Comparing hearing-aid algorithm performance using Simulated Performance Intensity Functions

ANDREW HINES AND NAOMI HARTE

Dept. of Electronic and Electrical Engineering, Trinity College Dublin, Ireland

Simulated performance-intensity functions were used to quantitatively discriminate speech intelligibility through phoneme discrimination assessment. Listener test results for subjects with a wide range of sensorineural hearing losses were simulated using an auditory nerve model and compared to real listeners' unaided and aided performance. Simulations of NAL-RP and DSL 4.0 fitting algorithms were compared. Auditory-nerve discharge patterns from the model were presented as neurograms. An automated ranking process was used to quantify neurogram degradation using a new measure, the Neurogram Similarity Index Measure (NSIM). The measure has previously been shown to correlate well in predictions of phoneme discrimination for normal hearing listeners in both quiet and noise. In this study, simulated responses to consonant-vowel-consonant word lists in a quiet environment at a range of presentation levels were used to produce phoneme discrimination scores. This represents a further step in validating the use of auditory-nerve models to predict speech intelligibility for different hearing-aid fitting methods in a simulated environment, allowing the potential for rapid prototyping and early design assessment of new hearing-aid algorithms.

INTRODUCTION

Developing improved hearing-aid algorithms is an intensive process in terms of test subjects, labour and time. A simulated test environment would allow rapid prototyping and basic assessment of new fitting algorithms. The ability to test and quantitatively compare the speech intelligibility improvements offered by different hearing-aid fitting methods would not replace listener tests, but could significantly reduce development costs and times.

The Simulated Performance Intensity Function (SPIF) test methodology developed by the authors, allows experimentation using an Auditory Nerve (AN) model to predict the phoneme recognition of listeners. This work seeks to reproduce the results for human listeners with a range of sensorineural hearing losses (SNHLs) by investigating whether the AN model yields comparable results with the same dataset. Experiments were carried out in unaided and aided scenarios. Prior work (Hines and Harte, 2011), showed that the SPIF test methodology produced a good prediction of Performance-Intensity (PI) functions for normal hearing listeners.

Proceedings of ISAAR 2011: Speech perception and auditory disorders. 3rd International Symposium on Auditory and Audiological Research. August 2011, Nyborg, Denmark. Edited by T. Dau, M. L. Jepsen, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN 87-990013-3-0. EAN 9788799001330. The Danavox Jubilee Foundation, 2012.

BACKGROUND

The Zilany *et al*. (2009) AN model is used to produce neurograms. Neurograms represent the auditory nerve discharge patterns in a time-frequency plot of intensity and are analogous to a signal spectrogram. The methodology used to create neurograms is described in detail in prior work (Hines and Harte, 2010). Neurograms for each phoneme are assessed as an image comparison metric, described below, between the test neurogram and a reference neurogram from a normal hearing AN model for the same input signal. The Neurogram Similarity Index Measure (NSIM), used here to compare neurograms, is a simplified version of the Structural Similarity (SSIM; Wang *et al*., 2004) index and is defined as

$$Nr,d=lr,d\cdot sr,d=2\mu r\mu d+C1\mu r2+\mu d2+C1\cdot\sigma rd+\ C2\sigma r\sigma d+C2 \quad \text{(Eq. 1)}$$

The NSIM between two neurograms, the reference (r), and the degraded (d), is constructed as a weighted function of intensity (l), and structure (s) as in eqn. (1). Intensity looks at a comparison of the mean (μ) values across the two neurograms. The structure uses the standard deviation (σ) and is equivalent to the correlation coefficient between the two neurograms. As with SSIM, each component contains constant values (C_1=0.01L and C_2=$(0.03L)^2$), where L is the intensity range, as per Wang *et al*. (2004), which have negligible influence on the results but are used to avoid instabilities at boundary conditions. A simulated PI function is produced by using NSIM to rank a large number of neurogram comparisons, over a range of intensity levels.

Simulated Performance Intensity Function (SPIF)

A PI function is used to plot phoneme discrimination against speech intensity. Evaluation of a test subject's Speech Reception Threshold (SRT) and word recognition in lists of phonetically balanced words allows validation of pure tone thresholds and estimation of auditory resolution respectively. The PI function has been shown to be useful for comparative tests of aided and unaided speech recognition results and it has been proposed as a useful method of evaluation of the performance improvement of subjects' speech recognition under different hearing-aid prescriptions or settings (Boothroyd, 2008).

The test corpus used came from the Computer Aided Speech Perception Assessment (CASPA; Boothroyd, 2006) software package which was developed to simplify the data recording and analysis for performance intensity listener tests. It contains 20 word lists of 10 phonemically balanced Consonant-Vowel-Consonant (CVC) words. Words are not repeated within 10 word lists and lists are designed to be isophonemic, i.e. to contain one instance of each of the same 30 phonemes. In a standard performance intensity listener test, CVC words are presented to the test subject who listens and repeats the words. The tester manually scores the open-set results, per phoneme correctly identified. This is repeated at a progressive range of intensity levels and a PI function is produced. To create a SPIF, the listener is

replaced by the AN model and scoring is based on automated NSIM comparisons of the neurograms produced by the nerve firing simulations from the model.

Hearing Profiles and Hearing Aid Algorithms

Three hearing impaired listeners were tested by Boothroyd (2008) with *flat moderate*, *flat severe* and *high frequency severe* impairments. These hearing impairments are simulated here for comparison with the reported results. The levels of hearing losses were simulated using the AN model using percentage inner and outer hair cell losses for the audiograms calculated using estimates provided in the Zilany *et al*. (2009). Two linear hearing aid fitting algorithms were tested: NAL-RP (National Acoustics Laboratory - Revised, Profound) and DSL 4.0 (Desired Sensation Level). The formulae for calculating insertion gains for these fitting algorithms are described in Dillon (2001).

SIMULATED TESTS

SPIF listener tests were carried out using the AN model to simulate listeners with SNHLs in unaided and aided scenarios. For this experiment, software implementations of the NAL-RP and DSL 4.0 algorithms were developed to apply the required insertion gains to the input signals. The hearing loss thresholds for the modelled test subjects are presented in Fig. 1. The thresholds are a mean of the left and right ear values for the human listener test subject where there were slight differences in the left/right ear thresholds (Boothroyd, 2008).

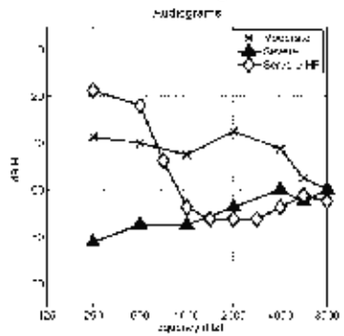

Fig. 1: Audiograms for hearing SNHLs tested

Hearing Type	Unaided PRT (dB SPL)	Aided PRT (dB SPL)
Unimpaired	15	-
Flat Moderate	54	42
Flat Severe	82	41
High Frequency Severe	70	-

Table 1: Phoneme Recognition Threshold (PRT) levels, unaided and aided, by hearing loss from Boothroyd (2008).

The SPIF procedure mimics that of a real listener test. The human listener is substituted with the AN model and the NSIM scores are used to assess neurogram degradation and to predict phoneme discrimination. Timing label files marking the phoneme boundaries were created for the 200 words from the CASPA corpus.

For normal hearing listeners, the phoneme recognition threshold (PRT; that is, the level in dB SPL at which the listener scores 50% of their maximum) was set at 15

dB SPL as per Boothroyd (2008). A level of 65 dB SPL was taken as the standard level to generate reference neurograms to test against.

The similarity measurement between a reference neurogram and a degraded neurogram at the PRT level measured over a large sample of phonemes gives a neurogram PRT (NPRT). The NPRT was evaluated at the PRT level measured in the real listener test, per phoneme position, using 10 lists of CVC words, as the median NSIM score at the PRT level. The word lists were then presented to the AN model at ten speech intensity levels in 5 dB increments covering sub-threshold to peak intelligibility levels. The same procedure that was used for evaluation of the NPRT was repeated at each speech intensity level using 5 other word lists (150 phonemes). The results were recorded and a phoneme discrimination score was calculated by counting the number of phonemes scoring above the NPRT value and a SPIF was plotted from the results. This procedure was repeated for each hearing loss in unaided and aided scenarios using the PRT values in Table 1.

HEARING LOSSES TESTED

A Flat Moderate Sensorineural Hearing Loss

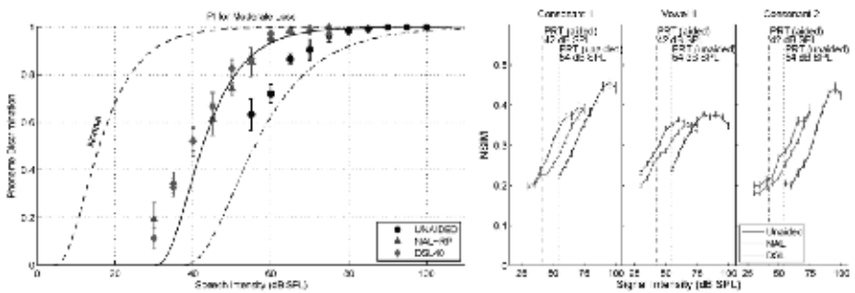

Fig. 2: Simulated PI functions and NSIM results for a *flat moderate* loss

The real listener test was carried out by Boothroyd on an adult with a *flat moderate* SNHL. Binaural phoneme recognition scores were obtained using headphones. The results were fitted to a PI function curve and are presented as the lines on the simulated PI function in Fig. 2. The NSIM scores for the simulations are also presented, broken down by phoneme position (i.e. initial consonant, vowel, final consonant). The bars mark one standard error.

The SPIF presents a normal listener result, for reference, which has been normalised to a PRT of 15 dB SPL and is plotted as a dashed line. The next two curves are the aided and unaided curves fitted to the results from the listener test. The triangle and diamond points mark the NAL-RP and DSL 4.0 aided simulations and the circles show the unaided simulation. The hearing aid shifts the PI curve by around 15-20 dB for the *flat moderate* hearing loss tested, which, from the audiogram in Fig. 1, can be seen to have a threshold loss ranging from 35 to 60 dB HL. The unaided results are a close match to the trend but are offset and over-predict the phoneme

recognition. Overall, the results track within the error bounds of psychoacoustic tests.

A Flat Severe Sensorineural Loss

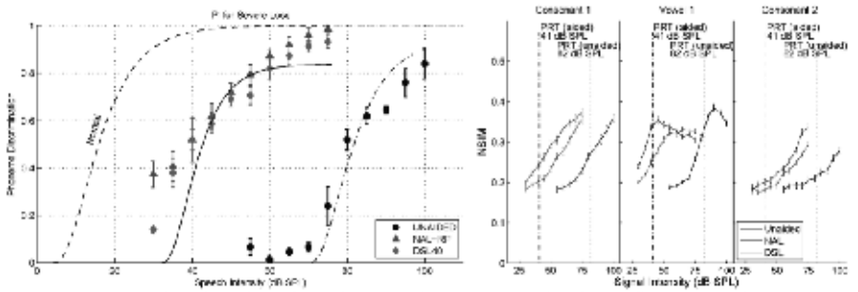

Fig. 3: Simulated PI functions and NSIM results for a *flat severe* loss

The results for an adult with a *flat severe* SNHL are presented in Fig. 3. They show that both the unaided and aided PI functions are steeper than those in the *flat moderate* case with phoneme recognition - peaking between below 90% for unaided listening through headphones. The intensity range for optimal scoring is narrower and a difference either way results in lower scoring due to audibility or discomfort (Boothroyd, 2008). The unaided NSIM scores show a sharp tail-off in similarity scores at high presentation levels for vowels. This is in contrast to the aided case where the vowel plateaus at a similarity level close to the unaided maximum. The NSIM results predict the range of optimal listening being extended from a few dB to around 25 dB. This feature is visible in the PI function for the listener test but is not replicated in the SPIF results where the aided phoneme recognition scores do not plateau. It is likely that this is due to the influence of the consonants where the NSIM trends continuously upwards over the range tested. The simulated results closely fit the listener test for the unaided case and show similar improvements in dB necessary for comparable phoneme discrimination when aided, but do not predict the maximum recognition tail-off in the aided case.

A High-Frequency Severe Sensorineural Hearing Loss

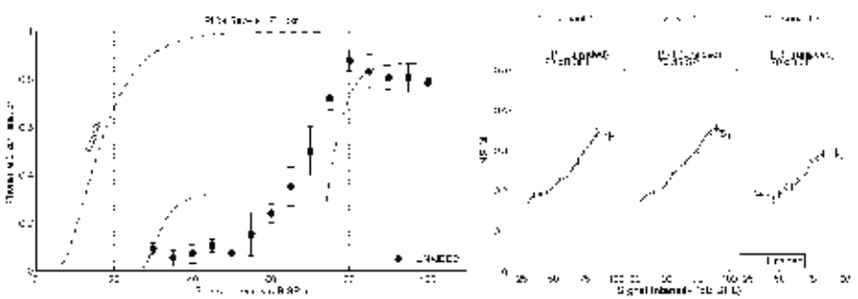

Fig. 4: Simulated PI function and NSIM results for a *high frequency severe* loss

Boothroyd (2008) only presented results for an adult with a *high frequency severe* SNHL in unaided conditions so aided simulations were not simulated. People with an audiogram similar to the one tested typically have a PI function composed of two sections (Boothroyd, 1968), as illustrated in the PI function in Fig. 4. The lower section is an initial threshold where a poor phoneme discrimination can be attained from the low frequency speech components alone. As vowel formants and the higher frequencies which make up consonants are not available, the low threshold for this listener is around 35%. When speech intensity increases, higher frequency speech cues become audible and the PI function begins to climb again in the second section of the PI function. The NSIM results show a trend that plateaus at a maximum similarity for both consonants and vowels. The simulated PI fails to predict the first section of the PI function where it predicts almost no recognition. The second section follows the listener PI as the higher frequencies become audible but it underestimates the maximum phoneme discrimination level - though it does match the speech intensity at which the PI curve reaches a maximum.

DISCUSSION

Simulation and Clinical Test Comparison

Comparing the results in Fig. 2 from the simulated test with the real listener results (points are the simulation results, lines are the PI functions fitted to the real listener results), the overall correlation is very promising. The key area of interest is between the 50% phoneme discrimination (%P.D.) and the maximum level. The results for the *flat moderate* SNHL (unaided) follow the shape of the listener curve quite closely but are over predicting the %P.D. and have shifted by 5-10 dB. This will be looked at in more detail below. The aided SPIF results closely fit the predicted listener PI function.

The error bars (representing +/- 1 standard error) for the simulated results are smaller than those for the real listener tests. The reported real listener tests refer to individuals rather than group means and used fewer word lists to test phoneme recognition than in the simulations, so from a purely statistical perspective such smaller error bars would be expected as there is not as much data available to establish the range and outliers. The size of the error bars highlight the variance in results from a clinical environment.

At high presentation levels the NSIM scores begin to drop, which may be a representation of rollover effects decreasing phoneme discrimination. A very small increase in the NPRT level would cause a significant change to the %P.D. The fit for the unaided *flat moderate* SNHL would improve the fit significantly, by applying a shift of the PRT by 1dB, suggesting that for good correlation, the methodology is heavily dependent on an accurate PRT measurement. This highlights the importance of an accurate PRT levels, together with an audiogram, as prerequisites for a reliable simulation. It should be noted that the tests simulated were for individuals tested and

reported by Boothroyd (2008) and hence experimental conditions could only matched to the details reported.

Fitting algorithm comparisons

Work has been done by others to investigate hearing aid fitting algorithms using AN models. Bruce *et al*. (2007) tested NAL-R and DSL 4.0 to find optimal single-band gain adjustments based on the response of auditory-nerve fibres to speech. They examined a range of dB adjustments above and below the prescribed target insertion gains. A mean absolute error measure was used to establish minimum neurogram differences. The results showed optimal gain adjustments for the NAL-R prescription were somewhat higher than those for DSL, and were consistent with the generally-lower insertion gains of NAL-R.

Here, the SPIFs for both fitting algorithms predicted negligible differences in phoneme recognition. However, the NSIM showed that neurogram similarities were higher for DSL than for NAL-RP. This can be explained by examining the procedure used in calculating the predicted PI scores. The percentage phoneme discrimination at any given intensity is calculated as the number of phonemes with NSIM greater than the NPRT. The magnitude of the NSIM above the NPRT threshold is not taken into account, so, the NSIM scores for NAL and DSL may display differences which do not translate into a significant difference in intelligibility when the SPIF is plotted. The hearing aid used for the real listener test was not specified, so the same aided PRT value was used for both the NAL and DSL simulations to calibrate their NPRT levels. This accounts for their results at 50% discrimination matching, but not for other intensity levels. Tests of hearing impaired listeners with PRT levels measured individually for each hearing aid algorithm would benefit further study. SPIFs created from NSIM measures of neurograms demonstrate that a correlation exists between neurogram similarity and speech intelligibility. However, it is possible that maximising the similarity is unnecessary as long as a threshold similarity level exits. Conversely, the neurogram similarity may be a good indicator of other factors beyond intelligibility such as speech quality, as has been investigated by Kates (2010).

Other research, carried out by Bondy *et al*. (2004) used their neurocompensation technique to model a range of SNHLs. Their results predicted optimal target insertion gains for hearing aids and the results predicted optimal gains which were close to those of NAL-R. This work shows that although NAL-RP and DSL 4.0 predict significantly different targets, the overall PI functions remain very similar. This could mean that for a given SNHL the optimal prescribed target insertion gains are not a single prescription but that a range of values, including those empirically found and used for NAL-RP and DSL-4.0 will work sufficiently well to give comparable PI functions. This was seen in a recent study by Ching *et al*. (2010) which tested the newer versions of NAL (NAL-NL1) and DSL (DSL 4.1) on a group of 48 children and showed both intelligibility judgments and preferences were equally split between prescriptions on average.

CONCLUSIONS

This study demonstrated that a SPIF can predict speech intelligibility for a range of hearing impairments. These results are promising, indicating that using the AN model, predict speech intelligibility results, even for aided listeners with SNHL. The NAL-RP and DSL 4.0 linear hearing-aid fitting algorithms were compared using simulated performance intensity functions. The results showed that, while for both a *flat moderate* and *flat severe* SNHL the simulated results matched those for real listeners, there was little to differentiate the results for the fitting algorithms. From a speech intelligibility perspective, the simulations predicted that both algorithms provide similar intelligibility gains which reinforces the empirical findings of Ching *et al*.

REFERENCES

Bondy, J., S. Becker, *et al*. (**2004**). "A novel signal-processing strategy for hearing-aid design: neurocompensation" Signal Processing **84** (7), 1239-1253.

Boothroyd, A. (**1968**). "Developments in Speech Audiometry" Sound **2** (1): 3 - 10.

Boothroyd, A. (**2006**). Computer-Aided Speech Perception Assessment (CASPA) 5.0 Software Manual. San Diego, CA. San Diego, CA.

Boothroyd, A. (**2008**). "The Performance/Intensity Function: An Underused Resource" Ear and Hearing **29** (4), 479-491.

Bruce, I. C., F. Dinath, *et al*. (**2007**). Insights into optimal phonemic compression from a computational model of the auditory periphery. *Auditory Signal Processing in Hearing-Impaired Listeners, Int. Symposium on Audiological and Auditory Research (ISAAR)*. eds T. Dau, J. Buchholz, J. M. Harte and T. U. Christiansen. Danavox Jubilee Foundation Denmark: 73-81.

Ching, T. Y. C., H. Dillon, *et al*. (**2010**). "Evaluation of the NAL-NL1 and the DSL v.4.1 prescriptions for children: Paired-comparison intelligibility judgments and functional performance ratings" International Journal of Audiology **49** (S1).

Dillon, H. (**2001**). *Hearing Aids*. Thieme Medical Publishers, New York.

Hines, A. and N. Harte (**2010**). "Speech Intelligibility from Image Processing" Speech Communication **52** (9), 736-752.

Hines, A. and N. Harte (**2011**). "Speech Intelligibility prediction using a Neurogram Similarity Index Measure" Speech Communication, [doi:10.1016/j.specom. 2011.09.004].

Kates, J. M. and K. H. Arehart (**2010**). "The Hearing-Aid Speech Quality Index (HASQI)" J. Audio Eng. Soc **58** (5), 363--381.

Zilany, M. S. A., I. C. Bruce, *et al*. (**2009**). "A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics" J. Acoust. Soc. Am. **126** (5), 2390-2412.

Predictive measures of the intelligibility of speech processed by noise reduction algorithms

KAROLINA SMEDS[1], FLORIAN WOLTERS[1,2], ARNE LEIJON[3], ANDERS NILSSON[1,3], SARA BÅSJÖ[1], AND SOFIA HERTZMAN[1]

[1]*Widex A/S, ORCA Europe, Stockholm, Sweden; karolina.smeds@orca-eu.info*
[2]*Univeristy of Applied Sciences, Oldenburg, Germany*
[3]*KTH, Stockholm, Sweden*

A number of predictive measures were evaluated in terms of their ability to predict the effect on speech intelligibility of different types of noise reduction (NR). Twenty listeners with hearing impairment and ten listeners with normal hearing participated in a blinded laboratory study. An adaptive speech test was used. The speech test produce results in terms of physical signal-to-noise ratios that correspond to equal speech recognition performance with and without the NR algorithms, which facilitates a direct statistical test of how well the predictive measures agree with the experimental results. Three NR algorithms and a reference condition were compared. The experimental results were used to evaluate a number of predictive measures, including a standard Speech Intelligibility Index (SII) method, two time-variable SII methods, and one coherence-based SII method. Further, one measure based on the correlation between band envelope magnitudes of clean and processed noisy speech was evaluated. The measures that make short-time analyses of both speech and noise did best in the comparison.

BACKGROUND

Noise reduction (NR) is commonly used in modern hearing aids (HAs). Previous measurements (Smeds *et al*., 2009) have shown that hearing aid NR algorithms function in very different ways. It would be of great value if predictive measures could be used to indicate the effect of various NR algorithms prior to laboratory or field testing with listeners. The now reported work was part of a larger study, where both speech intelligibility and sound quality of NR processed speech were evaluated. The sound-quality work has been reported by Smeds *et al*. (2010).

GENERAL METHOD

Twenty listeners with hearing impairment (HI) and ten listeners with normal hearing (NH) participated in an adaptive speech test. The listeners with impaired hearing were provided with individualized gain using tightly fitted linear hearing aids. Three NR algorithms and a reference condition were compared using pre-processed sound files. The experimental results were used to evaluate five predictive measures of speech intelligibility.

Proceedings of ISAAR 2011: Speech perception and auditory disorders. 3rd International Symposium on Auditory and Audiological Research. August 2011, Nyborg, Denmark. Edited by T. Dau, M. L. Jepsen, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN 87-990013-3-0. EAN 9788799001330. The Danavox Jubilee Foundation, 2012.