioural results.

Using an iterative method to find coefficients which minimise the overall mean-squared error between behavioural SIN scores and the SIN predictions yields an improvement in both the mean-squared error and the overall correlation coefficient with the behavioural scores in the cases of Hearing Aid Speech Perception Index (HASPI), Short-Time Objective Intelligibility (STOI) and Coherence Speech Intelligibility Index (CSII), as shown in Figure 6.1.

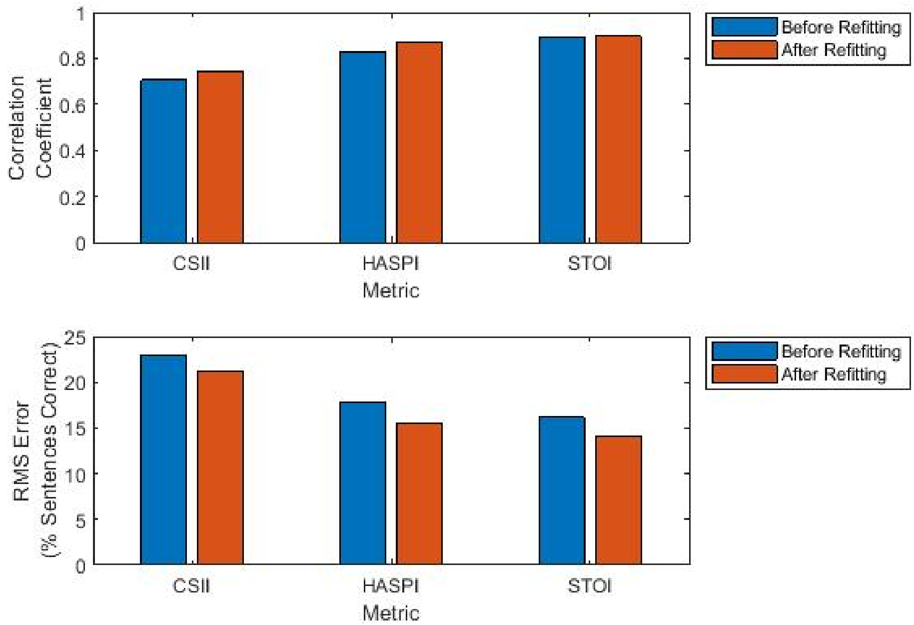


FIGURE 6.1: A bar graph showing the improvement in Pearson correlation coefficient and RMS error before and after remapping the components of each objective metric to fit the current data.

The improvements to the correlation coefficients between the behavioural scores and the demonstrated metrics are all small; the maximum increase in the correlation coefficient is 0.04, seen in HASPI. The reduction in the R.M.S. error between the behavioural scores and automated predictions in each case is also small at 2-3% sentences correctly recalled. The changes to individual coefficients reflect this - in most cases changes to the weighting coefficients used in the prediction calculation are small, as seen in Table 6.1.

One notable exception to this rule is the weightings of the low and mid-intensity coherence features in HASPI, which both change substantially with opposite signs but very similar absolute values. However, the change to these weightings has little impact on the overall HASPI value since the calculated values for low and mid-intensity coherence are highly correlated (as seen in Figure 6.2) and as such, the almost equal and opposite weightings have the effect of cancelling out both measures. Other weightings for HASPI

Parameter	HASPI		CSII		STOI	
	Original	Refitted	Original	Refitted	Original	Refitted
Offset	-9.05	-6.71	-2.62	-3.37	9.36	11.8
d	-	-	-	-	-13.5	-16.4
Low-Level Coherence	0	-43.0	0	-0.471	-	-
Mid-Level Coherence	0	43.1	9.26	4.00	-	-
High-Level Coherence	4.62	0.912	0.470	5.56	-	-
Cepstral Correlation	14.8	10.5	-	-	-	-

TABLE 6.1: A table of the original weighting coefficients for each of the components of HASPI, CSII and STOI compared with the weightings when refitted to behavioural data collected as described in Section 4.1.

and those for STOI change very little with refitting. Balance in weighting of CSII shifts a little from mid-level to high-level coherence, but, in a similar way to HASPI, since these features are so highly correlated (see Figure 6.2), little difference is seen in the overall accuracy of the metric.

6.1.2 Component Correlations

Figure 6.2 demonstrates the relationships between HASPI, its four individual components (low-, mid- and high-level coherence and cepstral correlation, described in Section 5.1) and the behavioural scores gained from testing in Section 4.1.

As expected, the high weighting of the cepstral correlation component in HASPI is immediately evident in the fourth plot on the top line, with a clear logistic pattern and high correlation and rank coefficients. High-level coherence, the only coherence component to have a non-zero weighting in the final calculation of HASPI, also displays a high correlation coefficient with HASPI and logistic shape. The rank correlation coefficient, however, is much lower between HASPI and high-level coherence compared to cepstral correlation.

Interestingly, both the Pearson correlation and Spearman's Rank correlation coefficients between the behavioural scores and HASPI are only slightly improved compared to the correlations between the behavioural scores and the cepstral correlation alone. A similar but less clear relationship is also seen between the high-level coherence and the behavioural scores. The correlation between the cepstral correlation and the high-level coherence is very high; this suggests that only one of these two features may be necessary for predicting SI for this particular data set, but it is unclear as to whether this is the case for other data sets.

Correlation, particularly rank correlation, between the cepstral correlation feature and the low-level coherence is lower than the correlation between cepstral correlation

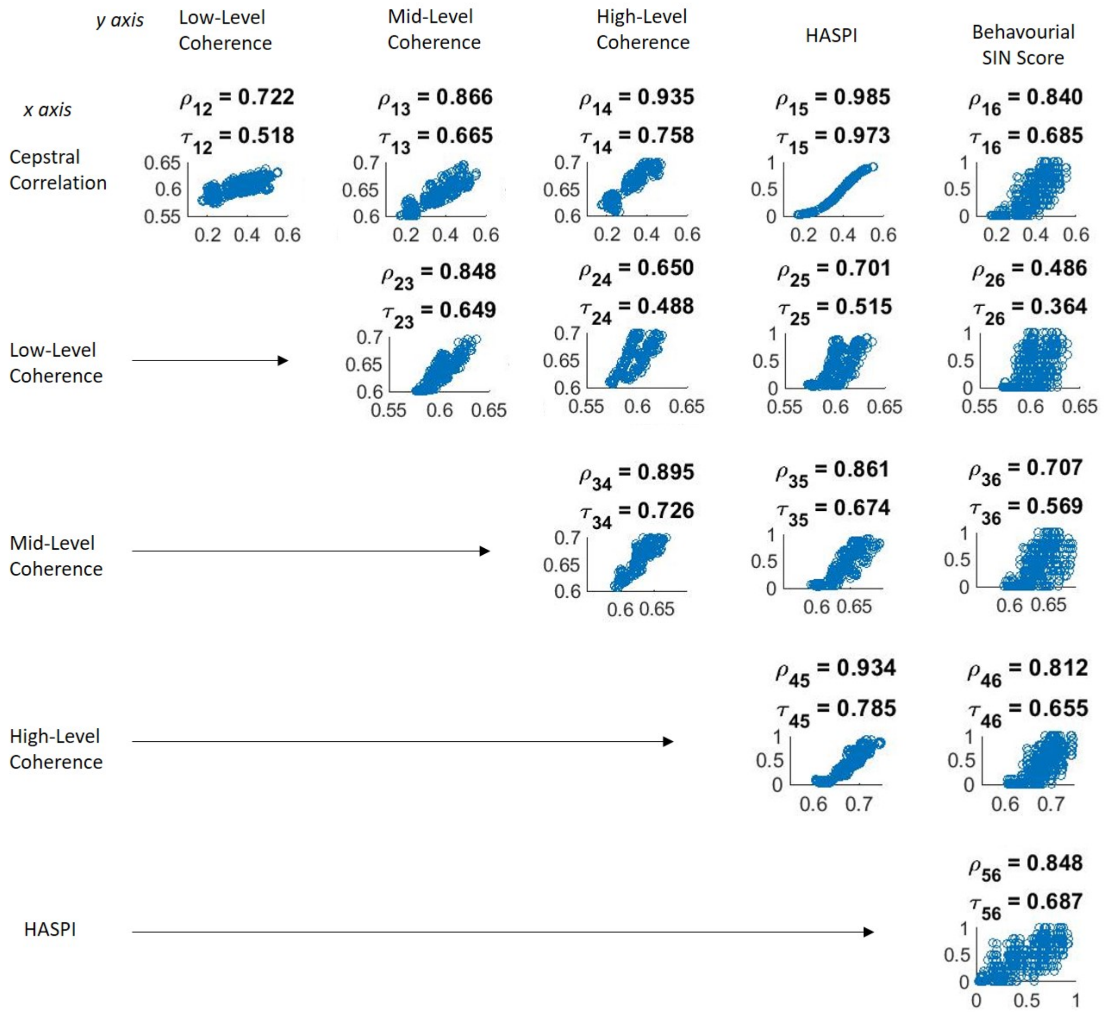


FIGURE 6.2: The relationships between the various component features used in HASPI, as well as the corresponding behavioural SIN test scores from the experiment detailed in Section 4.1. The Pearson correlation coefficient, ρ, and the Spearman's Rank coefficient, τ, are given between each pair of components.

and any of the other features, so it may be of interest as an additional feature for predicting SI in combination with cepstral correlation. However, correlation between the behavioural SI and the low-level coherence is particularly poor, and as such the low-level coherence is unlikely to be a good feature to use to predict behavioural SI here.

6.1.3 Machine Learning Approaches with Current Features

Several machine learning techniques were used to remap and combine the features currently available from HASPI and STOI to behavioural SIN scores with a view of improving correlation and reducing error between automated predictions and behavioural

SIN scores. These included:

- Linear regression
- Logistic regression
- Ensemble learning
- Decision tree
- Support Vector Machine Regression (SVMR)
- Neural network

The initial steps for all machine learning approaches were the same. The data was formatted in a matrix, X, with each row i representing one hearing aid condition per signal-to-noise ratio (SNR) per participant (8 x 4 x 21 = 672 rows). Each column represents one feature for prediction; initially, these features were: cepstral correlation, low-, mid- and high-level coherence, and STOI intelligibility correlate, d. For the later methods (ensemble, decision tree, neural networks and SVMR), each of the features was then normalised such that the distribution has a mean of 0 and a standard deviation of one; normalising features ensures that magnitude and range differences between features are removed and so all features are treated equally when determining weights. After feature normalisation (if applicable), an additional column of ones was added to enable calculation of an offset. A column vector y of equal length to X contained the corresponding behavioural SIN test scores. The relationship between the features and the expected result of a behavioural SIN test can then be written as Equation 6.1. In Equation 6.1, the feature coefficients are represented by a vector θ and any random (non-systematic) error in the results is given by e.

$$y_i = f(X_i, \theta) + e_i \tag{6.1}$$

6.1.3.1 Dimension Reduction

Since some of the correlations between HASPI features and behavioural SIN scores were low (see Figure 6.2), dimension reduction methods were considered to remove features with low importance in predictions before machine learning methods were applied. Dimension reduction can simplify and speed up the training process for machine learning algorithms and can provide clarity to the researcher on which parameters are likely to be good predictors, which have little effect and which may be duplicating one another.

Feature importance was determined using two different methods: Univariate feature ranking for regression using F-tests and feature selection using neighborhood component analysis for regression using the MATLAB functions fsrftest and fsrnca respectively (The MathWorks Inc., 2021), shown in Figure 6.3.

Since the fsrnca function incorporates regularization, the features were normalised before processing. For feature selection only, the regularisation parameter was tuned such that the loss over the dataset with all features was minimised. Normalisation of features does not affect univariate feature ranking using F-tests.

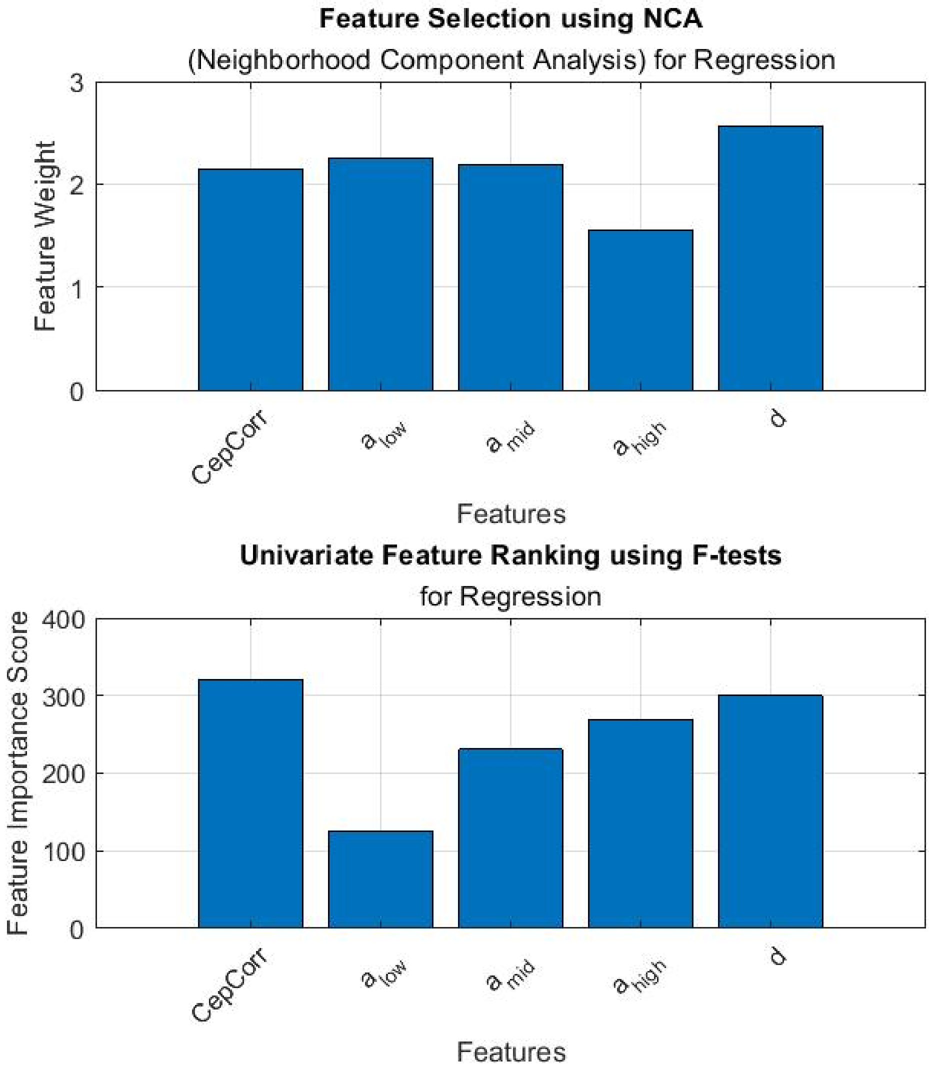


FIGURE 6.3: Feature importance for predicting behavioural SIN scores using two methods: Univariate feature ranking for regression using F-tests and feature selection using neighborhood component analysis for regression. The features analysed are the cepstral correlation, low-, mid- and high-level coherence from HASPI and the SI correlate from STOI.

Both methods highlight cepstral correlation as important for predicting the behavioural SIN score, as is reflected by its high weighting in HASPI and high correlation in Figure 6.2. Similarly, the STOI covariate is also indicated as an important feature by both methods of feature selection. However, the importance of each coherence feature differs between the two methods of feature selection. Neighbourhood Component Analysis identifies all features as relevant to prediction of behavioural SIN score, with high-level coherence being the least important of the three coherence measures and low-level coherence being the most important. The univariate feature ranking method shows the opposite. Since the correlations between behavioural SIN score and both mid- and low-level coherence measures are low (see Figure 6.2), the results from the univariate method with F-tests (prioritising cepstral correlation and high-level coherence as well as the SI parameter from STOI) have been taken forward for use in machine learning techniques.

6.1.3.2 Linear and Logistic Regression

Following the outcome of the feature reduction analysis, regression was used to refit the new subset of SI correlate features to the behavioural SIN scores. Linear and logistic regression methods were used, as shown in Figures 6.4 and 6.5.

The linear regression mapping was performed using a basic iterative least-squares approach in the same manor as that used previously in Section 6.1.1 when retraining the features of individual metrics.

The logistic regression approach finds the local minimum of the cost function, J (given by Equation 6.2) using gradient descent. By minimising the cost function, J, optimum combination of feature coefficients can be found such that the error between the predicted and behavioural SIN scores is as small as possible. For additional information on how the cost function is derived and used, see Ng (2020). The constant m is equal to the length of the vector of behavioural test scores, y, as described earlier by Equation 6.1. The matrix $\boldsymbol{X}$ and vector θ, as in Equation 6.1, represent the matrix of predictive features (a constant term, cepstral correlation, high-level coherence and the feature from STOI) and their coefficients, respectively. The sigmoid function, h, is identical to that given in Equation 2.5. λ, the regularisation coefficient, was set to zero, and the coefficients θ were initialised using those found to be optimal by linear regression.

$$J = -\frac{1}{m}\left[\vec{y}^T \ln(h) + (1-\vec{y})^T \ln(1-h)\right] + \frac{\lambda}{2m}\sum_{k=2}^{m}\theta_k^2; \tag{6.2}$$

$$h = \frac{1}{1 + \exp(-\boldsymbol{X}\vec{\theta})}$$

The results of both of these methods are shown in Figures 6.4 and 6.5.

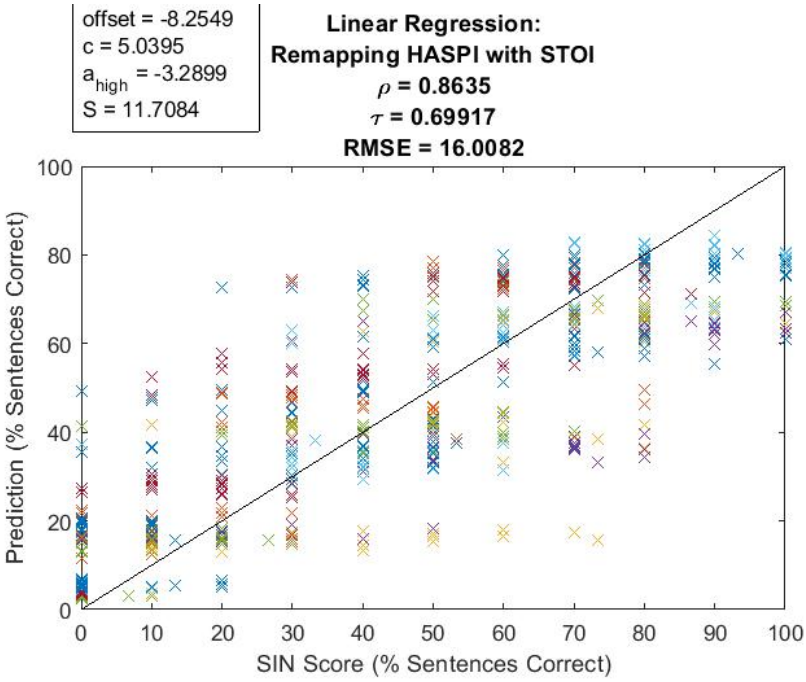


Figure 6.4: Linear regression using the most important features indicated by feature reduction analysis - cepstral correlation, high-level coherence and the STOI SI correlate.

6.1.3.3 More Complex Machine Learning Methods

A number of more complex, 'black-box' type machine learning methods were also tested to improve correlation and reduce RMS error between predicted and observed SIN scores, including decision trees, neural networks, SVMR and ensemble learning. Machine learning methods require fewer user-defined limitations and therefore provide a much more flexible framework for finding relationships between datasets. These methods can thus be used to detect patterns which may not easily translate into traditional mathematical relationships, for example complex non-linear relationships. Examples of the best outcomes of each of these are shown in Figures 6.6 to 6.9.

The decision tree, neural network and SVMR methods didn't produce results with high correlation or low RMS error compared to the simpler linear regression methods. The ensemble learning approach, however, did give promising results with high Pearson correlation, high Spearman's rank correlation and low RMS error. As such, ensemble learning methods were investigated further.

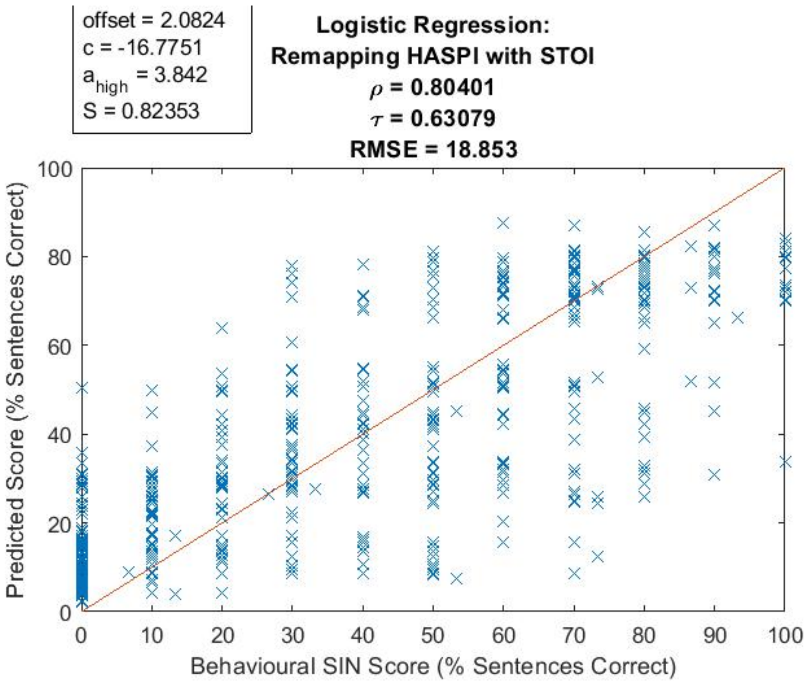


FIGURE 6.5: Logistic regression using the most important features indicated by feature reduction analysis - cepstral correlation, high-level coherence and the STOI SI correlate.

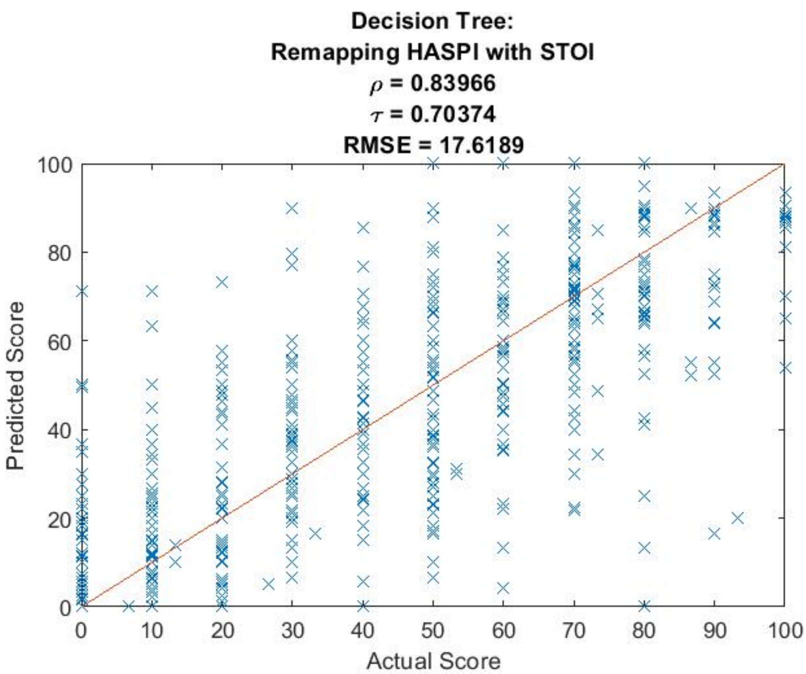


FIGURE 6.6: An illustrative example of prediction of behavioural SIN score using cepstral correlation, high-level coherence and the STOI SI correlate using decision trees.

To further test the efficacy and potential of the ensemble learning method, the data was split into eight sets - one per hearing aid condition. In order to assess whether a

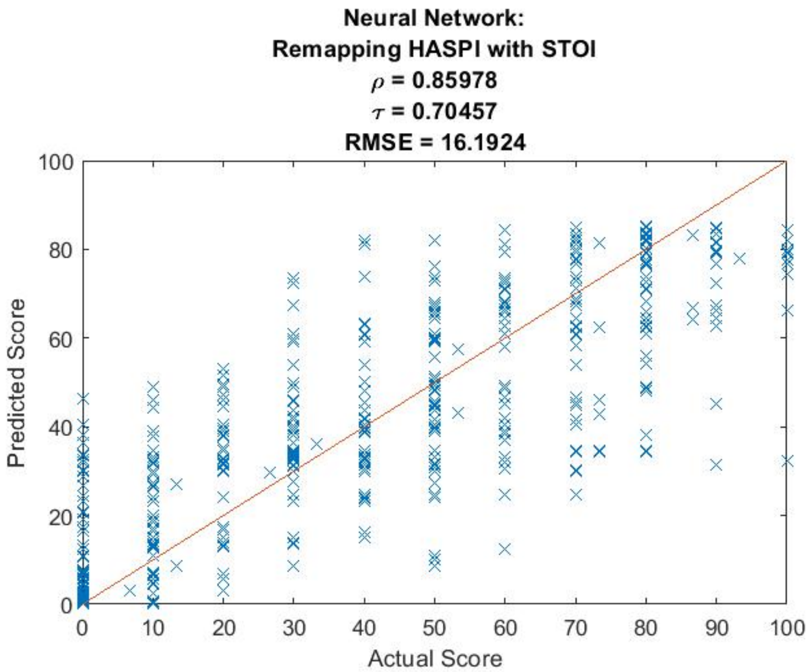


FIGURE 6.7: An illustrative example of prediction of behavioural SIN score using cepstral correlation, high-level coherence and the STOI SI correlate using neural networks. This particular example had ten hidden networks, each of size 10.

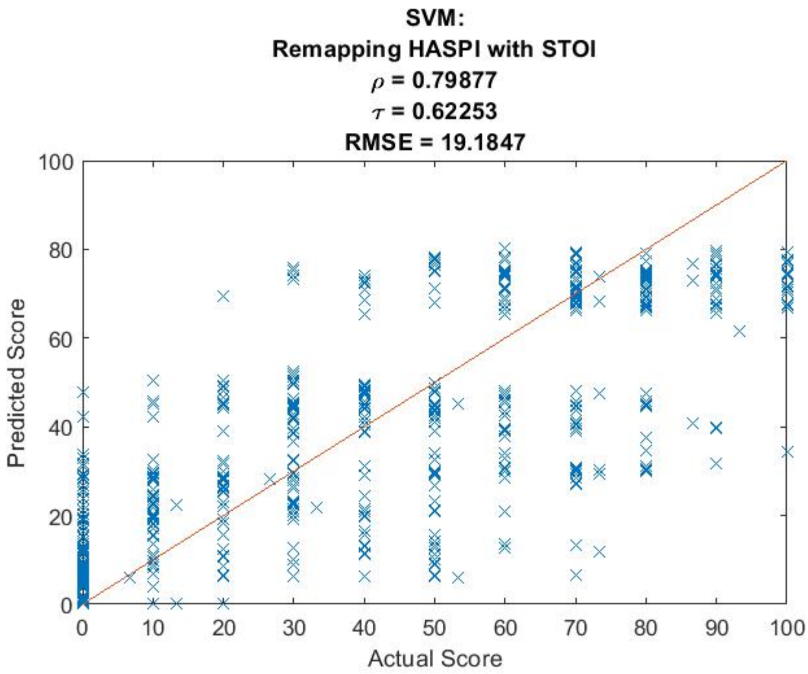


FIGURE 6.8: An illustrative example of prediction of behavioural SIN score using cepstral correlation, high-level coherence and the STOI SI correlate using SVMR.

trained model can generate good predictions for unseen conditions (e.g. a new hearing

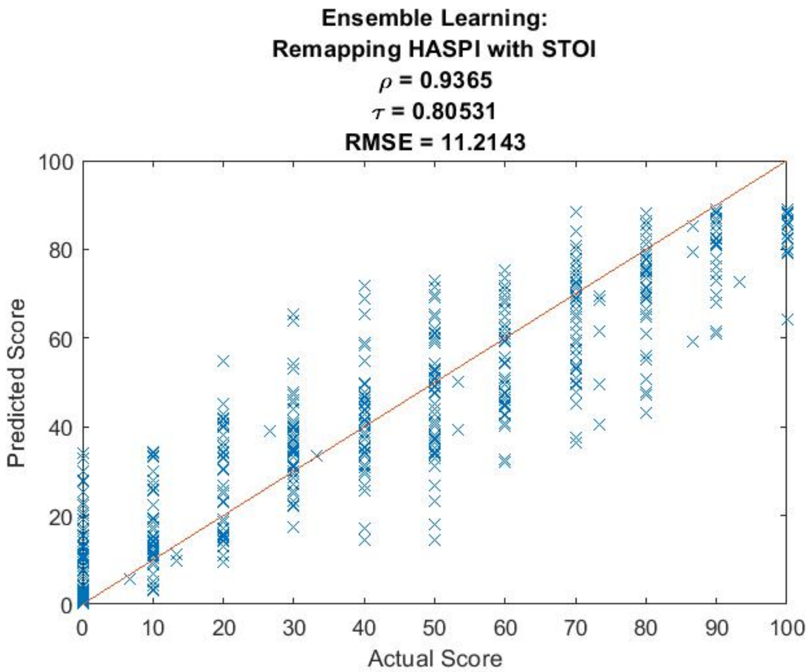


FIGURE 6.9: An illustrative example of prediction of behavioural SIN score using cepstral correlation, high-level coherence and the STOI SI correlate using ensemble learning.

aid or noise reduction algorithm), each model was trained on data from seven of the conditions at a time, and tested on the eighth. The results are shown in Figure 6.10.

Using ensemble learning, prediction of SI associated with untested conditions such as the low-cost amplifier appear to be much more consistent, with a high correlation, low RMS error and less skewed distribution than those produced by the original HASPI and STOI algorithms. However, there are still some conditions which appear to be difficult to predict using the existing features, for example, the noise reduction conditions. This can be seen most clearly in the case of the third hearing aid, where several outliers in the test set can be seen and the predicted scores are much lower than the equivalent behavioural SIN scores for the majority of cases.

6.1.4 Discussion and Conclusions

The key insights discussed in Section 6.1 are summarised below:

- Retraining of coefficients for individual metrics (HASPI, CSII and STOI) to match the behavioural SIN scores for current dataset does not substantially improve correlation or reduce RMS error in the differences between predicted and behavioural scores.

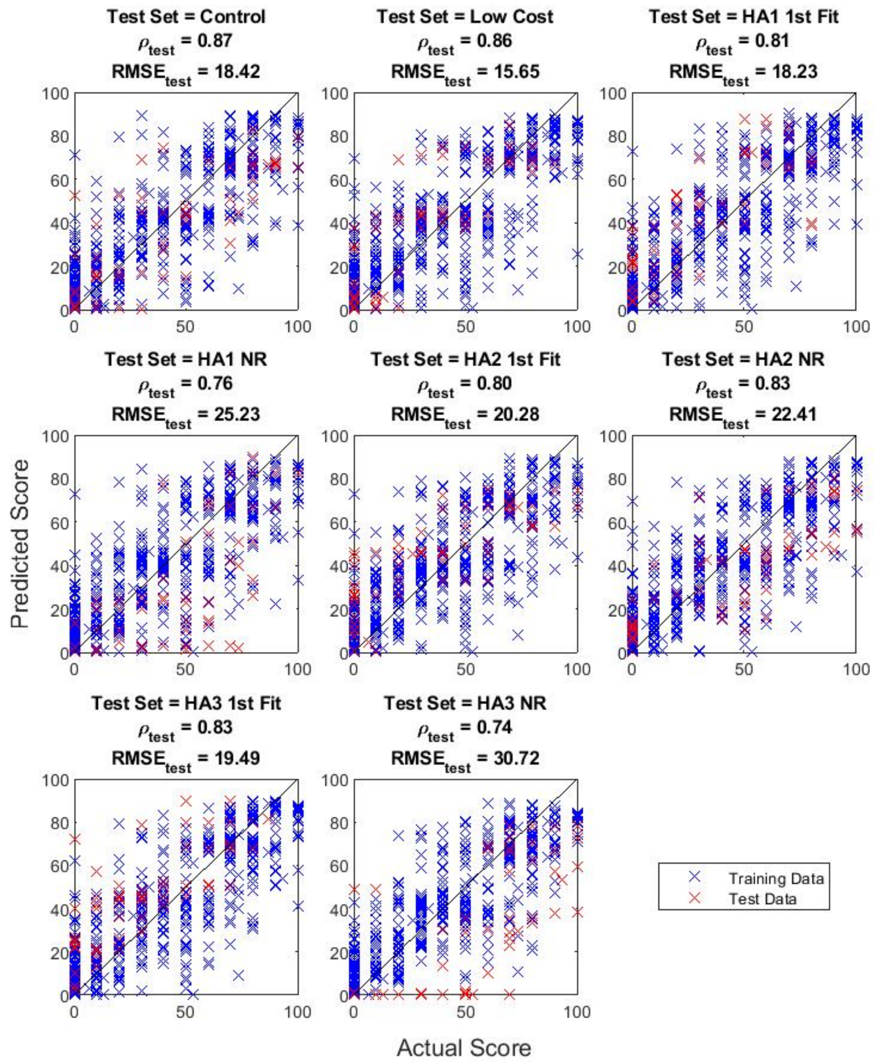


Figure 6.10: Prediction of behavioural SIN score using cepstral correlation, high-level coherence and the STOI SI correlate using ensemble learning, trained using seven hearing aid conditions and tested on the eighth.

- Component correlations and univariate feature ranking methods indicate that cepstral correlation, high-level coherence and the STOI SI correlate are important for behavioural SI prediction, consistent with previous findings (Kates and Arehart, 2014). Low- and mid-level coherence do not correlate strongly with behavioural SIN score.

- Neighbourhood component analysis with regularisation for minimum loss indicates that all features used have high importance in model prediction, with high-level coherence being the least important in predicting behavioural SIN scores.

- However, improvements to correlation coefficients and RMS error between predictions and behavioural SIN scores were minimal when the STOI SI correlate was introduced as an additional feature to those used in HASPI (cepstral correlation and high-level coherence).
- No improvements were seen using logistic regression, decision tree, SVMR or neural network methods to predict behavioural speech-in-noise scores over simple linear regression.
- Ensemble learning appears to substantially increase correlation and reduce RMS error in the original fitted predictions of behavioural SIN scores.

Whilst the ensemble learning method showed promise, machine learning methods like these are not as transparent as more traditional methods such as regression, where the relationship between input features and outputs can be clearly expressed in mathematical form. This attribute makes machine learning models much less easy, perhaps not even possible, to unpick into individual components and are therefore much more difficult to analyse to suggest improvements in the future.

The ensemble learning approach appears to work less well when predicting unseen behavioural SIN scores for particular hearing aid (HA) processing conditions, particularly two of the three noise reduction conditions. This may suggest that some aspects of the noisy speech signals which affect SI are not adequately represented in the features discussed in this thesis and further work is necessary identify additional features which can provide the information necessary for accurate prediction of the effects of these types of signal processing on SI.

Preliminary testing not shown here indicated that fitting the same parameters based on percentage of keywords correctly recalled rather than full sentences increases the performance of all methods tested. However, since words were not tested in isolation, effects relating to contextual cues and other cognitive factors are difficult to determine. For this reason, in conjunction with those discussed in Section 2.2.2 and to maintain continuity with previous research in Kates and Arehart (2014) and Taal et al. (2010) among others, scores for sentences rather than words correct have been retained.

6.2 Further Investigative Tests

6.2.1 Variations in HASPI with Changes to the Stimulus

In order to investigate how various characteristics of the noisy signal, such as speech corpus, speaker and noise type, effect the HASPI SI calculation, a number of recordings were made using the same setup as that described in Chapter 3.

A total of 15,600 samples of noisy speech were collected as part of this investigation. The first ten sentences were taken from the Institute of Electrical and Electronics Engineers (IEEE), Bamford-Kowal-Bench (BKB) and Institute of Hearing Research Sentence List (IHRSL) corpora, in addition to a subset of ten matrix sentences such that every possible word in the matrix is used once. The IEEE sentences are spoken by a male voice, whereas the IHRSL and matrix sentences are spoken in a female voice. In the case of the BKB sentences, recordings of both male and female voices were used, increasing the total number of clean sentences to be recorded to fifty. Next, a number of noise types were constructed. Each of the speech subsets described above was used to create iterations of matched Speech-Shaped Noise (SSN), and, with the exception of the matrix sentences, four- and six-talker babble, giving a total of thirteen types of additive noise. Each noise implementation was added to the sentences at SNRs of -3, 0 and 3 dB. Finally, the noisy sentences were recorded using the same eight HA conditions as those described in Chapter 3 - a control condition, the low-cost amplifier and three National Health Service (NHS) HAs with both standard and noise-reduction settings implemented.

6.2.1.1 Speech Corpora

Figure 6.11 shows the variation in the mean HASPI prediction of SI for each speech corpus (across all noise types) with SNR for the unaided control condition.

Immediately, it is easy to see from Figure 6.11 that HASPI predicts an increase in SI for male-voiced speech compared to female-voiced speech; this contradicts previous studies which suggest female voices may be more intelligible (Bradlow et al., 1996; Hazan and Markham, 2004), but is in support of some other studies (McCloy et al., 2015). It could be hypothesized that results might be skewed by the fact that there are more instances of noise based on female speech than on male speech; however, similar trends are seen when the samples which are combined to give an average for each corpus are limited to, for example, only those with the additive noise type being SSN from natural

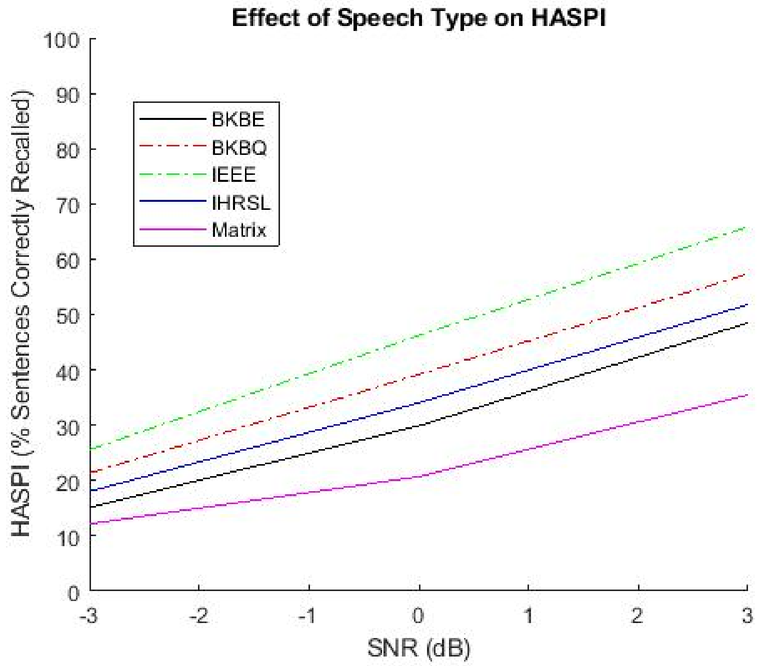


FIGURE 6.11: The mean across all additive noise types of the HASPI prediction of SI for each speech corpus using recordings of the unaided control condition only. Solid lines indicate corpora spoken by a female voice and dot-dashed lines represent male-voiced sentences.

voices (i.e. with matrix SSN removed, therefore balancing the number of samples with female- and male-voice-based noise types).

6.2.1.2 Noise Types

Figure 6.12 shows the variation in the mean HASPI prediction of SI for each noise type (across all speech corpora) with SNR for the unaided control condition.

From the bottom right subplot of Figure 6.12, it can be seen that the gender of the speaker used to create the noise samples does not appear to have a distinct effect on the HASPI SI prediction, in contrast to the trends seen with gender of the target sentences seen Figure 6.11. The sentence contents also appears not to affect the prediction of SI; the noise made with male-spoken BKB sentences has a much lower associated prediction of SI than the female-spoken equivalent.

The type of noise, however, does appear to have an effect on predicted SI. The bottom left subplot of Figure 6.12 indicates that babble noise has more of a detrimental impact on predicted SI than SSN; this result is similar to those seen in previous work such as Tang et al. (2018).

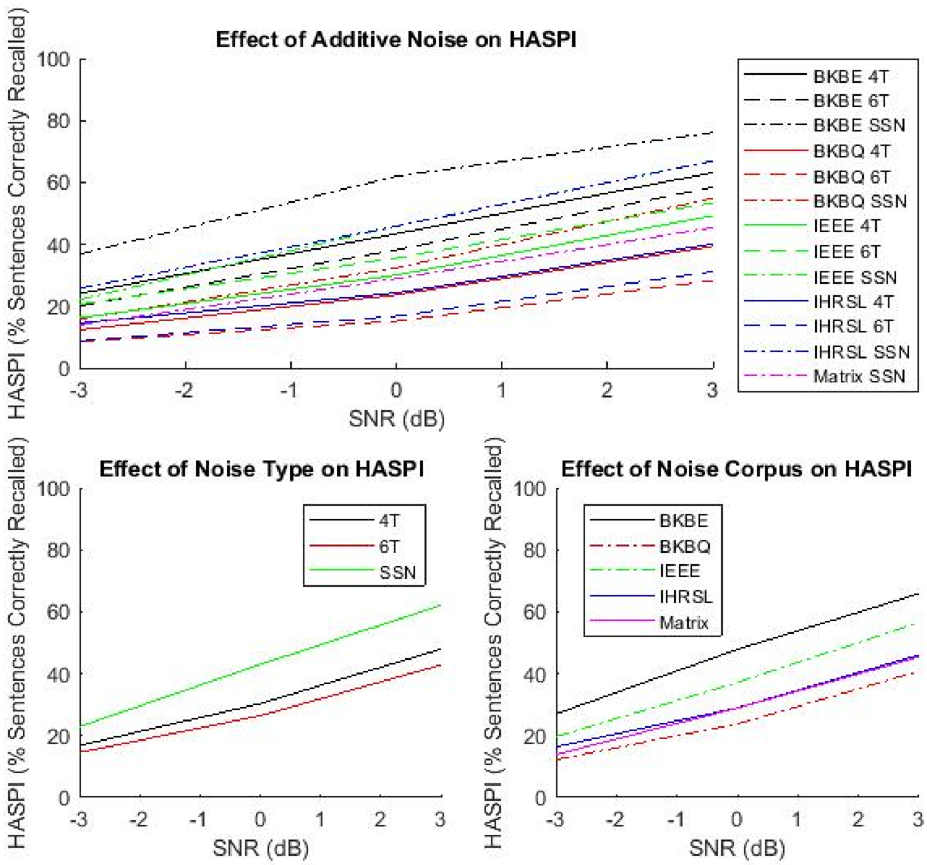


FIGURE 6.12: The mean across all speech types of the HASPI prediction of SI for each noise type and corpus using recordings of the unaided control condition only. In the upper plot, noise subtypes are grouped by corpus and type (4-talker babble (4T), 6-talker babble (6T) and SSN). The lower plots group the noise by type only (left) and corpus only (right). Solid lines in the lower right plot indicate corpora spoken by a female voice and dot-dashed lines represent male-voiced sentences.

6.2.2 How Robust are the Components of HASPI to Small Signal Changes?

In order to assess the robustness of HASPI to small variations in the degraded signal, a simulation was conducted where a single sentence (from the original set of unprocessed anechoically recorded samples) was combined with 100 realisations of the same spectrum speech-shaped noise at an SNR of 0 dB. The STOI predictions, HASPI predictions and internal HASPI parameters (cepstral correlation and three-level covariance) were then calculated for each of the 100 resulting noisy sentences. Since these changes to the signal are perceptually very similar, the predicted intelligibility should vary little between iterations if the metric is robust.

The distributions of STOI and HASPI scores across the 100 samples were not found to be significantly different at the 5% level from a normal distribution using Lilliefors test of normality, skewness was between -0.2 and 0.1 indicating no significant skew, and kurtosis was slightly below zero in both cases, indicating very few outliers.

The predicted HASPI values were found to be inconsistent; 6% of iterations differed from the mean value by more than the clinical noticable difference of 15% given in Section 2.3.5. STOI proved considerably more robust to these small variations than HASPI, with a total range of 12% and no cases in the sample which differed from the mean by more than 7%.

Analysis of the components of HASPI reveals that the main source of variance is the cepstral correlation. This is perhaps unsurprising since the cepstral correlation has the highest weighting and therefore most influence on the final HASPI prediction. Deeper investigation into why this variation occurs may be key to improving SI predictions in the future.

Chapter 7

Conclusions and Recommendations for Further Work

The present work aims to analyse and review currently available metrics for automated prediction of speech intelligibility in relation to hearing-aid processed sound.

A review of the literature concerning hearing aid use in the United Kingdom, along with common hearing aid processing and the problems faced by hearing aid users, was first conducted. The 'gold standard' for assessment of speech intelligibility, behavioural speech-in-noise tests, were discussed in detail, followed by an examination of current automated metrics which exist which attempt to predict the results of the behavioural tests. An experiment was conducted in which a behavioural speech-in-noise test was carried out using recorded outputs from assorted hearing aids with multiple settings in varying levels of background noise. The results and recordings from this speech-in-noise test were used to assess the accuracy of predictions from a number of automated speech intelligibility prediction metrics. Following this experiment, further work was executed to attempt to improve upon the existing metrics and analyse existing shortcomings.

The primary objective of this work was to evaluate the accuracy of various automated speech intelligibility prediction metrics in indicating the outcome of speech-in-noise tests in a number of listening conditions, specifically in response to noisy, hearing-aid processed speech. Of the three automated metrics tested, only Hearing Aid Speech Perception Index (HASPI) was able to detect statistically significant differences between conditions which mirrored those seen in behavioural speech in noise (SIN) testing. However, HASPI could not accurately predict the behavioural SIN scores associated with particular listening conditions, such as a low-cost amplifying device and some noise reduction conditions at particular signal-to-noise ratios (SNRs). Various investigative methods were used to attempt to identify the source of the issues with HASPI including feature importance and usage of a variety of mapping and machine learning methods, the effects of changing stimulus types and the robustness of HASPI and Short-Time Objective Intelligibility (STOI)'s component features. Ensemble learning using the aforementioned component features was shown to produce promising results in prediction of behavioural speech-in-noise scores but further investigation is needed with a much larger dataset for conclusions to be drawn.

The ideal outcome of this and future work would be to determine how, if at all possible, an automated procedure might be implemented to give hearing aid manufacturers and consumers (e.g., the NHS) a good indication of speech intelligibility associated with hearing assistive devices without the need for long, costly participant trials in the early stages of testing, development and approval. Work presented in this thesis indicates that currently available metrics are not adequate for use as such a tool without substantial developments. Several recommendations for additional work which could help to further progress towards the ideal outcome are listed below:

1. Further investigations into the causes of large fluctuations in cepstral correlation in response to small signal changes, in order to locate issues with and improve robustness of HASPI to very small changes in the stimulus.
2. Additional research into supplementary features which could be used in speech intelligibility (SI) prediction, particularly in relation to conditions for which behavioural results differ significantly from predicted values (such as the low-cost amplifying device).
3. It has been shown that performance of automated metrics may be improved with remapping of features to behavioural SIN scores using ensemble learning, but more work is needed to fully assess the performance of such an approach.

Appendix A

Hearing Aid Specifications

Attached documentation consists of the following technical specification sheets:

1. Signia Contrast S+, obtained directly from manufacturer
2. Oticon Spirit Synergy, obtained directly from manufacturer
3. Danalogic Ambio 77
4. GlobalCareMarket/FoKe FK-162 amplifier

BTE13 85
OTICON SPIRIT SYNERGY

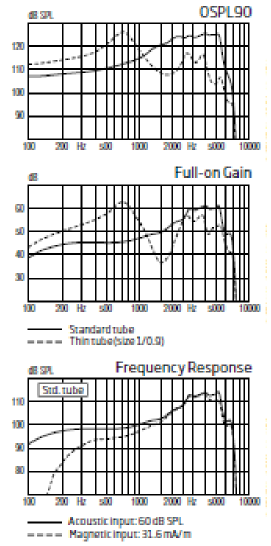
Oticon | Spirit Synergy

Scale 1:1

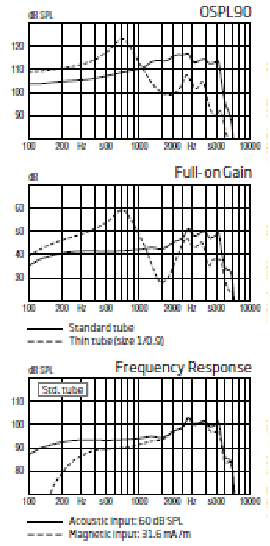
Technical information
Omnidirectional mode is used unless otherwise stated.

85

EAR SIMULATOR
Measured according to IEC 60118-0 (1983) and 60711 (1981) and DIN 45605.



2CC COUPLER
Measured according to ANSI S3.22 (2003) and S3.7 (1995), IEC 60118-7 (2005) and IEC 60318-5 (2006).



		Ear Simulator	2cc Coupler
OSPL90	Peak	126 (126*) dB SPL	117 (123*) dB SPL
	1600 Hz	121 (108*) dB SPL	114 (100*) dB SPL
	Average	116 (116*) dB SPL	113 (106*) dB SPL
Full-on gain	Peak	61 (63*) dB	51 (59*) dB
	1600 Hz	50 (36*) dB	43 (28*) dB
	Average	49 (52*) dB	44 (41*) dB
Reference test gain		43 dB	36 dB
Frequency range		100-7200 Hz	100-7000 Hz
Telecoil output (1600 Hz)	1 mA/m field	79 dB SPL	-
	10 mA/m field	99 dB SPL	-
	SPLITS L/R	-	94/94 dB SPL
Total harmonic distortion	500 Hz	<2 %	<2 %
(Input 70 dB SPL)	800 Hz	<2 %	<2 %
	1600 Hz	<2 %	<2 %
Equivalent input noise level (A)	Omni	23 dB SPL	18 dB SPL
	Dir	32 dB SPL	27 dB SPL
Battery consumption	Quiescent	1.1 mA	1.1 mA
	Typical	1.1 mA	1.1 mA

Battery life, calculated, hours**	240
Size 13 (IEC PR48)	
IRIL (IEC 60118-13-2011)	800/1400/2000 MHz: 24/48/45 dB SPL

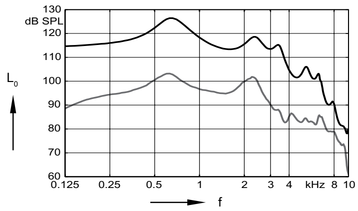
(*) For instruments fitted with Corda miniFit
**) Based on the standardised battery consumption measurement (IEC 60118-0.) The actual battery life depends on battery quality, use pattern, active feature set, hearing loss and sound environment

Contrast S+ | Technical Data

Type	Earhook damped		ThinTube	
				
Output sound pressure level	**2 ccm coupler**	**Ear simulator**	**2 ccm coupler**	**Ear simulator**
at 1.6 kHz	–	135 dB SPL	–	121 dB SPL
Peak	130 dB SPL	138 dB SPL	126 dB SPL	130 dB SPL
HFA-OSPL 90	127 dB SPL	–	116 dB SPL	–
Gain				
Full on gain (FOG) at 1.6 kHz	–	59 dB	–	54 dB
Full on gain (Peak)	60 dB	68 dB	53 dB	61 dB
HFA-FOG	53 dB	–	47 dB	–
Reference test gain	50 dB	52 dB	39 dB	46 dB
Frequency, noise and directivity				
Frequency range	110-7700 Hz	620-8100 Hz	100-8100 Hz	100-8100 Hz
Equivalent input noise	16 dB SPL	16 dB SPL	18 dB SPL	18 dB SPL
Total harmonic distortion at 500 / 800 / 1600 Hz	2 / 2 / 1 %	2 / 2 / 1 %	1 / 1 / 2 %	1 / 1 / 2 %
Tinnitus noiser broadband	70 dB SPL	–	70 dB SPL	–
AI-DI	4.0 dB		4.0 dB	
Inductive coil sensitivity				
MASL (1 mA/m) at 1.6 kHz	–	85 dB SPL	–	76 dB SPL
HFA MASL (1 mA/m)	80 dB SPL	–	69 dB SPL	–
HFA SPLITS (left/right)	109 / 109 dB SPL	–	98 / 98 dB SPL	–
RSETS (left/right)	-1 / -1 dB	–	-1 / -1 dB	–
Battery				
Battery voltage	1.3 V		1.3 V	
Battery current drain	1.0 mA		1.1 mA	
Battery life (cell zinc air)	~220 h		~200 h	
Battery life (rechargeable)	–		–	
IRIL IEC 118-13:2011 (bystander)				
800-960 MHz	<-43 dB SPL		<-43 dB SPL	
1400-2000 MHz	<-45 dB SPL		<-45 dB SPL	
ANSI C63.19	M4 / T4		M4 / T4	

Contrast S+ (ThinTube) | Basic Data

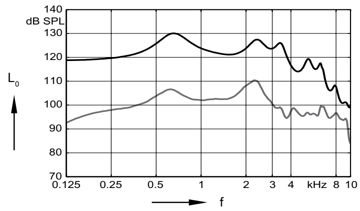
2 ccm coupler



Output sound pressure level (L_I = 90 dB)

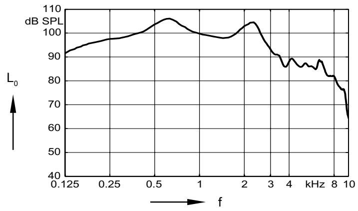
Full on gain (L_I = 50 dB)

Ear simulator

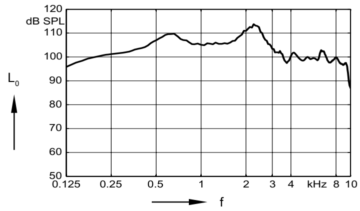


Output sound pressure level (L_I = 90 dB)

Full on gain (L_I = 50 dB)

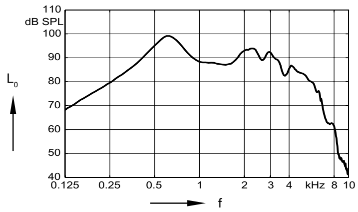


Frequency response (L_I = 60 dB)

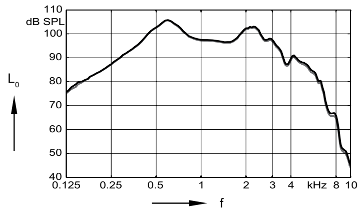


Basic acoustic response (L_I = 60 dB)

Inductive response



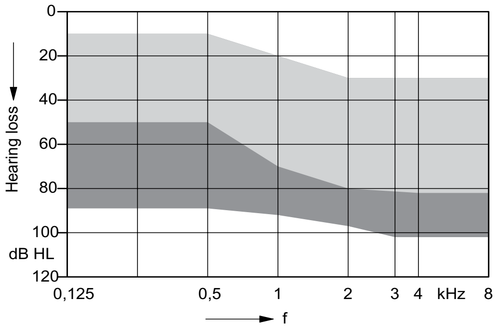
Inductive response (H = 10 mA/m)



SPLITS curve left (H = 31.6 mA/m)

SPLITS curve right (H = 31.6 mA/m)

Contrast S+ | Fitting Range



ThinTube double tip

+ Earhook damped

Contrast S+, HP+, SP+ | Features and Accessories

	S+
Audiology	
Signal processing (channels) **/ Gain/MPO** (handles)	32 / 16
Hearing programs	6
SpeechMaster[1]	●
HD Music (presets)	1
Wireless CROS/BICROS[3]	●
Directionality (channels)	32
Narrow Directionality[2]	■■■■
Directional microphone	■■■■
SpeechFocus	■■■■
TruEar™	■■■■
Frequency compression	●
Feedback cancellation	●
eWindScreen binaural[2]	●
eWindScreen™ (steps)	3
Noise Reduction (channels / steps)	32 / 5
Speech and noise management (steps)	5
SoundSmoothing™ (steps)	3
Directional speech enhancement (steps)	1
SoundBrilliance™[4]	●
Sound equalizer (classes)	3
Spatial Configurator[2]	●
Span[5]	●
Direction[6]	●
SoundBalance	●
Fitting	
Insitugram	●
Learning (classes) **/ Data logging**	3 / ●
Acclimatization manager	●
Tinnitus	
Tinnitus noiser	
Static therapy signal (handles / presets)	16 / 5
Ocean Waves therapy signal (presets)	4
Notch therapy	●

Contrast S+, HP+, SP+ | Features and Accessories

	S+
Style Specific Features	
Ingress Protection Rating	IP67
Telecoil	●
Battery Size	13
Battery door on/off function	●
Nanocoated housing	●
e2e wireless™ 3.0	●
Audio streaming with easyTek	●
User controls coupling via e2e	●
Wireless programming	●
Instrument configurations	
Push button	—
Rocker switch	●
Battery door – direct audio input	○
Battery door – child lock	—
LED status indicator	—
Small earhook	—
Programming Accessories	
ConnexxAir, ConnexxLink™	●
Programming adapter / cable	size 13
Accessories	
miniPocket	○
CROS Pure	○
easyPocket™	○
easyTek	○
TV Transmitter (req. easyTek)	○
Transmitter (req. easyTek)	○
VoiceLink™ (req. easyTek)	○
App	
easyTek App (req. easyTek)	○
touchControl App	○

● available ■■■■■ highest feature performance ○ optional — not available

1) primax fit only
2) req. bilateral fitting and e2e™ 3.0
3) req. CROS Pure accessory
4) streaming only, req. easyTek™
5) req. easyTek & easyTek App, touchControl App or rocker switch
6) req. easyTek & easyTek App or touchControl App

danalogic Ambio

AM577-DWT AM577-DW

Product Description

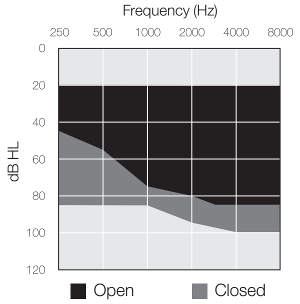
Model 77 Behind-the-Ear (BTE) hearing aids support closed and open configurations.

The 77 BTE model comes standard with Push Button, Volume Control, Telecoil, and Direct Audio Input (DAI) functionality.

This hearing aid is iSolate™ nanotech coated for optimum durability and meet the IP58 classification for ingress protection.

Model	AM577-DWT AM577-DW
Device Configurations	
Battery size	13
Audiological Features	
WARP compression (WDRC) - number of channels	17
Binaural Directionality	●
Directional Mix Processor	●
-Adjustable directional mix	●
Synchronized Soft Switching	●
Autoscope Adaptive Directionality	●
Multiscope Adaptive Directionality	●
Environmental Classifier	●
Binaural Environmental Optimizer II	●
Noise reduction	●
Expansion	●
Wind Guard	●
Sound Shaper	●
DFS Ultra II	●
-Music Mode	●
Synchronized Acceptance Manager	●
Low Frequency Boost	●
Amplification strategy (WDRC / Semi-Linear/Linear)	●
Tinnitus Sound Generator	●
Functional Features	
Synchronized Push Button	●
Synchronized Volume Control	●
Smart Start	●
Phone Now	●
Comfort Phone	●
Ear to Ear Communication	●
BeMore app	●
Fitting Features	
Fitting Software Danalogic 1.0	●
Fully Flexible Programs	4
Auto DFS	●
Onboard Analyzer II	●
Wireless Fitting with Noahlink Wireless	●
In-Situ Audiometry	●
Gain handles	17

Fitting Range



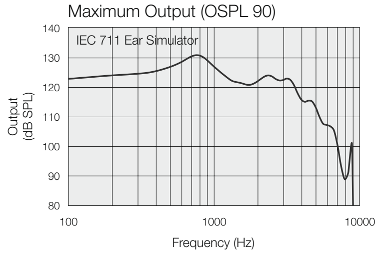
Technical Specifications

		AM577-DWT		
		IEC 60118-0 2nd IEC 711 Ear simulator	IEC 60118-0 3rd IEC 60118-7 ANSI S3.22 2cc coupler	
Reference test gain (60 dB SPL input)	1600 Hz/HFA	45	38	dB
Full-on gain (50 dB SPL input)	Max. 1600 Hz/HFA	62 54	51 48	dB
Maximum output (90 dB SPL input)	Max. 1600 Hz/HFA	131 121	127 116	dB SPL
Total harmonic distortion	500 Hz 800 Hz 1600 Hz	0.5 0.5 0.9	0.2 0.2 0.6	%
Telecoil sensitivity (1 mA/m input) HFA - SPLIV @ 31.6 mA/m (ANSI) Full-on telecoil sensitivity @ 1mA/m	Max. HFA 1600 Hz/HFA	94 87	 100 80	dB SPL
Equivalent input noise		25	22	dB SPL
Frequency range (DIN 45605/ANSI)		100-6920	100-6810	Hz
Current drain		1.2	1.2	mA

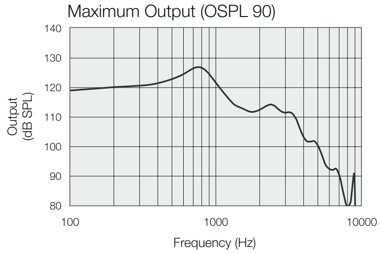
Patents pending

All specifications are subject to change without notice

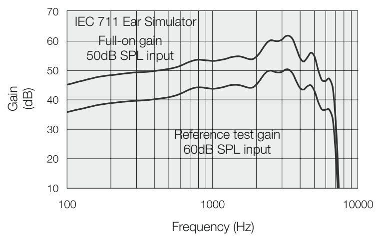
Maximum Output (OSPL 90)



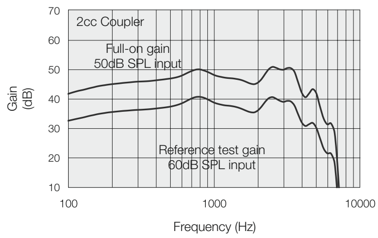
Maximum Output (OSPL 90)



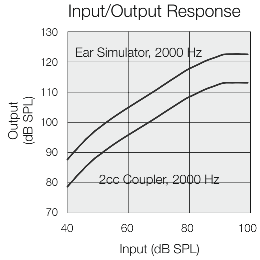
Full-On and Reference Test Gain



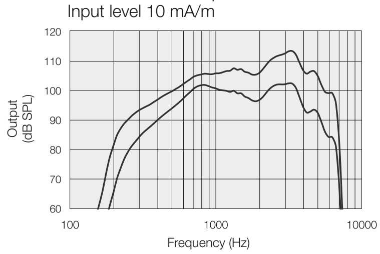
Full-On and Reference Test Gain



Input/Output Response



Full-On Telecoil Response
Input level 10 mA/m



Notes:
O.E.S. = Occluded Ear Simulator
2cc = 2 cm³ coupler

Basic fitting software settings:
Full-on Gain, Reference Test Gain,
MPO = Maximum Power Output,
Maximum Band Width

Measured according to IEC 60118-0, Edition 3.0 2015-06 at 1.3V, impedance 6.2 ohms and 23°C on O.E.S. according to IEC60318-5 2006.
Response on 2cc according to IEC 60118-7 2nd edition, 2005-10 and ANSI S3.22-2009.
(HFA average calculated at 1,000Hz, 1,600Hz, and 2,500Hz); 0 dBSPL sound pressure equals 20µPa).
All measurements without DSP and features activated unless indicated otherwise.
Measurement on O.E.S according to IEC711 1981 According to IEC60118-0 Edition 2 1983 and amendment 1 1994.

A.1 GlobalCareMarket/FoKe FK-162 amplifier

Max sound output	126 ± 5 dB
Gain	44 ± 5 dB
Total harmonic distortion	$\leq 8\%$
Input noise	≤ 30 dB
Frequency range	200 Hz - 3500 Hz
Voltage	≤ 1.5 V
Current	≤ 6 mA

Appendix B

Ethics Documentation

Attached documentation consists of the following:

1. ERGO Application Form
2. Data Management Action Plan
3. Risk Assessment
4. Participant Information Sheet
5. Consent Form
6. Participant Questionnaire

ERGO application form – Ethics form

All mandatory fields are marked (M*). Applications without mandatory fields completed are likely to be rejected by reviewers. Other fields are marked "if applicable". Help text is provided, where appropriate, in italics after each question.

1. APPLICANT DETAILS

1.1 (M*) Applicant name:	Robyn Hunt
1.2 Supervisor (if applicable):	Dr. S. Bell & Prof. D. Simpson
1.3 Other researchers/collaborators (if applicable): *Name, address, email, telephone*	N/A

2. STUDY DETAILS

2.1 (M*) Title of study:	Speech Intelligibility of Hearing Aids
2.2 (M*) Type of study (*e.g. Undergraduate, Doctorate, Masters, Staff*):	Doctorate
2.3 i) (M*) Proposed start date:	24/09/2018
2.3 ii) (M*) Proposed end date:	01/10/2020

2.4 (M*) What are the aims and objectives of this study?
The primary aim is to determine whether objective speech intelligibility metrics (such as HASPI) accurately predict subjective speech intelligibility in terms of 'percentage correct' as measured by a fixed-SNR (speech to noise ratio) speech-in-noise test. The study also aims to compare the speech intelligibility capabilities of different hearing aids and settings.

2.5 (M*) Background to study (*a **brief** rationale for conducting the study*):
Subjective speech intelligibility in noise trials (using speech in noise tests) are widely accepted as time-consuming and therefore expensive to conduct. It is therefore desirable to find an objective metric which can accurately predict the subjective speech intelligibility associated with a particular hearing aid or setting. Several of these metrics already exist but few have been tested using real hearing aids.

2.6 (M*) Key research question (*Specify hypothesis if applicable*):
Can subjective speech intelligibility can be predicted by objective metrics such as HASPI, STOI and CSII? Is there a difference in speech intelligibility between different hearing aids and settings? It is expected that a significant difference in speech intelligibility is observable between different hearing aids, as well as different hearing aid settings (noise reduction algorithms, directional microphones etc.).

UNIVERSITY OF Southampton

2.7 (M*) Study design (*Give a **brief** outline of basic study design*) *Outline what approach is being used, why certain methods have been chosen.*
Up to 30 normal hearing and 20 hearing impaired participants over the age of 18 will perform speech-in-noise tests to evaluate the speech intelligibility of a variety of hearing aids and hearing aid programs. The study involves several iterations of a fixed-SNR speech-in-noise test. The participant will listen to and repeat back a variety of sentences presented at several signal to noise ratios. These sentences have been prerecorded through hearing aids in various settings using a model of a head (KEMAR) with microphones in the ears. The stimulus material will consist of sentences combined with noise. The participant will be asked to repeat the sentence material and the percentage of keywords correctly repeated will be used as a measure of speech intelligibility for the hearing aid or setting in question. Each hearing aid or condition will be tested once. Consent will be gained from all participants before commencing with testing and they can ask to stop at any time. Prior to testing each participant will undergo otoscopy, to ensure they have no contraindications to testing and a hearing test to determine their audiometric thresholds (conducted by a qualified researcher as per BSA practice guidelines (2016, 2012). Each test will not exceed 82 dBA. Each participant will be tested for no longer than one hour at a time. According to the NIOSH (1998) guidelines, participants should only be exposed to sounds at 85 dBA for a maximum time of 8 hours, therefore my proposed testing level and duration is well below this. The presented levels will be the same for both hearing-impaired and normal hearing participants. The upper limit of 82 dBA includes any additional effects of hearing aid amplification. The speech material will be presented to the participants via insert earphones. Testing will be conducted in a sound treated listening room. Each participant will also be asked if they would like to be added to a contact database for further research in Audiology. As per ERGOII advice, a separate participant information sheet and consent form has been provided for this. If the participant consents, their name, gender, date of birth and a summary of their hearing health will be recorded and stored as detailed in the DPA.

3. SAMPLE AND SETTING

3.1 (M*) How are participants to be *approached*? *Give details of what you will do if recruitment is insufficient. If participants will be accessed through a third party (e.g. children accessed via a school) state if you have permission to contact them and **upload any letters of agreement to your submission in ERGO**.*
Students and staff members at the University of Southampton as well as local friends and family will be approached directly and asked if they would be willing to participate in the study. The option of participating in the trial will be advertised by the research sponsors, the charity Action on Hearing Loss, via newsletters or blog posts. Posters will also be used to recruit participants around the Engineering and Audiology departments and on the University of Southampton website, with an email address attached for contacting the researcher.

3.2 (M*) Who are the proposed sample and where are they from (e.g. fellow students, club members)? *List inclusion/exclusion criteria if applicable. NB The University does not condone the use of 'blanket emails' for contacting potential participants (i.e. fellow staff and/or students).*

It is usually advised to ensure groups of students/staff have given prior permission to be contacted in this way, or to use of a third party to pass on these requests. This is because there is a potential to take advantage of the access to 'group emails' and the relationship with colleagues and subordinates; we therefore generally do not support this method of approach. *If this is the only way to access a chosen cohort, a reasonable compromise is to obtain explicit approval from the Faculty Ethics Committee (FEC) and also from a senior member of the Faculty in case of complaint.*
Students, staff members, friends and family will be asked if they would be willing to participate in the study. Responders to any advertisement will also be included. Written consent will be obtained. Participants will be excluded from the study if: - Written consent is not obtained - Otoscopy detects abnormalities such as excessive wax, foreign bodies in the ear, signs of infection, abnormal anatomy (mastoid cavity, perforated tympanic membrane) etc. - The participant has been exposed to loud noises in the past 24 hours - The participant suffers with tinnitus - The participant has hearing thresholds greater than 15dBHL AND does not have a hearing loss matching the preprescribed limits at frequencies between 250Hz and 4000Hz.

3.3 (M*) Describe the relationship between researcher and sample (*Describe any relationship e.g. teacher, friend, boss, clinician, etc.*)
Normal hearing participants will predominantly consist of family, friends, fellow students and staff members known personally to the researcher.

3.4 (M*) Describe how you will ensure that fully informed consent is being given: (*include how long participants have to decide whether to take part*)
Consent forms and participant information sheets (for both the current study and the contact database) will be sent via email or a hard copy will be given to the participant at least 24 hours before testing. The participant will have a minimum of 24 hours to decide whether to take part. A signed consent form from each participant will be retained before testing begins and the participant will be able to withdraw their consent at any time without justification. This also applies to consent for their contact details to be added to the database.

4. RESEARCH PROCEDURES, INTERVENTIONS AND MEASUREMENTS

4.1 (M*) Give a brief account of the procedure as experienced by the participant *(Make clear who does what, how many times and in what order. Make clear the role of all assistants and collaborators. Make clear total demands made on participants, including time and travel).* ***Upload any copies of questionnaires and interview schedules to your submission in ERGO*.**
- Testers will give participants instructions about what the testing involves and gain written consent from the subjects before testing commences. - Participants will be asked to fill out a short questionnaire to ensure suitability for the trial (e.g. no tinnitus, recent ear infections etc.) and be asked if they would like to be included on the audiology contact database. If this is the case, written consent will be obtained and details will be recorded. - Otoscopy will be conducted by the testers on all participants to check for any

UNIVERSITY OF Southampton

contra-indications to testing (please see above). If any are seen testing will not continue. - Participants will undergo a hearing screen or Pure Tone Audiometry (PTA) to determine hearing thresholds. Testing hearing thresholds should not exceed 20 minutes and testing will be conducted by the researcher. The researcher is fully trained to perform otoscopy and PTA according to the BSA (British Society of Audiology, 2016 and 2012 respectively) guidelines. - Participants will listen to sentences combined with noise at various signal-to-noise ratios (as recorded through a hearing aid) presented through inserts (earphones). They will be asked to repeat as many of the words as possible from the sentences presented. The number of words that they relay correctly to the tester will act as the measure of speech intelligibility. The speech intelligibility tests alone should take no more than 1 hour to conduct. - Participants can request a break when needed and can stop testing at any time if they wish. - The level of the test material will not exceed 82 dBA at any test condition, as measured at the dummy head (KEMAR), from which the stimuli for the subjects will be recorded, including presentation of hearing-aid-amplified sounds to hearing impaired participants. This corresponds to normal exposure according to the NIOSH (1998) guidelines. This will be controlled for by the use of a sound level meter (SLM) and will never exceed 85 dBA under any circumstance.

5. STUDY MANAGEMENT

5.1 (M*) State any potential for psychological or physical discomfort and/or distress?
Since sound levels presented will be controlled using a sound level meter and kept below 85dBA at all times (in accordance with NIOSH (1998) guidelines), no potential sources of stress, psychological discomfort or physical discomfort have been identified.

5.2 (M*) Explain how you intend to alleviate any psychological or physical discomfort and/or distress that may arise? (if applicable)
Sessions (including all checks, PTA and speech testing) will last approximately 1 and a half hours in total, will be limited to a maximum of two hours in any one session and breaks may be requested at any time during the testing. Participants may also withdraw from testing at any time. Noise levels will be kept below 85dBA in accordance with NIOSH (1998) guidelines.

5.3 Explain how you will care for any participants in 'special groups' *(i.e. those in a dependent relationship, vulnerable or lacking in mental capacity)* **(if applicable)?**
N/A

5.4 Please give details of any payments or incentives being used to recruit participants (if applicable)?
N/A

5.5 i) How will participant anonymity and/or data anonymity be maintained (if applicable)? *Two definitions of anonymity exist:*

UNIVERSITY OF Southampton

i) Unlinked anonymity - *Complete anonymity can only be promised if questionnaires or other requests for information are not targeted to, or received from, individuals using their name or address or any other identifiable characteristics. For example if questionnaires are sent out with no possible identifiers when returned, or if they are picked up by respondents in a public place, then anonymity can be claimed. Research methods using interviews cannot usually claim anonymity – unless using telephone interviews when participants dial in.* *ii)* Linked anonymity - *Using this method, complete anonymity cannot be promised because participants can be identified; their data may be coded so that participants are not identified by researchers, but the information provided to participants should indicate that they could be linked to their data.*
All data will be anonymised by participant identification numbers, which link data to the consent form (linked anonymity). Contact details on the contact database will not be anonymised but the file will be password-protected and only accessible to researchers and supervisors. Please see the DPA Plan for further details.

5.5 ii) How will participant confidentiality be maintained (if applicable)? *Confidentiality is defined as the non-disclosure of research information except to another authorised person. Confidential information can be shared with those who are already party to it, and may also be disclosed where the person providing the information provides explicit consent.*
Participant names will be recorded only on consent forms. All other data will be anonymised and linked to personal identification numbers for the consent form only. In the case of the contact database, only researchers and supervisors of relevant studies will be given access to the password-protected file containing personal information.

5.6 (M*) How will personal data and study results be stored securely during and after the study? *Researchers should be aware of, and compliant with, the Data Protection policy of the University. You must be able to demonstrate this in respect of handling, storage and retention of data.*
All data will be stored on a secure, password-protected university computer. Consent forms will be kept in a lockable cabinet separate from any other data.

5.7 (M*) Who will have access to these data?
The researcher only will have access to the data but anonymized data may be made public in accordance with open-access publication guidelines.

N.B. – Before you upload this document to your ERGO submission remember to:

1. Complete ALL mandatory sections in this form

2. Upload any letters of agreement referred to in question 3.1 to your ERGO submission

3. Upload any interview schedules and copies of questionnaires referred to in question 4.1

DPA Plan

Ethics reference number: **ERGO/**FEPS**/41012**	Version: 1.1	Date: 2018-09-06
Study Title: Speech Intelligibility of Hearing Aids; Further Research in Audiology (request to join participant database)		
Investigator: Robyn Hunt		

The following is an exhaustive and complete list of all the data that will be collected (through questionnaires, interviews, extraction from records, etc): Gender, date of birth, hearing thresholds, basic summary of hearing health (from questionnaire) and speech-in-noise test results (anonymised). Additionally, if consent has been given for data to be stored on a participant database: Name, email and/or telephone number and/or address, summary of hearing health.

The data is relevant to the study purposes because contact and personal details are required if the participant wishes to be part of further research projects, and the study will examine the link between speech-in-noise test results (which gender and age may influence) and hearing loss. The data is adequate because enough data will be kept to enable contact of the participants (if consent is given) and enough data will be kept to enable the results from the speech in noise tests to be analysed accurately, and the data is not excessive because only basic personal and contact information is kept (if consent is given), as well as information relevant to further studies in Audiology.

The data will be processed fairly because the participants will have given explicit consent.

The data's accuracy is ensured because the participants have provided their data and as such, the data is accurate to the best of the researcher's knowledge.

Data will be stored on the University server. The data will be held in accordance with University policy on data retention.

Data files will be protected by password access only.

The data will be destroyed after ten years, by the researcher or supervisor, or if a participant withdraws their data.

The data will be processed in accordance with the rights of the participants because they will have the right to access, correct, and/or withdraw their data at any time and for any reason. Participants will be able to exercise their rights by contacting the investigator (e-mail: rmh1g13@soton.ac.uk) or the project supervisor (e-mail: s.l.bell@soton.ac.uk or ds@isvr.soton.ac.uk).

Consent forms will be linked to the data by a participant identification number and stored in a locked cabinet.

Data Protection Act 2018 (DPA) best practice

If the study involves personal or sensitive data, explicitly explain how data will be collected, stored, analysed, held securely, and in turn destroyed. The DPA does not apply to anonymous data and a DPA Plan is not required in the case of such data.

The principles of the DPA are that personal data must be:

1. Processed fairly and lawfully.
2. Processed for specified purposes and in an appropriate way.
3. Adequate, relevant and not excessive for the purposes.
4. Accurate and up-to-date.
5. Not kept for longer than necessary.
6. Processed in accordance with the rights of data subjects (participants).
7. Protected by appropriate security, both practical and organisational.
8. Not transferred outside the European Economic Area (EEA) without adequate data protection controls.

Data is recorded information, whether stored electronically on computer or in paper-based filing systems. Personal data is information about an identifiable living individual. It can be factual, such as the date of a person's interview, or an opinion, such someone's view on how the person has performed on a task. It obviously includes individuals' contact addresses or telephone numbers. (Less obviously, note that personal data is being processed where information is collected and analysed with the intention of distinguishing one individual from another and to take a particular action in respect of an individual. This can take place even if no obvious identifiers, such as names or addresses, are held.) Processing is any activity that involves data, including collecting, recording or retrieving, using, disclosing, organising, adapting, changing, updating, or destroying it.

The DPA identifies Sensitive Personal Data as:

a) the racial or ethnic origin of the data subject (participant);
b) his political opinions;
c) his religious beliefs or other beliefs of a similar nature;
d) whether he is a member of a trade union;
e) his physical or mental health or condition;
f) his sexual life;
g) the commission or alleged commission by him of any offence or
h) any proceedings for any offence committed or alleged to have been committed by him, the disposal of such proceedings and the sentence of court in such proceedings.

The processing of sensitive data must meet at least one of the 10 stricter conditions laid down in Schedule 3 of the DPA. It may be useful to know that condition 1 of this schedule permits processing of such data if the data subject has given his explicit consent, and condition 5 if the information has been made public as a result of steps deliberately taken by the data subject.

Keep in mind that the Police have a right of access to personal data held by the study for the purpose of safeguarding national security; preventing or detecting crime; prosecuting or apprehending offenders; assessing or collecting tax; or protecting the vital interests of the data subject or another.

Researchers are exempted: from the second data protection principle, meaning that personal data can be processed for purposes other than for which they were originally obtained; from the fifth data protection principle, meaning that personal data can be held indefinitely; and from the data subject's right of access to his personal data provided the data is processed for research purposes and the results do not identify data subjects. In addition, the Data Protection (Processing of Sensitive Personal Data) Order 2000 para.9 provides that processing in the course of maintaining archives for research purposes is permissible where the sensitive personal data are not used to take decisions about any person without their consent and no substantial damage or distress is caused to any person by the keeping of those data. These exemptions do NOT give a blanket exemption from all the Data Protection Principles to data provided and/or used for research purposes. Researchers wishing to use personal data should be aware that the Data Protection Principles still generally apply, notably the requirement to keep data secure1.

A study may seek to anonymise the data it keeps. Anonymisation involves the removal of participants' personal information (names; e-mail address; whatever data it is that might permit identification; etc) from the data such that what remains cannot be used to identify them. Note that audio and video recordings (and often transcriptions too) cannot easily be anonymised, so they should normally be treated as non-anonymous data. Anonymised data can usually be kept without security and can easily be passed to other investigators for specialist analysis.

The DPA requires access to be granted to participants to all of their data, if any part of that data allows their identification. If the data has been anonymised, two issues arise.

1. If the personal information has been removed from the data AND DESTROYED, then the DPA is no longer applicable, and the data can be kept without security. However, investigators should note that they will be unable to follow up or subsequently contact participants in any way, or associate individuals with particular data, and should not attempt to suggest they might do so.

2. If the personal information has been removed from the bulk of the data, but NOT destroyed (ie, is kept separately), then the DPA remains applicable. In this situation, the personal information needs to be (a) kept both separately and securely from the anonymised data, and (b) to be linked or 'keyed' to the anonymised data, such keys to be similarly kept securely (and often kept with the personal information).

If personal data is collected, in the ‘Participant Information’, inform the participant of:

· the processes the study will take to ensure data security;
· their right to access and correct their data and their right to request removal of their data;
· the authority which will give them access to their data (provide the contact information).

If sensitive data is collected, or the study involves clinical studies, human tissue samples, invasive procedures, or young or vulnerable people, provide additional detail. In the ‘Participant Information’, inform the participant of:

· the separation of identifying data and the anonymisation process;
· the method of linking the consent form (if any) to the participant’s data;
· the processes for the destruction of all study data (if appropriate).

The study should conform to the University policy on data management applicable:

http://www.southampton.ac.uk/library/research/researchdata/

Investigators may find the University’s survey platform useful:

https://www.isurvey.soton.ac.uk/

[1] http://www.jisc.ac.uk/publications/generalpublications/2001/pub_dpacop_0101.aspx

Contacts

risethic@soton.ac.uk.

University of Southampton Health & Safety Risk Assessment

Version: 1.4 06/09/18

Risk Assessment			
Risk Assessment for the activity of	**Speech Intelligibility of Hearing Aids – Speech in Noise tests with participants**	**Date**	**06/09/18**
Unit/Faculty/Directorate	***ISVR (FEPS)***	**Assessor**	**Robyn Hunt**
Line Manager/Supervisor	***Steve Bell***	**Signed off**	

PART A											
(1) Risk identification			**(2) Risk assessment**				**(3) Risk management**				
Hazard	**Potential Consequences**	**Who might be harmed (user; those nearby; those in the vicinity; members of the public)**	**Inherent**			**Control measures (use the risk hierarchy)**	**Residual**			**Further controls (use the risk hierarchy)**
			Likelihood	**Impact**	**Score**		**Likelihood**	**Impact**	**Score**	
Loud Sounds	Temporary hearing damage from too high a sound level being presented through insert headphones	Study participants	1	3	3	Turn on and calibrate equipment and check sound level before inserting earphones. Begin with volume at minimum and increase gradually to avoid loud sounds. Participants encouraged to remove earphones if sound is uncomfortably loud.	1	1	1	N/A

University of Southampton Health & Safety Risk Assessment

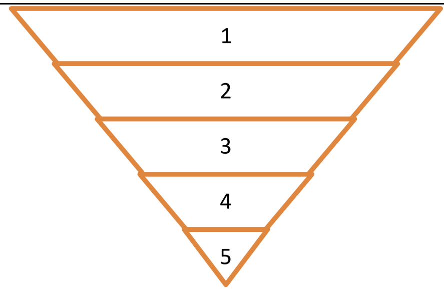
PART A										
(1) Risk identification			**(2) Risk assessment**				**(3) Risk management**			
Hazard	**Potential Consequences**	**Who might be harmed (user; those nearby; those in the vicinity; members of the public)**	**Inherent**				**Residual**			**Further controls (use the risk hierarchy)**
			Likelihood	**Impact**	**Score**	**Control measures (use the risk hierarchy)**	**Likelihood**	**Impact**	**Score**	
Trailing cables	Trips and falls	All	**3**	**2**	**6**	Move cables to the edges of the room or behind obstacles. Secure any remaining trailing cables with coloured tape and visible warning signs.	**2**	**1**	**2**	N/A
Fatigue	Minor psychological discomfort, stress, tiredness	Study participants	**4**	**1**	**4**	Allow regular breaks to participants. Make clear that withdrawal from the study is possible at any time.	**2**	**1**	**2**	N/A
Ear damage from inserted earphones into obstructed canal	Damage to the ear canal or ear drum resulting in temporary or permanent hearing loss	Study participants	**3**	**3**	**9**	Carry out otoscopy to check for any obstructions (e.g. excessive wax) before inserting earphones.	**1**	**3**	**3**	N/A
Transfer of infection between participants	Ear infection resulting in discomfort, illness, temporary hearing loss	Study participants	**3**	**3**	**9**	Carry out otoscopy to check for any signs of infection before inserting earphones. Use clean tips on earphones and otoscope for each participant and each ear. Clean equipment thoroughly after each use.	**1**	**3**	**3**	N/A

University of Southampton Health & Safety Risk Assessment

Version: 1.4 06/09/18

Assessment Guidance

1. Eliminate	Remove the hazard wherever possible which negates the need for further controls	If this is not possible then explain why
2. Substitute	Replace the hazard with one less hazardous	If not possible then explain why
3. Physical controls	Examples: enclosure, fume cupboard, glove box	Likely to still require admin controls as well
4. Admin controls	Examples: training, supervision, signage	
5. Personal protection	Examples: respirators, safety specs, gloves	Last resort as it only protects the individual



LIKELIHOOD					
5	5	10	15	20	25
4	4	8	12	16	20
3	3	6	9	12	15
2	2	4	6	8	10
1	1	2	3	4	5
	1	2	3	4	5
			IMPACT		

Risk process

1. Identify the impact and likelihood using the tables above.
2. Identify the risk rating by multiplying the Impact by the likelihood using the coloured matrix.
3. If the risk is amber or red – identify control measures to reduce the risk to as low as is reasonably practicable.
4. If the residual risk is green, additional controls are not necessary.
5. If the residual risk is amber the activity can continue but you must identify and implement further controls to reduce the risk to as low as reasonably practicable.
6. If the residual risk is red do not continue with the activity until additional controls have been implemented and the risk is reduced.
7. Control measures should follow the risk hierarchy, where appropriate as per the pyramid above.
8. The cost of implementing control measures can be taken into account but should be proportional to the risk i.e. a control to reduce low risk may not need to be carried out if the cost is high but a control to manage high risk means that even at high cost the control would be necessary.

Impact		Health & Safety
1	Trivial - insignificant	Very minor injuries e.g. slight bruising
2	Minor	Injuries or illness e.g. small cut or abrasion which require basic first aid treatment even in self-administered.
3	Moderate	Injuries or illness e.g. strain or sprain requiring first aid or medical support.
4	Major	Injuries or illness e.g. broken bone requiring medical support >24 hours and time off work >4 weeks.
5	Severe – extremely significant	Fatality or multiple serious injuries or illness requiring hospital admission or significant time off work.

Likelihood	
1	Rare e.g. 1 in 100,000 chance or higher
2	Unlikely e.g. 1 in 10,000 chance or higher
3	Possible e.g. 1 in 1,000 chance or higher
4	Likely e.g. 1 in 100 chance or higher
5	Very Likely e.g. 1 in 10 chance or higher

UNIVERSITY OF Southampton

Participant Information Sheet

Study Title: Speech Intelligibility for Hearing Aids

Researcher: Robyn Hunt
ERGO number: 41012

You are being invited to take part in the above research study. To help you decide whether you would like to take part or not, it is important that you understand why the research is being done and what it will involve. Please read the information below carefully and ask questions if anything is not clear or you would like more information before you decide to take part in this research. You may like to discuss it with others but it is up to you to decide whether or not to take part. If you are happy to participate you will be asked to sign a consent form.

What is the research about?
This study forms part of a PhD research project to develop and assess current objective methods (calculations performed by a computer) designed to predict speech intelligibility for hearing aid users. Implementation of objective measures is cheaper and less time-consuming than subjective trials (those involving study participants). Therefore, if a well-performing objective predictor of speech intelligibility for hearing aids can be found, the assessment process of hearing aids (for the NHS, for example) could be greatly improved in terms of monetary and time cost efficiency. We are therefore collecting data on human listeners' assessment of speech intelligibility to ensure that the objective measures can predict listeners' performance. This research project is funded by a charity called Action on Hearing Loss.

Why have I been asked to participate?
We are recruiting any English-speaking adults (over the age of 18) in this study.

What will happen to me if I take part?
You will be asked to attend one session, which will last approximately an hour and a half.

Firstly, ear checks will be conducted to ensure it is safe for you to continue with the study. If contraindications are found (e.g. excessive wax or an infection), you may not be able to continue safely with the study.

Next, a hearing screening will be conducted to check whether you have normal hearing. If you fail the screening, your hearing thresholds (the quietest sounds you can hear) will be tested and recorded. You may be excluded from the study if details of your hearing loss do not fit with the study criteria.

For the main part of the study, you will be asked to do a 'Speech-in-Noise test'. For this, you will listen to several short, noisy speech samples through insert earphones. After each sentence, you should repeat back to the researcher what you think was said in the recording. The accuracy of your response will be recorded for each sentence.

Comfort breaks can be taken at any time.

Are there any benefits in my taking part?
A free hearing screen is included as part of the study. If a hearing loss is found, you will be advised to visit your GP for further investigation. If you already have a hearing loss, this will be checked and confirmed by a full hearing test.

Are there any risks involved?

UNIVERSITY OF Southampton

There are no significant risks associated with this study. There is a small risk of excessive noise exposure, but all sounds to be presented have been carefully calibrated to minimise this risk. If any sounds are uncomfortably loud, remove the earphones, let us know and we will stop the experiment immediately. If you become tired during the study, you can take a comfort break or retire from the study at any time.

What data will be collected?
Your name will only be present on the consent form and will only be linked to your data by coded participant numbers.
The researcher will record your age, gender and your hearing thresholds, as well as your responses to the Speech-in-Noise test. These are the only details which will be retained by the researcher for further analysis. This data will only be analysed after anonymising.

If you consent to being contacted for further research, your name and contact details and a brief description of your hearing ability will be recorded and kept on a password-protected computer. A separate consent form and participant information sheet will be provided for this. This is optional and you do not have to consent to the storage of your contact details to take part in the study.

Will my participation be confidential?
Your participation and the information we collect about you during the course of the research will be kept strictly confidential.

Only members of the research team and responsible members of the University of Southampton may be given access to data about you for monitoring purposes and/or to carry out an audit of the study to ensure that the research is complying with applicable regulations. Individuals from regulatory authorities (people who check that we are carrying out the study correctly) may require access to your data. All of these people have a duty to keep your information, as a research participant, strictly confidential.

Do I have to take part?
No, it is entirely up to you to decide whether or not to take part. If you decide you want to take part, you will need to sign a consent form to show you have agreed to take part.

If you wish to take part in the study, please contact Robyn Hunt at rmh1g13@soton.ac.uk.

What happens if I change my mind?
You have the right to change your mind and withdraw at any time without giving a reason and without your participant rights being affected.

If you withdraw from the study, we will keep the information about you that we have already obtained for the purposes of achieving the objectives of the study only.

If you wish to remove your details from the contact database, or withdraw from the study, contact rmh1g13@soton.ac.uk.

What will happen to the results of the research?
Your personal details will remain strictly confidential. Research findings made available in any reports or publications will not include information that can directly identify you without your specific consent. Your anonymised data may be published as part of scientific journal articles and/or the researcher's PhD thesis submission.

You may ask for a copy of your hearing screening results at the end of the session.

Your anonymised data may be made available for future research projects. The data will be stored for a minimum of 10 years, as per University of Southampton policy. Publications and anonymised data relating to the research will be made available through the institutional repository.

Where can I get more information?
For more information, please do not hesitate to contact the researcher, Robyn Hunt, at rmh1g13@soton.ac.uk.

What happens if there is a problem?
If you have a concern about any aspect of this study, you should speak to the researchers who will do their best to answer your questions. You can contact the research team via rmh1g13@soton.ac.uk.
If you remain unhappy or have a complaint about any aspect of this study, please contact the University of Southampton Research Integrity and Governance Manager (023 8059 5058, rgoinfo@soton.ac.uk).

Data Protection Privacy Notice
The University of Southampton conducts research to the highest standards of research integrity. As a publicly-funded organisation, the University has to ensure that it is in the public interest when we use personally-identifiable information about people who have agreed to take part in research. This means that when you agree to take part in a research study, we will use information about you in the ways needed, and for the purposes specified, to conduct and complete the research project. Under data protection law, 'Personal data' means any information that relates to and is capable of identifying a living individual. The University's data protection policy governing the use of personal data by the University can be found on its website (https://www.southampton.ac.uk/legalservices/what-we-do/data-protection-and-foi.page).

This Participant Information Sheet tells you what data will be collected for this project and whether this includes any personal data. Please ask the research team if you have any questions or are unclear what data is being collected about you.

Our privacy notice for research participants provides more information on how the University of Southampton collects and uses your personal data when you take part in one of our research projects and can be found at http://www.southampton.ac.uk/assets/sharepoint/intranet/ls/Public/Research%20and%20Integrity%20Privacy%20Notice/Privacy%20Notice%20for%20Research%20Participants.pdf

Any personal data we collect in this study will be used only for the purposes of carrying out our research and will be handled according to the University's policies in line with data protection law. If any personal data is used from which you can be identified directly, it will not be disclosed to anyone else without your consent unless the University of Southampton is required by law to disclose it.

Data protection law requires us to have a valid legal reason ('lawful basis') to process and use your Personal data. The lawful basis for processing personal information in this research study is for the performance of a task carried out in the public interest. Personal data collected for research will not be used for any other purpose.

For the purposes of data protection law, the University of Southampton is the 'Data Controller' for this study, which means that we are responsible for looking after your information and using it properly. The University of Southampton will keep identifiable information about you for 10 years after the study has finished after which time any link between you and your information will be removed.

To safeguard your rights, we will use the minimum personal data necessary to achieve our research study objectives. Your data protection rights – such as to access, change, or transfer such information - may be limited, however, in order for the research output to be reliable and accurate. The University will not do anything with your personal data that you would not reasonably expect.

If you have any questions about how your personal data is used, or wish to exercise any of your rights, please consult the University's data protection webpage (https://www.southampton.ac.uk/legalservices/what-we-do/data-protection-and-foi.page) where you can make a request using our online form. If you need further assistance, please contact the University's Data Protection Officer (data.protection@soton.ac.uk).

Thank you.

Thank you for taking the time to read the information sheet and considering taking part in the research.

CONSENT FORM

Study title: Speech Intelligibility of Hearing Aids

Researcher name: Robyn Hunt
ERGO number: 41012
Participant Number:

Please initial the box(es) if you agree with the statement(s):

I have read and understood the information sheet (06/09/18, v1.4) and have had the opportunity to ask questions about the study.	
I agree to take part in this research project and agree for my data to be used for the purpose of this study.	
I understand my participation is voluntary and I may withdraw at any time for any reason without my participation rights being affected.	
I understand that if I withdraw from the study that it may not be possible to remove the data once my personal information is no longer linked to the data.	
I understand that I will not be directly identified in any reports of the research.	
I give permission for my data (my hearing test results and speech understanding) that I provide to be deposited to the University of Southampton's institutional repository as described in the participant information sheet so it can be used for future research and learning into speech intelligibility and hearing aid technology.	

Name of participant (print name)..

Signature of participant...

Date...

Name of researcher (print name)............... Robyn Hunt ....

Signature of researcher ....

Date..

Participant Questionnaire Sheet

Study Title: Speech Intelligibility of Hearing Aids

Researcher: Robyn Hunt **Ethics Number:** 41012

Participant ID:

Please fill in the following questionnaire to determine your eligibility for this experiment. If yes to any of the following questions please give details.

1. Are you less than 18 years of age?

Yes / No

2. Do you have a known hearing impairment?

Yes / No

3. Have you ever had any recent pain, tenderness, infections, discharge, surgery or bleeding from either of your ears?

Yes / No

4. Do you experience tinnitus (ringing, buzzing, whistling or any other sounds in either of your ears)?

Yes / No

5. Do you suffer from hyperacusis (sensitivity to loud sounds)?

Yes / No

6. Have you been exposed to loud sounds in the past 24 hours?

Yes / No

Appendix C

Statistical Details

Attached documentation consists of the following:

1. Descriptive statistics relating to subjective data
2. ANOVA between hearing aid make, noise reduction setting, age, presentation ear and SNR for subjective data
3. Significance of differences between individual hearing aid conditions for subjective data
4. Significance of differences between individual hearing aid conditions for HASPI
5. Significance of differences between individual hearing aid conditions for CSII
6. Significance of differences between conditions for STOI
7. Significance of differences between subjective data and objective predictions
8. Significance of differences in subjective data due to other covariates

Descriptive Statistics - Subjective Data

SNR = 3	N Statistic	Minimum Statistic	Maximum Statistic	Mean Statistic	Std. Deviation Statistic	Skewness Statistic	Skewness Std. Error	Kurtosis Statistic	Kurtosis Std. Error
Cheap	21	0.20	0.90	0.6095	0.19211	-0.617	0.501	-0.632	0.972
HA1, FF	21	0.30	0.80	0.6476	0.12891	-0.871	0.501	1.088	0.972
HA1, NR	21	0.40	1.00	0.7587	0.14640	-0.665	0.501	0.894	0.972
NoHA	21	0.30	1.00	0.8079	0.18071	-1.707	0.501	2.879	0.972
HA2, FF	21	0.40	1.00	0.7063	0.17179	-0.111	0.501	-0.437	0.972
HA2, NR	21	0.60	1.00	0.8571	0.11212	-0.432	0.501	-0.188	0.972
HA3, FF	21	0.30	0.87	0.6413	0.12949	-0.764	0.501	1.086	0.972
HA3, NR	21	0.60	1.00	0.8206	0.13395	-0.116	0.501	-1.303	0.972

SNR = 0	N Statistic	Minimum Statistic	Maximum Statistic	Mean Statistic	Std. Deviation Statistic	Skewness Statistic	Skewness Std. Error	Kurtosis Statistic	Kurtosis Std. Error
Cheap	21	0.00	0.60	0.3381	0.15961	-0.291	0.501	-0.202	0.972
HA1, FF	21	0.20	0.80	0.3968	0.17854	0.936	0.501	0.592	0.972
HA1, NR	21	0.10	0.80	0.4968	0.18765	-0.264	0.501	-0.272	0.972
NoHA	21	0.30	0.80	0.5778	0.16544	-0.182	0.501	-1.312	0.972
HA2, FF	21	0.00	0.70	0.4095	0.16705	-0.450	0.501	0.407	0.972
HA2, NR	21	0.30	0.90	0.6460	0.17272	-0.305	0.501	-0.054	0.972
HA3, FF	21	0.10	0.70	0.3714	0.16169	0.122	0.501	-0.397	0.972
HA3, NR	21	0.40	1.00	0.7016	0.13803	-0.163	0.501	0.556	0.972

SNR = -3	N Statistic	Minimum Statistic	Maximum Statistic	Mean Statistic	Std. Deviation Statistic	Skewness Statistic	Skewness Std. Error	Kurtosis Statistic	Kurtosis Std. Error
Cheap	21	0.00	0.20	0.0683	0.07414	0.567	0.501	-0.948	0.972
HA1, FF	21	0.00	0.30	0.1429	0.11212	0.197	0.501	-1.287	0.972
HA1, NR	21	0.00	0.73	0.3683	0.20856	-0.021	0.501	-0.895	0.972
NoHA	21	0.00	0.50	0.1762	0.12209	1.050	0.501	1.408	0.972
HA2, FF	21	0.00	0.30	0.1175	0.10034	0.108	0.501	-1.334	0.972
HA2, NR	21	0.00	0.60	0.3540	0.14662	-0.454	0.501	0.758	0.972
HA3, FF	21	0.00	0.30	0.1667	0.09129	-0.113	0.501	-0.634	0.972
HA3, NR	21	0.00	0.70	0.3444	0.18479	-0.168	0.501	-0.792	0.972

SNR = -8	N Statistic	Minimum Statistic	Maximum Statistic	Mean Statistic	Std. Deviation Statistic	Skewness Statistic	Skewness Std. Error	Kurtosis Statistic	Kurtosis Std. Error
Cheap	21	0.00	0.00	0.0000	0.00000				
HA1, FF	21	0.00	0.10	0.0048	0.02182	4.583	0.501	21.000	0.972
HA1, NR	21	0.00	0.10	0.0095	0.03008	2.975	0.501	7.562	0.972
NoHA	21	0.00	0.00	0.0000	0.00000				
HA2, FF	21	0.00	0.10	0.0079	0.02561	3.201	0.501	9.642	0.972
HA2, NR	21	0.00	0.10	0.0048	0.02182	4.583	0.501	21.000	0.972
HA3, FF	21	0.00	0.00	0.0000	0.00000				
HA3, NR	21	0.00	0.20	0.0540	0.08265	1.072	0.501	-0.633	0.972

2.800	Significantly skewed at 1% level
2.000	Significantly skewed at 5% level
0.000	Not significantly skewed

Appendix A - Statistics

Subjective Data - ANOVA

Multivariate Tests[a]

Effect		Value	F	Hypothesis df	Error df	Sig.
HA	Pillai's Trace	0.017	.684[b]	2.000	79.000	0.507
	Wilks' Lambda	0.983	.684[b]	2.000	79.000	0.507
	Hotelling's Trace	0.017	.684[b]	2.000	79.000	0.507
	Roy's Largest Root	0.017	.684[b]	2.000	79.000	0.507
HA * SNR	Pillai's Trace	0.167	1.822	8.000	160.000	0.077
	Wilks' Lambda	0.838	1.819[b]	8.000	158.000	0.077
	Hotelling's Trace	0.186	1.815	8.000	156.000	0.078
	Roy's Largest Root	0.140	2.797[c]	4.000	80.000	0.031
Noise Reduction Setting	Pillai's Trace	0.282	31.365[b]	1.000	80.000	0.000
	Wilks' Lambda	0.718	31.365[b]	1.000	80.000	0.000
	Hotelling's Trace	0.392	31.365[b]	1.000	80.000	0.000
	Roy's Largest Root	0.392	31.365[b]	1.000	80.000	0.000
Noise Reduction Setting * SNR	Pillai's Trace	0.428	14.961[b]	4.000	80.000	0.000
	Wilks' Lambda	0.572	14.961[b]	4.000	80.000	0.000
	Hotelling's Trace	0.748	14.961[b]	4.000	80.000	0.000
	Roy's Largest Root	0.748	14.961[b]	4.000	80.000	0.000
HA * Noise Reduction Setting	Pillai's Trace	0.018	.711[b]	2.000	79.000	0.494
	Wilks' Lambda	0.982	.711[b]	2.000	79.000	0.494
	Hotelling's Trace	0.018	.711[b]	2.000	79.000	0.494
	Roy's Largest Root	0.018	.711[b]	2.000	79.000	0.494
HA * Noise Reduction Setting * SNR	Pillai's Trace	0.195	2.161	8.000	160.000	0.033
	Wilks' Lambda	0.809	2.207[b]	8.000	158.000	0.030
	Hotelling's Trace	0.231	2.253	8.000	156.000	0.026
	Roy's Largest Root	0.207	4.138[c]	4.000	80.000	0.004

a. Design: Intercept + SNR
Within Subjects Design: HA + Noise Reduction Setting + HA *Noise Reduction Setting
b. Exact statistic
c. The statistic is an upper bound on F that yields a lower bound on the significance level.

Tests of Within-Subjects Effects

Measure:

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
HA	Sphericity Assumed	0.021	2	0.011	0.713	0.492
	Greenhouse-Geisser	0.021	1.994	0.011	0.713	0.491
	Huynh-Feldt	0.021	2.000	0.011	0.713	0.492
	Lower-bound	0.021	1.000	0.021	0.713	0.401
HA * SNR	Sphericity Assumed	0.229	8	0.029	1.913	0.061
	Greenhouse-Geisser	0.229	7.975	0.029	1.913	0.062
	Huynh-Feldt	0.229	8.000	0.029	1.913	0.061
	Lower-bound	0.229	4.000	0.057	1.913	0.116
Error(HA)	Sphericity Assumed	2.398	160	0.015		
	Greenhouse-Geisser	2.398	159.501	0.015		
	Huynh-Feldt	2.398	160.000	0.015		
	Lower-bound	2.398	80.000	0.030		
Noise Reduction Setting	Sphericity Assumed	0.455	1	0.455	31.365	0.000
	Greenhouse-Geisser	0.455	1.000	0.455	31.365	0.000
	Huynh-Feldt	0.455	1.000	0.455	31.365	0.000
	Lower-bound	0.455	1.000	0.455	31.365	0.000
Noise Reduction Setting * SNR	Sphericity Assumed	0.869	4	0.217	14.961	0.000
	Greenhouse-Geisser	0.869	4.000	0.217	14.961	0.000
	Huynh-Feldt	0.869	4.000	0.217	14.961	0.000
	Lower-bound	0.869	4.000	0.217	14.961	0.000
Error(Noise Reduction Setting)	Sphericity Assumed	1.161	80	0.015		
	Greenhouse-Geisser	1.161	80.000	0.015		
	Huynh-Feldt	1.161	80.000	0.015		
	Lower-bound	1.161	80.000	0.015		

Tests of Within-Subjects Effects - continued

Measure:

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
HA * Noise Reduction Setting	Sphericity Assumed	0.019	2	0.010	0.856	0.427
	Greenhouse-Geisser	0.019	1.893	0.010	0.856	0.421
	Huynh-Feldt	0.019	2.000	0.010	0.856	0.427
	Lower-bound	0.019	1.000	0.019	0.856	0.358
HA * Noise Reduction Setting * SNR	Sphericity Assumed	0.228	8	0.029	2.554	0.012
	Greenhouse-Geisser	0.228	7.574	0.030	2.554	0.014
	Huynh-Feldt	0.228	8.000	0.029	2.554	0.012
	Lower-bound	0.228	4.000	0.057	2.554	0.045
Error(HA*Noise Reduction Setting)	Sphericity Assumed	1.788	160	0.011		
	Greenhouse-Geisser	1.788	151.472	0.012		
	Huynh-Feldt	1.788	160.000	0.011		
	Lower-bound	1.788	80.000	0.022		

Tests of Between-Subjects Effects

Measure:

Transformed Variable:

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	11.412	1	11.412	290.755	0.000
SNR	38.055	4	9.514	242.385	0.000
Error	3.140	80	0.039		

Appendix A - Statistics Subjective

1E-07	Significant Difference with Bonferroni Correction
0.01	Significant Difference with no Bonferroni Correction (t-test only)
0.4	No Significant Difference

p values from t-test

3	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.427599	0.004565	6.12E-06	0.040169	2.22E-05	0.412135	0.000284
HA1, F	0.427599	1	0.003839	0.001313	0.199285	7.81E-06	0.864139	0.000125
HA1, N	0.004565	0.003839	1	0.277771	0.24759	0.017609	0.008003	0.136707
NoHA	6.12E-06	0.001313	0.277771	1	0.048431	0.260071	0.00024	0.75916
HA2, F	0.040169	0.199285	0.24759	0.048431	1	0.001139	0.052562	0.012363
HA2, N	2.22E-05	7.81E-06	0.017609	0.260071	0.001139	1	1.69E-06	0.048231
HA3, F	0.412135	0.864139	0.008003	0.00024	0.052562	1.69E-06	1	1.47E-05
HA3, N	0.000284	0.000125	0.136707	0.75916	0.012363	0.048231	1.47E-05	1

0	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.26927	0.000642	5.35E-07	0.196965	1.83E-07	0.348658	2.49E-08
HA1, F	0.26927	1	0.080596	0.000477	0.779506	9.14E-05	0.578863	9E-08
HA1, N	0.000642	0.080596	1	0.156911	0.072097	0.002861	0.018155	0.000149
NoHA	5.35E-07	0.000477	0.156911	1	0.002576	0.121925	8.37E-07	0.004745
HA2, F	0.196965	0.779506	0.072097	0.002576	1	2.79E-06	0.450622	2.85E-06
HA2, N	1.83E-07	9.14E-05	0.002861	0.121925	2.79E-06	1	1.84E-05	0.254782
HA3, F	0.348658	0.578863	0.018155	8.37E-07	0.450622	1.84E-05	1	7.61E-09
HA3, N	2.49E-08	9E-08	0.000149	0.004745	2.85E-06	0.254782	7.61E-09	1

-3	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.006878	4.22E-07	0.00275	0.026392	5.42E-08	0.000547	2.03E-06
HA1, F	0.006878	1	4.82E-06	0.1842	0.291624	1.34E-05	0.329257	0.000199
HA1, N	4.22E-07	4.82E-06	1	0.000397	6.21E-06	0.779632	4.76E-05	0.649447
NoHA	0.00275	0.1842	0.000397	1	0.078729	0.000294	0.715186	0.001335
HA2, F	0.026392	0.291624	6.21E-06	0.078729	1	4.57E-07	0.072833	1.91E-05
HA2, N	5.42E-08	1.34E-05	0.779632	0.000294	4.57E-07	1	9.26E-05	0.822586
HA3, F	0.000547	0.329257	4.76E-05	0.715186	0.072833	9.26E-05	1	0.000695
HA3, N	2.03E-06	0.000199	0.649447	0.001335	1.91E-05	0.822586	0.000695	1

-8	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.329257	0.162298	1	0.171036	0.329257	1	0.007198
HA1, F	0.329257	1	0.576382	0.329257	0.680445	1	0.329257	0.013038
HA1, N	0.162298	0.576382	1	0.162298	0.86252	0.329257	0.162298	0.042286
NoHA	1	0.329257	0.162298	1	0.171036	0.329257	1	0.007198
HA2, F	0.171036	0.680445	0.86252	0.171036	1	0.680445	0.171036	0.024318
HA2, N	0.329257	1	0.329257	0.329257	0.680445	1	0.329257	0.019258
HA3, F	1	0.329257	0.162298	1	0.171036	0.329257	1	0.007198
HA3, N	0.007198	0.013038	0.042286	0.007198	0.024318	0.019258	0.007198	1

p values from Wilcoxon ranks test

3	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.455038	0.008099	0.000229	0.024902	0.000197	0.459595	0.000646
HA1, F	0.455038	1	0.003713	0.005154	0.184277	0.000139	0.965576	0.000364
HA1, N	0.008099	0.003713	1	0.156344	0.258755	0.015869	0.016906	0.165896
NoHA	0.000229	0.005154	0.156344	1	0.03827	0.215088	0.0013	0.752741
HA2, F	0.024902	0.184277	0.258755	0.03827	1	0.00295	0.035649	0.017059
HA2, N	0.000197	0.000139	0.015869	0.215088	0.00295	1	0.000163	0.076172
HA3, F	0.459595	0.965576	0.016906	0.0013	0.035649	0.000163	1	0.0003
HA3, N	0.000646	0.000364	0.165896	0.752741	0.017059	0.076172	0.0003	1

0	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.390363	0.001101	0.00012	0.188354	0.000123	0.401245	5.59E-05
HA1, F	0.390363	1	0.098694	0.0023	0.758076	0.000647	0.631462	7.61E-05
HA1, N	0.001101	0.098694	1	0.122884	0.075933	0.00626	0.02846	0.000944
NoHA	0.00012	0.0023	0.122884	1	0.005658	0.18054	0.000119	0.007045
HA2, F	0.188354	0.758076	0.075933	0.005658	1	0.000184	0.281372	0.000125
HA2, N	0.000123	0.000647	0.00626	0.18054	0.000184	1	0.000213	0.372243
HA3, F	0.401245	0.631462	0.02846	0.000119	0.281372	0.000213	1	7.85E-05
HA3, N	5.59E-05	7.61E-05	0.000944	0.007045	0.000125	0.372243	7.85E-05	1

-3	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.008667	0.000125	0.00293	0.034912	7.7E-05	0.00205	0.000272
HA1, F	0.008667	1	0.000173	0.219727	0.289673	0.000405	0.325195	0.001229
HA1, N	0.000125	0.000173	1	0.001321	0.000278	0.773888	0.000542	0.701948
NoHA	0.00293	0.219727	0.001321	1	0.085693	0.001135	0.882813	0.002197
HA2, F	0.034912	0.289673	0.000278	0.085693	1	0.000122	0.062988	0.000318
HA2, N	7.7E-05	0.000405	0.773888	0.001135	0.000122	1	0.000581	0.829055
HA3, F	0.00205	0.325195	0.000542	0.882813	0.062988	0.000581	1	0.002106
HA3, N	0.000272	0.001229	0.701948	0.002197	0.000318	0.829055	0.002106	1

-8	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	1	0.5	1	0.5	1	1	0.015625
HA1, F	1	1	1	1	1	1	1	0.03125
HA1, N	0.5	1	1	0.5	1	1	0.5	0.046875
NoHA	1	1	0.5	1	0.5	1	1	0.015625
HA2, F	0.5	1	1	0.5	1	1	0.5	0.039063
HA2, N	1	1	1	1	1	1	1	0.03125
HA3, F	1	1	0.5	1	0.5	1	1	0.015625
HA3, N	0.015625	0.03125	0.046875	0.015625	0.039063	0.03125	0.015625	1

Appendix A - Statistics HASPI

1E-07	Significant Difference with Bonferroni Correction
0.01	Significant Difference with no Bonferroni Correction (t-test only)
0.4	No Significant Difference

p values from t-test

3	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.279076	1.53E-11	2.04E-10	0.00087	1.92E-05	4.8E-05	1.13E-09
HA1, F	0.279076	1	1.62E-12	1.37E-09	0.0067	7.96E-06	5.27E-06	3.8E-09
HA1, N	1.53E-11	1.62E-12	1	0.000157	8.26E-13	2.36E-08	1.24E-08	0.001891
NoHA	2.04E-10	1.37E-09	0.000157	1	1.37E-13	4.68E-05	1.67E-08	0.703304
HA2, F	0.00087	0.0067	8.26E-13	1.37E-13	1	4.67E-10	3.88E-07	5.5E-11
HA2, N	1.92E-05	7.96E-06	2.36E-08	4.68E-05	4.67E-10	1	0.022131	0.00089
HA3, F	4.8E-05	5.27E-06	1.24E-08	1.67E-08	3.88E-07	0.022131	1	4.43E-09
HA3, N	1.13E-09	3.8E-09	0.001891	0.703304	5.5E-11	0.00089	4.43E-09	1

0	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	5.84E-11	8.82E-13	0.045819	1.47E-08	0.710999	6.96E-08	8.42E-08
HA1, F	5.84E-11	1	2.42E-13	8.06E-10	0.17074	1.78E-09	1.66E-06	3.93E-13
HA1, N	8.82E-13	2.42E-13	1	1.69E-07	2.64E-12	2.07E-10	3.07E-12	0.010341
NoHA	0.045819	8.06E-10	1.69E-07	1	7.61E-09	0.133706	9.37E-10	9.85E-07
HA2, F	1.47E-08	0.17074	2.64E-12	7.61E-09	1	2.06E-07	0.000211	1.21E-12
HA2, N	0.710999	1.78E-09	2.07E-10	0.133706	2.06E-07	1	5.35E-06	6.31E-06
HA3, F	6.96E-08	1.66E-06	3.07E-12	9.37E-10	0.000211	5.35E-06	1	8.8E-12
HA3, N	8.42E-08	3.93E-13	0.010341	9.85E-07	1.21E-12	6.31E-06	8.8E-12	1

-3	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	1.4E-06	5.56E-10	0.00228	1.7E-07	0.284269	0.000648	6.88E-09
HA1, F	1.4E-06	1	3.77E-11	3.66E-09	0.007877	2E-06	1.24E-05	6.62E-10
HA1, N	5.56E-10	3.77E-11	1	6.48E-09	2.25E-11	2.27E-07	4.69E-12	6.54E-08
NoHA	0.00228	3.66E-09	6.48E-09	1	2.12E-08	0.287831	1.5E-06	1.56E-06
HA2, F	1.7E-07	0.007877	2.25E-11	2.12E-08	1	1.4E-09	4.85E-07	1.14E-09
HA2, N	0.284269	2E-06	2.27E-07	0.287831	1.4E-09	1	0.000992	0.000225
HA3, F	0.000648	1.24E-05	4.69E-12	1.5E-06	4.85E-07	0.000992	1	7.34E-10
HA3, N	6.88E-09	6.62E-10	6.54E-08	1.56E-06	1.14E-09	0.000225	7.34E-10	1

-8	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	7.32E-06	8.23E-10	4.27E-05	1.92E-08	0.179861	0.001693	2.42E-09
HA1, F	7.32E-06	1	2.22E-12	2.49E-11	0.000585	1.49E-05	1.57E-05	4.25E-11
HA1, N	8.23E-10	2.22E-12	1	1.01E-09	3.02E-12	3.65E-09	3.38E-11	7.08E-09
NoHA	4.27E-05	2.49E-11	1.01E-09	1	4.93E-09	4.92E-05	8.54E-10	0.002144
HA2, F	1.92E-08	0.000585	3.02E-12	4.93E-09	1	3.43E-05	3.42E-06	5.81E-12
HA2, N	0.179861	1.49E-05	3.65E-09	4.92E-05	3.43E-05	1	0.001346	9.91E-05
HA3, F	0.001693	1.57E-05	3.38E-11	8.54E-10	3.42E-06	0.001346	1	9.99E-10
HA3, N	2.42E-09	4.25E-11	7.08E-09	0.002144	5.81E-12	9.91E-05	9.99E-10	1

p values from Wilcoxon ranks test

3	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.162964	0.000122	0.000122	0.000244	0.000244	0.000122	0.000122
HA1, F	0.162964	1	0.000122	0.000122	0.004517	0.000244	0.000122	0.000122
HA1, N	0.000122	0.000122	1	0.000488	0.000122	0.000122	0.000122	0.002808
NoHA	0.000122	0.000122	0.000488	1	0.000122	0.000244	0.000122	0.380493
HA2, F	0.000244	0.004517	0.000122	0.000122	1	0.000122	0.000122	0.000122
HA2, N	0.000244	0.000244	0.000122	0.000244	0.000122	1	0.022827	0.000122
HA3, F	0.000122	0.000122	0.000122	0.000122	0.000122	0.022827	1	0.000122
HA3, N	0.000122	0.000122	0.002808	0.380493	0.000122	0.000122	0.000122	1

0	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.000122	0.000122	0.03186	0.000122	0.895752	0.000122	0.000122
HA1, F	0.000122	1	0.000122	0.000122	0.163086	0.000122	0.000122	0.000122
HA1, N	0.000122	0.000122	1	0.000122	0.000122	0.000122	0.000122	0.002197
NoHA	0.03186	0.000122	0.000122	1	0.000122	0.097656	0.000122	0.000122
HA2, F	0.000122	0.163086	0.000122	0.000122	1	0.000122	0.001587	0.000122
HA2, N	0.895752	0.000122	0.000122	0.097656	0.000122	1	0.000122	0.000122
HA3, F	0.000122	0.000122	0.000122	0.000122	0.001587	0.000122	1	0.000122
HA3, N	0.000122	0.000122	0.002197	0.000122	0.000122	0.000122	0.000122	1

-3	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.000122	0.000122	0.00061	0.000122	0.518677	0.000854	0.000122
HA1, F	0.000122	1	0.000122	0.000122	0.003052	0.000122	0.000366	0.000122
HA1, N	0.000122	0.000122	1	0.000122	0.000122	0.000122	0.000122	0.000122
NoHA	0.00061	0.000122	0.000122	1	0.000122	0.529785	0.000122	0.000122
HA2, F	0.000122	0.003052	0.000122	0.000122	1	0.000122	0.000122	0.000122
HA2, N	0.518677	0.000122	0.000122	0.529785	0.000122	1	0.003784	0.000122
HA3, F	0.000854	0.000366	0.000122	0.000122	0.000122	0.003784	1	0.000122
HA3, N	0.000122	0.000122	0.000122	0.000122	0.000122	0.000122	0.000122	1

-8	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.000122	0.000122	0.00061	0.000122	0.287964	0.000488	0.000122
HA1, F	0.000122	1	0.000122	0.000122	0.001953	0.000366	0.00061	0.000122
HA1, N	0.000122	0.000122	1	0.000122	0.000122	0.000122	0.000122	0.000122
NoHA	0.00061	0.000122	0.000122	1	0.000122	0.000122	0.000122	0.015869
HA2, F	0.000122	0.001953	0.000122	0.000122	1	0.000366	0.000122	0.000122
HA2, N	0.287964	0.000366	0.000122	0.000122	0.000366	1	0.002563	0.00061
HA3, F	0.000488	0.00061	0.000122	0.000122	0.000122	0.002563	1	0.000122
HA3, N	0.000122	0.000122	0.000122	0.015869	0.000122	0.00061	0.000122	1

1E-07	Significant Difference with Bonferroni Correction
0.01	Significant Difference with no Bonferroni Correction (t-test only)
0.4	No Significant Difference

p values from t-test

3	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	1.17E-05	1.13E-08	0.521659	4.85E-13	2.46E-09	1.9E-11	0.000152
HA1, F	1.17E-05	1	1.1E-10	0.004051	1.73E-10	0.005388	3.75E-08	6.13E-09
HA1, N	1.13E-08	1.1E-10	1	1.05E-08	2.32E-15	5.98E-12	5.44E-14	2.05E-06
NoHA	0.521659	0.004051	1.05E-08	1	5.27E-11	1.62E-08	6.27E-12	8.47E-05
HA2, F	4.85E-13	1.73E-10	2.32E-15	5.27E-11	1	8.29E-09	0.395101	5.06E-15
HA2, N	2.46E-09	0.005388	5.98E-12	1.62E-08	8.29E-09	1	5.25E-08	2.55E-09
HA3, F	1.9E-11	3.75E-08	5.44E-14	6.27E-12	0.395101	5.25E-08	1	1.74E-12
HA3, N	0.000152	6.13E-09	2.05E-06	8.47E-05	5.06E-15	2.55E-09	1.74E-12	1

0	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.011409	3.94E-10	0.20948	1.58E-13	1.88E-11	1.1E-10	0.017525
HA1, F	0.011409	1	2.02E-11	0.000249	2.23E-09	0.009297	1.32E-06	5.24E-05
HA1, N	3.94E-10	2.02E-11	1	1.22E-09	3.08E-18	1.29E-14	2.35E-15	4.99E-08
NoHA	0.20948	0.000249	1.22E-09	1	4.63E-15	6.56E-08	4.99E-14	0.018233
HA2, F	1.58E-13	2.23E-09	3.08E-18	4.63E-15	1	3.78E-11	1.48E-08	2.58E-13
HA2, N	1.88E-11	0.009297	1.29E-14	6.56E-08	3.78E-11	1	3.67E-07	4.88E-08
HA3, F	1.1E-10	1.32E-06	2.35E-15	4.99E-14	1.48E-08	3.67E-07	1	1E-12
HA3, N	0.017525	5.24E-05	4.99E-08	0.018233	2.58E-13	4.88E-08	1E-12	1

-3	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	2.73E-08	3.05E-07	0.238931	4.1E-13	4.08E-10	3.13E-11	0.014085
HA1, F	2.73E-08	1	2.07E-16	1.59E-07	1.8E-11	2.15E-05	1.11E-10	3.96E-12
HA1, N	3.05E-07	2.07E-16	1	6.73E-09	1.37E-16	5.44E-15	2.54E-15	1.04E-13
NoHA	0.238931	1.59E-07	6.73E-09	1	4.14E-15	3.12E-08	3.44E-14	0.150015
HA2, F	4.1E-13	1.8E-11	1.37E-16	4.14E-15	1	5.28E-08	0.006306	1.72E-15
HA2, N	4.08E-10	2.15E-05	5.44E-15	3.12E-08	5.28E-08	1	4.25E-07	1.68E-09
HA3, F	3.13E-11	1.11E-10	2.54E-15	3.44E-14	0.006306	4.25E-07	1	8.25E-13
HA3, N	0.014085	3.96E-12	1.04E-13	0.150015	1.72E-15	1.68E-09	8.25E-13	1

-8	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	1.14E-10	9.76E-06	0.001631	4.91E-17	5.75E-15	7.84E-16	1.39E-10
HA1, F	1.14E-10	1	1.59E-11	1.82E-07	1.3E-15	1.25E-12	3.68E-13	0.049619
HA1, N	9.76E-06	1.59E-11	1	8.76E-07	1.2E-18	1.11E-17	2.11E-17	2.85E-13
NoHA	0.001631	1.82E-07	8.76E-07	1	7.16E-18	9.44E-13	4.65E-13	5.83E-11
HA2, F	4.91E-17	1.3E-15	1.2E-18	7.16E-18	1	1.79E-12	5.06E-09	7.07E-17
HA2, N	5.75E-15	1.25E-12	1.11E-17	9.44E-13	1.79E-12	1	1.79E-05	1.02E-11
HA3, F	7.84E-16	3.68E-13	2.11E-17	4.65E-13	5.06E-09	1.79E-05	1	3.55E-12
HA3, N	1.39E-10	0.049619	2.85E-13	5.83E-11	7.07E-17	1.02E-11	3.55E-12	1

p values from Wilcoxon ranks test

3	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.000122	0.000122	0.281738	0.000122	0.000122	0.000122	0.00061
HA1, F	0.000122	1	0.000122	0.013916	0.000122	0.010254	0.000122	0.000122
HA1, N	0.000122	0.000122	1	0.000122	0.000122	0.000122	0.000122	0.000122
NoHA	0.281738	0.013916	0.000122	1	0.000122	0.000122	0.000122	0.00061
HA2, F	0.000122	0.000122	0.000122	0.000122	1	0.000122	0.208496	0.000122
HA2, N	0.000122	0.010254	0.000122	0.000122	0.000122	1	0.000122	0.000122
HA3, F	0.000122	0.000122	0.000122	0.000122	0.208496	0.000122	1	0.000122
HA3, N	0.00061	0.000122	0.000122	0.00061	0.000122	0.000122	0.000122	1

0	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.000122	0.000122	0.452148	0.000122	0.000122	0.000122	0.039063
HA1, F	0.000122	1	0.000122	0.000122	0.000122	0.023193	0.000244	0.000122
HA1, N	0.000122	0.000122	1	0.000122	0.000122	0.000122	0.000122	0.000122
NoHA	0.452148	0.000122	0.000122	1	0.000122	0.000122	0.000122	0.003662
HA2, F	0.000122	0.000122	0.000122	0.000122	1	0.000122	0.000122	0.000122
HA2, N	0.000122	0.023193	0.000122	0.000122	0.000122	1	0.000122	0.000122
HA3, F	0.000122	0.000244	0.000122	0.000122	0.000122	0.000122	1	0.000122
HA3, N	0.039063	0.000122	0.000122	0.003662	0.000122	0.000122	0.000122	1

-3	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.000122	0.000122	0.2323	0.000122	0.000122	0.000122	0.007324
HA1, F	0.000122	1	0.000122	0.000122	0.000122	0.000244	0.000122	0.000122
HA1, N	0.000122	0.000122	1	0.000122	0.000122	0.000122	0.000122	0.000122
NoHA	0.2323	0.000122	0.000122	1	0.000122	0.000122	0.000122	0.163696
HA2, F	0.000122	0.000122	0.000122	0.000122	1	0.000122	0.00293	0.000122
HA2, N	0.000122	0.000244	0.000122	0.000122	0.000122	1	0.000122	0.000122
HA3, F	0.000122	0.000122	0.000122	0.000122	0.00293	0.000122	1	0.000122
HA3, N	0.007324	0.000122	0.000122	0.163696	0.000122	0.000122	0.000122	1

-8	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.000122	0.000122	0.002197	0.000122	0.000122	0.000122	0.000122
HA1, F	0.000122	1	0.000122	0.000122	0.000122	0.000122	0.000122	0.038696
HA1, N	0.000122	0.000122	1	0.000122	0.000122	0.000122	0.000122	0.000122
NoHA	0.002197	0.000122	0.000122	1	0.000122	0.000122	0.000122	0.000122
HA2, F	0.000122	0.000122	0.000122	0.000122	1	0.000122	0.000122	0.000122
HA2, N	0.000122	0.000122	0.000122	0.000122	0.000122	1	0.000244	0.000122
HA3, F	0.000122	0.000122	0.000122	0.000122	0.000122	0.000244	1	0.000122
HA3, N	0.000122	0.038696	0.000122	0.000122	0.000122	0.000122	0.000122	1

Appendix A - Statistics

STOI

1E-07	Significant Difference with Bonferroni Correction
0.01	Significant Difference with no Bonferroni Correction (t-test only)
0.4	No Significant Difference

p values from t-test

3	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.000789	0.000704	0.007353	3.32E-07	0.00038	0.011061	2.02E-08
HA1, F	0.000789	1	2.34E-06	1.1E-05	0.226075	1.21E-06	0.051392	4.08E-09
HA1, N	0.000704	2.34E-06	1	0.117182	4.14E-08	0.560191	6.57E-07	0.133357
NoHA	0.007353	1.1E-05	0.117182	1	3.22E-06	0.165885	5.74E-06	0.024147
HA2, F	3.32E-07	0.226075	4.14E-08	3.22E-06	1	1.15E-08	0.001417	2.96E-14
HA2, N	0.00038	1.21E-06	0.560191	0.165885	1.15E-08	1	7.44E-08	0.050363
HA3, F	0.011061	0.051392	6.57E-07	5.74E-06	0.001417	7.44E-08	1	7.65E-08
HA3, N	2.02E-08	4.08E-09	0.133357	0.024147	2.96E-14	0.050363	7.65E-08	1

0	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	6.4E-06	1.59E-12	5.63E-07	0.784096	1.06E-11	0.000243	1.42E-15
HA1, F	6.4E-06	1	5.36E-11	0.001633	0.019526	3.06E-08	0.971885	8.95E-12
HA1, N	1.59E-12	5.36E-11	1	2.34E-07	3.13E-09	4.23E-06	1.59E-10	0.59482
NoHA	5.63E-07	0.001633	2.34E-07	1	3.85E-08	1.8E-06	1.25E-06	4.08E-09
HA2, F	0.784096	0.019526	3.13E-09	3.85E-08	1	1.63E-09	0.000358	6.95E-11
HA2, N	1.06E-11	3.06E-08	4.23E-06	1.8E-06	1.63E-09	1	7.19E-10	0.00136
HA3, F	0.000243	0.971885	1.59E-10	1.25E-06	0.000358	7.19E-10	1	3.92E-12
HA3, N	1.42E-15	8.95E-12	0.59482	4.08E-09	6.95E-11	0.00136	3.92E-12	1

-3	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.925383	7.03E-11	0.048712	0.406241	7.73E-10	0.961361	8.41E-09
HA1, F	0.925383	1	2.92E-10	6.09E-07	0.036966	2.6E-11	0.92139	3.28E-12
HA1, N	7.03E-11	2.92E-10	1	4.03E-08	3.07E-10	0.016256	4.19E-10	0.01102
NoHA	0.048712	6.09E-07	4.03E-08	1	0.000314	5.26E-09	0.001084	5.07E-11
HA2, F	0.406241	0.036966	3.07E-10	0.000314	1	1.1E-10	0.115469	7.85E-13
HA2, N	7.73E-10	2.6E-11	0.016256	5.26E-09	1.1E-10	1	5.33E-11	0.000642
HA3, F	0.961361	0.92139	4.19E-10	0.001084	0.115469	5.33E-11	1	2.36E-12
HA3, N	8.41E-09	3.28E-12	0.01102	5.07E-11	7.85E-13	0.000642	2.36E-12	1

-8	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	8.02E-08	7.15E-08	0.349042	0.002292	0.052498	0.0057	5.21E-06
HA1, F	8.02E-08	1	6.75E-11	0.033672	0.413461	6.71E-05	0.076516	6.69E-08
HA1, N	7.15E-08	6.75E-11	1	9.96E-13	2.73E-12	1.73E-10	1.7E-13	0.147895
NoHA	0.349042	0.033672	9.96E-13	1	0.017185	4.15E-06	0.041164	7.92E-09
HA2, F	0.002292	0.413461	2.73E-12	0.017185	1	1.95E-06	0.079024	7.52E-09
HA2, N	0.052498	6.71E-05	1.73E-10	4.15E-06	1.95E-06	1	1.02E-06	1.37E-06
HA3, F	0.0057	0.076516	1.7E-13	0.041164	0.079024	1.02E-06	1	1.64E-08
HA3, N	5.21E-06	6.69E-08	0.147895	7.92E-09	7.52E-09	1.37E-06	1.64E-08	1

p values from Wilcoxon ranks test

3	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.002808	0.000122	0.002808	0.000244	0.001709	0.012207	0.000122
HA1, FF	0.002808	1	0.000244	0.000488	0.345459	0.000244	0.028442	0.000122
HA1, N	0.000122	0.000244	1	0.111694	0.000122	0.348267	0.000244	0.320923
NoHA	0.002808	0.000488	0.111694	1	0.000244	0.099365	0.000244	0.017578
HA2, FF	0.000244	0.345459	0.000122	0.000244	1	0.000122	0.001099	0.000122
HA2, N	0.001709	0.000244	0.348267	0.099365	0.000122	1	0.000122	0.053833
HA3, FF	0.012207	0.028442	0.000244	0.000244	0.001099	0.000122	1	0.000122
HA3, N	0.000122	0.000122	0.320923	0.017578	0.000122	0.053833	0.000122	1

0	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.000122	0.000122	0.000122	0.844238	0.000122	0.000854	0.000122
HA1, FF	0.000122	1	0.000122	0.00061	0.008301	0.000122	0.985107	0.000122
HA1, N	0.000122	0.000122	1	0.000244	0.000122	0.000122	0.000122	0.613525
NoHA	0.000122	0.00061	0.000244	1	0.000122	0.000244	0.000122	0.000122
HA2, FF	0.844238	0.008301	0.000122	0.000122	1	0.000122	0.000488	0.000122
HA2, N	0.000122	0.000122	0.000122	0.000244	0.000122	1	0.000122	0.000854
HA3, FF	0.000854	0.985107	0.000122	0.000122	0.000488	0.000122	1	0.000122
HA3, N	0.000122	0.000122	0.613525	0.000122	0.000122	0.000854	0.000122	1

-3	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.891235	0.000122	0.113159	0.700928	0.000122	0.891602	0.000122
HA1, FF	0.891235	1	0.000122	0.000122	0.039063	0.000122	0.744873	0.000122
HA1, N	0.000122	0.000122	1	0.000122	0.000122	0.008179	0.000122	0.004761
NoHA	0.113159	0.000122	0.000122	1	0.000122	0.000122	0.000244	0.000122
HA2, FF	0.700928	0.039063	0.000122	0.000122	1	0.000122	0.051147	0.000122
HA2, N	0.000122	0.000122	0.008179	0.000122	0.000122	1	0.000122	0.000488
HA3, FF	0.891602	0.744873	0.000122	0.000244	0.051147	0.000122	1	0.000122
HA3, N	0.000122	0.000122	0.004761	0.000122	0.000122	0.000488	0.000122	1

-8	Cheap	HA1, FF	HA1, NR	NoHA	HA2, FF	HA2, NR	HA3, FF	HA3, NR
Cheap	1	0.000122	0.000122	0.287842	0.008179	0.037842	0.009521	0.000122
HA1, FF	0.000122	1	0.000122	0.039063	0.451416	0.000122	0.046509	0.000122
HA1, N	0.000122	0.000122	1	0.000122	0.000122	0.000122	0.000122	0.568848
NoHA	0.287842	0.039063	0.000122	1	0.045166	0.000244	0.046021	0.000122
HA2, FF	0.008179	0.451416	0.000122	0.045166	1	0.000122	0.084717	0.000122
HA2, N	0.037842	0.000122	0.000122	0.000244	0.000122	1	0.000122	0.000122
HA3, FF	0.009521	0.046509	0.000122	0.046021	0.084717	0.000122	1	0.000122
HA3, N	0.000122	0.000122	0.568848	0.000122	0.000122	0.000122	0.000122	1

Appendix A - Statistics

Comparing Subjective with Objective Predictions

1E-07	Significant Difference with Bonferroni Correction
0.01	Significant Difference with no Bonferroni Correction (t-test only)
0.4	No Significant Difference

p values from t-test

HASPI	3	0	-3	-8
Cheap	0.117795691	0.005757584	9.69335E-08	1.89062E-19
HA1, FF	0.001350851	0.021002095	0.828825385	1.32089E-05
HA1, NR	0.601634636	0.016309876	0.745002974	2.86941E-09
NoHA	0.050871611	0.004985659	0.101395944	7.73343E-22
HA2, FF	9.25249E-06	0.011061156	0.806554238	0.002106407
HA2, NR	1.14874E-07	0.000162304	0.000290068	8.95503E-09
HA3, FF	0.263519001	0.67262438	0.744075184	2.88349E-27
HA3, NR	0.003916101	0.000254889	0.176730046	0.534260014

STOI	3	0	-3	-8
Cheap	0.003383402	0.000604307	3.4741E-12	9.95285E-22
HA1, FF	0.027444921	0.002435317	0.000133366	2.64401E-11
HA1, NR	0.264377727	0.000314339	0.098865818	4.86067E-13
NoHA	0.565663842	0.845249326	0.000222924	7.89917E-20
HA2, FF	0.926767048	0.040586529	1.38358E-06	2.88796E-09
HA2, NR	0.017007974	0.891829347	0.026883277	7.46817E-12
HA3, FF	0.002983405	0.000369211	0.00014509	7.27763E-23
HA3, NR	0.626103041	0.43056973	0.004632576	0.004239418

CSII	3	0	-3	-8
Cheap	0.000702531	4.80565E-06	4.1698E-12	4.56848E-25
HA1, FF	3.44618E-05	0.000250092	1.34984E-10	1.83473E-21
HA1, NR	0.241275387	0.123163645	0.028743971	3.34189E-16
NoHA	0.538332044	0.261146443	2.0545E-06	8.41934E-25
HA2, FF	0.001392177	7.58375E-08	1.29109E-14	1.37495E-25
HA2, NR	0.101659323	0.461077146	0.00575066	2.6249E-24
HA3, FF	4.01497E-07	5.98032E-08	8.88742E-14	1.10442E-31
HA3, NR	0.04383053	2.25613E-05	0.77120752	3.8527E-09

p values from Wilcoxon ranks test

HASPI	3	0	-3	-8
Cheap	0.130225025	0.00781955	9.08611E-05	5.49674E-05
HA1, FF	0.001296365	0.032457851	0.613681613	0.00098487
HA1, NR	0.689100809	0.027178383	0.767623677	8.55232E-05
NoHA	0.022736689	0.00509519	0.049380486	5.49674E-05
HA2, FF	0.000121951	0.010575384	0.663445442	0.00680135
HA2, NR	7.91101E-05	0.001015896	0.000688797	0.00013151
HA3, FF	0.180445809	0.614164258	0.688877677	5.49674E-05
HA3, NR	0.017219908	0.000894986	0.192367016	0.374831319

STOI	3	0	-3	-8
Cheap	0.004106137	0.001652272	5.86566E-05	5.49674E-05
HA1, FF	0.027224883	0.00411742	0.000467075	6.54828E-05
HA1, NR	0.589726521	0.001291975	0.121875532	5.50394E-05
NoHA	0.105859795	0.794124607	0.001644226	5.49674E-05
HA2, FF	0.903138413	0.032431654	7.77397E-05	8.54168E-05
HA2, NR	0.011693456	0.821180307	0.038339147	6.38237E-05
HA3, FF	0.004103319	0.001466228	0.000409667	5.49674E-05
HA3, NR	0.875654421	0.304868881	0.014254529	0.001137843

CSII	3	0	-3	-8
Cheap	0.001143774	0.000140599	5.86566E-05	5.49674E-05
HA1, FF	0.000319051	0.001015001	5.79039E-05	5.64944E-05
HA1, NR	0.180445809	0.105754267	0.053688549	5.50394E-05
NoHA	0.079056008	0.258131018	0.000160615	5.49674E-05
HA2, FF	0.004123068	5.88839E-05	5.79789E-05	5.49674E-05
HA2, NR	0.192199036	0.566129396	0.00565027	5.50394E-05
HA3, FF	6.79472E-05	6.86394E-05	5.79039E-05	5.49674E-05
HA3, NR	0.105886175	0.000243095	0.821233426	5.82793E-05

Appendix A - Statistical Details

Additional Statistics

1E-07	Significant Difference with Bonferroni Correction or Significant Correlation
0.01	Significant Difference with no Bonferroni Correction (t-test only)
0.4	No Significant Difference or correlation

Gender (M/F) - p values from uncorrelated t-test

	3	0	-3	-8
Cheap	0.818582	0.831092	0.780066	
HA1, FF	0.462311	0.426501	0.518108	0.306184
HA1, NR	0.104377	0.93765	0.35815	0.172398
NoHA	0.976438	0.887729	0.895632	
HA2, FF	0.677041	0.439504	0.416969	0.834845
HA2, NR	0.628745	0.987672	0.760177	0.353452
HA3, FF	0.860885	0.823564	0.439164	
HA3, NR	0.932516	0.720129	0.371254	0.63284

Ear Tested (L/R) - p values from uncorrelated t-test

	3	0	-3	-8
Cheap	0.915564	0.051678	0.5972	
HA1, FF	0.092629	0.521929	0.185155	0.639785
HA1, NR	0.099594	0.971142	0.944855	0.000916
NoHA	0.925172	0.173213	0.64549	
HA2, FF	0.816295	0.237965	0.360131	0.505065
HA2, NR	0.733035	0.719396	0.149839	0.035448
HA3, FF	0.885339	0.338883	0.322679	
HA3, NR	0.741479	0.679737	0.821861	0.151148

Correlation coefficient between Age and Score (Significance at 1% > 0.55 or <-0.55, for 21 pairs)

	3	0	-3	-8
Cheap	-0.05651	-0.17892	-0.27356	
HA1, FF	-0.03312	0.016904	0.042481	-0.02163
HA1, NR	-0.12594	0.152948	-0.05836	0.373054
NoHA	0.377122	-0.11741	-0.1246	
HA2, FF	0.219266	0.307227	0.054595	-0.20662
HA2, NR	0.230777	0.182278	0.243158	-0.02163
HA3, FF	0.250254	-0.31409	-0.11681	
HA3, NR	0.391744	-0.3145	0.212667	-0.01021

Bibliography

Acoustical Society of America. Speech audiometry, 2022. URL https://asastandards.org/asa-standard-term-database/. Accessed 5th March 2022.

Action on Hearing Loss. Cochlear implants - factsheet, 2016. URL https://www.actiononhearingloss.org.uk/how-we-help/information-and-resources/publications/hearing-aids-and-cochlear-implants/.

Action on Hearing Loss. Technology used to aid hearing in background noise: Insight from people with hearing loss. Report, Action on Hearing Loss, 2017a. URL https://www.actiononhearingloss.org.uk/how-we-help/information-and-resources/publications/research-reports/.

Action on Hearing Loss. Age-related hearing loss, 2017b. URL https://www.actiononhearingloss.org.uk/hearing-health/hearing-loss-and-deafness/types-and-causes/age-related-hearing-loss/. Accessed 21st March 2018.

Action on Hearing Loss. Hearing aids - factsheet, 2018a. URL https://www.actiononhearingloss.org.uk/how-we-help/information-and-resources/publications/hearing-aids-and-cochlear-implants/.

Action on Hearing Loss. Products and technology to help with hearing loss, 2018b. URL https://www.actiononhearingloss.org.uk/live-well/products-and-technology/products-and-technology-to-help-with-hearing-loss/. Accessed 16th November 2018.

Manal Alfakhri. Speech audiometry, 2012. URL http://fac.ksu.edu.sa/sites/default/files/Speech_Audiometry.pdf. Accessed 13th June 2018.

American National Standards Institute. S1.1-1960, acoustical terminology, 1960.

American National Standards Institute. Methods for calculation of the speech intelligibility index, 1997.

Marc Aubreville and Stefan Petrausch. Measuring directionality of modern hearing aids. *Canadian Audiologist*, 2(3), 2015.

John Bench, Ase Kowal, and John Bamford. The BKB (Bamford-Kowal-Bench) Sentence Lists for Partially-Hearing Children. *British Journal of Audiology*, 13(3):108–112, 1979.

R.A Bentler. Digital noise reduction: An overview. *Trends in Amplification*, 10(2): 67–82, 2006.

Virginia Best, Christine R. Mason, Jayaganesh Swaminathan, Elin Roverud, and Gerald Jr. Kidd. Use of a glimpsing model to understand the performance of listeners with and without hearing loss in spatialized speech mixtures. *Journal of the Acoustical Society of America*, 141(1):81–91, 2017.

Nikolai Bisgaard, Marcel S. M. G. Vlaming, and Martin Dahlquist. Standard audiograms for the iec 60118-15 measurement procedure. *Trends in Amplification*, 14(2):113–120, 2010.

Arthur Boothroyd, Laurie Hanin, and Theresa Hnath. A sentence test of speech perception: reliability, set equivalence, and short term learning. *CUNY Academic Works*, 1985.

A. R. Bradlow, G. M. Torretta, and D. B. Pisoni. Intelligibility of normal speech i: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20 (3-4):225–272, 1996.

T. Brand and B. Kollmeier. Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *Journal of the Acoustical Society of America*, 111(6):2801, 2002.

J. Bridges, A. Lataille, C. Buttorff, S. White, and J. Niparko. Consumer preferences for hearing aid attributes: A comparison of rating and conjoint analysis methods. *Trends in Amplification*, 16(1):40–48, 2012.

British Society of Audiology. Recommended procedure: Pure-tone air-conduction and bone-conduction threshold audiometry with and without masking, 2011.

British Society of Audiology. Practice guidance: Assessment of speech understanding in noise in adults with hearing difficulties, 2019.

Inge Brons, Rolph Houben, and Wouter A. Dreschler. Perceptual effects of noise reduction by time-frequency masking of noisy speech. *Journal of the Acoustical Society of America*, 132:26902699, 2012.

Wade W. Chien, Howard Francis, and John K. Niparko. *Chapter 49 - Hearing Prosthetics: Surgical Techniques*, pages 588–596. 2012.

R. F. Coelho and V. H. Nascimento. *Chapter 5 - Acoustic Signal Processing*. 2016.

H. Dillon. *Hearing Aids*. Thieme Publishers, 2001.

C. Doire, M. Brookes, and P. A. Naylor. Robust and efficient bayesian adaptive psychometric function estimation. *Journal of the Acoustical Society of America*, 141(4): 2501, 2017.

Etymōtic Research Inc. *BKB-SIN Speech-in-Noise Test - User Manual*, 2018a. URL https://www.etymotic.com/downloads/dl/file/id/260/product/160/bkb_sintm_user_manual.pdf. Accessed 27th February 2018.

Etymōtic Research Inc. *QuickSIN Speech-in-Noise Test - User Manual*, 2018b. URL https://www.etymotic.com/downloads/dl/file/id/259/product/159/quicksin_user_manual.pdf. Accessed 27th February 2018.

T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie. Objective quality and intelligibility prediction for users of assistive listening devices. *IEEE Signal Processing Magazine*, 32(2):114–124, 2015.

John R. Foster and Mark P. Haggard. The four alternative auditory feature test (faaf) - linguistic and psychometric properties of the material with normative data in noise. *British Journal of Audiology*, 21(3):165–174, 1987.

N. R. French and J.C. Steinberg. Factors governing the intelligibility of speech sounds. *The Journal of the Acoustical Society of America*, 19:90–119, 1947.

R. L. Goldsworthy and J. E. Greenberg. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. *Journal of the Acoustical Society of America*, 116(6):3679–3689, 2004.

Simon Graf, Tobias Herbig, Markus Buck, and Gerhard Schmidt. Features for voice activity detection: a comparative analysis. *EURASIP Journal on Advances in Signal Processing*, 2015(1):91, 2015.

G.R.A.S. Sound and Vibration. The life of KEMAR, 2013. URL http://kemar.us/. Accessed 8th May 2018.

V. Hazan and D. Markham. Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America*, 116(5):3108–3118, 2004.

Healthwise Inc. Styles of hearing aids, 2017. URL https://myhealth.alberta.ca/Health/pages/conditions.aspx?hwid=zm6102. Accessed 13th February 2018.

E. Healy and S. Yoho. An algorithm to improve speech recognition in noise for hearing-impaired listeners. *Journal of the Acoustical Society of America*, 134(4):3029–3028, 2013.

Hearing Link. *Facts About Deafness and Hearing Loss*, 2018. URL https://www.hearinglink.org/your-hearing/about-hearing/facts-about-deafness-hearing-loss/. First accessed 29th November 2017, last updated 2018.

Gaston Hilkhuysen. Effects of noise suppression on intelligibility: Dependency on signal-to-noise ratios. *Journal of the Acoustical Society of America*, 131:531, 2012.

S. Hochmuth, T. Brand, M. Zokoll, F. Z. Castro, N. Wardenga, and B. Kollmeier. A spanish matrix sentence test for assessing speech reception thresholds in noise. *International Journal of Audiology*, 51(7):536–544, 2012.

HörTech gGmbH. *International Matrix Tests - Reliable Speech Audiometry in Noise*, 2018. URL http://www.hoertech.de/images/hoertech/pdf/mp/Internationale-Matrixtests.pdf. Accessed 27th February 2018.

Yi Hu and Philipos C. Loizou. A comparative intelligibility study of single-microphone noise reduction algorithms. *Journal of the Acoustical Society of America*, 122(3), 2007.

International Telecommunication Union. ITU-T P.56 12/2011 - Objective measurement of active speech level, 2011.

Jesper Jensen and Cees H. Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24(11), 2016.

Arehart K. H. Anderson M. A. Muralimanohar R. K. Kates, J. M. and L. O. Harvey Jr. Using objective metrics to measure hearing aid performance. *Ear and Hearing*, 39(6): 1165–1175, 2018.

J. Kates. Digital hearing aids, 2008.

J. M. Kates. An auditory model for intelligibility and quality predictions. In *Proceedings of Meetings on Acoustics, Acoustical Society of America*, volume 19, page 3560, 2013a.

J. M. Kates. Users' Guide: HASPI Version 1, HASQI Version 2, HAAQI Version 1, 2013b.

J. M. Kates. Modeling the effects of single-microphone noise-suppression. *Speech Communication*, 90:15–25, 2017.

J. M. Kates and K. H. Arehart. Coherence and the speech intelligibility index. *Journal of the Acoustical Society of America*, 117(4):2224–2237, 2005.

J. M. Kates and K. H. Arehart. The Hearing-Aid Speech Perception Index (HASPI). *Speech Communication*, 65:75–93, 2014.

G. Keidser, H. Dillon, M. Flax, T. Ching, and S. Brewer. The nal-nl2 prescription procedure. *Audiology Research*, 1(1):88–90, 2011.

M. C. Killion. Comparison of vu-meter-based and rms-based calibration of speech levels. *Journal of the Acoustical Society of America*, 126(3):EL97–9, 2009.

M. C. Killion, P. A. Niquette, G. I. Gudmundsen, L. J. Revit, and S. Banerjee. Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 119(3):1888–1888, 2006.

B. Kollmeier, A. Warzybok, S. Hochmuth, M. Zokoll, V. Uslar, T. Brand, and K. Wagener. The multilingual matrix test: Principles, applications, and comparison across languages: A review. *International Journal of Audiology*, 54(Suppliment 2):3–16, 2015.

Francis Kuk, Chi-Chuen Lau, Petri Korhonen, Bryan Crose, Heidi Peeters, and Denise Keenan. Development of the orca nonsense syllable test. *Ear and Hearing*, 31(6): 779–795, 2010.

Ying-Hui Lai and Wei-Zhong Zheng. Multi-objective learning based speech enhancement method to increase speech quality and intelligibility for hearing aid device users. *Biomedical Signal Processing and Control*, 48:35–45, 2019.

Ning Li and Philipos C. Loizou. Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction. *Journal of the Acoustical Society of America*, 123(3):1673, 2008.

Philipos C. Loizou. *Speech Enhancement - Theory and Practice, 2nd Ed.* Taylor and Francis, 2007.

Jianfen Ma, Yi Hu, and Philipos C. Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *Journal of the Acoustical Society of America*, 125:3387–3405, 2009.

Andreas Markides. *Chapter 9 - Speech tests of hearing for children*, pages 1–33. 1997.

D. R. McCloy, R. A. Wright, and P. E. Souza. Talker versus dialect effects on speech intelligibility: A symmetrical study. *Language and Speech*, 58(3):371–386, 2015.

H. Meister, I. Lausberg, J. Kiessling, M. Walger, and H. von Wedel. Determining the importance of fundamental hearing aid attributes. *Otology and Neurotology*, 23(4): 457–462, 2002.

Il Joon Moon and Sung Hwa Hong. What is temporal fine structure and why is it important? *Korean Journal of Audiology*, 18(1):1–7, 2014.

Farid Moshgelani, Vijay Parsa, Chris Allan, Sangamanatha Ankmnal Veeranna, and Prudence Allen. Objective and subjective assessment of envelope enhancement algorithms for assistive hearing devices. *Biomedical Signal Processing and Control*, 47: 16–25, 2019.

H. G. Mueller, J. Weber, and B. W. Y. Hornsby. The effects of digital noise reduction on the acceptance of background noise. *Trends in Amplification*, 10(2):83–93, 2006.

Patrick Naylor. Personal communication on development of matrix tests, 2018.

Andrew Ng. Machine learning - coursera course material, week 2 - gradient descent, 2020. URL https://www.coursera.org/learn/machine-learning/home/info. Accessed 5th March 2022.

M. Nilsson, S. D. Soli, and J. A. Sullivan. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America*, 95(2):1085–1099, 1994.

P Niquette, J Arcaroli, L Revit, A Parkinson, S Staller, M Skinner, and M Killion. Development of the BKB-SIN Test. 2003.

R. Plomp. The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation transfer function. *Journal of the Acoustical Society of America*, 83:2322–7, 1988.

J. Ramirez, J. M. Gorriz, and J. C. Segura. *Voice Activity Detection. Fundamentals and Speech Recognition System Robustness.* INTECH Open Access Publisher, 2007.

P Reinhart, P Zahorik, and P Souza. Interactions between digital noise reduction and reverberation: Acoustic and behavioral effects. *Journal of the American Academy of Audiology*, 2020.

Urmila Shrawankar and Vilas Thakare. Noise estimation and noise removal techniques for speech recognition in adverse environment. In *International Conference on Intelligent Information Processing V. IIP 2010.*, pages 336–342, Berlin, 2010. Springer.

H. Steeneken and T. Houtgast. A physical method for measuring speech transmission quality. *Journal of the Acoustical Society of America*, 67:318–326, 1980.

C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 4214–4217, New York, 2010. IEEE.

Yan Tang, Christopher Arnold, and Trevor Cox. A study on the relationship between the intelligibility and quality of algorithmically-modified speech for normal hearing listeners. *Journal of Otorhinolaryngology, Hearing and Balance Medicine*, 1(5), 2018.

The Ear Foundation. The real cost of adult hearing loss. Report, The Ear Foundation, 2014. URL https://www.earfoundation.org.uk/research/adult-strategy-reports/the-real-cost-of-adult-hearing-loss-2014.

The MathWorks Inc. Matlab r2021a software, 2021.

M. Valente. *Hearing Aids: Standards, Options and Limitations. 2nd Ed.* Thieme Publishers, 2002.

S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks. An evaluation of intrusive instrumental intelligibility metrics. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11), 2018.

Saeed V. Vaseghi. *Chapter 11 - Spectral Subtraction*, pages 749–752. John Wiley & Sons, Ltd., New York, 2001.

Andrew J. Vermiglio. The american english hearing in noise test. *International Journal of Audiology*, 47(6):386–387, 2008.

R. H. Wilson. Development of a speech-in-multitalker-babble paradigm to assess word-recognition performance. *Journal of the American Academy of Audiology*, 14:465, 2003.

R. H. Wilson, T. S. Bell, and J. A. Koslowski. Learning effects associated with repeated word-recognition measures using sentence materials. *Journal of Rehabilitation Research and Development*, 40(4):329–336, 2003.

R. H. Wilson, R. A. McArdle, and S. L. Smith. An evaluation of the bkb-sin, hint, quicksin, and win materials on listeners with normal hearing and listeners with hearing loss. *Journal of Speech Language and Hearing Research*, 50(4):844–856, 2007.

A. Winkler, M. Latzel, and I. Holube. Open versus closed hearing-aid fittings: A literature review of both fitting approaches. *Trends in Hearing*, 20:1–13, 2016.

Jan Wouters, Luc Litière, and Astrid van Wieringen. Speech intelligibility in noisy environments with one- and two-microphone hearing aids. *International Journal of Audiology*, 38(2):91–98, 2009.

Richard Wright. *Chapter 1 - Basic properties of speech*, pages 1–33. 1997.

S. Yoo, R. Boston, J. D. Durrant, K. Kovacyk, S. Karn, S. Shaiman, A. El-Jaroudi, and C-C Li. Relative energy and intelligibility of transient speech components. In *Proceedings of EUSIPCO*, Vienna, 2004.

Stephen A. Zahorian and Martin Rothenberg. Principal-components analysis for low-redundancy encoding of speech spectra. *Journal of the Acoustical Society of America*, 69:832–845, 1981.