Using Objective Metrics to Measure Hearing-Aid Performance

James M. Kates,
University of Colorado, 409 UCB, Boulder, CO 80309

Kathryn H. Arehart,
University of Colorado, 409 UCB, Boulder, CO 80309

Melinda C. Anderson,
University of Colorado School of Medicine, Aurora, CO 80045

Ramesh Kumar Muralimanohar, and
University of Colorado, 409 UCB, Boulder, CO 80309

Lewis O. Harvey Jr.
University of Colorado, 345 UCB, Boulder, CO 80309

Abstract

Objectives—The performance of hearing aids is generally characterized by a small set of standardized measurements. The primary goals of these procedures are to measure basic aspects of the hearing-aid performance and to ascertain that the device is operating properly. A more general need exists for objective metrics that can predict hearing-aid outcomes. Such metrics must consider the interaction of all the signal processing operating in the hearing aid, and must do so while also accounting for the hearing loss for which the hearing aid has been prescribed. This paper represents a first step in determining the clinical applicability of the HASPI intelligibility and HASQI speech quality metrics. The goals of this paper are to demonstrate the feasibility of applying these metrics to commercial hearing aids, and to illustrate the anticipated range of measured values and identify implementation concerns that may not be present for conventional measurements.

Design—This paper uses the HASPI intelligibility and HASQI speech quality metrics to measure the performance of commercial hearing aids. These metrics measure several aspects of the processed signal, including envelope fidelity, modifications of the temporal fine structure, and changes in the long-term frequency response, all in the context of an auditory model that reproduces the salient aspects of the peripheral hearing loss. The metrics are used to measure the performance of basic and premium hearing aids from three different manufacturers. Test conditions include the environmental factors of signal to noise ratio and presentation level, and the fitting configurations were varied to provide different degrees of processing from linear to aggressive nonlinear processing for two different audiograms.

Results—The results show that the metrics are capable of measuring statistically-significant differences across devices and processing settings. HASPI and HASQI measure both audibility

Corresponding author: James.Kates@colorado.edu, Telephone: 720-226-1266.

and nonlinear distortion in the devices, and conditions are identified where predicted intelligibility is high but predicted speech quality is substantially reduced. The external signal properties of signal-to-noise ratio (SNR) and presentation level are both statistically significant. Hearing loss is significant for HASPI but not for HASQI, and degree of processing is significant for both metrics. A quadratic model for manufacturer showed large effect sizes for HASPI and HASQI, but basic vs. premium hearing-aid model is not significant.

Conclusions—The results presented in this paper represent a first step in applying the HASPI and HASQI metrics to commercial hearing aids. Modern hearing aids often use several different processing strategies operating simultaneously. The proposed metrics provide a way to predict the total impact of this processing, including algorithm interactions that may be missed by conventional measurement procedures. The measurements in this paper show significant differences between manufacturers, processing settings, and adjustment for different hearing losses. No significant differences were found between basic and premium hearing-aid models. Further research will be needed to determine the clinical relevance of these measurements and to provide target values appropriate for successful fittings.

Keywords

hearing loss; hearing aids; speech intelligibility metrics; speech quality metrics

1. Introduction

There are many reasons for using a metric or standardized procedure to measure hearing-aid behavior. Metrics can be used to establish the baseline performance of a device, to determine if a device is operating properly, or to estimate the potential improvement in speech intelligibility or sound quality that will occur when using the hearing aid. That is, determine what a hearing aid does, whether it does what it is supposed to do, and whether it will do any good. The emphasis in this paper is on the third question, estimating the changes in intelligibility and speech quality expected when a hearing aid is used, and specifically on some of the issues associated with translating laboratory metrics for predicting algorithm effectiveness into practical measurement procedures suitable for use in assessing commercial hearing aids. The results presented in this paper illustrate the range of intelligibility and speech quality predictions obtained from six commercial hearing aids. Further research will be needed to determine the clinical relevance of these measurements and to provide target values appropriate for successful fittings.

Several measurement procedures have been developed and standardized. The primary goals of these procedures are to characterize basic aspects of the hearing-aid performance and to ascertain that the device is operating properly. For example, American National Standards Institute (ANSI) standards describe hearing-aid measurements such as pure-tone frequency response and gain (ANSI S3.22, 2014), gain as a function of frequency for speech-like stimuli (ANSI S3.42, 2012), and real-ear gain as a function of frequency (ANSI S3.46, 2013). A limitation of these standards is that they lag behind product development by several years; for example, there are no standards for measuring noise suppression or frequency lowering despite the availability of these features in many commercial hearing aids. An

additional limitation is that the standards do not indicate how the measured hearing-aid performance affects speech perception.

In addition to the existing ANSI standards, it is possible to develop metrics that predict the perceptual impact of using hearing aids. The primary goal of these metrics is to objectively measure the signal changes caused by the hearing-aid processing and to relate these changes to perceptual outcomes; the link between the signal measurements and the perceptual judgments is generally an auditory model. Examples of this approach include the speech intelligibility index (SII) (ANSI S3.5, 1997) applied to hearing loss and hearing aids (Pavlovic *et al.*, 1986; Amlani *et al.*, 2002), the speech transmission index (STI) (Houtgast and Steeneken, 1985) applied to hearing loss (Humes *et al.*, 1986; Festen and Plomp, 2002), and the hearing-aid speech perception index (HASPI) (Kates and Arehart, 2014a) and hearing-aid speech quality index (HASQI) (Kates and Arehart, 2014b).

The SII is based on performing a frequency analysis of the noisy speech, calculating the speech signal-to-noise ratio (SNR) in each frequency band, and then forming a weighted average of the SNR dB values across frequency. Hearing loss is incorporated as a corresponding increase in the background noise level. The SII emphasizes the audibility of speech (Souza and Turner, 1999) while ignoring any nonlinear distortion introduced by the hearing-aid processing. The SII is limited, however, in that there is little correlation with hearing-aid fitting effectiveness and the SII value (Souza *et al.*, 2000), and the SII can be very inaccurate in predicting the benefit of hearing-aid processing algorithms such as noise suppression (Hilkhuysen and Huckvale, 2010; Taal *et al.*, 2011; Smeds *et al.*, 2014). The STI also uses an auditory frequency analysis, but measures envelope modulation depth rather than the SNR. The STI has also been found to have limitations when applied to the types of signal processing found in hearing aids, including dynamic-range compression (Hohmann and Kollmeier, 1995; Rhebergen *et al.*, 2009) and noise suppression (Ludvigsen *et al.*, 1993; Dubbelboer and Houtgast, 2007).

To overcome the limitations of the SII and STI, metrics are needed that measure the effects of both audibility and distortion in the context of a model of impaired hearing. The HASPI intelligibility (Kates and Arehart, 2014a) and HASQI speech quality (Kates and Arehart, 2014b) metrics considered in this paper have been designed for hearing-aid applications. Both metrics incorporate a model of the auditory periphery that includes auditory frequency analysis that depends on signal level and audiogram, auditory dynamic-range compression, the shift in auditory threshold that corresponds to the hearing loss, and neural firing-rate adaptation. The metrics have been validated for a wide range of hearing-aid processing algorithms and signal degradation conditions, including frequency-response shaping, wide dynamic-range compression (WDRC), noise suppression, feedback cancellation, frequency lowering, additive noise and babble, and nonlinear distortion such as amplitude quantization and peak clipping (Kates and Arehart, 2010; Houben *et al.*, 2011; Kressner *et al.*, 2013; Suelzle *et al.*, 2013; Kates and Arehart, 2014a; Kates and Arehart, 2014b; Huber *et al.*, 2014; Falk *et al.*, 2015; Kendrick *et al.*, 2015, Van Kuyk *et al.*, 2017).

It is a large step to go from laboratory hearing-aid simulations to measurements of actual devices. Because HASPI and HASQI are accurate for a wide variety of hearing-aid

processing conditions and incorporate a model of impaired hearing, they are promising for characterizing hearing aids (Houben *et al.*, 2011; Suelzle *et al.*, 2013; Huber *et al.*, 2014). However, a systematic investigation of the application of these metrics to benchmarking hearing aids is currently lacking. This paper provides such a systematic investigation; the environmental and processing conditions and their justifications are presented in Table I. The measurements varied the external SNR, signal level, hearing loss, HA processing setting, and the manufacturer and model of the hearing aids tested. The test conditions can be divided into those associated with the external signal properties (SNR and level), those associated with the hearing-aid programming (audiogram and degree of processing), and those associated with the individual device (manufacturer and model).

The remainder of this paper presents the procedures used to select and program the hearing aids used in the study. The stimuli are described, followed by how the hearing-aid recordings were made. The HASPI and HASQI metrics are explained in more detail along with how the metric calculations were implemented. The resultant HASPI and HASQI values are then presented as a function of the test conditions and hearing-aid settings. The challenges in moving the measurements from the laboratory to the clinic are then discussed, and the paper concludes with recommendations for practical implementations.

2. Methods

2.1 Stimuli

The speech stimuli were adjusted to give a range of signal intensity and SNR values. Speech stimuli at low, moderate conversational, and high intensity were used as inputs to the hearing aids. A pair of HINT sentences, one spoken by a male talker (Nilsson *et al.*, 1994) and the other by a female talker (Nilsson *et al.*, 2005) were used. The sentence spoken by the male talker was "The boy got into trouble." The sentence spoken by the female talker was "Two cats played with yarn." The sentences were set to the same RMS level before being concatenated, with a gap of 200 ms between sentences. Speech was presented in quiet and in a background of six-talker babble (Cox and Gilmore, 1987) used as the noise. For the noisy speech stimuli, the noise was on for 10 seconds prior to the onset of the sentences, which was found to be sufficient time to activate the noise suppression in the hearing aids. In quiet conditions, there was 10 seconds of silence preceding the sentences. For both the loudspeaker output signal and the hearing-aid outputs, twenty repetitions of the signals were collected and averaged prior to applying the metrics to reduce the impact of the test booth background noise.

The combinations of speech intensity and signal-to-noise ratio (SNR) used for the measurements are summarized in Table II. The three speech levels were 55, 65, and 75 dB SPL. Speech was presented in quiet and at SNRs of 10 and 0 dB.

2.2 Audiograms and Hearing-Aid Settings

The hearing aids were programmed to fit each of two different audiograms with four different processing configurations. Two audiograms were chosen from the set of International Electrotechnical Commission (IEC) standard audiograms (Bisgaard *et al.*,

2010; ANSI, 2012). The hearing losses for the audiograms are presented in Fig 1. The S2 audiogram is representative of a mild sloping to severe hearing loss. The N4 audiogram is representative of a flat moderately-severe hearing loss. These audiograms give common losses for which a hearing aid is often prescribed.

Hearing aid signal processing algorithms were programmed to a control setting and three experimental settings. As a baseline control condition, the devices were programmed to the NAL-R linear fitting (Byrne and Dillon, 1986) as provided by each manufacturer's fitting software. For the linear condition, all other hearing aid features were disabled. These features included directional microphones, noise reduction, frequency lowering, and feedback management. The manufacturers' default NAL-R gain settings were used as this was the only way to ensure that all other nonlinear processing options were turned off. The hearing-aid output was measured using an Audioscan Verifit 2 audiometer with the hearing aid connected to the 2-cc coupler. All six hearing aids were compared to the NAL-R targets. No device was underfit by more than 7 dB at any frequency at or below 4 kHz. However, several frequencies were overfit by the manufacturers' NAL-R fitting by 4 – 19 dB at 1.5 and 2 kHz for both the S2 and N4 audiograms.

In each of the three experimental conditions, the gain settings and compression ratios were determined by the NAL-NL2 (Keidser *et al.*, 2011) WDRC fitting rationale for each of the audiograms. Active feedback management was disabled for all devices. In addition, directional microphones were disabled, and all microphones were set to omnidirectional mode. The remaining two processing features of noise suppression and frequency lowering were set in the following way:

- Mild Processing: noise suppression and frequency lowering were disabled.

- Moderate Processing: noise suppression and frequency lowering were programmed to the manufacturers recommended first fit for the audiogram.

- Maximum Processing: noise suppression and frequency lowering were programmed to the most aggressive processing to the extent allowed by the programming software for that audiogram.

For the NAL-NL2 conditions, the fittings were verified using the Verifit 2. Response verification was performed for the hearing aids set to the mild processing condition. For both the S2 and N4 audiograms, targets were matched within 3 dB through 4 kHz for all but six points (four frequencies with a 4-dB difference, one having a 7-dB difference, and one having a 12-dB difference). Targets above 4 kHz were matched within 10 dB, reflecting limitations in the hearing aids providing the higher amounts of gain desired at the highest frequencies.

2.3 Devices Tested

A total of six behind-the-ear (BTE) hearing aids were measured. Three major manufacturers were represented. Within each manufacturer a premium level and a basic level device from the same model line were selected for analysis. The devices represent the most current commercial devices available from each manufacturer as of December 2015.

2.4 Recording Apparatus and Procedures

The measurement system is shown in the block diagram of Fig 2. The configuration complies with the measurement procedure for coupler gain specified in ANSI S3.42-2012/Part 2 (ANSI, 2012). To record the hearing-aid output, the ear hook was connected to a click-on PCB AEC202 2-cc coupler attached to a G.R.A.S. 40AO ½-inch prepolarized pressure microphone and G.R.A.S 26CA ½-inch CCP preamplifier. The coupler was placed inside a Brüel and Kjær (B&K) anechoic test box (model 4232) that included a reference microphone (G.R.A.S. 40AO microphone and 26CA preamplifier). Stimuli were presented through the loudspeaker in the test box. The microphone outputs were amplified using a G.R.A.S. 12AQ signal conditioner which provided the inputs to a National Instruments USB4431 4-channel data recorder. The reference and hearing-aid coupler output signals were digitized at a sampling rate of 48 kHz in separate channels before being stored in the measurement computer. Custom MATLAB routines were used to control the stimulus presentation and to coordinate recording of the coupler microphone and reference microphone signals for offline analysis. The complete recording system was placed in a sound-isolation booth to minimize the amount of ambient noise.

2.5 Signal Analysis

The hearing-aid performance was quantified using the HASPI speech intelligibility metric (Kates and Arehart, 2014a) and the HASQI speech quality metric (Kates and Arehart, 2014b). Both metrics compare the output of an auditory model for the processed signal (e.g. output from the hearing-aid receiver loaded with the 2-cc coupler) to the output for a reference signal (e.g. input at the hearing-aid microphone). For HASPI, the reference signal was an average of twenty repetitions of the loudspeaker output for the noise-free speech at 65 dB SPL, with the average passed through a model of normal hearing. For HASQI, the reference was the averaged noise-free loudspeaker output at the selected test level (55, 65, or 75 dB SPL). NAL-R compensation was provided for the hearing loss for the HASQI reference, and the amplified signal was then passed through the model of the impaired periphery associated with the test audiogram.

The metrics are based on a computer model of the auditory periphery (Kates, 2013). The model starts with a middle ear filter that reduces the response below 350 Hz and above 5000 Hz. Auditory frequency analysis is provided by a 32-band gammatone filter bank with center frequencies spanning 80 to 8000 Hz. The model incorporates the auditory dynamic-range compression mediated by the outer hair cells (OHC), with compression ratios decreased and auditory filter bandwidths increased with increasing hearing loss and increasing signal intensity. The final stage in the model is inner hair cell (IHC) firing-rate adaptation, which provides a greater output level for signal onsets than for the steady-state portions of a signal. Impaired hearing in the model is represented by shifted auditory thresholds, wider bandwidths of the auditory filters, lower OHC compression ratios, greater upward spread of masking, and reduced two-tone suppression.

The metrics compare the envelope and temporal fine structure (TFS) of the hearing-aid output signal to that of the reference to estimate the changes in predicted intelligibility and speech quality. HASPI uses the time-frequency envelope modulation pattern and the TFS

associated with the more-intense portions of the speech signal. HASQI also uses the envelope modulation, and combines this with changes in the TFS at all signal intensities and changes in the signal long-term spectrum. The metrics have been validated for a wide range of laboratory processing conditions. The correlation coefficient comparing HASPI predictions to subject sentence intelligibility scores, when averaged across all stimuli and both NH and HI listeners, is 0.97; the tested processing included speech with additive noise and nonlinear distortion, frequency compression, and ideal binary mask noise suppression (Kates and Arehart, 2014a). The correlation coefficient comparing HASQI predictions to subject sentence quality ratings, when averaged across all stimuli and both NH and HI listeners, is 0.96; the tested processing included speech with additive noise and nonlinear distortion, frequency compression, ideal binary mask noise suppression, noise vocoder output, and acoustic feedback and feedback cancellation (Kates and Arehart, 2014b).

HASPI and HASQI values for normal hearing are plotted in Fig 3 as a function of SNR for long-term average speech spectrum (LTASS) noise added to the pair of sentences at 65 dB SPL. To gain an intuitive interpretation of HASPI and HASQI, one can relate the metric values to the SNR. The HASPI scores reproduce the classic performance-intensity (PI) function for speech in noise, going from 0 to 1 over the SNR range of -10 to 10 dB. Because HASPI combines the effects of nonlinear distortion and noise into a single number, a HASPI value of 0.8 due to distortion corresponds to the same intelligibility as a value of 0.8 due to additive noise. HASQI is much more sensitive to the noise. At a SNR of 10 dB, where HASPI is nearly 1, HASQI is only about 0.3, and even at a SNR of 40 dB, HASQI is only 0.9. If one were to define a 30-dB SNR as yielding "acceptable" quality, for example, the corresponding HASQI threshold would be about 0.75. Like HASPI, HASQI is based on results from experiments incorporating many different noise and distortion conditions, so the overall speech quality related to a given HASQI value will be the same no matter what has caused the underlying signal degradation.

2.6 Hearing Aid Measurement Concerns

There are several measurement concerns present in hearing-aid measurements that were not present in the laboratory measurements on which the metrics are based. One consideration is the receiver frequency response. For the listener experiments that were used to create HASPI and HASQI, participants heard stimuli over headphones. The reference and processed files that were compared to produce the metrics were the computer WAV files, which did not have any response equalization. However, many receivers used in hearing aids have a response peak around 3 kHz that is designed to reproduce the ear-canal resonance that is lost when a hearing aid is used. The hearing-aid output, when a receiver response peak is present, thus has a shaped frequency response that disagrees with the flat sound-file responses used in deriving HASPI and HASQI. To match the hearing-aid test conditions to the metric derivation conditions, the receiver response between 1.5 and 4.5 kHz was measured. If a response peak was found, it was removed by applying a 2-pole/2-zero notch filter (Procházka and Landau, 2003) to the hearing-aid output.

A second concern in the hearing-aid measurements is the presence of background noise in the test room. It was not possible to obtain complete sound isolation despite using a test box

in a sound-isolation booth. Much of the ambient noise was removed by applying a 7-pole Butterworth high-pass filter having a cutoff frequency of 80 Hz to both the reference and processed signals. The auditory spectra produced for the sentence stimuli at 65 dB SPL and for the ambient noise, both after highpass filtering, are plotted in Fig 4. The filtered residual test booth noise gave a SNR of 33 dB, which limited the maximum HASQI value assuming normal hearing to 0.81. Hearing-aid performance better than this maximum HASQI value would have been impossible to measure because of the background noise. To improve the baseline HASQI value, twenty hearing-aid output files were averaged to produce the reference and processed files used as the inputs to the HASPI and HASQI calculations. The averaging reduced the room noise by an additional 13 dB and raised the baseline HASQI score to 0.93 for stimuli at 65 dB SPL. For speech stimuli at 75 dB SPL, the 10-dB increase in speech level relative to the noise increased the baseline HASQI score to 0.96. Thus averaging multiple stimuli and responses greatly improves the measurement sensitivity.

There is a trade-off, however, in averaging the outputs. HASPI and HASQI are reduced by the random noise present in the hearing aid, such as the thermal noise from the microphone, and averaging the responses reduces the level of this internal noise as well as the level of the external room noise. However, the internal noise levels in hearing aids tend to be close to auditory threshold (Lee and Geddes, 1998; Lewis *et al.*, 2010), so the internal noise will either be inaudible or dominated by external noise and the nonlinear distortion caused by the hearing-aid circuitry and processing. Thus the measurements presented in this paper emphasize the impact of the spectral changes and nonlinear distortion produced by the hearing aids.

A final concern is the impact of the level of presentation on the HASQI calculations. One of the considerations in this paper is effect of increasing or decreasing the input signal level. However, all of the stimuli used to derive HASQI were at 65 dB SPL, and the authors are not aware of any data giving hearing-aid speech quality ratings as a function of stimulus intensity. HASQI compares the auditory model output for the processed signal to the output for a reference signal. For the HASQI calculations, it was assumed that the reference signal for speech presented to the hearing aid at 55 dB SPL was also at 55 dB SPL, and the reference signal level was set to 75 dB SPL for speech stimuli presented at 75 dB SPL.

3. Results

The purpose of this paper is to explore the use of objective metrics in evaluating hearing aids. As shown in Table I, the hearing-aid measurements involved several experimental parameters: SNR (3), signal intensity (3), audiogram (2), degree of processing (4), hearing-aid manufacturer (3), and model (2), for a total of 432 different combinations of the test variables.

Analyses of variance (ANOVA) were applied to linear mixed-effects statistical models of the intelligibility and quality data. The six hearing aids (three manufacturers × two models each) were considered as subjects drawn randomly from the population of all commercial hearing aids. The between-subject factors of manufacturer and model could then be examined, but

there were insufficient degrees of freedom to evaluate their interaction. Separate ANOVAs were conducted for the HASPI and HASQI scores.

The ANOVA results for HASPI are presented in Table III and the results for HASQI are presented in Table IV. There are significant main effects of SNR, level of presentation, and degree of processing for both HASPI and HASQI. Hearing loss is significant for HASPI but not for HASQI, and manufacturer is significant for HASQI but not for HASPI. Hearing-aid model has no significant effect on either predicted intelligibility or quality. Two-way interactions between SNR and level, SNR and loss, SNR and processing, and loss and processing are significant for both HASPI and HASQI, and the interaction between level and processing is significant for HASPI but not for HASQI.

The standardized effect sizes for the main effects in the linear models were estimated via bootstrapping and are presented in Table V. The effect sizes for Cohen's d are generally classified as $d = 0.2$ being a small effect, $d = 0.5$ a medium effect, $d = 0.8$ a large effect, and $d = 1.3$ a very large effect (Sullivan and Feinn, 2012). For HASPI, very large effects were found for SNR, large effects for hearing loss and processing, and moderate-to-large effects for level and manufacturer. For HASQI, very large effects were found for SNR, level, and processing.

3.1 External Variables: SNR and Signal Level

The effects of SNR and presentation level on HASPI and HASQI are plotted in Fig 5. The data are averaged over the hearing loss, degree of processing, manufacturer, and model. The error bars give the standard deviation. The SNR and presentation level are statistically significant and have medium to very large effect sizes. Both HASPI and HASQI increase with increasing SNR, with the increase when going from 0 to 10 dB greater for HASPI than for HASQI. Presentation level for HASPI has a very small impact for the 0-dB and 10-dB SNRs, but exhibits an increase in the scores with increasing intensity for speech in quiet. The HASQI scores, on the other hand, show a consistent pattern of decreasing predicted quality as the signal level increases for all three noise levels.

3.2 Hearing-Aid Programming: Audiogram and Degree of Processing

The effects of hearing loss and degree of processing are plotted in Fig 6 for speech in quiet and at a SNR of 10 dB at an input level of 65 dB SPL. The data are averaged over manufacturer and model, and the error bars give the standard deviation. Loss is significant for HASPI but not for HASQI, and has a large effect size for HASPI. Degree of processing is significant for both HASPI and HASQI, and has a large effect size for HASPI and a very large effect size for HASQI.

HASPI in quiet shows nearly perfect predicted intelligibility for the S2 audiogram for all the processing settings, and decreasing values for the N4 audiogram as the aggressiveness of the processing is increased. At the 10-dB SNR, HASPI shows a general trend of decreasing scores with increasing degree of processing for both audiograms, but the moderate processing for the N4 audiogram gives a higher average score than for the mild processing condition. HASQI in quiet also shows a trend of decreasing scores with increasing processing aggressiveness for both audiograms, although the differences between the mild

and moderate processing conditions are small. The HASQI scores show a much larger difference between quiet and the 10-dB SNR than observed for HASPI, indicating a greater effect of this noise level on predicted quality than on predicted intelligibility. For both HASPI and HASQI in quiet and in noise, the linear NAL-R amplification tends to have the highest scores and the maximum processing condition the lowest.

3.3 Device: Manufacturer and Model

The effects of manufacturer and model are plotted in Fig 7 for speech in quiet and at a SNR of 10 dB at an input level of 65 dB SPL. The data are averaged over hearing loss and degree of processing, and the error bars give the standard deviation. Neither manufacturer nor model are significant in the HASPI linear model, while manufacturer is significant for HASQI. Manufacturer has a large effect size for HASPI but not for the linear model used for HASQI. To further evaluate the differences between manufacturers, the effect size was recomputed using a quadratic model, which compared Mfg2 against Mfg1 and Mfg3. The quadratic model effect size was found to be -1.09 for HASPI and 1.66 for HASQI. These large to very large effect sizes indicate that the hearing aids from Mfg2 rate higher on predicted intelligibility and quality than those from Mfg1 or Mfg3.

HASPI in quiet shows similar scores for Mfg1 and Mfg2, and a reduction in the scores for Mfg3. At an SNR of 10 dB, the HASPI values are higher for Mfg2 than for Mfg1 or Mfg3. The HASQI scores in quiet and at the 10-dB SNR are higher for Mfg2 than for Mfg1 or Mfg3, while the scores for Mfg1 and Mfg3 are similar to each other. The basic and premium models from the three manufacturers show very similar HASPI and HASQI values.

4. Discussion

4.1 External Variables: SNR and Signal Level

The HASPI values plotted in Fig 5 only show a benefit of increasing the signal intensity for speech in quiet. When noise is present to give a SNR of 0 or 10 dB the signal level has essentially no effect. This pattern indicates that when noise is present at moderate to high levels, the noise dominates the nonlinear distortion generated by the hearing aid as well as the loss of audibility associated with the hearing loss. Only when the external noise is eliminated can one see the effect of increasing the signal level on improved predicted intelligibility, which can be explained as the benefit of increased audibility of the speech signal at the higher levels of presentation.

The HASQI values in Fig 5 show a different pattern than HASPI. HASQI is more sensitive to small amounts of distortion than is HASPI (c.f. Fig 3), while HASPI gives greater importance to audibility. Thus even when HASPI is close to its asymptote of 1 for the 75-dB SPL speech in quiet, HASQI is still quite a bit lower due to the nonlinear distortion present in the hearing-aid outputs. In addition, the HASQI values decrease as the signal level increases, both in quiet and in noise. This behavior in HASQI could be caused by an increase in distortion with increasing signal level as more of the speech falls above the compression threshold in each frequency band. Another possible explanation could be

increased audibility of existing distortion that falls above the impaired auditory threshold for the more-intense signals.

4.2 Hearing-Aid Programming: Audiogram and Degree of Processing

The NAL-R setting generally produced the highest HASPI and HASQI scores for all the hearing aids and test conditions. For the NAL-R settings the hearing-aid gain-*vs*-intensity function was set to be as linear as possible and the other algorithms disabled. The difference in the metric scores between NAL-R and the Mild, Mod, and Max settings indicates the impact of the WDRC, noise suppression, and frequency compression algorithms. The metrics measure the behavior of the entire system, including all the processing interactions, as they change with the processing settings. In general, increasing the aggressiveness of the processing reduced the HASPI and HASQI scores, indicating that the increase in distortion introduced by the processing outweighed the potential increase in audibility.

It is important to note that the Mild, Mod, and Max settings may have different meanings and different effects when comparing devices across manufacturers and when applied to different hearing losses. The data plotted in Fig 6 have been averaged across manufacturers and models, and therefore illustrate the general hearing-aid behavior. However, the wide error bars, for example for HASPI at a SNR of 10 dB, indicate that there can be substantial deviations from the general pattern when a specific hearing aid is programmed for an individual hearing loss. The impact of the processing settings also is dependent on the audiogram, with greater differences observed for the N4 audiogram than for S2. The S2 audiogram has much less loss at low frequencies than N4, so this behavior is consistent with reduced processing strength (e.g. lower dynamic-range compression ratios) associated with the higher auditory thresholds. The relative performance of a hearing aid results from the interaction of all the processing in the hearing aid, and can change with the hearing loss and the specific configuration.

4.3 Device: Manufacturer and Model

The HASPI and HASQI values show noticeable differences among devices. Overall, the hearing aids from Mfg2 have the highest HASPI and HASQI scores, while those from Mfg3 have the lowest. However, this overall judgment must be tempered by the measured performance for specific combinations of hearing aid, audiogram, SNR, and processing setting. For example, in Fig 7 the hearing aids from Mfg2 and Mfg1 have similar HASPI scores in quiet, but differ in noise. Thus measuring hearing-aid performance only for speech in quiet may ignore differences in performance that are only noticeable in noisy situations.

The measured differences between hearing aids also depends on the metric. For speech in quiet, the HASPI scores for Mfg2 and Mfg1 are very similar and are near the asymptote of 1. However, under the same test conditions the HASQI scores show significant differences. Thus there is a potential danger in focusing on intelligibility without also considering speech quality, and the HASQI scores show differences between manufacturers that would not be apparent if predicted intelligibility were the sole criterion. The quality metric is more sensitive to small differences between devices, and provides a different set of weights for

assessing the relative importance of audibility and distortion than implemented in the intelligibility metric.

The measured differences between the premium and basic hearing aids indicate that premium is not always better, a result consistent with the clinical outcomes of Cox *et al.* (2014). The premium hearing aids may differ from the basic models from the same company in the number of processing bands, type of signal processing, and parameters associated with the processing settings. Thus the trade-offs between audibility and distortion may change going from a basic to an premium hearing aid from the same manufacturer. However, the similarity in performance between the basic and premium devices suggests that increased processing complexity does not necessarily lead to improved performance.

4.4 Implementation Concerns

Conventional hearing aid measurements (e.g. ANSI S3.22, 2014) are primarily concerned with speech audibility. While the dynamic range of speech depends on language and measurement procedure (Jin *et al.*, 2014), it is normally assumed to be 30 dB (ANSI S3.5, 1997). Given this knowledge of the dynamic range, one needs only to measure the peak or RMS intensity of the speech to estimate the audibility of the entire signal (ANSI S3.5, 1997). Such measurements require that the speech signal be sufficiently above the background noise level to get a reasonable intensity estimate, which corresponds to an SNR of just 10 dB (ANSI S3.42, 2012). This SNR can easily be achieved in a quiet office or clinic.

HASQI, however, uses measurements of signal changes over the entire range of signal intensities. For example, HASQI includes the effects of noise during low-intensity portions of the speech to estimate the amount of audible noise and distortion in the hearing aid output signal, as opposed to just estimating the low-level speech levels from the speech RMS or peaks. HASQI is therefore much more sensitive to low levels of audible background noise than conventional hearing aid measurements. The sensitivity of HASPI and HASQI to background noise was plotted in Fig 3 for speech in a background of speech-shaped noise as a function of SNR. At a SNR of 10 dB, HASPI is 0.99, but HASQI is only 0.34. At a SNR of 30 dB, similar to the measurement environment used for this paper, HASQI is 0.76 and even at 40 dB only increases to 0.91. Thus the specification of a 10-dB SNR that works for conventional measurements is inadequate for HASQI, and the background noise level for the test system determines the upper limit for the highest HASQI value that can be measured.

As explained in Section 2.6, averaging improves the sensitivity of HASQI to hearing aid distortion, but reduces the sensitivity to the internal microphone and circuit noise. Because averaged hearing aid outputs have been used in this paper, the HASQI measurements show the effects of the babble added to the speech and indicate the impact of the amplification and nonlinear processing, but effectively ignore the hearing aid internal noise. If the distortion dominates the internal noise, HASQI computed for the linear NAL-R processing condition should remain nearly constant as the signal level is varied because the amount of nonlinear distortion tracks the signal level. On the other hand, if the internal noise is dominant, HASQI should increase as the input signal level increases due to the improved SNR within the hearing aid. The HASQI scores for NAL-R processing were compared for the three input

levels of 55, 65, and 75 dB SPL, and there was no observed increase in HASQI as the input signal level was increased. The averaging used to improve the HASQI sensitivity to low levels of distortion is therefore justified.

A further consideration is that the hearing-aid response may change over the course of the averaging if processing having very long time constants has been implemented in the device. While no long-term response changes were observed in the hearing aids tested for this paper, the averaging procedure may need to be adjusted (e.g. longer noise preamble prior to recording the processed sentences) if such behavior is found.

5. Conclusions

The results presented in this paper represent a first step in applying the HASPI and HASQI metrics to hearing aids. The results show that the metrics are capable of differentiating between hearing aids and processing settings, and that the measured differences can be substantial. HASPI and HASQI are based on perceptual results averaged over a large number of listeners and simulation studies, and are expected to be accurate in predicting average perceived hearing-aid responses. Individual factors, such as differences in cognitive abilities, loudness tolerance, or the preferred tradeoff between intelligibility and quality, are not included in the metrics. Thus comparisons between devices or processing settings are expected to be accurate on average, but there may still be differences in individual responses to the same hearing aid.

The trends in the hearing-aid measurements are consistent with those found previously for the laboratory processing simulations. The additive multi-talker babble reduced intelligibility and quality, and at the highest noise level the reductions in the metric values across the board were sufficient to make processing differences hard to discern. For most processing and noise conditions, the NAL-R setting for the hearing aids gave higher HASPI and HASQI values than any of the more-aggressive WDRC and frequency compression settings, and the Max setting tended to give lower values than the Mild setting. The HASQI scores were uniformly lower than HASPI, which is consistent with HASPI saturating at 1 at about a 10-dB SNR where HASQI is still approximately 0.34.

In comparing hearing aids, it was found that in general the premium models did not give higher scores than the basic devices. However, processing settings had a strong impact on the scores, with the least amount of processing, as represented by the NAL-R setting, often giving the highest HASPI and HASQI scores. Increased processing aggressiveness generally causes increased nonlinear distortion, which lowers the metric values. The observed differences for different processing settings for the same hearing aid illustrate the danger in generalizing too broadly from the limited set of measurements typically made in the clinic. A hearing aid that works well for one audiogram may not give the highest scores for a different audiogram. Furthermore, audibility is only one factor in fitting a hearing aid. The metrics also incorporate measurements of nonlinear distortion, with HASPI and HASQI weighing audibility and distortion in different proportions.

A potential concern in implementing the metrics is the sensitivity of HASQI to background noise when a hearing aid is being measured. However, the measurements in this paper indicate that nonlinear distortion in the hearing aids dominates the internal noise; signal averaging can therefore be used to reduce the impact of the ambient noise on the computed HASQI values. Even so, the ambient noise may set an upper bound on the maximum HASQI score that can be achieved when measuring hearing aids in practical situations.

The results presented in this paper are a first step in applying the HASPI and HASQI metrics to hearing aids. However, the results do not indicate the clinical relevance of the HASPI and HASQI scores or provide target values that might be appropriate for successful fittings. More data are needed to relate the metric scores to clinical outcomes. For example, experiments are needed to relate the HASPI score to intelligibility when listening through hearing aids as opposed to simulated processing. Experiments are also needed to establish the relationship between the HASPI and HASQI scores and listening preference and long-term listener satisfaction with their hearing aids.

Acknowledgments

References

Amlani AM, Punch JL, Ching TYC. Methods and applications of the audibility index in hearing aid selection and fitting. Trends Amplif. 2002; 6:81–129. [PubMed: 25425917]

ANSI. Methods for the calculation of the Speech Intelligibility Index. Am Nat Std Inst. 1997 S3.5-1997, Approved 6 June 1997.

ANSI. Testing Hearing Aids – Part 2: Methods for characterizing signal processing in hearing aids with a speech-like signal. Am Nat Std Inst. 2012 S3.42-2012/Pt. 2, Approved 12 sept 2012.

ANSI. Methods of measurement of real-ear performance characteristics of hearing aids. Am Nat Std Inst. 2013 S3.46-2013, Corrected and Republished, Sept 2014.

ANSI. Specification of hearing aid characteristics. Am Nat Std Inst. 2014 S3.22-2014, Approved 10 Nov 2014.

Bisgaard N, Vlaming MSMG, Dahlquist M. Standard audiograms for the IEC 60118-15 measurement procedure. Trends Amplif. 2010; 14:113–120. [PubMed: 20724358]

Byrne D, Dillon H. The national acoustics laboratories' (NAL) new procedure for selecting gain and frequency response of a hearing aid. Ear Hear. 1986; 7:257–265. [PubMed: 3743918]

Cox R, Alexander G, Gilmore C. Development of the Connected Speech Test (CST). Ear and Hear. 1987; 8(Suppl):119S–125S.

Cox RM, Johnson JA, Xu J. Impact of advanced hearing aid technology on speech understanding for older listeners with mild to moderate, adult-onset, sensorineural hearing loss. Gerontology. 2014; 60:557–568. [PubMed: 25139516]

Dubbelboer F, Houtgast T. A detailed study on the effects of noise on speech intelligibility. J Acoust Soc Am. 2007; 122:2865–2871. [PubMed: 18189576]

Falk TH, Parsa V, Santos JF, Arehart KH, Hazrati O, Huber R, Kates JM, Scollie S. Objective quality and intelligibility prediction for users of assistive listening devices. IEEE Sig Proc Mag. 2015; 32:114–124.

Festen JM, Plomp R. Application of the Speech Transmission Index to the Hearing impaired. In: Houtgast T, Steeneken H, editorsPast, Present and Future of the Speech Transmission Index. TNO Human Factors; Soesterberg, The Netherlands: 2002. 69–78.

Hilkhuysen G, Huckvale M. Signal properties reducing intelligibility of speech after noise reduction. 18th European Sig. Proc. Conf.; Aalborg. Aug 23–27, 2010; 2010.

Hohmann V, Kollmeier B. The effect of multichannel dynamic compression on speech intelligibility. J Acooust Soc Am. 1995; 97:1191–1195.

Houben R, Brons I, Dreschler WA. A method to remove differences in frequency response between commercial hearing aids to allow direct comparison of the sound quality of hearing-aid features. Trends Ampl. 2011; 15:77–83.

Houtgast T, Steeneken HJM. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J Acoust Soc Am. 1985; 77:1069–1077.

Huber R, Parsa V, Scollie S. Predicting the perceived sound quality of frequency-compressed speech. PLoS ONE. 2014:9. paper e110260 (13 pages).

Humes LE, Dirks DD, Bell TS, Ahlstrom C, Kincaid GE. Application of the Articulation Index and the Speech Transmission Index to the recognition of speech by normal-hearing and hearing-impaired listeners. J Speech Hear Res. 1986; 29:447–462. [PubMed: 3795887]

Jin IK, Kates JM, Arehart KH. The dynamic range for speech materials in Korean, English, and Mandarin: A cross-language comparison. J Speech Lang Hear Res. 2014; 57:2024–2030. [PubMed: 24826909]

Kates JM. An auditory model for intelligibility and quality predictions. Proc. Mtgs. Acoust. (POMA) 19, 050184: Acoust. Soc. Am. 165th Meeting; Montreal. June 2–7, 2013; 2013.

Kates JM, Arehart KH. The hearing-aid speech quality index (HASQI). J Audio Eng Soc. 2010; 58:363–381. (2010).

Kates JM, Arehart KH. The hearing-aid speech perception index (HASPI). Speech Comm. 2014a; 65:75–93.

Kates JM, Arehart KH. The hearing aid speech quality index (HASQI), version 2. J Audio Eng Soc. 2014b; 62:99–117.

Keidser G, Dillon H, Flax M, Ching T, Brewer S. The NAL-NL2 fitting procedure. Aud Res. 2011; 1(e24):88–90.

Kendrick P, Jackson IR, Li FF, Cox TJ, Fazenda BM. Perceived audio quality of sounds degraded by non-linear distortions and single-ended assessment using HASQI. J Audio Eng Soc. 2015; 63:698–712.

Kressner AA, Anderson DV, Rozell CJ. Evaluating the generalization of the hearing-aid speech quality index (HASQI). IEEE Trans Audio Speech Lang Proc. 2013; 21:407–415.

Lee LW, Geddes ER. Perception of microphone noise in hearing instruments. J Acoust Soc Am. 1998; 104:3364–3368. [PubMed: 9857504]

Lewis JD, Goodman SS, Bentler RA. Measurement of hearing aid internal noise. J Acoust Soc Am. 2010; 127:2521–2528. [PubMed: 20370034]

Ludvigsen C, Elberling C, Keidser G. Evaluation of a noise reduction method: Comparison between observed scores and scores predicted from STI. Scand Audiol. 1993; 22(Suppl. 38):50–55.

Nilsson M, Ghent RM, Bray V. Development of a test environment to evaluate performance of modern hearing aid features. J Am Acad Audiol. 2005; 16:27–41. [PubMed: 15715066]

Nilsson M, Soli SD, Sullivan J. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. J Acoust Soc Am. 1994; 95:1085–1099. [PubMed: 8132902]

Pavlovic CV, Studebaker GA, Sherbecoe RL. An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals. J Acoust Soc Am. 1986; 80:50–57. [PubMed: 3745665]

Procházka H, Landau ID. Pole placement with sensitivity function shaping using 2nd order digital notch filters. Automatica. 2003; 39:1103–1107.

Rhebergen KS, Versfeld NJ, Dreschler WA. The dynamic range of speech, compression, and its effect on the speech reception threshold in stationary and interrupted noise. J Acoust Soc Am. 2009; 126:3236–3245. [PubMed: 20000937]

Smeds K, Leijon A, Wolters F, Hammarstedt A, Båsjö S, Hertzman S. Comparison of predictive measures of speech recognition after noise reduction processing. J Acoust Soc of Am. 2014; 136:1363–1374. [PubMed: 25190409]

Souza PE, Yueh B, Sarubbi M, Loovis CF. Fitting hearing aids with the Articulation Index: Impact on hearing aid effectiveness. J Rehab Res and Dev. 2000; 37:473–481.

Souza PE, Turner CW. Quantifying the contribution of audibility to recognition of compression-amplified speech. Ear & Hear. 1999; 20:12–20.

Suelzle D, Parsa V, Falk TH. On a reference-free speech quality estimator for hearing aids. J Acoust Soc Am. 2013; 133:EL412–EL418. [PubMed: 23656102]

Sullivan GM, Feinn R. Using effect size – or why the *p* value is not enough. J Grad Med Ed. 2012; 4:279–282.

Taal CH, Hendriks RC, Heusdens R, Jensen J. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans Audio Speech and Lang Proc. 2011; 19:2125–2136.

Van Kuyk S, Kleijn WB, Hendriks RC. An evaluation of intrusive instrumental intelligibility measures. 2017 arXiv:1708.06027 (15 pages).

Author Manuscript

Author Manuscript

Author Manuscript
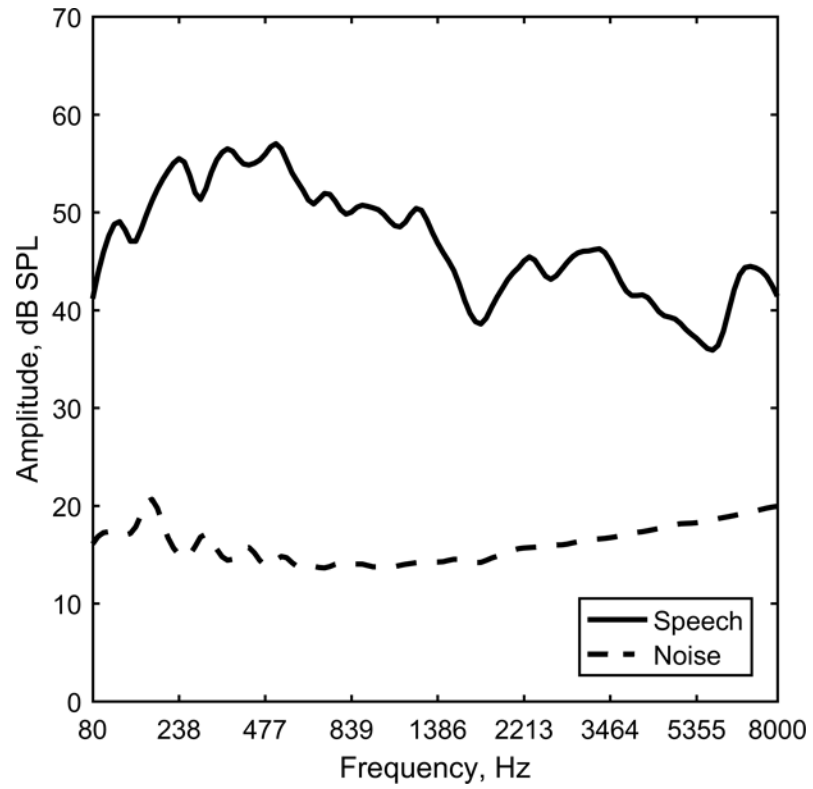
Author Manuscript
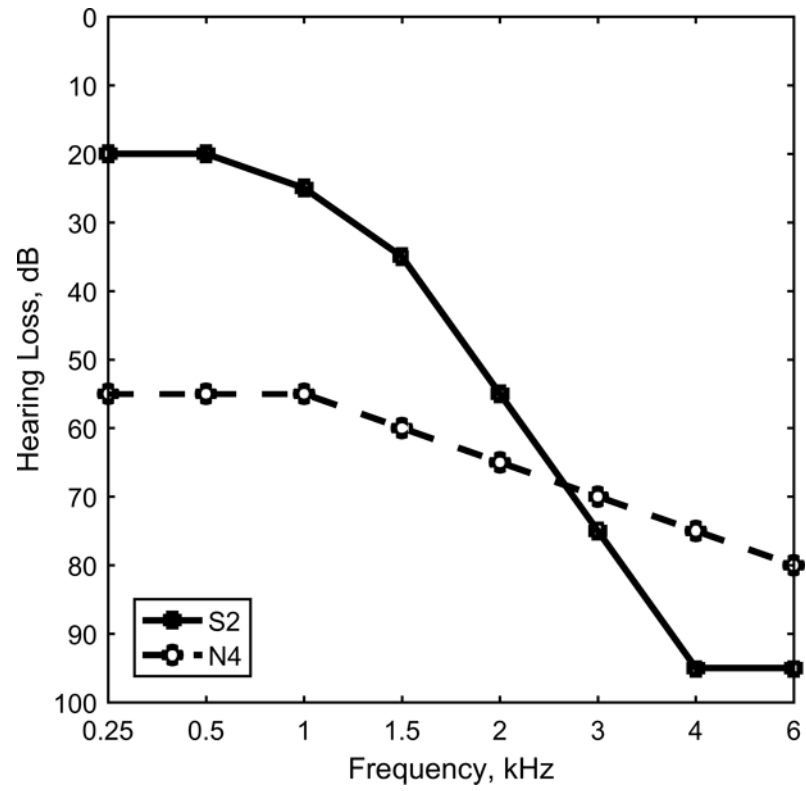
Fig 1.
Standard audiograms used for the hearing-aid settings and measurements.
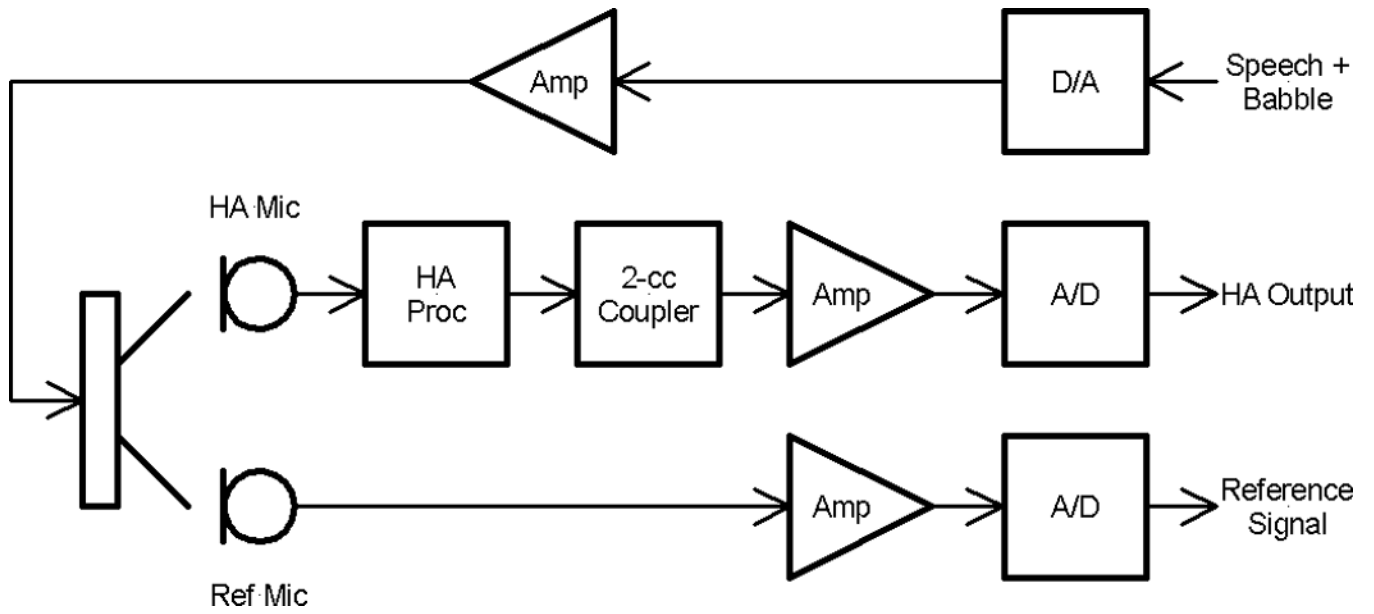
Fig 2.
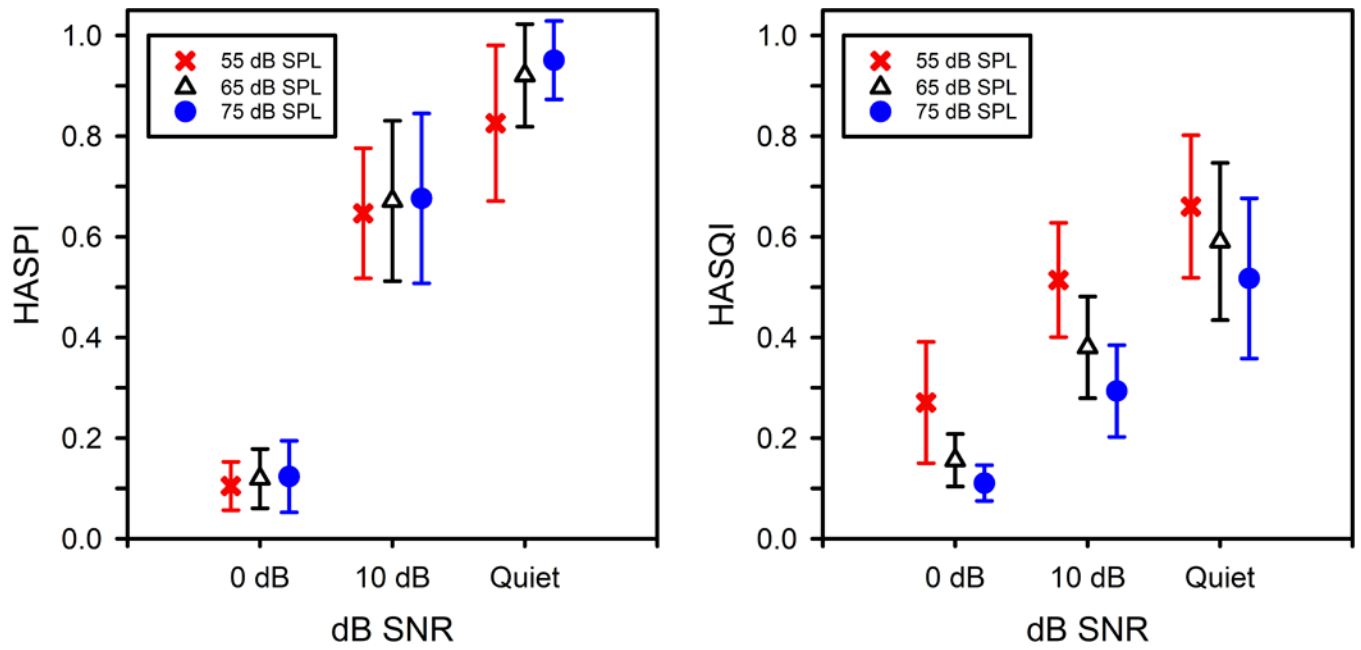Block diagram of the hearing-aid measurement system.
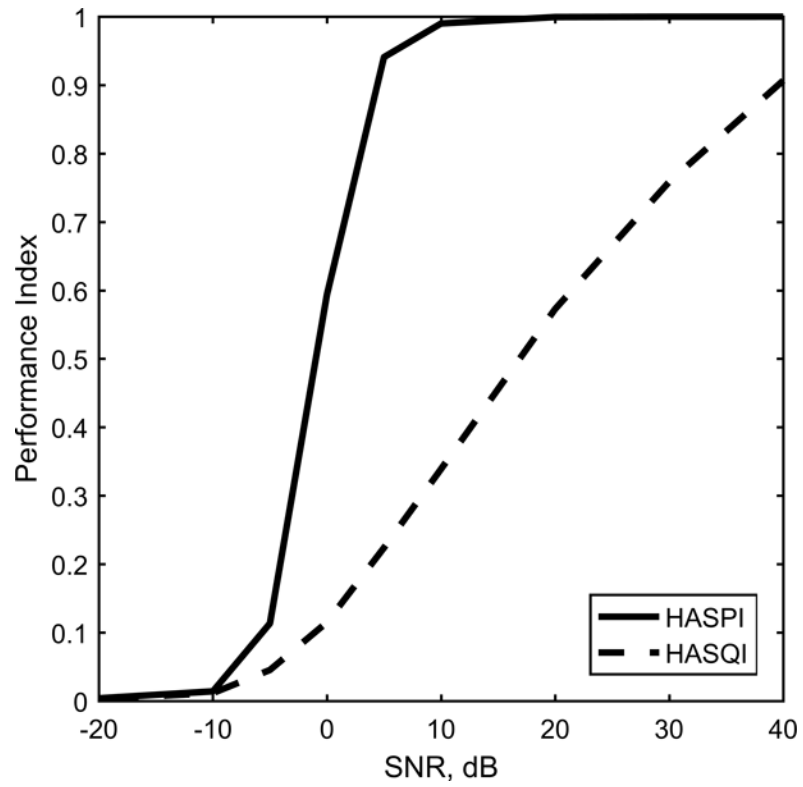
Fig 3.
HASPI and HASQI as a function of SNR for the pair of test sentences in speech-shaped random noise, normal hearing with no hearing-aid processing.
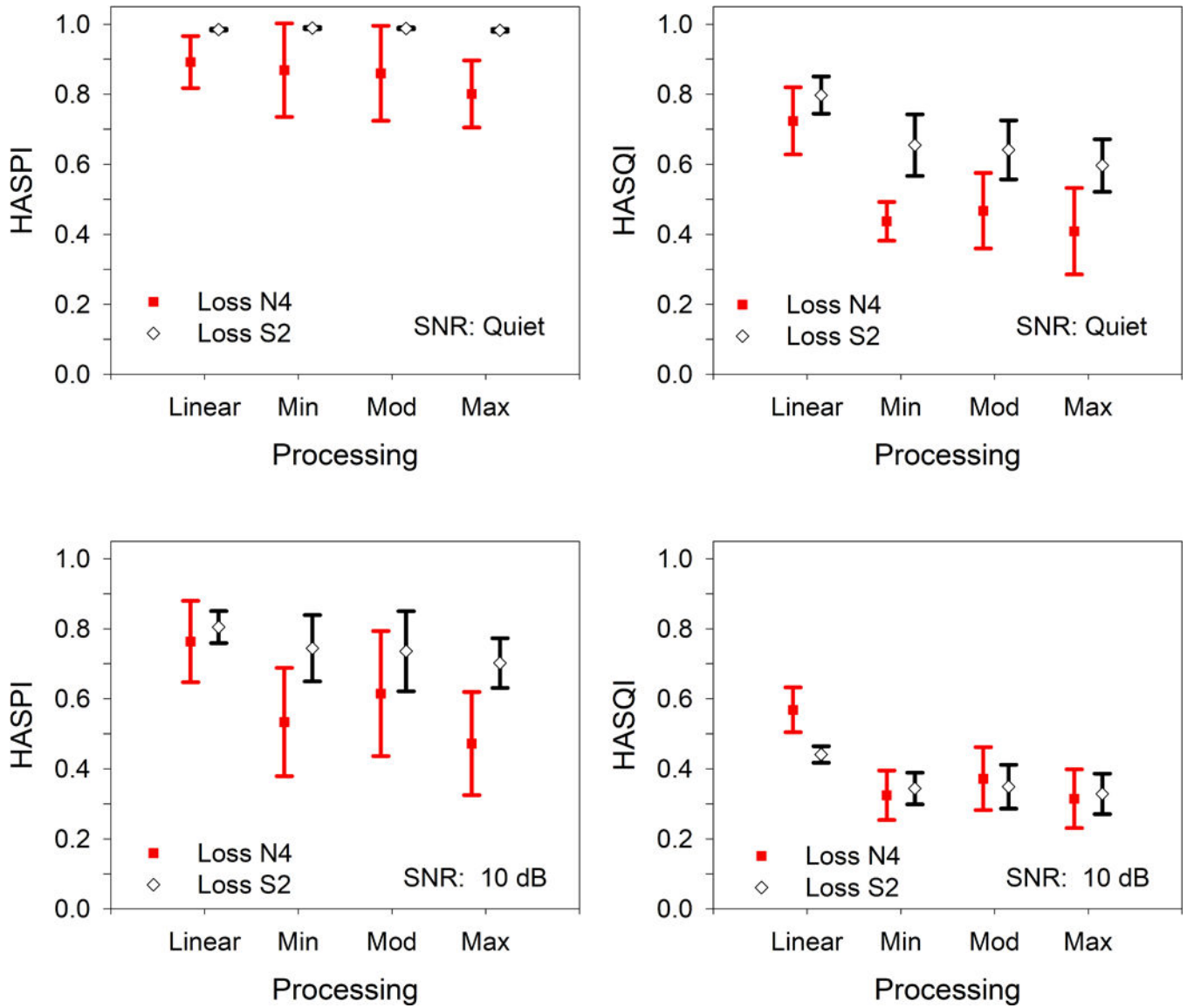
Fig 4.
Auditory spectra measured for the sentence pair output at 65 dB SPL at the reference microphone position in the hearing aid test box.
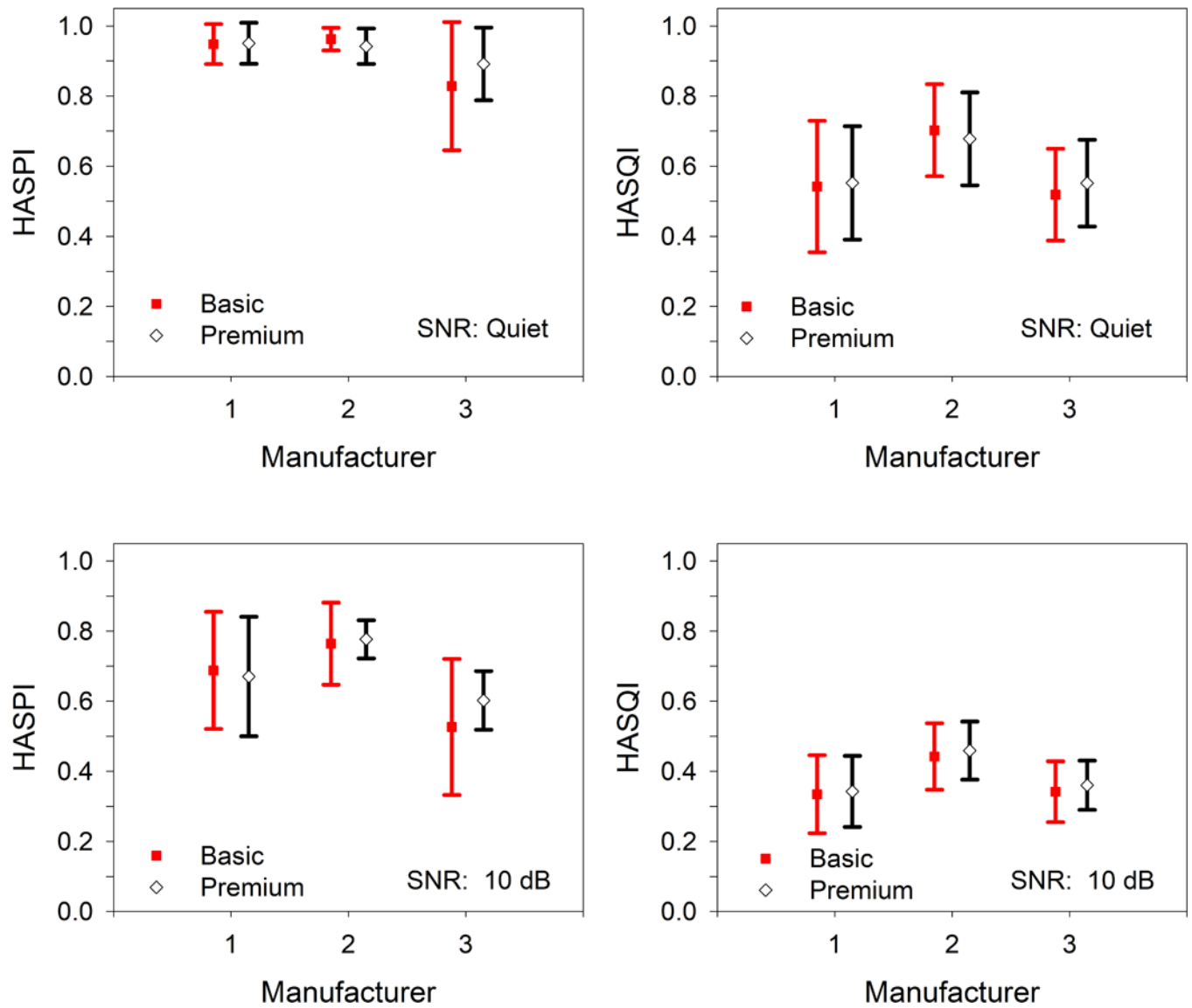
Fig 5.
HASPI and HASQI as a function of the SNR and presentation level, averaged over the hearing loss, degree of processing, manufacturer, and model.

Fig 6.
HASPI and HASQI as a function of degree of processing and hearing loss for stimuli in quiet and at an SNR of 10 dB at an input level of 65 dB SPL. The data are averaged over manufacturer, and model.

Fig 7.
HASPI and HASQI as a function of manufacturer and model in quiet and at an SNR of 10 dB at a signal level of 65 dB SPL. The data are averaged over hearing loss and degree of processing.

Table I

Range of noise and signal-processing conditions tested in the hearing-aid measurements.

Experimental Parameter	Hearing-Aid Processing Hypothesis
SNR: Quiet/10/0 dB	The nonlinear hearing-aid processing, such as noise suppression, is expected to depend on the speech SNR. The external noise is expected to be the strongest factor in the HASPI and HASQI scores, but the scores may also vary across hearing aids depending on how the devices process the noisy speech.
Signal level: 55/65/75 dB SPL	The hearing-aid processing, such as WDRC, is expected to change with the input signal level. Thus the measured amount of distortion may depend on input intensity.
Audiogram: S2/N4	The hearing-aid frequency response and nonlinear processing is dependent on the audiogram. Difference audiograms are expected to tax the hearing-aid amplification to different degrees and to cause different frequency-dependent patterns of nonlinear distortion, and thus produce different HASPI and HASQI values.
Degree of processing: NAL-R/ Mild/Mod/Max	The amount of nonlinear distortion produced by a hearing aid will depend on the type and aggressiveness of the signal processing. Linear processing is expected to give the highest HASPI and HASQI scores, and the maximum processing settings the lowest scores.
Hearing aid manufacturer and model	Different manufacturers use different microphones, receivers, circuitry, and algorithms in their hearing aids. In addition, the signal processing available in premium *versus* basic devices may differ. These implementation differences are expected to produce differences in measured performance.

Table II

Stimulus conditions used for the hearing-aid measurements.

Stimulus Condition	SNR, dB	Overall Level, dB SPL
Low-intensity speech quiet	No noise	55
Moderate-intensity speech quiet	No noise	65
High-intensity speech quiet	No noise	75
Low-intensity speech in noise	0, 10	55
Moderate-intensity speech in noise	0, 10	65
High-intensity speech in noise	0, 10	75

Table III

Analysis of variance (ANOVA) of the mixed-effects linear model for the HASPI intelligibility measurements.

	Num. *df*	Denom. *df*	*F*	*p*
SNR	2	355	5543.69	<0.001
Level	2	355	31.62	<0.001
Loss	1	355	319.54	<0.001
Proc	3	355	36.84	<0.001
Mfg	2	2	12.69	0.0730
Model	1	2	0.77	0.4737
SNR × Level	4	355	10.64	<0.001
SNR × Loss	2	355	42.72	<0.001
SNR × Proc	6	355	9.16	<0.001
Level × Proc	6	355	5.60	<0.001
Loss × Proc	3	355	7.77	<0.001
SNR × Level × Loss	4	355	14.20	<0.001
SNR × Level × Proc	12	355	0.84	0.6060
SNR × Loss × Proc	6	355	3.56	0.0019
Level × Loss × Proc	6	355	0.64	0.6965
SNR × Level × Loss × Proc	12	355	0.75	0.6974

Table IV

Analysis of variance (ANOVA) of the mixed-effects linear model for the HASQI quality measurements.

	Num. df	Denom. df	F	p
SNR	2	355	3410.12	<0.001
Level	2	355	625.46	<0.001
Loss	1	355	2.62	0.1066
Proc	3	355	365.88	<0.001
Mfg	2	2	76.27	0.0129
Model	1	2	1.82	0.3095
SNR × Level	4	355	14.47	<0.001
SNR × Loss	2	355	281.96	<0.001
SNR × Proc	6	355	31.36	<0.001
Level × Proc	6	355	1.44	0.1984
Loss × Proc	3	355	44.95	<0.001
SNR × Level × Loss	4	355	26.40	<0.001
SNR × Level × Proc	12	355	1.98	0.0253
SNR × Loss × Proc	6	355	1.32	0.2494
Level × Loss × Proc	6	355	4.26	<0.001
SNR × Level × Loss × Proc	12	355	0.18	0.9992

Table V

Standardized effect sizes (Cohen's *d*) for the HASPI and HASQI main effects. The lower and upper bounds of the 95-percent confidence intervals are given within the parentheses. Confidence intervals that do not include 0.0 effect size are marked with an asterisk.

	HASPI	HASQI
SNR	6.30 (5.41, 7.3)*	2.64 (2.37, 2.91)*
Level	0.56 (0.38, 0.71)*	-2.01 (-2.26, -1.78)*
Loss	0.91 (0.80, 1.01)*	0.05 (-0.08, 0.18)
Proc	-0.75 (-0.88, -0.62)*	-1.53 (-1.75, -1.33)*
Mfg	-0.71 (-0.83, -0.57)*	0.08 (-0.08, 0.25)
Model	0.24 (0.13, 0.34)*	0.30 (0.16, 0.42)*