To generate evidence on the effectiveness of wireless streaming technologies for deaf children, compared to radio aids

A project commissioned by the National Deaf Children's Society.

Specified personnel :

Michael Stone (PI, ManCAD)

Harvey Dillon (ManCAD)

Helen Chilton (ManCAD)

Helen Glyde (ManCAD)

James Mander (Ewing Foundation)

Additional contributing personnel :

Melanie Lough, (ManCAD)

Helen Whiston (ManCAD)

Keith Wilbraham (ManCAD)

Manchester Centre for Audiology and Deafness (ManCAD)

Background

In response to a request for proposal issued by the NDCS, based on the hypothesis that "use of wireless streaming technologies is equally effective for deaf children as a radio aid", Manchester Centre for Audiology and Deafness (ManCAD) was commissioned to execute a program of agreed work, the "work package".

The agreed work was a technical, as well as a user, assessment of the features of a group of five devices in common use with Deaf Children, as of late 2019. Just as the project was about to start (March 2020), the UK went into lockdown due to COVID-19, causing the project to be suspended. By the functional end of lockdown (July 2021), two of the devices proposed to be assessed had been rendered obsolete by their manufacturers: the Comfort Audio 'Digisystem' and the base Phonak 'Roger' system. By agreement with Ian Noon, NDCS, these two systems were substituted with the Oticon 'EduMic' and the Phonak 'Compilot II'. The remaining three systems were from the 'Roger' range; the Touchscreen, the Select and the Pen.

The work package offered was designed as an introduction to, and exploration of the feasibility of, assessing the devices with a view to setting up a routine test suite. Given the budget constraints, it was never intended as a complete investigation, more as a "proof of concept".

The package involved two aspects: (1) a set of technical measures of the electro-acoustic performance of each device and (2) a hands-on session with likely users of the devices.

Executive Summary

From a technical perspective, the five devices were all capable of delivering speech signals likely to produce high *intelligibility*. However, especially in background noise, larger differences were found between the devices in terms of speech *quality*, and hence potential user acceptability.

Large differences were also found in the ability of the systems to "focus" on the talker and reject background noise arriving from away from the talker.

Aside from the expected non-linear processing, other aspects of non-linear processing which could not be de-activated in these devices meant that some technical measures did not produce meaningful objective results. The use of open-source software models of speech intelligibility and quality appears to be a useful substitute for these measures since they can cope with the non-linear processing.

Feedback was canvassed from five different types of users of the devices on aspects of the device physical design, support such as ancillary equipment and instruction manuals, and "in-field" usability. Of the Phonak Roger systems, the Touchscreen was particularly commended, especially for its ability to convey clear information to the user rather than via a series of lights employing different colours, flash patterns or number alight. However, the in-field use received some negative comments, especially concerning component costs. The Oticon EduMic received positive comments on its design and ergonomics, and, although cheaper than the Roger systems, showed up well in the technical assessment, differing its much longer signal delay between microphone and listener. A similar delay was also observed with the Compilot.

On the basis of limited evidence from lip-reading studies, the measured delays between microphone and hearing aid are acceptable. However, in future, more devices will use versions of the Bluetooth data wireless transmission protocol, known for its introduction of long delays in some implementations. There is a scientific need to identify the upper limit of an acceptable delay, especially in the use case of children acquiring language via degraded audio cues. There is also a need for manufacturers to specify the delays introduced by their devices on their data sheets.

Aspects of the Compilot physical design were not appreciated for use with younger children.

Glossary/Abbreviations

DRC : Dynamic Range Compression

EIN : Equivalent Input Noise

HA : hearing aid

HASPI : Hearing Aid Speech Performance Index

HASQI : Hearing Aid Speech Quality Index

ISTS : International Speech Test Signal

MUT : microphone system under test

SNR : signal-to-noise-ratio

THD : Total Harmonic Distortion

Contents

Supplementary Content *(available as a separate .xlsx file)*

Tabulation of responses to individual questions, as well as the free-text feedback from likely device-users summarised in section 4 above, is available as a separate spreadsheet. Summary rankings based on these answers are also derived therein.

1. Introduction

Remote microphones are a very useful tool for where speech communication is difficult, such as with, but not limited to, deaf people. They enable the capture of an audio source close to the source, and enable delivery to a remote location. By doing so they mostly bypass the degrading effects of the acoustic conditions commonly present between source and listener. The delivery of a higher-fidelity signal at the remote location enables either higher intelligibility, or reduced listening effort, or a combination of both. This is especially useful to those whose hearing abilities are degraded.

Remote microphones are commonly used in educational settings, where large classroom sizes, large numbers of competing audio sources, remote location of the listener from the teacher, and often poor acoustics all contribute to the potential for degraded intelligibility.

There are multiple methods of transmission of the microphone signal to the remote listener. When first developed, there were wires, but the technology progressed to wireless, attaching the analogue audio signal to a 'FM' (Frequency Modulation) radio carrier. Although optical (e.g. infra-red) transmission of the microphone signal can be used, it is more common to use radio carriers. However, the ability to convert the analogue microphone signal into a digital data stream opens up a wide choice of data packaging techniques that each use different (incompatible) radio carriers. Examples of such "streaming" techniques include Bluetooth and WiFi, as well as manufacturers' proprietary systems. In practice, the packaging technique is irrelevant, provided that it can reliably deliver good quality audio to the listener with only a short delay between source and listener.

Part of this report therefore deals with a technical assessment of the ability to deliver good quality audio to the listener. The measurement technique involves the simulation of a talker standing in the middle of a moderate sized, low-reverberation, room while wearing a microphone which is transmitting to a remote listener. The listener consists of a decoder connected to a hearing aid (HA). The HA is delivering audio into a simulated human ear which itself contains a high quality microphone. Recording from this microphone can then be used to compare the loss of quality from the original signal.

Since remote microphones are a complex piece of technology, and are generally used by people unfamiliar with such technology, then good documentation, as well as good device design are important to ensure easy and persistent use of the systems. A second part of this report therefore details more subjective feedback from potential users of these systems, such as audiologists, teachers of the Deaf, parents and also deaf children.

The devices to be compared are the 'Roger Touchscreen' (abbreviated to 'Touch'), 'Roger Pen', 'Roger Select', all produced by Phonak, as well the Oticon EduMic and the Phonak Compilot II. This latter device is not a remote microphone, but a remote receiver system, which, for the tests reported here, converted the Roger X receiver output (transmitted from a Roger device) into audio data for transmission to a local HA. All the remaining systems have discreet microphones which transmit to manufacturer-specific receivers attached to HAs. Ideally, in order to provide a fair comparison between systems, one would intend to intercept the transmitted signal before it entered the HA, since this in itself will introduce distortions of its own. Unfortunately, this is not possible across manufacturers. Although Phonak offer a 'checker', which would obviate the need for a receiving HA, this was not compatible with the ComPilot, and we also could not get an Oticon 'checker'. Hence the radio microphone devices were compared by recording the output of a system-compatible HA in a 711 (ear simulator) coupler.

Collectively, we refer to the microphone systems under test as 'MUTs'.

2. Technical Measures : Methods

2.1 Equipment setup

The torso of a KEMAR manikin was set up in the middle of a large 'listening room'. This room has dimensions 3.5 W x 4.9 L x 2.8 H metres. The walls and ceiling are covered with sound-attenuating slabs, arranged in a semi-random pattern. Reverberation time is uniformly flat across the 125 to 8,000 Hz range, around 120 msec. See Appendices 1 and 2 for photographs of the setup.

An adaptor plate was made that could carry a "bookshelf" loudspeaker, the KEF Q150 (30 x 28 x 18 cm. This plate was attached to the neck of the torso, in place of the head. The shelf enabled the loudspeaker to be laid on its side, at approximately the same relative distance from the mouth to the (usually) high-chest-mounted remote microphone of the system under test. The Q150 has a "dual-concentric" loudspeaker design: it has two loudspeakers mounted on the same axis. One loudspeaker, the 'woofer', covers the range up to about 2.5 kHz, while the other loudspeaker, the 'tweeter' cover the range above 2.5 kHz. The entire frequency range of the sound therefore comes from a point source, similar to a mouth. The woofer diameter was 5.25", while the tweeter diameter was 1". The aim of a loudspeaker is to provide sound mostly in front of the cones so that mostly directly transmitted sound arrives at the listener. The diameter of the loudspeaker determines the frequency range at which the cone becomes very directional in its transmission. There are several (old) studies that report the relative frequency content of speech at different positions around,

compared to in front of, the mouth, and these show generally an increasing drop-off of high-frequency content the further off-axis the measurement point is compared to the axis of the mouth. This loss of content is similar to that which occurs around a loudspeaker compared to its on-axis response. Although 'head and torso simulators' can come with 'mouth simulators' (a) the mouth simulator is mainly intended for use with microphones mounted close to the mouth, such as is common with telephones or headsets (b) the loudspeaker in the mouth simulator is a compromise on size and quality, leading to distortion issues compared to the high-fidelity Q150, (c) they are very expensive (ca £15,000 full equipped), and (d) ManCAD does not own one.

The Q150 has a very-near-flat acoustic response from below 100 Hz to above 10 kHz, and, combined with low distortion, is therefore well suited for both speech and music reproduction.

Two other Q150s were setup: one at 2 metres in front of the manikin (AHEAD) and one at 1 metre behind (REAR) of the manikin. Normally one uses loudspeakers at 1 metre distance, but reflections from the AHEAD loudspeaker could come back at an unacceptable level at the chest-worn microphone, confounding the measure there. The AHEAD and REAR loudspeakers were set up to produce a sound level of 65 dB(A) SPL in the middle of the manikin(with the manikin absent). The manikin-mounted loudspeaker (MOUTH) produced 65 dB(A) SPL at a distance of one metre in front of the loudspeaker. The test signal for this measure was a speech-spectrum random noise, matched to the spectrum of the ISTS speech signal, a signal commonly found in hearing instrument test boxes. The horizontal axis of all three loudspeakers at their cone centres were aligned, 1.5 metres off the floor.

Loudspeakers AHEAD and REAR were intended for the delivery of interfering background noise, as well as a rough measure of relative directionality of the remote microphone when attached to a torso.

Pictures of the equipment set up in the listening room are given in Appendices 1 and 2.

The reference level of 65 dB(A) was also used for the speech level from the manikin loudspeaker, measured at 1 metre away from the centre of the manikin. This is equivalent to 68 dB SPL (unweighted), defined as "raised" in ANSI S3.5 (ANSI, 1997): slightly higher than is usually referenced in a hearing instrument testing, but similar to what a teacher may produce in a classroom when "projecting" their voice..

2.2 Acoustic measures

The proposed acoustic measures were based on the IEC 60118 standards used for the testing of HAs specifically 60118-0, (IEC, 2015). These measures comprised :

(1) Frequency response across the range 125 to 10,000 Hz using the ISTS speech test signal, as well as a random noise with the same spectrum as the ISTS signal. The frequency range of 400 to 4,500 Hz is especially important for speech intelligibility, but there is evidence that bandwidth up to 10 kHz is of importance to children.

(2) Throughput delay : the time taken for the signal to travel from the radio microphone to the ear of the receiver. For this measure the test signal was intended to be bipolar click pulses, but several of the systems heavily attenuated such brief signals. Hence measures are obtained by correlation between the signal recorded next to the radio microphone and the signal recorded in the receiver's ear.

(3) Total Harmonic Distortion (THD)

THD measures the amount of distortion generated by the system under test. In HAs, levels exceeding 2% are regarded as "poor" since they will degrade the likely resulting clarity, and hence, intelligibility. 60118-0 requires tests with pure tones at 500, 800, 1600 and 3200 Hz. These will work in a test box, but in a room, the acoustic "modes" of the room (despite its low reverberation) could lead to effects highly dependent on position of the microphone. Hence narrow bands of noise are preferred. We used 1/3rd-octave bands of low-noise noise centred at 707, 1,414 and 2,828 Hz, at relative levels as they would be found in the ISTS speech signal when presented at a level of 68 dB SPL. The frequency spans of the bans of noise were 630-800, 1260-1600, and 2520-3200 Hz respectively, i.e. with their upper edges close to the test frequencies of 60118-0.

(4) Dynamic Range Compression (DRC)

60118-0 requires measurement using tones at 2,000 Hz, and, optionally, 707 Hz. Again, due to the presentation in a room, rather than a test box, we used 1/3rd-octave-wide bands of noise.

(i) Compression Ratio

This was measured by using consecutive noise bursts, each of 5-secs duration, of level -10, -5, 0, +5 and +10 dB relative to the bandpower in the 65 dB SPL ISTS signal below or above 1,414 Hz, for the 707-Hz centred, or the 2,000 Hz centred, bands of noise, respectively.

(ii) Speed (attack and release times)

This was measured using the same signal types, and same reference SPL as for the compression ratio, but the signals were shaped into a sequence alternating between -10 to +25 dB in relative level. Compression ratio can also be determined from this test.

(5) EIN, Equivalent Input Noise : this is usually specified as the equivalent SPL at the input to the system when there is negligible acoustic activity. However, it often needs to be assessed with a low-level input (typically around 50 dB SPL) so as to avoid possible confound of low-level expansion operating in the system. The EIN is effectively the absolute level of a form of background noise that arises due to random electrical activity in the system electronics. Usually it is quoted as the level of a broad-band noise, i.e. a single figure-of-merit. However, this obscures if any particular frequency range was particularly badly affected by noise. We therefore simultaneously presented 4-off 0.28-octave wide bands of pink noise centred at 0.5, 1, 2 and 4 kHz, at the same relative levels as would be found in a 50 dB ISTS signal. This signal therefore had gaps of (1-0.28 = 0.72) octaves between bands. After performing a spectral analysis, we could then compare the level of signal in the gaps to either side of each band (effectively system noise) to the reproduced level within the bands, so as to obtain a frequency-localised signal-to-noise-ratio (SNR). At the location of the microphone on the chest, a non-ideal spectrum is present due to reflections and shadowing by the mouth, which may mean that some speech-frequency bands are very low in level. Rather than a single figure-of-merit for EIN, our 4-band method gives an idea of the likely ease of listening/disturbance due to the system background noise across the range of frequencies important for speech communication.

(6) Predicted speech intelligibility and quality: measured with speech from the talker, in either quiet or at a +6 dB SNR with 4-talker babble being presented frontally. These two conditions represent two extremes of likely real-world use : ideal versus difficult. We use the recently updated version of the Hearing Aid Speech Performance Index (HASPI version 2, Kates & Arehart, 2021), and the Hearing Aid Speech Quality Index (HASQI version 2, Kates & Arehart, 2014), software meters comparing the recorded signal to the original signal to produce perceptually-based measures of a processed speech signal. Other well-respected software intelligibility meters do exist, such as the STOI (Taal et al., 2011) or its update, eSTOI, which are open source. We do report some use of the STOI below.

(7) Directionality of the MUTs : There are two concepts of directionality that can be invoked here:

(1) The relative response of the MUT to interfering noises coming from locations around the body of the talker.

(2) The relative response of the (high-chest mounted) MUT as the talker rotates their head relative to their torso.

Measurement (1) normally requires a very laborious procedure (unless automated): measuring the relative response at the microphone as a noise source moves around the MUT. This produces a "polar response" pattern. Given the impractical length of time to measure such, we chose to measure the relative response between a loudspeaker in front and a loudspeaker behind the talker (AHEAD and REAR, respectively). Some of the difference will be due to the "shadowing" effect caused by the body of the talker, and some due to the test microphone response pattern. Again, this measure is more of a "real-world" indicator for comparison with other devices, rather than an abstract scientific measure.

Measurement (2) is necessary because, as the talker rotates their head relative to their chest, the microphone pick-up goes "off-axis" compared to the straight ahead positioning and therefore could lose some of its sensitivity. We therefore measured the frequency response of the microphone as the head was positioned at -45, 0 and 45 degrees relative to straight ahead.

The measures described so far require pre-calculated test signals for presentation in a controlled acoustic environment. The relevant signals were assembled in sequences into three separate computer files : (i) test-box like measures, (ii) speech in noise, and (iii) the signals for measuring directionality.

Two further tests were devised that were more informal but required more human intervention. These are described in (8) and (9) below.

(8) Transmission distance of the MUTs

We envisage two basic usage scenarios for these systems, a playground or a classroom. We therefore tested the distance at which radio signal could be received in an HA. Radio propagation, especially in the multi-GigaHertz region (e.g. 2.4 GHz with the EduMic) is very much "line of sight", and is heavily influenced by relative orientation of the wearers between MUT and receiver as well as objects in the line-of-sight pathway. Such objects, such as windows, doors, have differing "opacities" to radio signals. Metal signboards, for example, are functionally opaque.

(9) Susceptibility to other sources of radio interference

Modern life involves a plethora of wireless-linked devices. Some of these use low-power radio links, such as Wi-Fi and BlueTooth, but others, such as mobile phones use higher powers. These

transmissions are most likely to interfere with the receiver of radio signals, where, being remote from the transmitter, much lower powers are received.

A further characteristic of many of these devices is that they are transmitting and receiving even with no human intervention. We therefore assembled a group of devices that rely on radio transmit/receive in order to simulate a desktop environment where the system receiver was also located, and assessed whether audio link quality was affected.

2.3 Playback and Recording method

Playback and recording was performed by a single computer and a single soundcard. Playback required 3 channels; AHEAD, MOUTH and REAR. Recording required two channels, an omnidirectional measurement microphone adjacent to the MUT, and the MUT itself.

The single-soundcard approach, as well as recording via the measurement microphone permits a guard against the computer sometimes "dropping" small chunks of audio (a common problem with many computers these days) as well as preserving the exact timing relationship between the two recordings. The compromise was that only 16-bit recordings could be made.

The soundcard was a PreSonus Studio 26; a 2-input, 4-output-capable card. An ART SLA-4 power amplifier converted the soundcard signals to drive the loudspeakers. The measurement microphone was a DBX DriveRack RTA-M. The remote signal from the microphone under test was derived by decoding the signal in a relevant HA set to modest gain, delivering via a 4-mm Libby horn inserted through a foam earplug into the entrance of the 711 coupler in the Kemar head. Whereas the Phonak aids all permitted a "Radio-Microphone-input only" mode, (muting the HA microphones), the Oticon Engage HAs only permitted the HA microphone to be set at -12 dB relative to the input from the Oticon EduMic. In order to reduce leakage of the direct acoustic signal to the recorded output, either via the HA's own microphone, or leakage around the foam earplug, the head carrying the coupler was situated in a cardboard box covered in acoustic wadding. This permitted recording of the HA signal in the same room as that in which the acoustic presentations were being made to the microphone under test. The microphone signal from the 711 coupler was amplified sufficiently to drive the soundcard.

2.4 Hearing aid configurations

The HAs used were

(1) For the three Roger systems, Phonak Sky Q70 M13

(2) For the Oticon EduMic, Oticon Engage 1123.

(3) For the Compilot II, Phonak Nathos M.

Since the Compilot is not a true radio microphone, but a receiver/loop driver, we used the Roger Touch as its source for measures.

The HAs were programmed for a modest, flat-30 dB HL loss, using the NAL-NL2 prescription as a starting point. The compression ratios were all set to unity (linear) for levels between 50 and 80 dB SPL, and, using the Phonak Sky as the reference, the gains adjusted until they were within +/-2 dB of each other from 250 to at least 8k Hz (except for the Nathos, which could only achieve this balance up to 5k Hz). The test signal used was 45 secs of the International Speech Test Signal, presented at 65 dB to the aid sited in an Aurical HIT box. The recorded measure was the aided gain in a 2cc coupler. The choice of audiogram produced a gain configuration (low gain and linear) was such that the aids were unlikely to contribute much distortion to the measured results. Such distortions could arise from causes such as (1) too much electrical noise from the aid compared to the signal output, hence use of some gain, as could be expected in realistic use (2) DRC deactivated to prevent non-linearities associated with replay level, (3) low gains so as to prevent high input signal levels being distorted due to receiver limitations at the output of the aid, and (4) sustained high output levels causing temporary battery depletion. This manifests itself in test box measures as an apparent slow-acting DRC. It is an unintentional form of DRC since it arises due to a failure of the hearing aid circuitry to operate consistently under conditions of a reduced battery voltage.

The drawback of using aided responses on which to perform measures is that most measures will be influenced by the relative across-frequency gain prescribed. Hence all reported measures are for signals that have been filtered to :

(1) Remove ultra-low frequency sounds, such as building vibrations (high-pass at 50 Hz)

(2a) remove the aided gain response (Appendix 3) to produce a HA with flat 0-dB gain response referenced to the 2-cc coupler.

(2b) reference the 2-cc response to the "eardrum" of the 711 coupler (by use of a "2-cc to eardrum correction" response). This is the acoustic position at which we recorded the aid outputs to the received remote microphone signals.

(2c) reference the 711 eardrum response to the diffuse acoustic field by use of the inverse "diffuse-field to 711 coupler response".

For the measurement microphone recordings, only filtering stage (1) above was necessary.

This two-stage filtering process means that all measures of the radio microphone systems are referenced to a recording position at the microphone of the system-under-test, as if it were sitting in a diffuse acoustic field. This is similar to that which would be found in a room with modest reverberation, similar to our listening room. Therefore the HA gain is removed from the comparisons (although not necessarily any acoustic imperfections of the HA). In order to preserve the recording fidelity, the filtered signals were stored in 24-bit precision, still with the original sampling rate of 44.1 kHz.

3. Technical Methods : Results

3.1 Frequency response

(a) The influence of placing the microphone at the high-chest.

This measure does not rely on the processing in the system-under-test, but indicates the properties of the acoustic signal from the MOUTH loudspeaker, as recorded at the measurement microphone, adjacent to the radio microphone transmitter systems (see Appendix 1). The reduced high-frequency content, compared to the original signal, is due to the "acoustic shadowing" of the neck by the platform supporting the loudspeaker. In practice, this is probably greater than would be achieved by a real mouth.

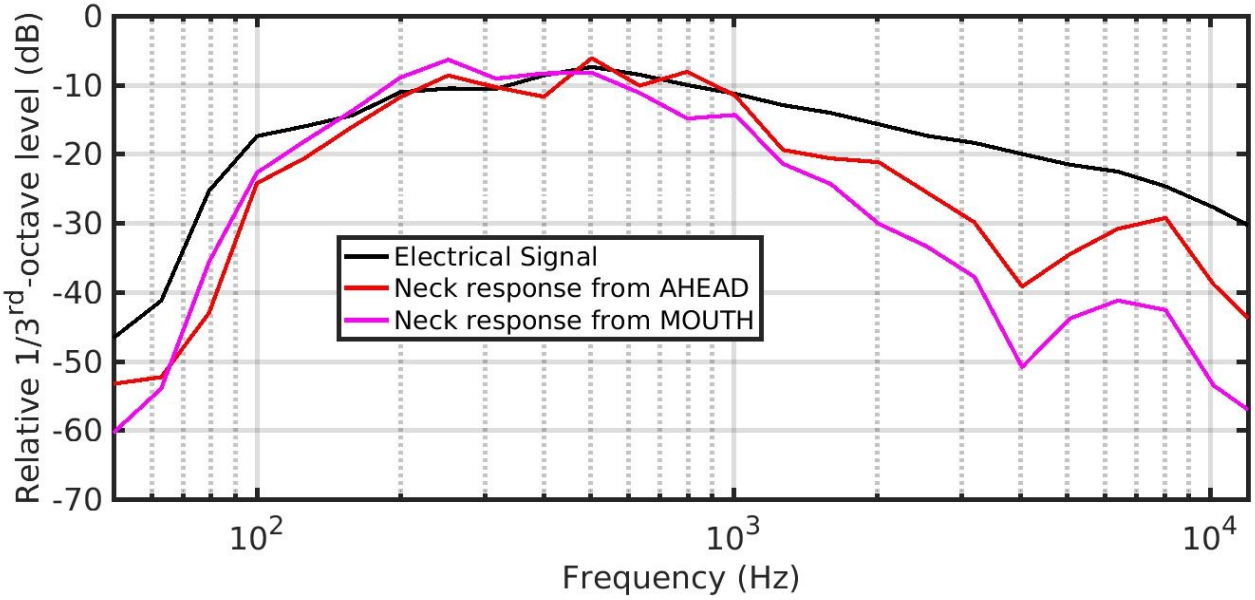
The overall wideband level at the measurement microphone was 7.4 dB (unweighted) higher than the level that the mouth-simulating loudspeaker was producing at 1 metre.

Figure 1 shows the relative frequency responses of the calibration noise signals in the digital files, either as sent electrically to the loudspeaker, or via the AHEAD and FRONT loudspeakers and recorded at the measurement microphone.

Compared to the electrical signal, the high-chest-recorded signals show a reduction at high-frequencies (> 1 kHz), varying between 5 and 15 dB, according to frequency. This is presumably

largely due to the acoustic signal interacting with structures around the neck. It does mean that the tone quality of a microphone signal at the neck will start off "muffled" compared to the original signal. (The Q150 loudspeakers have a very flat response between 100 and 10,000 Hz, so can be presumed to add no "colouration").

Figure 1. *Comparison of relative 1/3rd-octave power between the electrical drive signal to the loudspeakers, and the response at the measurement microphone from either the AHEAD or the MOUTH loudspeaker.*



It is this "muffling" that the radio microphone systems should be designed to overcome, otherwise it will result in lower audio quality and possibly reduced intelligibility.

(b) Frequency responses of the systems in response to the ISTS signal

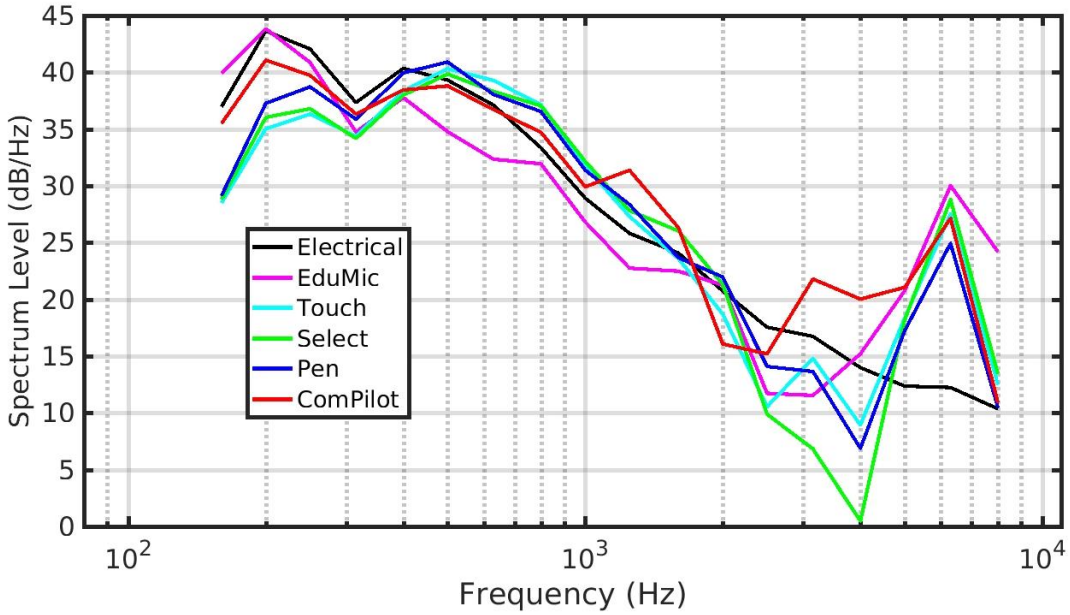
This measure is very similar to the method for measuring aid response in a test box. The average response to the 60-s duration of the ISTS is calculated. Figure 2 shows the average response for each of the systems. Notice that this is plotted in Spectrum Level, not 1/3rd-octave power, since it is the format necessary for input to one of the speech metrics, the Speech Intelligibility Index (SII, ANSI, 1997).

Note also that, for all recordings (i.e. except the electrical), there is a narrow peak around 6 kHz. This is not over-amplification of this frequency region but mostly system electrical noise. Since we cannot "break into the signal chain between MUT and HA, we cannot identify the origin of this noise.

For all systems, the response generally follows the reference (electrical response), except that the Select does attenuate the 3-4 kHz region noticeably more than for the other systems. There

is therefore generally no gross change to the "timbre" (tone quality) of the original signal. There is some low-frequency cut below 300 Hz in the Touch, Select and Pen systems, all made by Phonak, as part of their "Roger" systems. This would imply a general design decision at Phonak. This was less visible in the Compilot, where we used the Touch as its microphone input.

Figure 2. *Comparison of the average system responses to the 60-s duration ISTS signal, normalised so that each has the same overall power.*



3.2 Throughput time delay

This measures the delay between the acoustic signal entering the MUT and its being received at the microphone in the 711 coupler on Kemar's head. As such, it will also include any throughput delay of the HA. This is typically 6 ms (milli-seconds) maximum for a modern digital hearing aid. However, it is reasonable to include the delay in these figures since a real user would have to use this processing path. HAs with delays shorter than 6 ms are available but tend to be used to correct mild-to-moderate losses. HAs with a sub 1-ms delay are currently rare, certainly for digital implementations of HA functions.

Low system delays are desirable so as to facilitate speech reading. McGrath & Summerfield (1985) recommended that delays should be kept below 40 ms so as not to degrade intelligibility for the best lip readers. After accounting for hearing-aid delays, this should not be greater than around 35 ms. In practice, because of acoustic leakage around the moulds of all but the most severely

impaired, giving rise to the perception of the same signal but with two different delays (and consumer rejection), the preferred delay is lower than this.

We originally envisaged using clicks in the input signal to measure this delay, but, in practice, many of the MUT systems appear to use transient suppression, so that it was difficult to detect these in the recorded outputs. However, onsets of narrowband noises were much easier to track in "spectrogram" representations of the signals. Table 1 details the results.

Table 1. *Comparison of signal delays from system microphone though to the 711 coupler. These are delays that would be experienced in "real-world usage". I.e. including HA processing delay (typically 6 ms) but excluding equalisation filter used in preparation of recordings for analysis.*

Microphone system	Delay (ms)
EduMic	47
Touch	24
Select	23
Pen	24
Compilot II	43

The Roger systems are all comfortably within the 40-ms limit, but the EduMic and Compilot systems are slightly above this, but not unacceptably so. Both of these systems use a Bluetooth wireless data protocol which, in some implementations, introduces longer delays. We discuss this further in section *5.4* below.

3.3 Total harmonic distortion (THD)

This measures how the microphone adds in additional signal at higher frequencies due to internal distortions. With a hearing aid, it is normally measured in a test box and a sinusoid test signal. However, in a room acoustic, it is quite common for "modes" in the room response to boost or attenuate some frequencies. Hence a narrow band noise, here 1/3rd octave, is used.

We measured the band power in the output of the equalised 711 coupler, around the test signal, and at harmonic multiples, up to the 5th harmonic. A comparison between the first harmonic power and the sum of the power at the remaining harmonics gives the measure of THD. However we ran into several problems.

We measured at 707, 1414 and 2828 Hz, spanning the range used for HAs (500, 800, 1600 & 3200 Hz: this latter frequency is so high that distortion would be out of the audible frequency range for most HA fittings.). The slightly different span was to ensure that any measured distortion components would be within a range of frequencies important for speech intelligibility. The levels of the signals were adjusted to be similar to the mean level of the same 1/3rd octave band in the ISTS speech signal, [-9.6, -14.2 and -18.4 dB] respectively, relative to the signal power in the full bandwidth]. These lower levels are necessary in case the processing in the microphone systems deems that the same signals presented at a higher level were not representative of speech and performed non-linear processing on them.

(a) Several of the systems used slow-acting DRC, possibly also coupled with noise reduction. So although we used a 23-s duration constant-level test signal, after about 7 s, aggressive reduction of level was appearing in many of the outputs. Therefore we only measured the test signal for the first 7 s of presentation.

(b) There was so little distortion in all of the systems that the system electrical noise dominated in the measures. Once the measures were in the range spanning 5 kHz, as previously noted, the output was dominated by the high-frequency electrical noise.

The figures for THD are expressed in dB in Table 2. Since the frequency region exceeding 4 kHz was dominated by system random noise, power from harmonics in this region have not been included, except for the measures at 2828 Hz since the second harmonic lies in this region. In fact, apart from some second harmonic distortion with Roger-derived systems at 707 Hz, all other measures are dominated by system random noise, and so are comparatively meaningless in the sense that they are not primarily measuring distortion, especially the measures around 2828 Hz. It is because of these low distortion figures that we quote in dB, rather than the conventional %. For comparison, measures less than -34 dB represent a < 2% THD, which is acceptable for HAs.

Table 2. *Comparison of total harmonic distortion (in dB) using 1/3rd-octave noises around the nominal centre frequency, and harmonic numbers used in calculation of THD. Those for 2828 Hz are entirely dominated by system noise, do not reflect distortion of the input signal, and so are italicised.*

Microphone system	707 Hz (H2 to H5)	1414 Hz (H2 and H3)	2828 Hz H2
EduMic	< -40	< -37	< *-16*
Touch	< -29	< -43	< *-17*
Select	< -32	< -40	< *-9*
Pen	< -32	< -34	< *-7*
Compilot	< -35	< -44	< *-13*

The EduMic system therefore comes out as the quietest/least distorting of the systems, but only by a short lead.

3.4 Dynamic Range Compression

(a) Wideband attack and release times, including compression ratios.

We used the 60118-0 testing method of providing a 35 dB step-change in level from -10 to +25dB relative to the power levels in the long-term spectrum of the reference speech spectrum above and below 1414 Hz. The test signals were 1/3rd-octave bands of noise centred at 707 or 2000 Hz. The pulses had the same temporal pattern at both centre frequencies:

(a) A 2-s duration "low" period (-10 dB) to provide a chance for adaptation.

(b) A 2-s duration "high" period (+25 dB) to test "attack"

(c) A 5-s duration "low" period (-10 dB) to provide a "release" period

(d) A repeat of (b) and (c).

The total duration of each test signal was therefore 16 seconds. The attack and release times, as well as compression ratios were virtually impossible to specify in the method required by the standard. This was because of evidence of long-term (> 10s) adaptation to the test signals, leading to inconsistent performance between the first and second high-level portions, and their respective aftermaths. The behaviour is summarised and compared between systems in Table 3.

Table 3. *Summary of dynamic range control properties for the systems in response to the 60118-0 test signal of long-duration pulses alternating 35 dB in level. All CRs are quoted for the longer-term, i.e. observed during the second high-low level periods, (d) above.*

MUT	Test centred on 707 Hz	Test centred on 2 kHz
EduMic	**Adaptation stage:** From quiet, initially 800 ms of no action, then adapts over 400 ms **Attack** : multi-stage, approx 100ms, followed by 1-s delay then slower, longer than test signal could reveal **Release** : 1.5 s **CR** : ~3	**Adaptation stage:** Same as for 707 Hz **Attack** : Very fast << 100 ms, but secondary adaptation where level INCREASES 3 dB after 0.5 s delay at high level **Release** : Very fast, unmeasurable **CR** : ~1.2
Touch	**Adaptation stage** : gates on over 100 ms from quiet, then 600 ms delay **Attack** : 1.4 s (not fully defined by test signal due to 600 ms adaptation) **Release** : 200 ms initially then slows to 600ms, before third stage of release after 3 s. **CR** : ~3	**Adaptation stage** : 600 ms delay from quiet **Attack** : Very fast, unmeasurable **Release** : 100 ms initially then slows over 600 ms **CR** : ~1.2
Select	Similar to Touch at 707 Hz	Similar to Touch at 2 kHz
Pen	Similar to Touch at 707 Hz, but lower CR: ~2. Shows sign of two-stage, different-rate compression on attack, handing over from fast to slow, as well as that observed in release for Touch.	Similar to Touch at 2 kHz, but no long-term slow release.
Compilot	Similar to Touch at 707 Hz, but lower CR ~2.	Similar to Touch at 2 kHz.

None of the MUTs came with software to disable the non-linear processing aspects such as noise reduction, which would normally be done when testing HAs. The procedures of 60118-0 are therefore very hard to implement meaningfully in these systems.

(b) Compression ratio shape measured by noise bursts stepped in level

The results in (a) are very hard to quantify since there is evidence of several adaptations. These adaptations occur because of non-linear, algorithmic programming in the signal processing within these systems. It is not possible to discriminate as to whether this is due to the arrangement of the DRC stages, or whether noise reduction is operating, since both behave as a form of gain change. An alternative method of measuring compression characteristic is to determine the output level as a

function of input level, using signals that change in level. Here we use 5-s duration steps in different configurations of level changes between steps, to measure both general, and possibly different compression characteristics between low-frequencies (centred at 707 Hz) and high-frequency compression (centred at 2 kHz).

(b.1) Wideband compression

Although the EduMic showed waveform behaviour typical of DRC, as did the Roger systems for the first (-10 dB) burst, the remaining bursts showed odd behaviour: a level decrease (as expected with DRC) followed a short while later by an increase in level, similar to dynamic range expansion. All waveform changes were "slow, i.e. occurring over timescales greater than 1 sec.

Table 4 shows the level change between successive bursts, measured over the period 2 s to 5 s after the start of the burst. This was done to avoid the influence of the level change during the "attack" phase of any DRC. However, this approach cannot avoid the odd behaviour of the Roger systems. Except for the first step with the Select, all systems showed DRC of some form since the change in output levels were less than the change in input levels. For all of the systems except the Compilot, the compression ratio increased with increasing input level (figures in first column of numbers are larger than those in second column of numbers), implying some form of limiting compression. However, the Compilot, which was sourced by a "Touch" microphone did the opposite. The reason for this is unclear.

Table 4. *Level change between 5-s duration bursts of speech spectrum shaped noise stepped 10 dB in level between bursts. "Comments" column indicates general shape of envelope of output bursts.*

MUT	Step -10 to 0 dB	Step 0 to +10 dB	Comments
EduMic	9.8	2.3	1-s attack time is actually a sudden drop in level over 100 ms period after 1 s delay.
Touch	8.5	5.2	First burst looks like conventional DRC. Second and third bursts show decrease over first 3 s then then increase over remaining 2 s
Select	11	3.9	ditto
Pen	8.8	2.9	ditto
Compilot	4.2	6.8	ditto

It would be expected that the MUTs have some form of multi-channel DRC so as to provide compensation for any loss of relative level between vowel and consonant energy. We therefore included a further stage of investigation of response to stepped in puts, but with frequency-selective signals.

(b.2) Low-frequency compression (centred on 707 Hz).

The systems were tested in response to a 1/3rd-octave band of noise being stepped UP in level every 4.7 s, from -10 to +10 dB in steps of 5 dB, relative to the reference speech level from the "mouth" in the band from 100 to 1414 Hz.

Table 5. *Level change between 5-s duration bursts of 1/3rdoctave noise centred on 707 Hz stepped up 5 dB in level between bursts. "Comments" column indicates general shape of envelope of output bursts.*

MUT	Step -10 to -5 dB	Step -5 to 0 dB	Step 0 to +5 dB	Step +5 to +10 dB	Comments
EduMic	4.2	5.3	1.6	0	Same sudden decrease after 1 s as for wideband signal (Table 4)
Touch	-2.0	-3.3	-0.3	0	Nothing special observed
Select	5.3	2.6	1.5	1.1	ditto
Pen	-0.1	6.0	13.7	0.7	ditto
Compilot	-2.9	-6.8	-0.4	2.3	ditto

Table 5 shows the level change between adjacent steps. If DRC is active one would expect to see numbers less than 5. Numbers less than 0 denote negative DRC (absolute output level decreasing with increasing input level), which would be unusual to observe for a mid-level signal.

(b.3) High-frequency compression (centred on 2 kHz).

The systems were tested in response to a 1/3rd-octave band of noise being stepped DOWN in level every 4.7 s, from +10 to -10 dB in steps of 5 dB, relative to the reference speech level from the "mouth" in the band from 1414 to 10,000 Hz. Stepping down was used since this signal followed the high level presentation of the 707-Hz 1/3rd-octave noise band used in (c) above, but separated by a silent interval of 125 ms to permit reverberation to die away in the room.

Table 6 shows the level change between adjacent steps. (The sign of the change has been changed to make the table more consistent with Table 5). If DRC is active one would expect to see numbers less than 5. Numbers greater than 5 indicate dynamic range expansion, which would be

unusual to observe for a mid-level signal. For the Roger Pen, it appears that the output level "switches on" once the input level is greater than 10 dB below RMS.

Table 6. *Level change between 5-s duration bursts of 1/3rd-octave noise centred on 2 kHz stepped DOWN 5 dB in level between bursts. "Comments" column indicates general shape of envelope of output bursts.*

MUT	Step +10 to +5 dB	Step +5 to 0 dB	Step 0 to -5 dB	Step -5 to -10 dB	Comments
EduMic	5.2	4.7	5.0	4.6	Fast-attack compressor "overshoot" seen on waveform : conventional DRC.
Touch	4.7	6.7	6.1	6.1	Nothing special observed
Select	6.9	6.2	7.1	6.4	ditto
Pen	0.3	2.5	11.4	10.8	ditto
Compilot	5.2	1.7	3.0	1.5	ditto

3.5 Equivalent Input Noise (EIN)

This is defined as the broadband equivalent noise at the input to the system when no signal is present. It is therefore independent of the subsequent amplification that may be provided by the HAs, and is therefore comparable to real-world signal levels arriving at the microphone. . However, many devices incorporate low-level expansion, effectively turning off audio transmission when no input is present. Hence a low-level signal is often necessary to "condition" the device so that it is switched on, i.e. so that low-level expansion is not activated. The "low-level" signal is intended to be below the compression threshold of the subsequent system, so that the system gain is at maximum and can therefore be specified by the difference in level of the conditioning signal between input and output. It is usually performed by measuring the system output SPL and then measuring the system gain and subtracting this from the output SPL. However, this measure is not entirely meaningful for at least two reasons :

(1) The system gain will be different for the MUT and the internal HA microphones. These differences arise because of the ability to apply level changes in the relative mix, as well as absolute sensitivity of the MUT.

(2) EIN is a broadband measure, and therefore does not reflect the perceptual "damage" that the noise may be performing on localised frequency regions of the input signal.

We used a conditioning signal, presented at a low input level (50 dB SPL), and measured the Signal-to-Noise Ratio (SNR) at the system output. Rather than just measuring overall SNR, we use 4-off, 1/3rd-octave noise bands, octave-spaced with centre frequencies of 707, 1414, 2828 and 5656 Hz. There are therefore 2/3rd-octave gaps between the individual bands. The relative levels of the bands follows that of the powers in the same bandwidths of the reference speech spectrum. Spectrum analysis of the output signal then permits us to analyse the relative power between each signal band and its surrounding noise floor. The SNR is defined as the mean signal density (dB / Hz) in the noise band compared to the mean signal density (again in dB/Hz) in narrow bands close to either edge of the noise band. This then gives us an idea of a localised SNR, specified at up to 4 points spread across the entire range of frequencies useful for speech intelligibility. Table 7 shows these frequency-specific SNRs.

Table 7. *Frequency-specific SNRs for each system.*

MUT	Centre frequency (Hz)				Measurement duration of non-adapting signal (secs)	Comments
	707	1414	2828	5656		
EduMic	-30	-26	-17	-18	3.5	Stepped gain adaptation after first 1 second.
Touch	-34	-26	-20	-21	3.5	Gain adaptation more progressive through entire signal
Select	-32	-27	-14	-22	3.5	ditto
Pen	-20	-28	-17	-21	3.5	ditto
Compilot	-34	-30	-21	-8*	3.5	* beyond upper edge of hearing aid matched gains See graph in Appendix 3

Although the signal was of low input level, there were obvious signs in the waveforms of reducing gain (expansion) during the 5-s duration. Therefore all measures were performed over durations where most of this adaptation had occurred. These SNR figures are encouraging since, even for a quiet input (50 dB SPL, equivalent to very quiet speech), there is still a dynamic range below the mean level that is not occupied by noise. The microphone systems are therefore not applying a major limit on audibility before the signal reaches the hearing aid.

3.6 Objective intelligibility and quality metrics of the systems in response to speech signals

Three software models were used to compare speech intelligibility and quality between the MUTs operating in both quiet, and with a multi-talker babble ("noise") :

(1) The Speech Intelligibility Index (SII, ANSI, 1997) was only applicable for speech-in-quiet measures, so was replaced by STOI (Taal et al., 2011) for speech-in-noise measures.

(2) The Hearing Aid Speech Perception Index (HASPI, version 2, Kates & Arehart, 2021),

(3) The Hearing Aid Speech Quality Index (HASQI, also version 2, Kates & Arehart, 2014).

All three metrics produce a value that ranges from unity (excellent) to 0 (abysmal). Translation of these metrics into exact intelligibility or quality depends on many factors, such as complexity of speech material, context, and especially in this application, the residual neural capabilities of the listener's hearing system. Hence we quote the metrics so as to enable ranking of system performance.

(a) Speech in quiet (female and male)

The Compilot recording level was generally much lower that for the other MUTs; in a real-world scenario this could be adjusted by a change in volume control. Hence we have performed all of these measures for signals at the same input level to the models, a nominal 68 dB SPL. Although this approach is logical, it caused problems later when we considered the directionality of the MUTs, which will be addressed there.

Table 8 compares the model outputs for a 60-s duration female speech signal in quiet, the ISTS. SII is a simple model, but it indicates that there is sufficient audibility for speech to be highly intelligible. The HASPI, a more sophisticated intelligibility model than SII, agrees, with slightly larger (but insignificant) differences between the systems. Greater differences are observed for the speech quality. This will be interpreted more after Table 9.

Table 8. *Comparison of speech intelligibility and quality models using the (female) 60-s ISTS in quiet.*

MUT	SII	HASPI	HASQI
EduMic	0.993	0.999	0.524
Touch	0.995	0.994	0.377
Select	0.995	0.988	0.352
Pen	0.997	0.995	0.420
Compilot	0.994	0.996	0.428

Table 9. *Comparison of speech intelligibility and quality models using 60-s male continuous speech in quiet. F*(Non-linear, Linear) *denotes that the composite score is a function of the non-linear and linear subscores, a function performed internal to the HASQI software.*

MUT	SII	HASPI	HASQI		
			Non-linear subscore	Linear subscore	Composite score *F*(Non-linear, Linear)
EduMic	0.991	1.000	*0.590*	*0.862*	0.509
Touch	0.993	0.999	*0.359*	*0.848*	0.305
Select	0.993	0.995	*0.313*	*0.836*	0.262
Pen	0.994	0.999	*0.421*	*0.872*	0.367
Compilot II	0.989	1.000	*0.461*	*0.887*	0.409

Table 9 repeats the same three metrics as for Table 8, but this time using a 60-s duration excerpt of continuous male, rather than nonsense female, speech. Additionally we list the sub-components of the HASQI measure, relating to the quality of the "non-linear" and "linear" distortions detected by the measure. The results for all three measures follow the same pattern as for the ISTS. The main difference between the systems shows up in the quality metric, which reflects the increased influence of the non-linear processing in the MUTs, such as DRC.

(b) Speech in noise (female and male)

This measure was designed to simulate a classroom acoustic where there is multi-talker babble coming from ahead of the person wearing the microphone system. For this measure alone, the level from the AHEAD loudspeaker was reduced to 59 dB(A) measured at the Kemar torso position, representing a +6 dB SNR when referenced to MOUTH level 1 metre away from the torso. However as shown in 3.1 above, the SNR at the MUT will be higher because the MOUTH level was 7.4 dB higher at the MUT than at the 1-metre distance. For this measure we cannot report values for SII since that particular model needs an estimate of the babble signal alone at the microphone, as well as the (speech + babble) signal. However, given that the presence of the MOUTH speech signal will influence the level of the babble spectrum due to the non-linear processing used in the MUT, then babble-alone signal would not be the same as the babble component of the (speech + babble) recording. The SII measure would therefore be unrealistic. Hence we have replaced it with another intelligibility measure, the STOI, and also report for the HASPI and HASQI measures. All three of these models require only a reference signal of the clean speech as well a (near-) time –aligned version of the processed speech.

Table 10 was derived from use of a 60-s duration excerpt of continuous female speech and Table 11 for a 60-s duration excerpt of continuous male speech. The speech intelligibility metrics (STOI & HASPI) show the potential for very good intelligibility. The main difference between the systems shows up in the quality metric, which reflects increased non-linear processing, such as DRC.

Table 10. *Comparison of speech intelligibility and quality models using 60-s female continuous speech in with 4-talker babble from AHEAD at -6 dB relative to MOUTH level when measured at a distance of 1 metre.*

MUT	STOI	HASPI	HASQI		
			Non-linear subscore	Linear subscore	Composite score *F*(Non-linear, Linear)
EduMic	0.857	0.977	*0.397*	*0.865*	0.343
Touch	0.855	0.963	*0.415*	*0.861*	0.358
Select	0.848	0.891	*0.363*	*0.858*	0.312
Pen	0.848	0.892	*0.386*	*0.867*	0.335
Compilot	0.833	0.970	*0.362*	*0.892*	0.323

Table 11. *Comparison of speech intelligibility and quality models using 60-s male continuous speech in with 4-talker babble from AHEAD at -6 dB relative to MOUTH level when measured at a distance of 1 metre.*

MUT	STOI	HASPI	HASQI		
			Non-linear subscore	Linear subscore	Composite score *F*(Non-linear, Linear)
EduMic	0.830	1.000	*0.365*	*0.867*	0.316
Touch	0.811	0.999	*0.326*	*0.852*	0.278
Select	0.807	0.992	*0.266*	*0.847*	0.225
Pen	0.857	0.996	*0.315*	*0.874*	0.275
Compilot	0.818	1.000	*0.307*	*0.880*	0.270

3.7 Directionality of the MUTs

Directionality defines the spatial selectivity of the MUT, that is, a measure of the relative response of the MUT to sound sources arriving from different directions. There are two concepts of directionality that were mentioned earlier :

(1) The relative response of the MUT to interfering noises coming from locations around the body of the talker.

(2) The relative response of the (high-chest mounted) MUT as the talker rotates their head relative to their torso.

We recorded the response of the MUT to 5-s burst of ISTS speech-spectrum-shaped noise arriving from either FRONT, REAR or MOUTH. For the MOUTH measure, we started with the head pointed 45 degrees to the left (as viewed from behind), then, during 6-s silent pauses, we moved the head to point to either 0-degrees, or 45 degrees to the right, before a 5-s burst was played. Repeatability of adjustment was ensured by taping a straight edge to the middle of the MOUTH loudspeaker, and using that as a pointer to either markers on the wall (see Appendix 2) or to the middle of the FRONT loudspeaker.

With our set up, we therefore have two sets of measures for all 5 orientations, one set for the measurement microphone and one set for the MUT.

We start off by performing SII measures, assuming that the signals have all been equalised to equal level, as we did earlier with the software models of quality and intelligibility (Tables 8 and 9).

Table 12. *SII measures for the measurement microphone and MUTs as a function of presentation direction. Results printed in strike-through red are potentially misleading, see text for details.*

MUT	FRONT	REAR	MOUTH 0º	MOUTH 45º LEFT	MOUTH 45º RIGHT
Measurement microphone	0.999	0.993	0.983	0.966	0.975
EduMic	~~0.997~~	~~0.998~~	0.996	0.998	0.997
Touch	~~0.995~~	~~0.992~~	0.995	0.996	0.996
Select	~~0.990~~	~~0.990~~	0.995	0.995	0.995
Pen	~~0.998~~	~~0.991~~	0.996	0.999	0.998
Compilot	~~0.999~~	~~0.998~~	0.996	0.996	0.997

At first sight, Table 12 shows results that have very little discriminatory power. Despite the torso blocking the REAR signal at the MUT, it appears to have very little effect on likely intelligibility. However, these directions are “non-preferred” for the MUTs, so can be expected to be modified heavily in their response by the directionality of the MUT, hence their printing in red with a strike-

through. Additionally, as identified earlier in the software modelling of intelligibility and quality, because the MUT outputs varied in level when carrying the MOUTH signal, we had assumed to normalise the level at the input to the SII software as if a suitable volume control had been applied. This approach is not valid for signals that the MUTs are likely to want to reject. Conversely, the measurement microphone has an omnidirectional response, and no signal processing, and it appears to show some directionality when placed close to the torso, as expected.

The level-normalisation method is valid for comparing the effects of MOUTH directionality relative to the MUT. The "MOUTH 0º" measure is very similar to that obtained with the 60-s ISTS signal in Table 8 (rather than its spectrally matched noise used here), despite the non-linear signal processing on noise. The measurement microphone is the only microphone that shows any directionality to the head rotation. This is unsurprising given its asymmetric placement, as shown in Appendix 1.

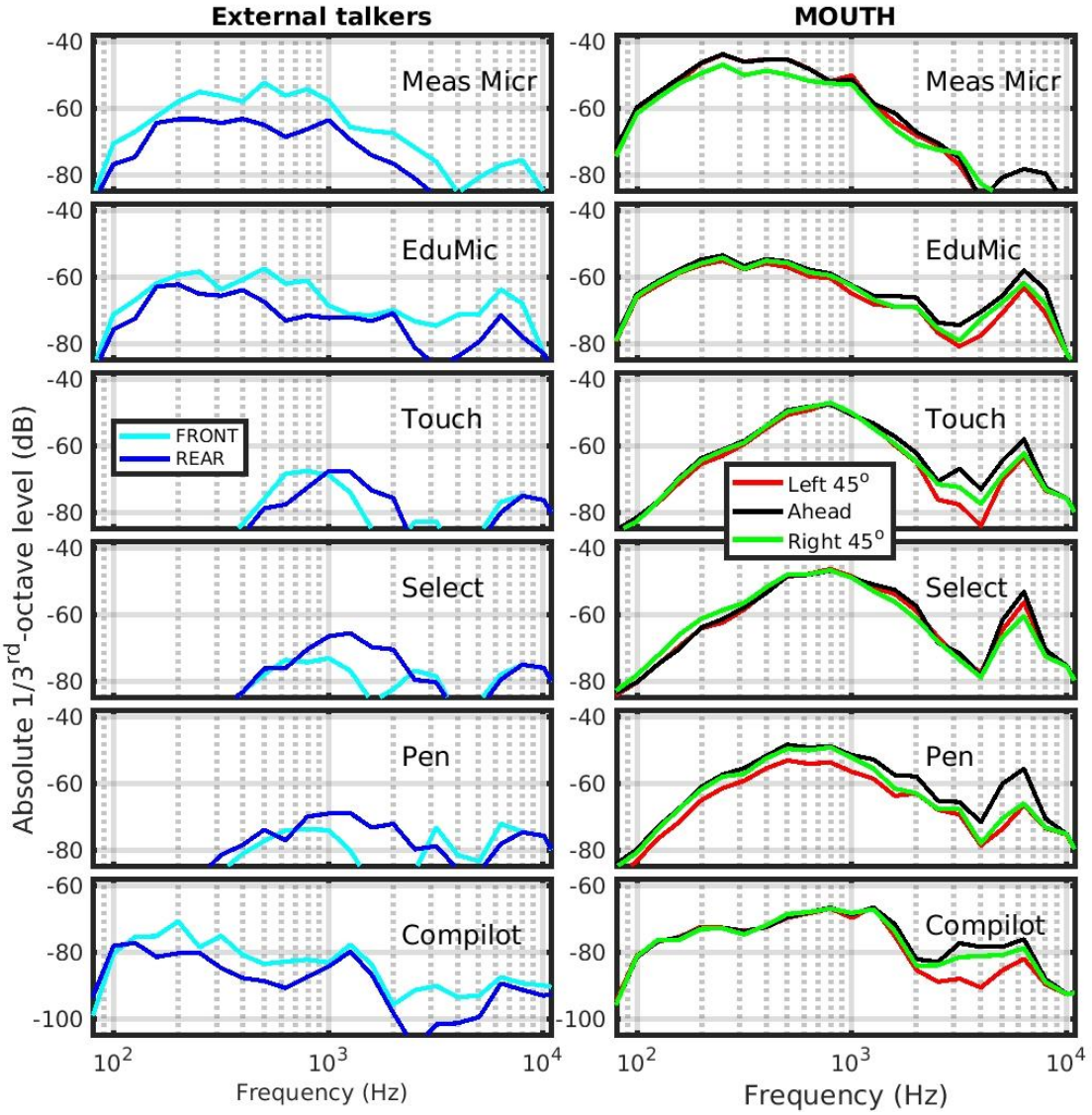
These mostly null results indicate that a different measure is necessary. The reason for these results becomes apparent when we plot the signal levels received at each MUT for the different source locations, as shown in Figure 3. The levels are plotted as absolute signal level, in dB, measured in 1/3rd-octave bands for each MUT, as well as including the measurement microphone (top panel of each column). Because the signal level in the measurement microphone recording channel had a different sensitivity from the channel recordings the MUTs, then we cannot plot in absolute dB SPL. This is a minor detail, since we are primarily concerned with the pattern of behaviour for each MUT between the left-hand (external talkers) and right-hand column (the desired signal to be handled well by the MUT). As mentioned earlier, since the Compilot signal was so low, (but could be overcome with a volume control), we have plotted the axes for all plots with the same span in dB, but with an offset of 20 dB for the Compilot.

Starting with the measurement microphone ("Meas Micr", top panel) we see that the torso is in fact performing a general 10-dB blocking across all frequencies to the REAR talker. For the MOUTH signal, there is only a small difference as the MOUTH rotates, apart from above 4 kHz, where the left and right 45-degree signals are much lower than the Ahead signal.

For the MUTs, especially for the Touch, Select and Pen, we see a slightly higher mid-frequency (700-2200 Hz) response for the REAR source, than for the FRONT source, but overall, the response is much lower than the response for the MOUTH source. For all three systems, the Ahead response is better than both of the 45-degree orientation responses, but there are only small differences, (slightly larger for the Pen than for the other two) which is a desirable feature.

For the external talker signals, the EduMic and Compilot receive the FRONT signal higher than the REAR signal, but with a lower sensitivity than for the MOUTH signals. Again, there is only a small change in sensitivity for the orientation of the MOUTH source (larger for the orientation of the MOUTH with the Compilot. This latter finding is surprising since we used the Touch as the source microphone, so might have expected similar results as for that MUT. We can offer no explanation, other than that the Touch can switch between different directionality patterns automatically, such as may be useful when placed on a table, rather than body-worn. We did not change this directionality pattern, allowing it to auto-select.

Figure 3. *The absolute signal level, in dB, in 1/3rd-octave bands for each MUT, including the measurement microphone (top panel of each column) in the directionality measures. The left-hand column shows the directionality for FRONT and REAR (external) sources, while the right-hand column shows the directionality for the MOUTH source at the three orientations relative to facing ahead*.



It should be pointed out that these levels were calculated across the 5-s duration of the noise bursts used. Given that all MUTs, except for the measurement microphone, exhibited adaptive behaviour, it is therefore likely that the responses are an average behaviour. Since auditory perception occurs over timescales of hundreds of milliseconds, meduium-to-long-duration averages are not always accurate predictors of perceptual effects. However, the principle of hearing instrument test boxes is also based on medium-term averages. (durations of seconds)

3.8 Transmission distance of the MUTs : "link distance"

(a) Open space e.g. playground

For the "playground" scenario, we used a large grassy open space with a few concrete paths, as well as a few low-height obstructions such as benches, signage and rubbish bins, but nothing in the direct line of sight. The person wearing the HA was stood in a fixed place with their arm raised, while the talker walked backwards while talking continuously. At the point where the link failed, the wearer dropped their arm. The talker then moved forward until the link was re-established. The position at which the link failed, while moving away from the wear was noted. Measurements were obtained by use of GPS readings on a smartphone in the "playground", and a tape measure in the classroom. It was noted that there was hysteresis between the distance at which the link failed, and that at which the link was re-established: the talker had to walk back towards the wearer in order for the link to be re-established. Two measures of the link distances were taken for each system, and so represent a range of expected performance.

The link distances are given in Table 13 for the playground scenario, and Table 14 for the classroom scenario. With some of the MUTs, it was possible to change the transmit power, which would increase the range. For these baseline measures, we left the (radio signal) power level at the default for all MUTs.

Table 13. *Link distances in simulated playground scenario, (in units of metres).*

MUT	Range minimum	Range maximum
EduMic	23	32
Touch	29	62
Select	10	27
Pen	5	9
Compilot + Touch ditto + Select ditto + Pen	103 30 21	

The Touch came out well ahead in terms of link distance, in which a hysteresis effect was more noticeable (the difference in distance between losing and re-establishing the link), followed by the Compilot. The Pen was noticeably weak in its transmission distance. The slightly greater distances achieved with the Compilot is probably attributable to the more direct line-of-sight of the receiver attached to the bottom of the chest-worn Compilot, than for the receivers located on HAs where the pinna of the receiver (at least partially) obstructs this line.

These distances will vary depending on a combination of factors (such as weather, orientation of talker and receiver, presence of metal objects nearby), so one should consider them more by rank ordering, i.e. relative, rather than absolute values.

(b) Closed space : e.g. classroom

The classroom dimensions were 5.5 W x 7.5 L x 2.75 H metres. It contained metal framed desks and chairs, but no students. The walls were concrete block set in a reinforced concrete structure. Metal panel radiators are set along the length of the classroom, but not extending above desk height. Beyond the end of the classroom was a similarly furnished separate room with dimensions 5.2 W x 3.7 L x 2.5 H metres. The connection was via a wide doorway set in a wood and glass stud wall. The direct pathway from the transmitter position to the furthest corner of this room would pass through a breeze-block wall.

Table 14 shows that, for nearly all the systems, it was possible to maintain a link between the two rooms over the combined room lengths provided that a direct pathway existed between the transmitter and the person wearing the receiver, even if they were facing away from the transmitter. In order to stress-test the transmission it was possible to change the line of transmission so that a glass and stud partition wall would be in the path. Even then, the EduMic and Touch maintained the link even when partially obstructed by the partition wall. Compared to the playground, it would

appear that internal reflections are maintaining the signal level at the receivers, certainly evidenced by the performance of the Roger Pen.

Table 14. *Link distances in simulated classroom scenario, (in units of metres).*

MUT	Pathway unobstructed by furniture		Transmission limit measured with pathway obstructed by partitions and brickwork into separate room	Comments
	Facing	Facing away		
EduMic	>10	>10	>10	14.5 m, after transmission through two open doorways, one set in partition wall and one in breeze-block wall.
Touch	>10	>10	>10	15 m for same conditions as for EduMic above.
Select	> 10	> 10	10	Link regained at 9.0 m
Pen	> 10	10	N/A	
Compilot + Touch ditto + Select ditto + Pen	> 10 > 10 > 10	> 10 > 10 > 10	10 10 10	Link regained at 10.5 m Link regained at 9.5 m Link regained at 9.5 m

3.9 Susceptibility to radio interference

On a desk sited in a doorway facing a long corridor, we assembled two laptops with Wi-Fi enabled, two mobile phones (Android and Apple iPhone), a wireless mouse connected to one of the laptops, a Noah-Link (using wireless BlueTooth) and an iPad. The iPad was set to receive the wireless "hotspot" generated by a third mobile phone receiving 4G signals. Additionally, the desk was deliberately sited to be close to a WiFi hotspot affixed to the ceiling in the corridor, 3.5 m from the desk. There was therefore a high density, and variety, of radio protocols in use local to the receiver therefore expected to compromise reception.

The talker walked the MUT down the corridor while simultaneously sending text messages to the iPhone until he reached a distance further than the link distance reported in 3.8 above. Additionally, for the Roger systems, the talker wore an activated EduMic, and for the EduMic, an activated Touch. For none of the MUTs was the link distance or reception compromised.

4. User feedback "PPI"

The assessment was performed in order to gain user opinion on each device regarding areas such as software programming, features, ergonomics, robustness and ability to interconnect with other manufacturers' products. We originally planned that each device would be assessed by a patient representative, a Teacher of the deaf (ToD), a classroom teacher, an Audiologist, a representative from the Ewing foundation. The patient representative was a 7-year-old with severe sensorineural hearing loss remediated by bilateral hearing-aid-and-FM, accompanied by their father, enabling a wider user experience to be captured with them both. A classroom teacher was not available at the time of the sessions, but the ToD had significant hands-on experience in the classroom setting, and so doubled up in that role. The Ewing foundation representative was also a qualified clinical Audiologist. The likely experiences of the panel with these devices therefore covered a wide range of perspectives.

4.1 Session design

Following initial consultation with the study team, which also included an experienced paediatric Audiologist and a technical officer with extensive knowledge of radio aid/wireless streaming technologies, a draft hearing aid evaluation document was assembled. From this, separate questionnaires were constructed for each intended member of the PPIE (Patient and Public Involvement and Engagement) panel, with questions appropriate for their perspectives on the devices.

For many of the panel, the questions could be sub-grouped covering different aspects of use:

(a) Child and parent: a single group of nine questions were provided covering cosmetics, usability of controls, and the clarity of the instruction manual.

(b) ToD: 22 questions sub-grouped into Features, Instruction Manual, and DAI shoe or ear level receiver.

(c) Audiologist and Ewing foundation representative: 18 questions sub-grouped into Features, Instruction manual, Hearing aid software, and DAI shoe or ear level receiver.

For the child and parent, the answers to the questions were a rating between 1, for "Strongly Disagree", to 10, for "Strongly Agree". For the remainder of the panel, Yes/No responses were sufficient, and meant that we were able to assign a score of 1 or 0 for the respective response. We then summed the scores from each user in order to obtain a ranking for the systems. A higher score indicated that the MUT was more preferable. Apart from the structured questions, feedback on each device included a free-text box for open comments.

The PPIE sessions was held in a spacious classroom at the University of Manchester (the same as used for measures of the classroom link distances). All five devices were laid out, fully charged and with the appropriate instruction manuals, at separate "stations" around the room. This enabled the members of the panel to inspect, handle and even try out the audio, where hearing aid compatibility allowed. The devices were labelled with the letters A to E. However it was not possible to anonymise the contents of the instruction manuals and therefore we could not conduct the session fully blinded.

The child and parent attended after school hours and spent approximately 2.5 hours thoroughly assessing and handling each of the five devices, including pairing with their hearing aids when there was compatibility. The ToD also attended this after-school session. The Audiologist and Ewing representative required less time to review the devices due to their existing familiarity with the technology generally, if not the specific devices included in the study. Each panel member was offered remuneration for their time.

4.2 Summary of responses

The full record of the panel responses is included in a supplementary spreadsheet distributed along with this report. The scores collated from the questionnaires overwhelmingly indicated that the Touch was rated as the most user friendly, had the additional features required and was the most accessible of the five devices.

Table 15. *Rating scores from individual panel members.*

MUT	Child and Parent *max = 90*	Teacher of the Deaf *max = 22*	Audiologist *max = 18*	Ewing representative *max = 18*
EduMic	50	13	11	12
Touch	**78**	**18**	**14**	**18**
Select	65	17	11	16
Pen	54	13	12	16
Compilot	48	4	10	15

Table 15 shows a summary of these scores, according to panel-member-type. However there are caveats to the scoring contained in the additional comments noted from all the participants, especially the parent and child, as well as the ToD's well-informed insights into using (and buying) them 'in-the-field'. These should be taken into account in order to fully represent the participant opinions. All panel members specifically appreciated the display screen on the Touch, unique among all the devices; it provided a good clear visual indication as to whether the device was

connected and working, and allowed easy checks of the settings, while other devices relied mainly on LED colour indicators which, without manuals or guides to hand, panel members felt may lead to confusion. While the Touch scored highly, other devices also received positive feedback. Panel members liked the design of the Select and Pen devices and their docking stations. Good and attractive design was important to all parties in our feedback but particularly the user, both the child and parent.

The Select's connectivity to a whiteboard in the classroom setting was recognised as a useful addition and it was felt that the Pen would be "great for older children/young people".

The EduMic was considered a "clear simple design not too small to lose, feels robust" and "very affordable & offers same quality as others on the market". All devices were considered to be easy to charge, had some audio input function and were thought to be easy to pair by the panel members. The instructional manuals were rated as accessible and easy to follow in all cases. They were also all able, to varying degrees to provide interconnection with other manufacturers hearing aids by the additional of an extra receiver.

General comments included how much and how rapidly this field is changing in terms of streaming devices and their inter-connectivity. Integrated receiver technology is replacing ear level and DAI shoes and the ability or lack thereof, to pair the MUT to the hearing aid becomes paramount in the decision making when choosing a wireless streaming device. Additional costs, such as extra receiver components, are likely to be incurred to facilitate compatibility. Consequently, depending on funding arrangements, price rather than functionality is still potentially a deciding factor in the selection of a device.

As previously noted, the Compilot II does differentiate itself from the rest of the devices and it was not possible for our panel members to view it in entirely the same way as the other devices. As it is a wireless streamer, rather than a transmitter, it either streams an audio signal using a Bluetooth connection or can act as an FM receiver system, but only with the addition of a universal Roger or FM receiver plugged into it (such as the Phonak Roger X). The neckloop was commented on by all panel members as being undesirable and not suitable at all for younger children.

In all the devices, no activation in the hearing aid fitting software was required to allow connection, but the ability to adjust the gain of the device varied. The Touch allowed the easiest access with the gain controls being available to adjust on the screen while the Select and Pen require an additional device, either the Phonak Inspiro or the Touch to adjust through the 'Easygain' setting. These changes could still potentially be done in the educational setting. The EduMic however can only be adjusted in the Oticon Genie fitting software by a professional with access to it, which would generally be the Audiologist based in a Healthcare setting: adjustment therefore

requires an additional visit to an Audiology department. The Compilot II has a volume control on it, available for the user, but it is unclear in the instruction manual as to whether this is still functional when used in the FM condition.

As part of their questionnaire, the ToD was asked to identify whether the device had a range of design features such as a natural resting position on the table, a docking station, a lanyard and a belt clip option. All these were felt pertinent in appraising the flexibility of the design for the teacher, parent or other. Only the Pen offered all of these options, although the Roger Select and Touch provided all but the belt clip attachment. The EduMic comes with an integrated clip to attach to clothing, but only a lanyard option was an option. The Compilot II, being a streamer, has to be worn by the user (not the talker) on a neckloop.

In summary, the panel rated the Phonak Touchscreen as the most acceptable and user friendly device from the range of devices used in this comparison: specifically its features, ergonomics, robustness, programming and interconnection. However the other devices, with the exception of the Compilot II, were all rated well overall. The Compilot II's rating was due, in the main, to it being a somewhat different device to the others in the evaluation.

In selecting to purchase and use one of these devices, the make and type of hearing aids a child wears and the cost still appear to be the determining factors in the field.

5. Lessons learnt for future studies

As a pilot study we met several problems, some anticipated and some not.

5.1 Anticipated problems

(1) In trying to compare microphones we wished to intercept the remote signal as soon as it was decoded. Although Phonak provide a "checker" for the Roger transmitters, this is not always the case with other systems. We therefore used the output of a suitable hearing aid that was programmed to have modest gain, and linear, rather than frequency-, or dynamic-range, compressed response. All other signal processing features of the hearing aids were de-activated (to the degree possible) so that the output signal was a near-faithful reproduction of the decoded remote microphone signal, as well as not constrained by the hearing aid internal noise.

(2) Hearing aid responses across different manufacturers should be near-identical, as verified in a hearing-instrument test box. Choosing a multi-channel programming model for each hearing aid permitted a fair degree of frequency-specific fine tuning, so that this could be achieved.

(3) Verifying the hearing aid response. This was performed via an acoustic test box, and so required the HA microphone to be active. The aided-gain settings were copied between this program and a separate program which would be "remote-microphone only" so that the recorded HA response was solely from the remote microphone. However, one then has to trust that the program-copy function works as intended, and that the HA manufacturer has not applied any extra processing (frequency shaping) to the remote-microphone signal compared to that of the microphone internal to the HA.

(4) Ensuring that the recorded HA response was solely from the remote microphone. Most modern HAs are fitted with earmould of various degrees of venting, from "open domes" to acrylic moulds. The venting provides a bypass path so that the acoustic field near the ear affects the recording. Sound delivery from all HAs was via a 4-mm Libby horn inserted through a foam earplug into the entrance of the 711 coupler on the Kemar head. The head was placed in a radio-transparent cardboard box with wadding over the top to reduce ingress of the acoustic field.

5.2 Unanticipated problems

(1) COVID impact. This project was suspended due to the necessary invocation of a "force majeure" clause in the original contract with NDCS when UK-wide lockdown was imposed in March 2020. Even after the lockdown restrictions were lifted in July 2021, availability of staff was compromised by the "pingdemic", e.g. requirements for occasional home schooling, summer holidays, and supply-chain availability.

(2) Working with hearing aid fitting software from multiple manufacturers. This raised multiple issues :

(i) small changes in parameter settings caused non-linear processing features to become re-activated, requiring perpetual checking of settings between the two nominally identical programs in each HA, as alongside validation in the HA test box.

(ii) buggy or 'esoteric' behaviour in fitting software where small changes to the software settings produced unpredictable changes in the HA response, requiring multiple iterations of fine-tuning and checking (such as gain changes in one frequency region drastically affecting one that was two octaves away, and large changes in gain and DRC being introduced when switching the tubing option).

(iii) HAs from the same manufacturer, but with the same programming, had very different responses in the test box. This was eventually sorted and verified (Appendix 3)

(3) The fast-moving nature of the technology.

Between the original definition of the contract and its execution, several changes occurred, such as two devices becoming obsolescent, as well as integration of radio receivers into the HAs becoming more commonplace. Therefore parts of the assessment became difficult or irrelevant. Moving forward, a test suite needs to establish a base set of "essential" features in terms of functionality and perceptual utility, which are the core reasons for the technology. The "non-essential" aspects of tests could still be useful in establishing a more rounded picture of the device.

5.3 Utility of test signals and methods chosen

Not all signals resulted in measures that discriminated between the different systems. A general problem, which has also been encountered in the testing of digital hearing aids, is that the MUTs incorporate proprietary algorithms that cannot be de-activated by a "fitting" software. Therefore all measures are prone to a bias because of adaptations within the MUTs, which occur over multiple different timescales. Any differences may only become apparent with more sophisticated (and longer-duration) test methods.

Of particular promise are software models of quality and intelligibility. We used the HASPI and HASQI because they are familiar to people within this field. Experience in Manchester has shown that these measures need a long sample of input, in excess of 20 s in order to obtain stable values. Given the observed adaptation times, it is suggested that around 60 s is a suitable duration. We introduced the use of one other software model, the STOI, which is encountered more in the telecommunications field, but which has an overlap with the radio communication aspect of the devices. All three software models require a copy of the original signal as well as the processed signal, so are more of a laboratory, than a "field", measure. "Blind" software models, which only need a copy of the output signal do exist, but they are not usually open source, so would require (expensive) licensing which is probably not justified/supportable by the size of this market at this stage.

5.4. Specifying the system delay

Correspondence by the PI with the European Hearing Instruments Manufacturers' Association (EHIMA) reveals that low system delay has been a consideration in the design of the recently specified (early 2020) BlueTooth Low Energy standard (BT-LE) which contains the ability to broadcast from one microphone to many receivers, essential for a classroom-usage scenario. Confidentiality agreements forbid the detailing of the delays involved. The BT-LE system, where data is streamed across radio waves to the receiver, is likely to become a common component of

remote microphone technology. This streaming introduces delays much longer than the more historic method of an audio signal directly changing the characteristics of a radio wave (e.g. the "FM" systems). Apart from the BT-LE delay, there is also a delay introduced by the remote-microphone manufacturer as they attempt to reduce background noise from the microphone signal. This latter component of delay is also company confidential.

For this report, we are not concerned with what exactly contributes to the delay. The nearest applicable evidence is from McGrath & Summerfield (1985), on the basis of which we suggest adopting a precautionary approach and recommend that delays should not be much greater than their 40 ms limit. The principle of a remote microphone is to enable better access to speech cues which otherwise would get degraded in background noise and reverberation. It is not unreasonable for this principle to be extended to include preserving the perceptual link between visual and audio speech cues. We would therefore like to see the system delays routinely specified in data sheets

6. Conclusion

A pilot study using test methods similar to those already in use in the HA field (60118-0), as well as some adapted measures has contrasted the technical performance of five different remote microphone systems. These systems exhibit sophisticated automatic digital signal processing which cannot be de-activated by the user. Combined with the necessity to test the systems in a room rather than a test box, we are constrained as to which test signals are suitable: tone signals are best avoided. Other compromises are necessary, such as testing the systems via the acoustic output of a (presumed) low-distortion HA due to the absence of a direct electrical output. This constraint does mean that some basic checks are performed on the HA to ensure that it is not the limiting factor in the signal chain.

The measures showed that the five systems were broadly similar in their performance. The main area where differences were exhibited were:

(a) Despite the difficulties of characterising the DRC systems, further, meaningful, differences were observed were in the software models of intelligibility and quality, which are broadly transferrable into likely acceptability from a technical performance angle.

(b) The Roger systems, when used on their own, showed the best directionality in terms of ability to reject sounds arriving from away from the talker while also exhibiting only a small variability in output with moderate "head" rotation of the talker.

(c) In the system delay, which was on the edge of acceptable for two of the systems, but well within acceptable for the other three systems. We urge manufacturers to report their system delays in their data sheets.

User feedback highlighted some issues with usability which varied between systems. The Roger Touch was particularly commended for its clear feedback via its screen as to device setup and connectivity.

This report should be seen as an opening in a drive to standardise test methodologies across multiple technology types used in remote microphone systems. Future work should especially concentrate on the use of signals that are more speech-like in their spectro-temporal properties, so as to represent more real-world usage, as well as characterising the boundaries of acceptable system delays in a paediatric population with hearing-impairment.

Acknowledgements

The authors gratefully acknowledge :

(1) Tony Murphy of Phonak UK for loan of some of the microphone system components.

(2) The members of our PPI team for contributing their time and feedback.

References

ANSI (1997). ANSI S3.5-1997, "Methods for the calculation of the Speech Intelligibility index" (American National Standards Institute, New York).

IEC (2015). IEC 60118-0. Electroacoustics – Hearing aids – Part 0: Measurement of the performance characteristics of hearing aids. IEC, Geneva, Switzerland.

Kates, J.M. and Arehart, K.H. (2014). "The hearing-aid speech quality index (HASQI) version 2," J. Audio Eng. Soc. 62, 99-117.

Kates, J.M. and Arehart, K.H. (2021), "The hearing-aid speech perception index (HASPI) version 2," Speech Communication, 131, 35-46.

McGrath, M. & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. Journal of the Acoustical Society of America, 77, 678-685.

Taal, C. H., Hendriks, R. C., Heusdens, R., and Jensen, J. (2011). "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio. Speech. Lang. Process. 19, 2125–2136.

Appendix 1

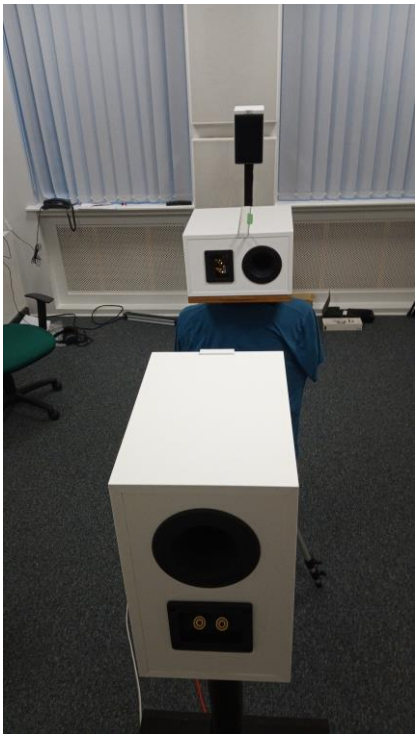
Photographs of equipment setup around Kemar torso showing (a) the loudspeaker centred on a platform so as to simulate the mannikin mouth, (b) black acoustic foam forming a “beard” under the loudspeaker platform (to reduce reflections off platform under-surface), (c) the particular radio microphone under test (labelled under the picture) (d) a PreSonus (omnidirectional) measurement microphone to record soundfield at the radio microphone high-chest position (e) the Kemar torso covered with a double layer t-shirt so as to be more representative of acoustic damping around the torso.

Roger Touch

Roger Select Roger Pen Oticon EduMic

Appendix 2

Photograph (right) showing alignment of loudspeaker sources from (a) nearest, 1 metre behind torso (“REAR”), (b) torso-mounted “mouth simulator” (“MOUTH”), and (c) frontal loudspeaker, 2 metres in front of torso ("FRONT”).

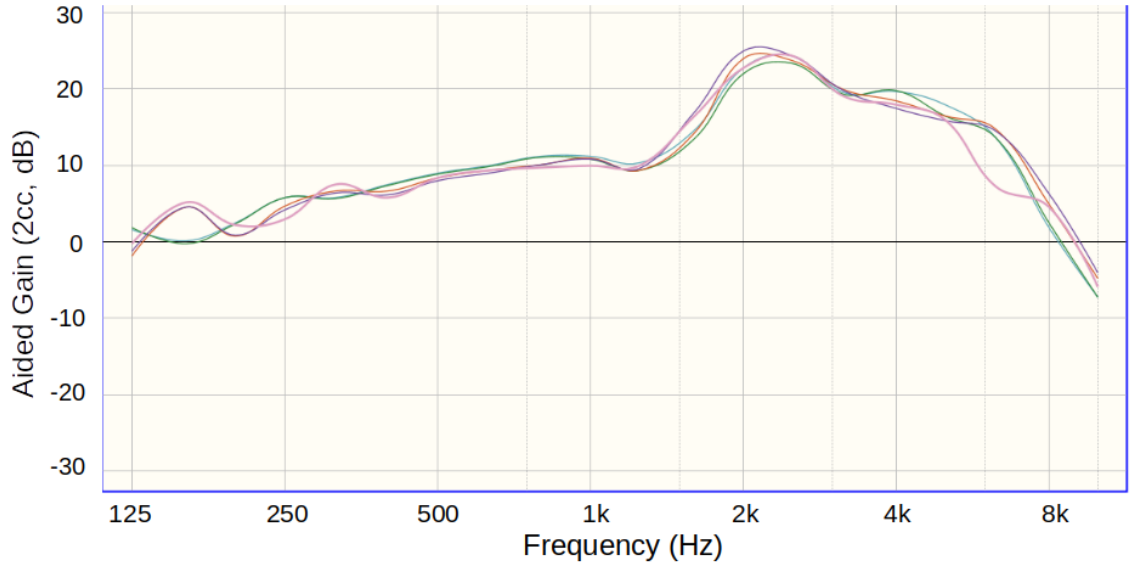


Appendix 2 (cont)

Photograph (below) showing ruler pointer on top of rotating torso-mounted loudspeaker, to point at marks on the walls (ringed in red), set at +/- 45^{o} relative to direction of frontal loudspeaker.

Appendix 3

Comparison of HIT box aided gain measures, in a 2cc coupler, of the different aids used in this study. Generally gain were within +/- 2 dB of each other in the range 250 to 8k Hz, except for the Phonak Nathos M (pink trace), which had a more restricted bandwidth, rolling off at about 5 kHz. The Nathos was necessary only for testing of the Compilot II system.


30
20
10
0
-10
-20
-30
Aided Gain (2cc, dB)
125
250
500
1k
2k
4k
8k
Frequency (Hz)