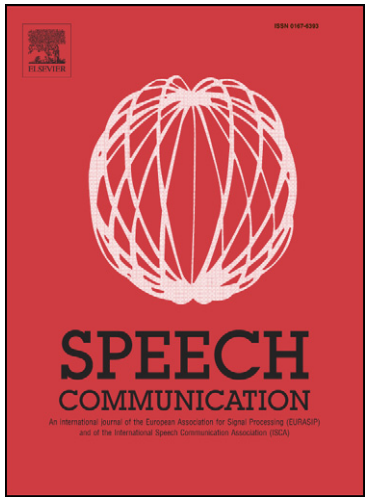

Speech Intelligibility prediction using a Neurogram Similarity Index Measure

Andrew Hines, Naomi Harte

PII: S0167-6393(11)00136-1
DOI: 10.1016/j.specom.2011.09.004
Reference: SPECOM 2026

To appear in: *Speech Communication*

Received Date: 24 November 2010
Revised Date: 12 September 2011
Accepted Date: 18 September 2011

Please cite this article as: Hines, A., Harte, N., Speech Intelligibility prediction using a Neurogram Similarity Index Measure, *Speech Communication* (2011), doi: 10.1016/j.specom.2011.09.004

Speech Intelligibility prediction using a Neurogram Similarity Index Measure

Andrew Hines, Naomi Harte

Department of Electronic & Electrical Engineering, Sigmedia Group, Trinity College Dublin, Ireland

Abstract

Discharge patterns produced by fibres from normal and impaired auditory nerves in response to speech and other complex sounds can be discriminated subjectively through visual inspection. Similarly, responses from auditory nerves where speech is presented at diminishing sound levels progressively deteriorate from those at normal listening levels. This paper presents a Neurogram Similarity Index Measure (NSIM) that automates this inspection process, and translates the response pattern differences into a bounded discrimination metric.

Performance Intensity functions can be used to provide additional information over measurement of speech reception threshold and maximum phoneme recognition by plotting a test subject's recognition probability over a range of sound intensities. A computational model of the auditory periphery was used to replace the human subject and develop a methodology that simulates a real listener test. The newly developed NSIM is used to evaluate the model outputs in response to Consonant-Vowel-Consonant (CVC) word lists and produce phoneme discrimination scores. The simulated results are rigorously compared to those from normal hearing subjects in both quiet and noise conditions. The accuracy of the tests and the minimum number of word lists necessary for repeatable results is established and the results are compared to predictions using the speech intelligibility index (SII). The experiments demonstrate that the proposed Simulated Performance Intensity Function (SPIF) produces results with confidence intervals within the human error bounds expected with real listener tests. This work represents

Email address: hinesa@tcd.ie (Andrew Hines)

an important step in validating the use of auditory nerve models to predict speech intelligibility.

Key words:
auditory periphery model, simulated performance intensity function, NSIM, SSIM, Speech Intelligibility

1. Introduction

It has been shown that auditory nerve (AN) discharge patterns in response to complex vowel sounds can be discriminated using a subjective visual inspection, and how impaired representations from those with sensorineural hearing loss (SNHL) differ from the normal (Sachs et al., 2002). If this subjective visual inspection can be replaced by a quantitative automated inspection, rapid prototyping of hearing aid algorithms could become possible. This would, however, require another question to be answered - how to directly link the quantitative measure of degradation in neural patterns to speech intelligibility. If this were achieved, it could allow simulated speech intelligibility tests, where the human listener would be substituted with a computational model of the auditory periphery and measured outputs would correlate with actual listener test results. This concept is illustrated in Fig. 1.

Recent work by the authors (Hines and Harte, 2010) developed a technique for assessing speech intelligibility by using image similarity assessment techniques to rank the information degradation in the modelled output from impaired AN models. It demonstrated effective discrimination of progressively deteriorating hearing losses through analysis of the spectro-temporal outputs and showed that hearing losses could be ranked relative to their scores using the structural similarity index (SSIM). A review and discussion of other techniques (Elhilali et al., 2003; Bondy et al., 2004) using AN models was also presented. Example outputs from the AN model for progressive hearing losses are displayed in Fig. 2 along with their SSIM scores.

In this paper the inspection process is extended to translate the SSIM measure from ranking AN discharge pattern differences into an actual phonemic recognition metric. This involved developing a test procedure that can simulate a real human listener test, with the person in the test being substituted with the AN model. The objective of the test is to determine the percentage of words or phonemes correct by using an image similarity as-

sessment technique to analyse the AN model output, and a transfer function to produce an objective measure of speech discrimination. The methodology has been developed to allow testing over a wide range of SNHLs and speech intensity levels. While the ultimate goal of this work is to assess hearing loss and hearing aid assessment, this paper focuses on validating the methodology with normal hearing at low signal levels in a quiet environment. Preliminary tests in steady state background noise are also presented, however, testing could be extended in future to include other signal distortions.

It was necessary to develop a simulated listener test methodology that would map to a human listener test and scoring system. The methodology needed to use the same dataset and produce results that were formatted in a comparable way to real listener test. In addition, the methodology needed to be validated to ensure that results were consistent and repeatable. The accuracy of the tests and the minimum number of word lists necessary for repeatable results were also measured. To demonstrate that the AN model was an essential element in the system, an end-to-end test was also carried out with an adaptation of the methodology excluding the AN model.

Section 2 describes the AN model and how image similarity metrics can assess the model's output neurograms. It presents the structural similarity index and details how it can be adapted to neurogram degradation assessment as well as introducing performance intensity listener testing. Section 3 details the general experimental setup. Section 4 describes the specific experiments undertaken to develop and assess the simulated performance intensity tests and presents results. Section 5 reviews these results in the context of other work and uses the Simulated Performance Intensity Function (SPIF) method presented to compare the results in quiet and noise with the Speech Intelligiblity Index (SII) standard (ANSI, 1997).

2. Background

2.1. Auditory Nerve Models

Previous work (Hines and Harte, 2010) used the AN model of Zilany and Bruce (2006) that is derived from empirical data matched to cat auditory nerves. The model has since been extended and improved. In this study, their new model (Zilany et al., 2009) was used which includes power-law dynamics as well as exponential adaptation in the synapse model. The AN model is the latest version based on ongoing research and has been extended and enhanced over the last decade (Zhang et al., 2001; Bruce et al., 2003).

It has been developed through extensive testing and matching with physiological data to a wide variety of inputs. This study focused on using the AN model to develop a simulated listener intelligibility metric and the model was treated as a black box. Changes to the AN model to incorporate human cochlear tuning (e.g. Ibrahim and Bruce (2010)) were not implemented as currently a difference in tuning between the human cochlea and that of common laboratory animals has not been definitively shown (Young, 2008).

2.2. Neurograms

A neurogram is analogous to a spectrogram. It presents a pictorial representation of a signal in the time-frequency domains using colour to indicate activity intensity.

As in prior work (Hines and Harte, 2010), neurograms with 30 characteristic frequencies (CFs) were used, spaced logarithmically between 250 and 8000 Hz. The neural response at each CF was created from the post stimulus time histogram (PSTH) of 50 simulated AN fibres varying spontaneous rates. Two types of neurograms were used, an average discharge or envelope (ENV) and a fine timing or temporal fine structure (TFS). The TFS and ENV responses were smoothed by convolving them with a 50% overlap, 32 and 128 sample Hamming window respectively.

When referring to neurograms, the terms ENV and TFS are distinct from, although related to, the corresponding signal terms. Although the ENV and TFS neurograms allow auditory nerve firing rates to be investigated at different time resolutions they are not the strict isolating metrics of acoustic ENV and TFS (Rosen, 1992; Smith et al., 2002). As the ENV neurogram is a smoothed average discharge rate, only slow temporal modulations will be available, which allows the envelope information that is embedded to be assessed. TFS neurograms preserve spike timing information and the synchronisation to particular stimulus phase, or phase-locking phenomenon (Young, 2008), allow TFS cues to be examined.

An example signal, the word "ship" presented to the AN model, is shown in Fig. 3. The top row shows the time domain signal. Below it, the spectrogram presents the sound pressure level of a signal for frequency bands in the y-axis against time on the x-axis. Three ENV neurograms, created from AN model outputs for signals presented at progressively lower presentation levels (65, 30 and 15 dB SPL), are then shown. The colour represents the neural firing activity for a given CF band in the y-axis over time in the x-axis. The neural activity is binned into time bins (TFS=10μs; ENV=100μs)

to create PSTH information. The fine timing information of neural spikes is retained and presented in TFS neurograms (not illustrated) while the ENV neurogram smoothes the information and presents an average discharge rate using a larger bin and a wider Hamming window.

2.3. Structural Similarity Index (SSIM)

The neurograms created from the AN model output can be treated as images. The output created by presenting words at a conversational level to a model of a normal hearing listener can be used as a reference. Segregating the neurogram into images for each phoneme and comparing the reference to degraded versions allows an image similarity metric to assess the level of degradation.

Prior work (Hines and Harte, 2010) demonstrated that the structural similarity index (SSIM) (Wang et al., 2004) could be used to discriminate between a reference and degraded neurogram of a given phoneme. SSIM was developed to evaluate JPEG compression techniques by assessing image similarity relative to a reference uncompressed image. It exhibited better discrimination than basic point to point measures, i.e. relative mean squared error (RMSE) and relative mean absolute error (RMAE), for image similarity evaluations carried out between neurograms of the reference and degraded versions of phonemes. Unlike these measures, SSIM "looks" at images over a patch or windowed area rather than just using a simple point-to-point pixel comparison. The optimal window size was found to be 3x3 pixels for both TFS and ENV neurograms (covering three CF bands on the y-axis and a time duration on the x-axis of approximately 0.5ms and 20ms respectively).

SSIM uses the overall range of pixel intensity for the image along with a measure of three factors on each individual pixel comparison. The factors: luminance, contrast and structure, give a weighted adjustment to the similarity measure that look at the intensity (luminance), variance (contrast) and cross-correlation (structure) between a given pixel and those that surround it versus the reference image.

The SSIM between two neurograms, the reference, r, and the degraded, d, is constructed as a weighted function of luminance (l), contrast (c) and structure (s) as in eqn. (2). Luminance looks at a comparison of the mean (μ) values across the two neurograms. The contrast is a variance measure, and structure equivalent to the correlation coefficient between the neurograms (r) and (d). Let k be the CF band index, and m the index for the sub-sampled smoothed auditory nerve output. As per Wang et al. (2004), for each

phoneme neurogram, the local statistics (μ_r,σ_r,σ_{xy}) are computed within a 3x3 square window, which moves pixel by pixel ($k = 1..K$, $m = 1..M$) over the entire neurogram. At each point, the local statistics and SSIM are calculated within the local window, producing an SSIM map. The mean of the SSIM map is used as the overall similarity metric. Each component also 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. The weighting coefficients, α, β and γ, can be used to adjust the relative importance of the components, expressing SSIM as in eqn. (2). See Hines and Harte (2010) for further information on neurogram ranking with SSIM and Wang et al. (2004) for a full description of the metric.

$$S(r,d) = l(r,d)^{\alpha} \cdot c(r,d)^{\beta} \cdot s(r,d)^{\gamma} \tag{1}$$

$$S(r,d) = \left(\frac{2\mu_r\mu_d + C_1}{\mu_r^2 + \mu_d^2 + C_1}\right)^{\alpha} \cdot \left(\frac{2\sigma_r\sigma_d + C_2}{\sigma_r^2 + \sigma_d^2 + C_2}\right)^{\beta} \cdot \left(\frac{\sigma_{rd} + C_3}{\sigma_r\sigma_d + C_3}\right)^{\gamma} \tag{2}$$

The SSIM is calculated for each point on a neurogram. The overall SSIM similarity index for two neurograms is computed as the mean of the SSIM index values computed for all patches of the two neurograms.

A cursory investigation of the component weightings in SSIM was undertaken in prior work, where the weightings proposed by Kandadai et al. (2008) for auditory signal analysis were compared to the un-weighted results. As phoneme discrimination was significantly poorer using the suggested weightings when compared to the un-weighted SSIM results, undertaking a full investigation was deemed necessary. This work seeks to establish the component weights for SSIM that give the best correlation with human listener test results when being used to compare phoneme neurograms.

2.4. The Performance Intensity Function

A useful way of presenting listener test results is the performance versus intensity (PI) function. It describes recognition probability as a function of average speech amplitude, showing the cumulative distribution of useful speech information across the amplitude domain as speech rises from inaudibility to full audibility (Boothroyd, 2008). Boothroyd uses phoneme scoring of responses to Consonant-Vowel-Consonant (CVC) words to obtain PI functions and argues that the potentially useful information provided by the

PI function over a basic *speech reception threshold test* and *maximum word recognition test* with CVC word lists is worth the extra time and effort.

According to Mackersie et al. (2001) PI evaluation can provide a more comprehensive estimation of speech recognition. Before computerised versions of the test, such as the Computer-Aided Speech Perception Assessment (CASPA; Boothroyd (2006)), automated the procedure, calculating a PI function with phoneme scoring was a significantly more time consuming test.

There are a number of advantages to phonemic scoring tests over similar word scoring tests (Markides, 1978; Gelfand, 1998). From a statistical perspective, the simple increase in the number of test items improves test-retest reliability by decreasing variability (Boothroyd, 1968a; Gelfand, 1998). Phoneme scores are less dependent on a listeners vocabulary as they can be instructed to repeat the sounds that they hear, not the word, even if they believe it to be a nonsense word. Results are less influenced by the listener's vocabulary knowledge than whole-word scoring and provide a well-grounded measure of auditory resolution (Boothroyd, 1968b; Olsen et al., 1997). This factor is important in testing with children, who would have a more limited vocabulary than adults (McCreery et al., 2010).

The PI test 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). It has also been used in testing for rollover effect at high intensities (Jerger and Jerger, 1971).

The test corpus used here contains 20 word lists of 10 phonemically balanced CVC words. It was developed by Boothroyd for use with the CASPA software for PI measurement. Words are not repeated within lists and lists are designed to be isophonemic, i.e. to contain one instance of each of the same 30 phonemes. There are 10 vowels and 20 consonants in each list and they are chosen to represent those that occur most frequently in English CVC words. The lists are balanced only for phonemic content - not for word frequency or lexical neighbourhood size. Word lists comprising 10 words are presented over a range of intensity levels. The tester records the subject's responses with the CASPA software. It scores results in terms of words, phonemes, consonants, and vowels correctly identified and generates separate PI functions for each analysis. A sample word list is illustrated in Fig. 1.

3. Simulation Method

Experiments using the AN model were designed to allow comparison of simulated listener test results with real listener data. The real listener tests, presented by Boothroyd (2008), were carried out dichotically via insert headphones on a group of normal hearing listeners in quiet at speech presentation levels between 5 and 40 dB SPL. The tests are reproduced here, substituting the human listener with the AN model and measuring neurogram degradation to predict phoneme discrimination.

First, different image similarity metrics were investigated to quantify the measurements' fitting accuracy to human listener data. Then PI functions were simulated for normal hearing listeners over a wide range of presentation levels in both quiet and noise conditions, using the newly refined metric and methodology.

3.1. Experimental Setup

Timing label files marking the phoneme boundaries were created for the 200 words in Boothroyd's dataset. For each word, the time was split into 5 portions: a leading silence, a trailing silence, and 3 distinct phonemes. All calculations were based on lists containing 10 words (30 phonemes). For actual listener tests Boothroyd (2008) made an assumption of 25 independent phonemes per list, due to the overlap of phoneme sounds within words.

The most comfortable level (MCL) for speech listening with normal hearing is generally around 40-50 dB above the initial speech reception threshold (Hochberg, 1975; Sammeth et al., 1989) and the mean sound field pressures of conversational speech is 65-70 dB SPL (Moore, 2007). A level of 65 dB SPL was taken as the standard level to generate reference neurograms for similarity comparisons. The word lists were presented to the AN model at speech intensity levels of 5 through to 50 dB SPL in 5 dB increments and neurograms were created from the simulated AN output.

Phoneme Recognition Threshold (PRT) is the level in dB SPL at which the listener scores 50% of their maximum. The modal PRT value for normal hearing listeners was 15 dB SPL in Boothroyd (2008) but was previously set at 20 dB SPL (Boothroyd, 1968a). The 15 dB value was used for these experiments.

The similarity measurement between a reference neurogram at 65 dB SPL (MCL level) and a degraded neurogram at 15 dB SPL (PRT level) measured over a large sample of phonemes gives a neurogram PRT (NPRT) for a given

image similarity metric (ISM). The NPRT for each ISM was evaluated per phoneme position ($p = \{C1, V1, C2\}$) using lists of CVC words. The NPRT values were calculated as the medians, $\tilde{\mu_p}$, of the subsets S_p, containing image similarity metric F for the 100 phonemes in each subset, between the PRT and MCL levels. Using the notation from eqn. (2), the MCL level is r and the PRT level is d_{PRT}, the NPRT value is $\tilde{\mu}$ for K phonemes of the set,

$$S_p = \{F(r(i), d_{PRT}(i)) | 1 \leq i \leq K\} \tag{3}$$

The threshold was calculated per phoneme position (C1,V1,C2) rather than across all phonemes together. While Boothroyd does not differentiate between recognition by phoneme type in calculating the PI function, the image similarity metrics are susceptible to differences in some circumstances, e.g. noise. This is discussed further in section 5.

The same procedure that was used for evaluation of the NPRT was repeated at each speech intensity level. The results for each image similarity metric were recorded and a phoneme discrimination score was calculated by counting the number of phonemes scoring above the NPRT value. Fig. 8 illustrates SSIM scores per phoneme position with the NPRT marked. The comparison measurement was carried out in the same manner for both ENV and TFS neurograms and allowed a PI function to be plotted from the results for both neurogram types.

4. Experiments and Results

4.1. Image Similarity Metrics

The first experiment compared the ability of three image similarity metrics (SSIM, RMAE and RMSE) to predict human listener test scores directly from neurograms. Ten lists (100 words) were presented at each presentation level. Phoneme discrimination scores were calculated for phonemes scoring above the NPRT for SSIM (as it is an ascending similarity metric) and below the NPRT value for RMSE and RMAE (as they are ascending error metrics).

The relative contribution from each of the SSIM components: luminance, contrast and structure was also investigated for both neurogram types. From eqn. (2), α, β and γ are the exponents associated with each component of the SSIM metric. Each combination of α, β and γ for .05 increments between .05 and 1 was tested.

Following the same methodology to calculate phoneme neurogram similarity, PI functions were created for each weighting combination. The curve

fitting error was calculated as the sum of the least square difference between the real listener PI function and the simulated PI function at each of the ten presentation levels. The minimum error score gave the best weighting combination to curve fit modelled results to the human listener tests.

The PI curves for each image similarity metric are presented in Fig. 4. There are two for SSIM, one with un-weighted components, Fig. 4B, and one using the optimal SSIM component weightings, Fig. 4D.

The 10 lists are isophonemic and should thus be comparable in terms of the PI scores yielded. The PI function for each list was calculated and the mean PI discrimination scores are presented. The error bars show standard error 95% confidence interval measurement between lists at each speech intensity level.

The highlighted area in the graph highlights the speech intensity range from 20-40 dB SPL which was used to evaluate the correlation between the PI function for each ISM and the actual listener data PI curve. Boothroyd (1968a) recommends that clinicians carry out tests at a minimum of three levels along the sloping part of the curve. The scores above 40 dB SPL were 100% for all ISMs tested and the threshold 15 dB level was used to anchor the 50% level. The intermediate 5 data points were used as the range to assess deviation from the actual listener test PI function.

The root mean square deviation (RMSD) between modelled PI results and listener data results over the 20-40 dB SPL range was calculated for both ENV and TFS neurogram types. This quantified how closely the modelled results (PI_{neuro}) followed the real listener PI function ($PI_{listener}$). The expected RMSD value was calculated between 20 and 40 dB SPL as:

$$RMSD = \sqrt{E(PI_{listener} - PI_{neuro})^2} \tag{4}$$

The superior PI function fit for SSIM can be seen in Fig. 4 where RMSE and RMAE have significantly poorer RMSD scores for both ENV and TFS.

The SSIM PI function, shown in Fig. 4B, tracks the listener PI curve significantly better than either the RMSE or RMAE. The root mean square deviation in the highlighted box shows the deviation from the actual listener test curve for the AN modelled results when calculated for ENV and TFS neurograms.

The optimised SSIM, where exponents α, β and γ were varied to find the factors contributing most to neurogram similarity measurement, are presented in Fig. 5. The curve fitting errors demonstrate that the measure is

fairly robust to changes in weightings with α and γ being the primary measures over β. Fixing α and β at their optimum values, the graph displays the error for weightings of γ over full range in 0.05 increment tests. Results for α and β are similarly shown. The results for both TFS and ENV neurograms were optimal with α and γ closer to full weighing and β as a minimal contribution. The optimal weightings for the SSIM components are in Table 1. It should be noted from Fig. 5 that while the error trends downwards as the α weighting increases, both β and γ are relatively flat with local minima, such that the difference between the TFS results for a γ value of .65 or 1 is negligible. The PI function for optimised SSIM is shown in Fig.4D. It can be seen that the results display an improvement in correlation to the listener test data over un-weighted SSIM for both ENV and TFS neurogram types.

	α	β	γ
TFS	1	0.05	.65
ENV	0.95	0.05	0.9

Table 1: SSIM component weighting test. The optimal weightings for α, β and γ exponentials when using SSIM to assess listener tests results with TFS and ENV neurograms.

4.2. Neurogram Similarity Index Measure (NSIM)

The optimally weighted SSIM results are better than those for the unweighted metric although the magnitude of the improvement is not as profound as the difference between SSIM and the other similarity metrics tested. Looking at the results in Fig. 5, there is a strong argument for dropping the contrast component β, which contributed minimal positive correlation, and setting α and γ at 1. Testing this proposal with 10 lists gave results comparable in accuracy and reliability to those measured using the optimum SSIM weightings. This would simplify the metric considerably and also create a uniform calculation for both ENV and TFS neurograms. It is proposed that this simplified adaptation of SSIM will be used and referred to as the Neurogram Similarity Index Measure (NSIM):

$$NSIM(r,d) = \frac{2\mu_r\mu_d + C_1}{\mu_r^2 + \mu_d^2 + C_1} \cdot \frac{\sigma_{rd} + C_2}{\sigma_r\sigma_d + C_2} \tag{5}$$

4.3. Accuracy and Repeatability

Tests were carried out using multiple different lists at each presentation level as well as with repetitions of a single list to assess the repeatability and

accuracy of the simulations.

The accuracy was assessed by measuring the root mean squared deviation between the real listener and simulated PI functions. The repeatability was measured by comparing the standard error variability at each presentation level.

A single word list (list #1) was presented to the model 10 times. PI functions were calculated and the confidence intervals were estimated using 3, 5, 8 and 10 iterations of a single list (Fig. 6). 95% confidence intervals (1.96 times the standard error) between iterations above and below the mean value are shown at each presentation level tested.

For iterations of the same list, the ENV and TFS PI functions do not follow as closely to the real listener PI function as for the same number of varied lists. A comparison of the RMSD values quoted in Fig. 6 show that the deviation remains consistent as the number of simulation iterations increased. More iterations did however decrease the variability, as the error bars illustrate.

Multiple word lists were presented to the model and PI functions were calculated and the confidence intervals were estimated using 3, 5, 8 and 10 lists at each presentation level (Fig. 7). The RMSD values show the deviation decreases for tests using 3 to 5 lists but is relatively consistent for 5,8 and 10 lists.

As with multiple presentations of the same list, the variability decreases as the number of lists increases, illustrated by the error bars decreasing in size in Figs. 7A-D.

These results show that repeating lists do not improve the accuracy but does improve the confidence interval in the simulated PI functions. Using 5 different lists improves the accuracy and the confidence interval over using 3 lists in the simulated PI functions, but more than 5 has little impact on either accuracy or reliability. This result coincides with the recommendations to present a minimum of 3 lists in the original PI listener test proposal (Boothroyd, 1968a).

4.4. Method and Model Validation

To rule out the potential of false-positive results, and to verify that the AN model was the principle factor influencing the PI function shape, PI functions were created using spectrograms of the input signal with comparable resolutions to neurograms. The number of frequency bands matched the 30 CF bands in the neurograms and the sampling and smoothing windows were

comparable to those used to created ENV and TFS neurograms from the AN model PSTH outputs. The spectrograms were created directly from the input dataset signals (i.e. the words at each intensity level). Using the same methodology that was used for neurogram assessment, SSIM was used with the spectrograms to calculate PI functions. The NPRT level was set at 15 dB SPL, although without the AN model present, there is no inherent reception threshold boundary at this level in the signal spectrogram.

Fig. 9 confirms that the AN model is the critical factor influencing the PI function shape. The RMSD values are an order of magnitude worse than those measured using neurograms from the AN model. This is primarily attributable to the 100% scores for 30 dB SPL and above. The reason for this is apparent when the SSIM results are examined. Although the range in the SSIM scores is much wider for the spectrograms than it is for the neurograms, the NPRT line is much closer to zero. The wider range and spread in SSIM values are indicative of the procedure purely measuring the increase in signal intensity from the spectrograms.

This validates the assumption that the accuracy of the simulated PI is primarily a function of the AN model and not just a function of the data or test parameters used in the methodology.

4.5. Simulated Performance Intensity Functions (SPIFs)

Further experiments were carried out to assess the prediction of normal hearing across a wider range of presentation levels in quiet and a range of signal to noise ratios in steady state noise. Based on the prior findings, 5 word lists were used and the neurograms were compared using the NSIM.

A test in quiet was carried out over 5 dB intervals from 5 to 100 dB SPL with the reference neurogram level set at 65 dB SPL. The results are presented in Fig. 10. The ENV results reached 100% phoneme recognition at 45 dB SPL and remain there through to 100 dB SPL. The TFS results begin to fall from 90 dB SPL.

A second test was carried out, with a steady state noise fixed at 55 dB SPL and the words were presented at 5 dB increments between -15 and +15 dB SNR. A reference +20 dB SNR was used for comparisons and a -11 dB SNR was used as the phoneme recognition threshold in line with results presented in Boothroyd (2008). The results are presented in Fig. 11.

In noise, NSIM provided a marginally superior fit to RMAE or RMSE. Further tests in a range of noise and reverberations may allow further refinement and assessment of the SPIF methodology. This basic test in noise

demonstrates the model is not limited to speech intelligibility assessment in quiet.

4.6. Comparison to SII

A comparison was carried out between the results presented for NSIM and the speech intelligibility index. The SII was calculated by the one-third octave procedure in ANSI (1997) using the long term spectrum for five CASPA word lists. SII was calculated in quiet over 5 dB steps between 5 and 100 dB SPL. SII is a measure, bounded between zero and one, that computes the amount of audible speech cues that are available to a listener. An SII score of 0.5 does not translate directly to a speech discrimination score of 50%. The frequency importance and transfer functions for NU6 words were used to convert SII to word recognition (Studebaker et al., 1993) followed by a word-to-phoneme recognition transfer function(Boothroyd, 2008). Fig. 10 shows the SII and the SII phoneme recognition predictions in quiet and Fig. 11 shows SII in noise. The SII input was adjusted to match the PRT of the listener test results.

In quiet, SII follows the listener PI function well but overestimates results in the 20-40 dB SPL range (RMSD=0.059). The linear correlation between modelled and listener phoneme discrimination is presented along with their RMSD values in Fig. 10.

SII and NSIM both underestimated the phoneme recognition in the preliminary tests in noise, with gradients more linear than the real listener PI function between 50% and 90% phoneme discrimination levels. Results for both ENV and TFS neurograms showed similar levels of accuracy but both underestimated phoneme discrimination more than SII.

5. Discussion

Using the Neurogram Similarity Index Measure to compare the neurogram outputs from an AN model has been shown to produce a PI function with statistically significant correlation accuracy to real listener data. This is an important step that not only validates the AN model as a tool for assessing speech intelligibility, but provides a mechanism for quantitatively assessing phoneme and word recognition at progressive speech intensity levels. It must be acknowledged that so far this has only been demonstrated for simulations of normal hearing in quiet and steady state noise. The methodology, having

been developed and validated, now has the potential to be extended to simulations in other environments such as speech shaped noise or reverberation and also for simulation of SNHL in aided and unaided scenarios.

Measuring the similarity of spectrograms instead of neurograms demonstrated that the AN model was essential to the overall accuracy of the simulated PI function. One limitation of the AN model is its computational requirements preclude real time simulation of even limited word lists. While this paper focused on the development of a methodology for using image similarity metrics in neurogram assessment, one could speculate that substituting an alternative, simpler AN model to that of Zilany et al. (2009), may yield comparable results. In its current form, the proposed methodology could ultimately prove effective as a measure for use in the assessment of hearing aid algorithms, but would be unsuitable for any real-time applications.

NSIM provides a simpler metric to SSIM while still giving comparable results that are superior to basic point-to-point similarity metrics in quiet conditions. The simulated PI functions demonstrate that modelled results for both ENV and TFS neurograms can be correlated with psychometric tests. One apparent weakness, is the poor correlation below the PRT level, where RMAE and RMSE performance was superior (see Fig. 4A). As testing at these levels has limited practical applications in hearing assessment or enhancement, it is not perceived as a major shortcoming.

The methodology presented is based on transforming an image similarity metric to an estimate of phoneme discrimination, by measuring the similarity between a reference and degraded neurogram. The premise is that, over a long run of phoneme neurogram comparisons, a threshold value (NPRT) for similarity can be matched to a psychoacoustic phoneme recognition level.

The NPRT is set based on the median levels for the leading consonant, vowel and trailing consonant (C1,V1,C2). For early experiments, the NPRT was set as the median across all phonemes regardless of position. This worked well in quiet conditions and the difference in value between the NPRT calculated across all phonemes versus the NPRT, calculated per phoneme position, was negligible (for ENV and TFS $\tilde{\mu}-\tilde{\mu}_p < 0.016$). In noise this was found not to be the case where the NPRT range was up to 0.056. It is illustrated in Figs. 10 & 11 where the NPRT lines are plotted on the NSIM boxplots. While the results in quiet show similar maximum, minimum and NPRT scores for C1,V1 and C2, the pattern is not repeated in noise. The trailing consonant, C2, has a lower maximum, minimum and NPRT than either C1 or V1. The likely reason for this is due to the higher occurrence of stop phonemes at

the end, rather than at the start of the test words. When analysed as an image, a time-frequency neurogram plot of a stop phoneme is predominantly an empty image followed by an vertical line of intensity across the frequency range and then trailing off (see Fig. 3 from approximately 0.55 seconds). Comparison of the stop in quiet will rank the silence portion of the image equally and the similarity ranking is dominated by differences in the intensity of the plosive burst. When comparing stop phonemes in noise, the absence of comparative features in the pre-plosive burst section of the neurogram results in a dominance of noise over spectro-temporal phoneme features in the similarity analysis and a consequent shift down in similarity scores.

Using a image similarity metric has a dependence on the spectro-temporal features within a phoneme's neurogram. While this causes problems when assessing the similarity of stop consonants in noise, an analogous problem is faced by real listeners decoding speech, where noise masks the expected silence and reduces the intensity difference at the start of the plosive burst. The full reference, time-aligned neurogram comparison, means that each phoneme is assessed based on its degradation in isolation. Practically, the measurement is devoid of any advantage of context, but it also means that slight misalignments will not critically impact the results as a vowel phoneme that is shorter, or longer, will still yield a comparable similarity score due to the periodic nature of the neurogram.

5.1. Comparison with other models

Approaches similar to those presented in this paper have been adopted by a number of authors in their work on the prediction of speech intelligibility using AN models.

Huber and Kollmeier (2006) used the Dau et al. (1996) auditory model to develop PEMO-Q, an audio quality assessment. While their goal was quality assessment, a strong correlation between quality and speech intelligibility has been shown (Preminger and Tasell, 1995). The PEMO-Q approach is based on a full reference comparison between "internal representations" of a high quality reference signal and distorted signals. The metric uses a correlation co-efficient and requires time-aligned signals and uniform band importance weightings that are applied across frequency bands. The envelope modulation from each band forms a weighted cross correlation of modulations to obtain the quality index.

Spectro-temporal modulation transfer functions (MTF) have been used to develop intelligibility indices (STI/STMI). The spectro-temporal modu-

lation index (STMI) was developed by Elhilali et al. (2003) to quantify the degradation in the encoding of spectral and temporal modulations due to noise, regardless of its exact nature.

Zilany and Bruce (2007) combined the use of STMI with their AN model (Zilany and Bruce, 2006) to measure intelligibility by presenting sentences and words in quiet and noise. They showed correlation between STMI scores and word recognition, but only tested with a limited number of presentation levels. They demonstrated that STMI would predict the same general trends as listener tests in quiet, noise and with SNHL. Quantitative prediction or mapping to word recognition via a transfer function was not demonstrated.

A key difference in this work is the quantitative link between neurogram similarity and phoneme recognition performance across a range of intensity levels. The measurement and scoring on a per phoneme basis aims to allow direct comparison between clinical testing techniques and simulated modelling. Phoneme based modelling was undertaken by Jurgens and Brand (2009) who correlated simulated recognition rates with human recognition rates and also looked at confusion matrices for vowels and consonants. In their Perception Model (named PeMo), the comparisons are made using a distance measurement between unseen, noise corrupted sounds and reference sounds. A dynamic time warp speech recogniser computes the distance for each reference and the reference with the smallest distance measurement is recognised. This means that recognition is based on guessing words from a limited vocabulary and that there is a threshold percentage correct that can be scored (random hit probability), which necessitated adjustments in the intelligibility scores. Their model showed similar prediction accuracy to SII. As in this paper, Gallun and Souza (2008) investigated the affect on intelligibility changes to the envelope at a phonemic level using a time-averaged modulation spectrum alone, without measuring phase components. They concluded that it could capture a "meaningful aspect" of information used in speech discrimination.

The results presented here show that, in both quiet and noise, neurogram similarity can be used to predict the phoneme recognition across a range of presentation levels or SNRs for a normal listener within the levels of accuracy expected from real listener tests. Jurgens et al. (2010) noted that observed speech reception thresholds in normal hearing individuals varied by about 5 dB. They note that inter-individual differences in SRT is an important and not adequately represented factor in modelled speech intelligibility, either using their model or the ANSI (1997) standard model, speech intelligibility

index (SII). Here, comparison of modelled data with real listener data necessitated calibrating the PI function to the phoneme reception threshold. In the results presented, the PRT was set according to the measured level from the psychoacoustic tests.

The NSIM results show a similar trend to SII. In quiet, the SII peaks just below 60 dB SPL and remains at a maximum through approximately 10 dB before beginning to degrade dropping to 0.84 by 100 dB SPL. The NSIM results plateau at 65 dB SPL, where they show a maximum similarity before tailing off at a faster rate than SII. It should be pointed out that the maximum NSIM value reached is not 1 as the 65 dB reference neurograms and the 65 dB test neurograms compared are from independent simulations with the AN model. A score of between .7 and .8 is the maximum similarity that occurs even for the same signal presented at the same level to the AN model. The fact that the ENV neurograms predict 100% phoneme discrimination all the way up to 100 dB SPL but that TFS predicts a sharp drop off beginning at 90 dB SPL is mainly due to the NSIM vowels scores dropping below the NPRT rather than the consonant similarity scores. It can be speculated that this behaviour in neurogram similarity may be linked to hearing phenomena, e.g. the rollover effect (Jerger and Jerger, 1971). However, this work only demonstrates that both ENV and TFS neurograms can be used to predict speech intelligibility in normal hearing listeners. Modelling sensorineural hearing loss will allow better insight in distinguishing the predictive qualities and factors influencing ENV and TFS neurograms.

The simulated performance intensity functions presented here compare favourably to predictions with SII and are a good validation of NSIM's potential. Like SII, it is not a direct measure of intelligibility. NSIM measures the difference between simulated auditory nerve firings at given intensities compared to a reference level. SII predicts the proportion of speech information that is available to the listener in given conditions. It does this by estimating the loss of information due to the masking of noise, audibility threshold or hearing impairment. A transfer function is required to predict speech intelligibility. Unlike SII, which is using importance weightings for general speech at each frequency band, NSIM is equally weighted across all neurogram CF bands measuring similarity per phoneme. As the NSIM scores are based on per phoneme neurograms, a direct comparison with results from real listener tests is possible. The methodology also opens up the possibility of examining other factors that may provide insight into cues used in speech intelligibility, such as different neurograms types (TFS or ENV) or individual

phoneme performance.

The superior performance of NSIM in quiet conditions compared to in noise is not surprising given the underlying methodology. In quiet, the AN activity decreases with presented sound intensity, and consequently there is less 'information' in the neurogram. Conversely, phonemes presented in noise contain additional erroneous information, in the form of AN activity due to the noise. The NSIM comparison between neurograms is not weighted by band or by time, so differences between noise patterns or between noise patterns and quiet patches are weighted equally with changes to actual phonemic features. This is an area where the metric could be further optimised, possibly even through the inclusion of features from SII such as frequency band importance weightings.

6. Conclusions and future work

The results presented for normal hearing listeners demonstrate that substituting an auditory nerve model for a real human listener can quantitatively predict speech intelligibility. The methodology and newly proposed Neurogram Similarity Index Measure (NSIM) have been shown to produce accurate and repeatable results. The confidence intervals for the simulated tests are within human error bounds expected with real listener tests. The simulated performance intensity functions in both quiet and in noise compared favourably with SII predictions of phoneme recognition of the CVC material tested with normal hearing listeners.

Work is ongoing to investigate simulating performance intensity functions for listeners with SNHL in unaided and aided scenarios. This opens up the potential to test and quantitatively compare the speech intelligibility improvements offered by hearing aid fitting algorithms in a simulated environment.

7. Acknowledgements

Thanks to Arthur Boothroyd for providing the word list WAV files from his CASPA software used in this research. Thanks to Edmund Lalor and an anonymous reviewer for comments on a previous revision of this paper.

References

References

ANSI, 1997. Ansi s3.5-1997 (r2007). methods for calculation of the speech intelligibility index.

Bondy, J., Bruce, I. C., Becker, S., Haykin, S., 2004. Predicting speech intelligibility from a population of neurons. In: S. Thrun, L. S., Schlkopf, B. (Eds.), NIPS 2003: Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA, p. 14091416.

Boothroyd, A., 1968a. Developments in speech audiometry. Sound 2 (1), 3 – 10.

Boothroyd, A., 1968b. Statistical theory of the speech discrimination score. The Journal of the Acoustical Society of America 43 (2), 362–367.

Boothroyd, A., 2006. Computer-aided speech perception assessment (caspa) 5.0 software manual. san diego, ca.

Boothroyd, A., 2008. The performance/intensity function: An underused resource. Ear and Hearing 29 (4), 479–491.

Bruce, I. C., Sachs, M. B., Young, E. D., 2003. An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses. J. Acoust. Soc. Am. 113, 369–388.

Dau, T., Puschel, D., Kohlrausch, A., 1996. A quantitative model of the "effective" signal processing in the auditory system. i. model structure. The Journal of the Acoustical Society of America 99 (6), 3615–3622.

Dillon, H., 2001. Hearing Aids. New York: Thieme Medical Publishers.

Elhilali, M., Chi, T., Shamma, S. A., 2003. A spectro-temporal modulation index (stmi) for assessment of speech intelligibility. Speech Communication 41 (2-3), 331–348.

Gallun, F., Souza, P., 2008. Exploring the role of the modulation spectrum in phoneme recognition. Ear and Hearing 29 (5), 800–813.

Gelfand, S. A., 1998. Optimizing the reliability of speech recognition scores. Journal of Speech, Language and Hearing Research 41 (5), 1088.

Hines, A., Harte, N., 2010. Speech intelligibility from image processing. Speech Communication 52 (9), 736–752.

Hochberg, I., 1975. Most comfortable listening for the loudness and intelligibility of speech. International Journal of Audiology 14 (1), 27–33.

Huber, R., Kollmeier, B., 2006. Pemo-q – a new method for objective audio quality assessment using a model of auditory perception. Audio, Speech, and Language Processing, IEEE Transactions on 14 (6), 1902–1911.

Ibrahim, R. A., Bruce, I. C., 2010. Effects of peripheral tuning on the auditory nerves representation of speech envelope and temporal fine structure cues. In: Lopez-Poveda, E. A., Meddis, R., Palmer, A. R. (Eds.), The Neurophysiological Bases of Auditory Perception. pp. 429–438.

Jerger, J., Jerger, S., 1971. Diagnostic significance of pb word functions. Arch Otolaryngol 93 (6), 573–580.

Jurgens, T., Brand, T., 2009. Microscopic prediction of speech recognition for listeners with normal hearing in noise using an auditory model. The Journal of the Acoustical Society of America 126 (5), 2635–2648.

Jurgens, T., Fredelake, S., Meyer, R. M., Kollmeier, B., Brand, T., 2010. Challenging the speech intelligibility index: Macroscopic vs. microscopic prediction of sentence recognition in normal and hearing-impaired listeners. In: INTERSPEECH-2010. Makuhari, Japan, pp. 2478–2481.

Kandadai, S., Hardin, J., Creusere, C., 2008. Audio quality assessment using the mean structural similarity measure. In: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. pp. 221–224.

Mackersie, C. L., Boothroyd, A., Minniear, D., 2001. Evaluation of the computer-assisted speech perception assessment test (caspa). Journal of the American Academy of Audiology 12 (8), 390.

Markides, A., 1978. Whole-word scoring versus phoneme scoring in speech audiometry. British Journal of Audiology 12 (2), 40–46.

McCreery, R., Ito, R., Spratford, M., Lewis, D., Hoover, B., Stelmachowicz, P. G., 2010. Performance-intensity functions for normal-hearing adults and children using computer-aided speech perception assessment. Ear and Hearing 31 (1), 95–101.

Moore, B. C. J., 2007. Cochlear Hearing Loss - Physiological, Psychological and Technical Issues, 2nd Edition. John Wiley and Sons.

Olsen, W. O., Tasell, D. J. V., Speaks, C. E., 1997. Phoneme and word recognition for words in isolation and in sentences. Ear and Hearing 18 (3), 175–188.

Preminger, J. E., Tasell, D. J. V., 1995. Quantifying the relation between speech quality and speech intelligibility. J Speech Hear Res 38 (3), 714–725.

Rosen, S., 1992. Temporal information in speech: Acoustic, auditory and linguistic aspects. Philosophical Transactions: Biological Sciences 336 (1278), 367–373.

Sachs, M. B., Bruce, I. C., Miller, R. L., Young, E. D., 2002. Biological basis of hearing-aid design. Annals of Biomedical Engineering 30, 157168.

Sammeth, C. A., Birman, M., Hecox, K. E., 1989. Variability of most comfortable and uncomfortable loudness levels to speech stimuli in the hearing impaired. Ear and Hearing 10 (2), 94–100.

Smith, Z., Delgutte, B., Oxenham, A., 2002. Chimaeric sounds reveal dichotomies in auditory perception. Nature 416 (6876), 87–90, 10.1038/416087a.

Studebaker, G. A., Sherbecoe, R. L., Gilmore, C., 1993. Frequency-importance and transfer functions for the auditec of st. louis recordings of the nu-6 word test. J Speech Hear Res 36 (4), 799–807.

Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E., 2004. Image quality assessment: from error visibility to structural similarity. Image Processing, IEEE Transactions on 13 (4), 600–612.

Young, E. D., 2008. Neural representation of spectral and temporal information in speech. Philosophical Transactions of the Royal Society B: Biological Sciences 363 (1493), 923–945.

Zhang, X., Heinz, M.G., Bruce, I., Carney, L., 2001. A phenomenological model for the responses of auditory-nerve fibers. i. non-linear tuning with compression and suppression. J. Acoust. Soc. Am. 109, 648–670.

Zilany, M., Bruce, I., Sept 2006. Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. J. Acoust. Soc. Am. 120 (3), 1446–1466.

Zilany, M. S. A., Bruce, I. C., 2007. Predictions of speech intelligibility with a model of the normal and impaired auditory-periphery.

Zilany, M. S. A., Bruce, I. C., Nelson, P. C., Carney, L. H., 2009. A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics. The Journal of the Acoustical Society of America 126 (5), 2390–2412.

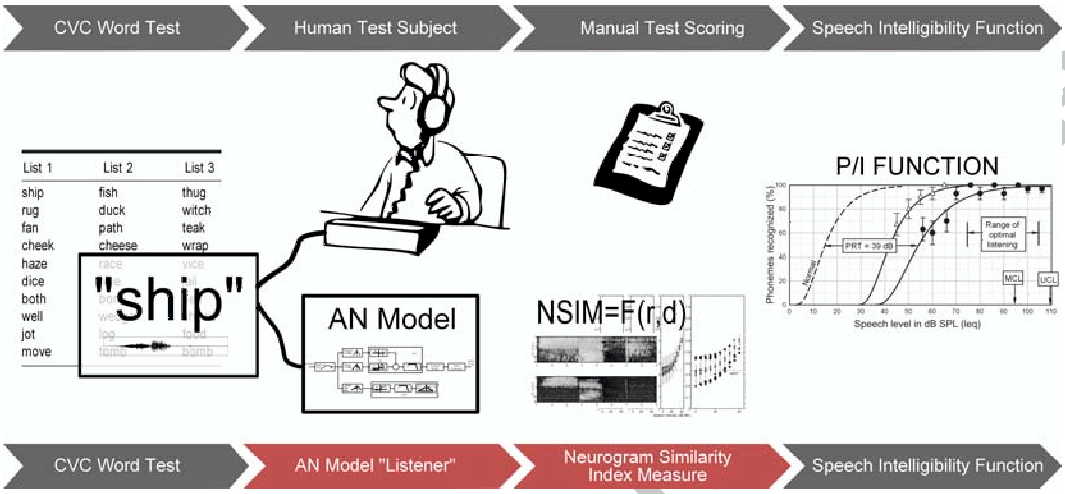

Fig. 1: The Simulated Performance Intensity Function. Above: In a standard listener test, word lists are presented to a human test subject who listens and repeats the words over a range of intensity levels. The words are manually scored per phoneme and a PI function is plotted. Below: the listener is replaced with the AN model and scoring is based on automated comparisons of simulated auditory nerve firing neurograms to quantify phoneme recognition. The results are quantifiable and are used to create a simulated PI function.

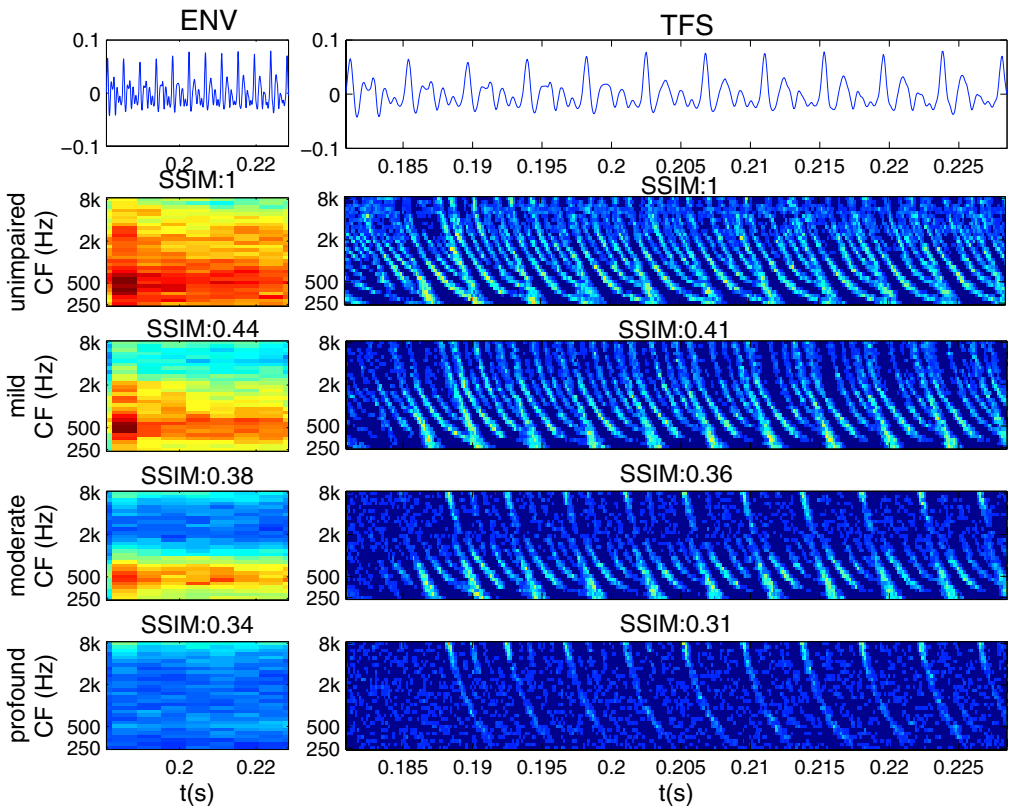

Fig. 2: Sample ENV and TFS neurograms of the vowel /aa/ presented at 65 dB SPL to progressive SNHLs simulated with an AN model. Sample hearing loss audiograms were used, described as mild (gentle sloping), moderate (steeply sloping) and profound (gently sloping) as per Dillon (2001). Neurograms were compared to the unimpaired neurogram shown and the unweighted SSIM scores are shown. A score of 1 indicates a matching image, with scores bounded from -1 to 1 with -1 being an inverse image.

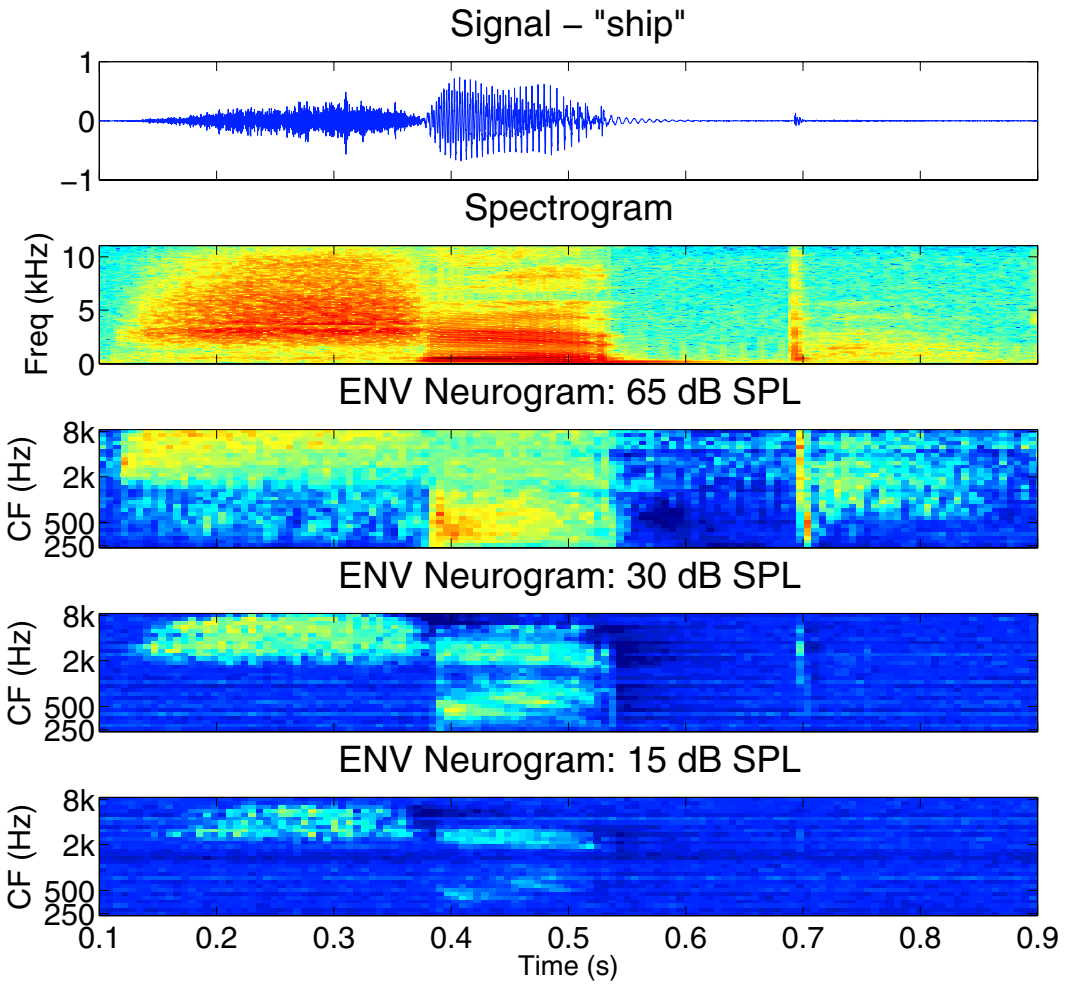

Fig. 3: A sample signal, the word "ship". The top row shows the time domain signal, with the time-frequency spectrogram below it. Three sample ENV neurograms for the same signal presented to the AN model at 65, 30 and 15 dB SPL signal intensities are presented.

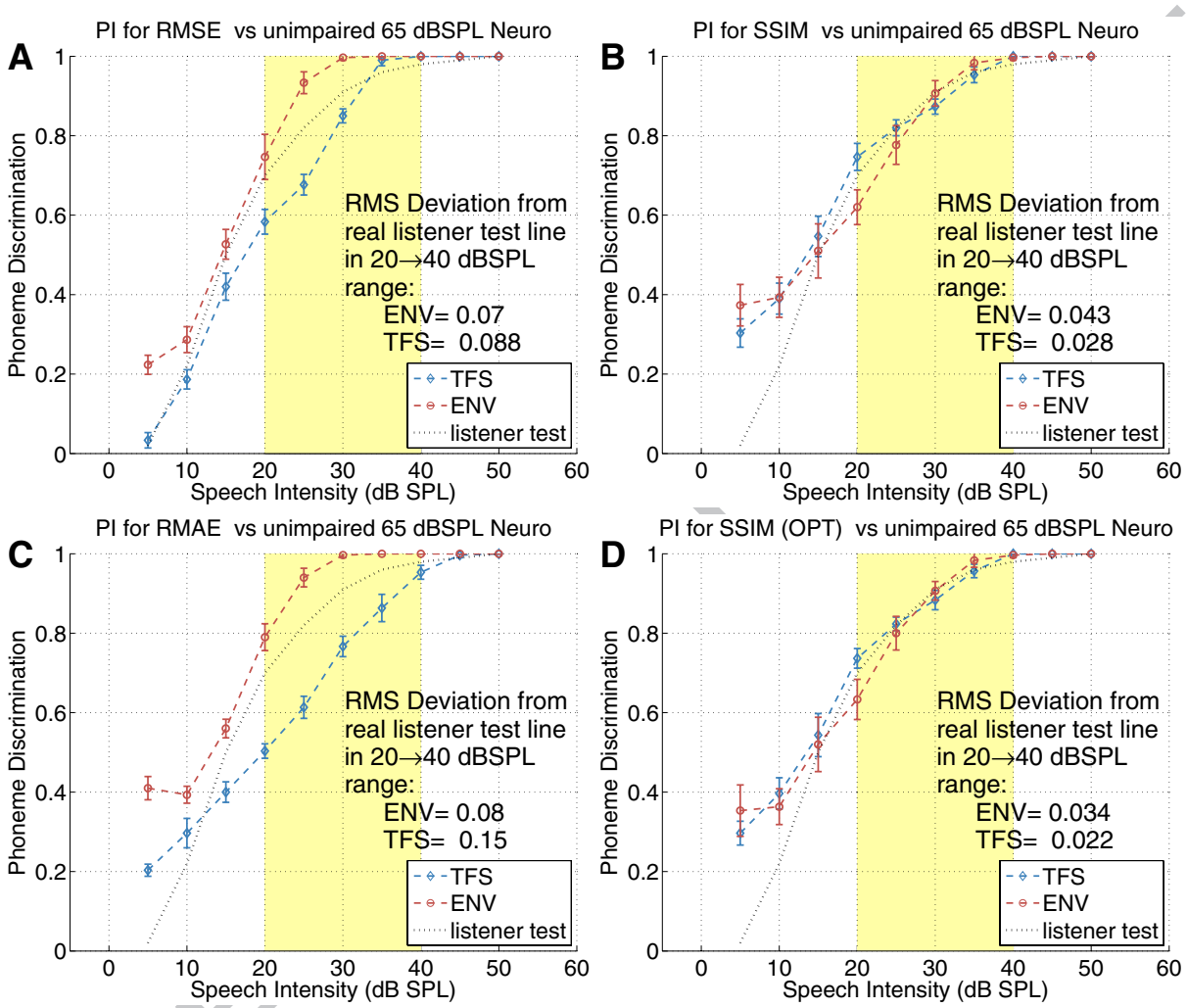

Fig. 4: PI functions simulated using AN model data from ten word lists. A: relative mean squared error (RMSE); B: SSIM with unweighted components; C: relative mean absolute error (RMAE); D: SSIM optimally weighted.

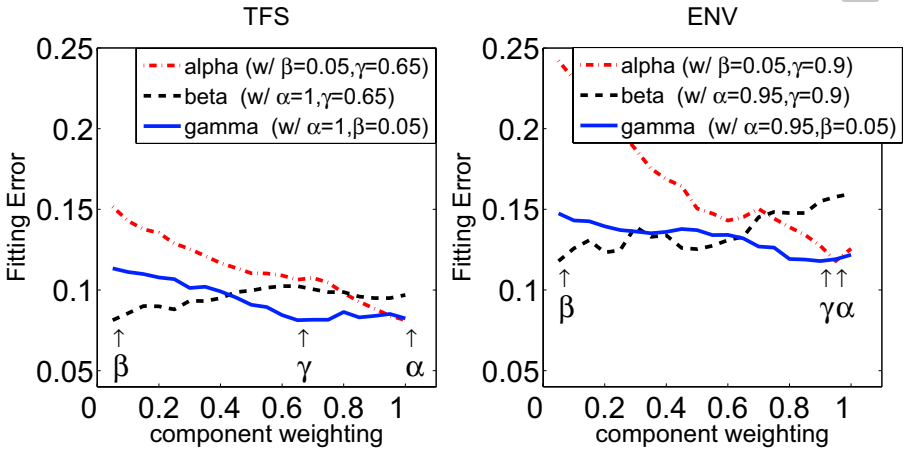

Fig. 5: Least Square Error for each SSIM component over the range of possible values .05→1 in .05 intervals, measured with the other components set at their optimum. It can be seen that SSIM is quite robust to changes in weightings. The β exponent, which controls the weighting on contrast, is of minimal value to neurogram assessment.

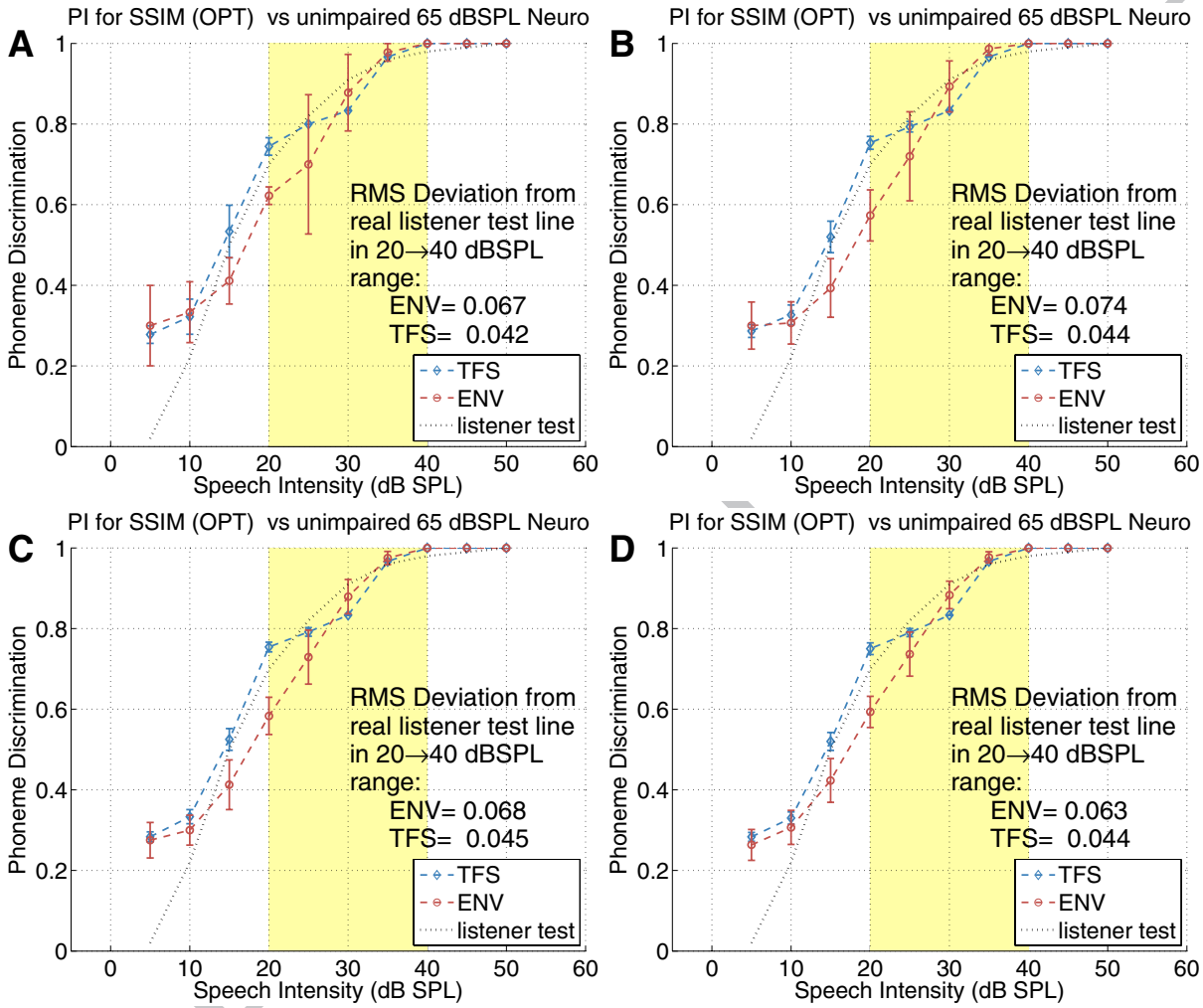

Fig. 6: AN Model variance test. PI functions calculated with SSIM (optimal weightings) using model data from 3, 5, 8 and 10 iterations of list no. 1.

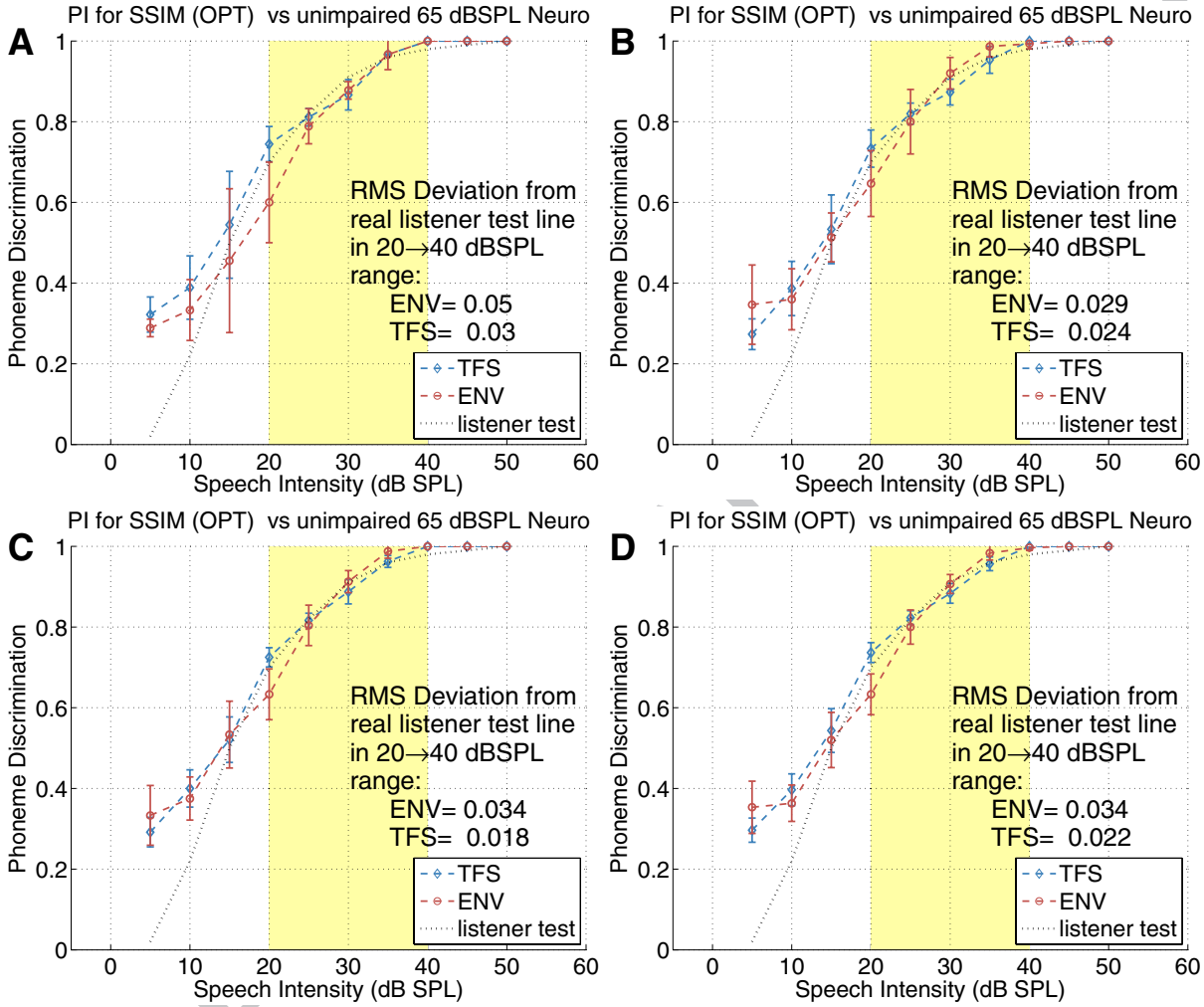

Fig. 7: Word List Test: PI functions calculated using model data from 3, 5, 8 and 10 lists.

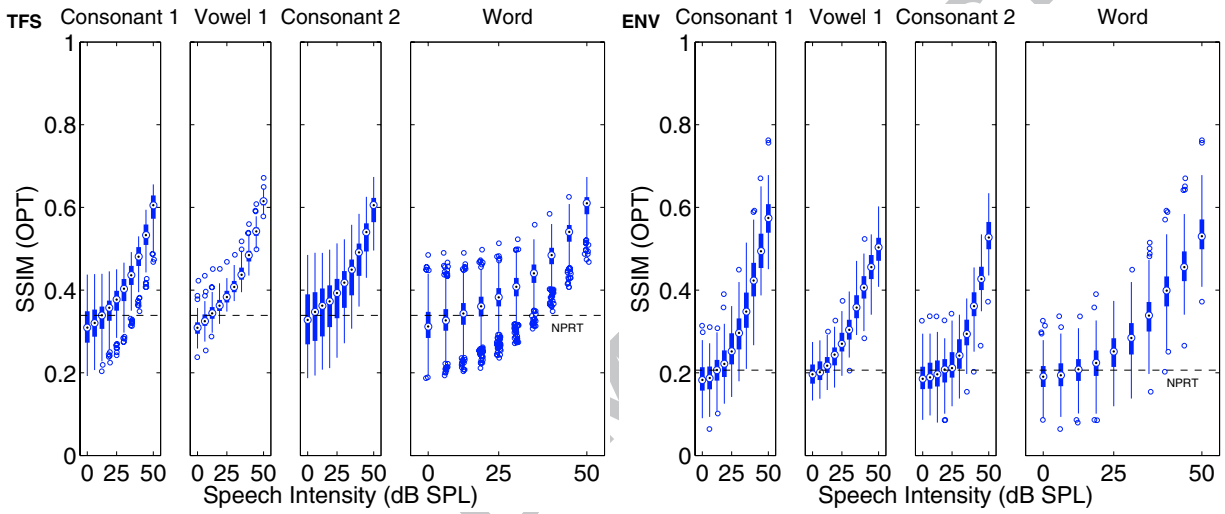

Fig. 8: SSIM scores for 10 lists. Broken down by phoneme (C1, V, C2) and a whole word plot combining the phoneme results in a single chart. The dashed line shows the neurogram phoneme recognition threshold level (NPRT).

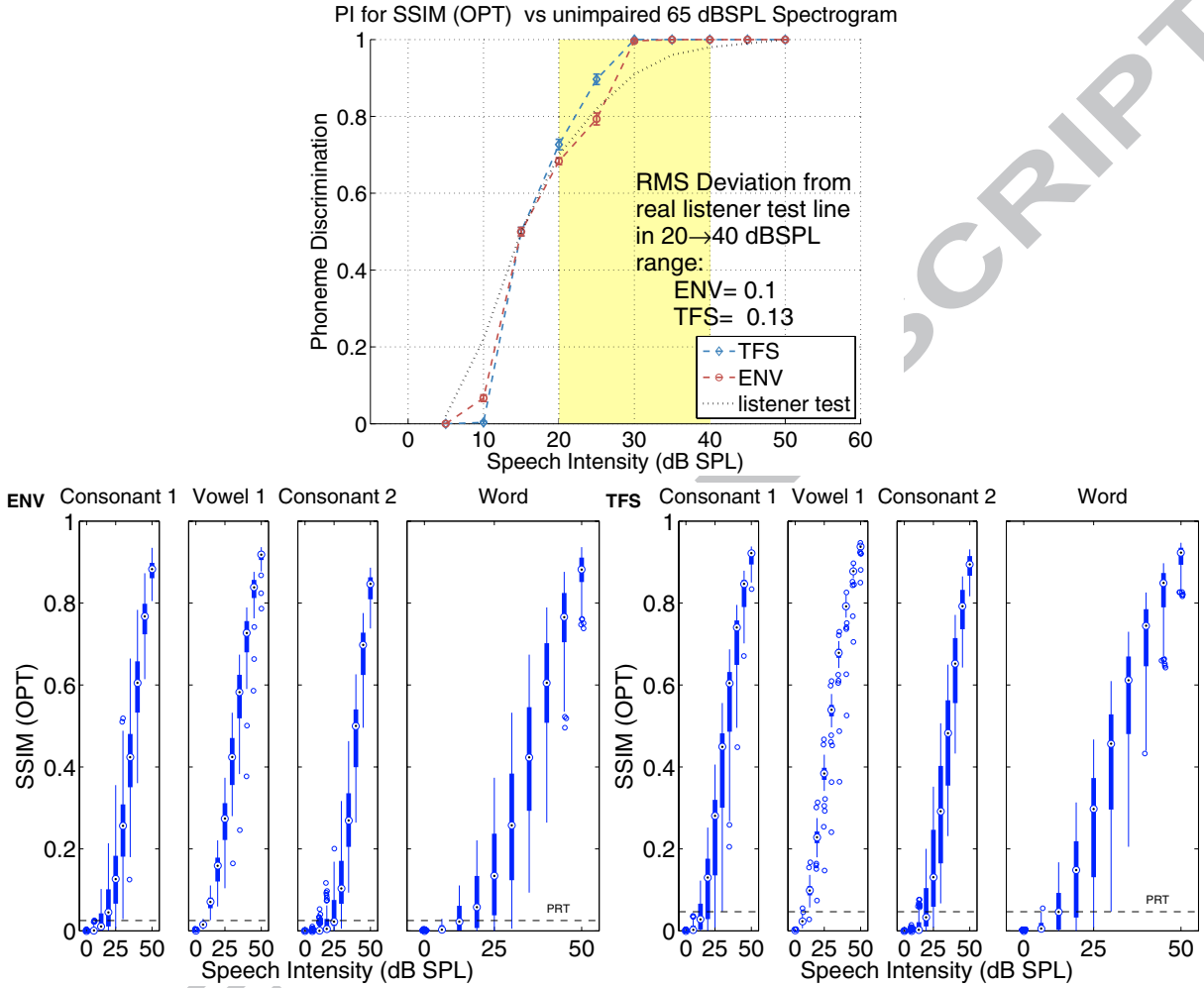

Fig. 9: Spectrogram tests. PI function generated using optimal SSIM weights without the use of the AN model. Raw SSIM data for spectrograms with resolutions equivelent to ENV and TFS neurograms.

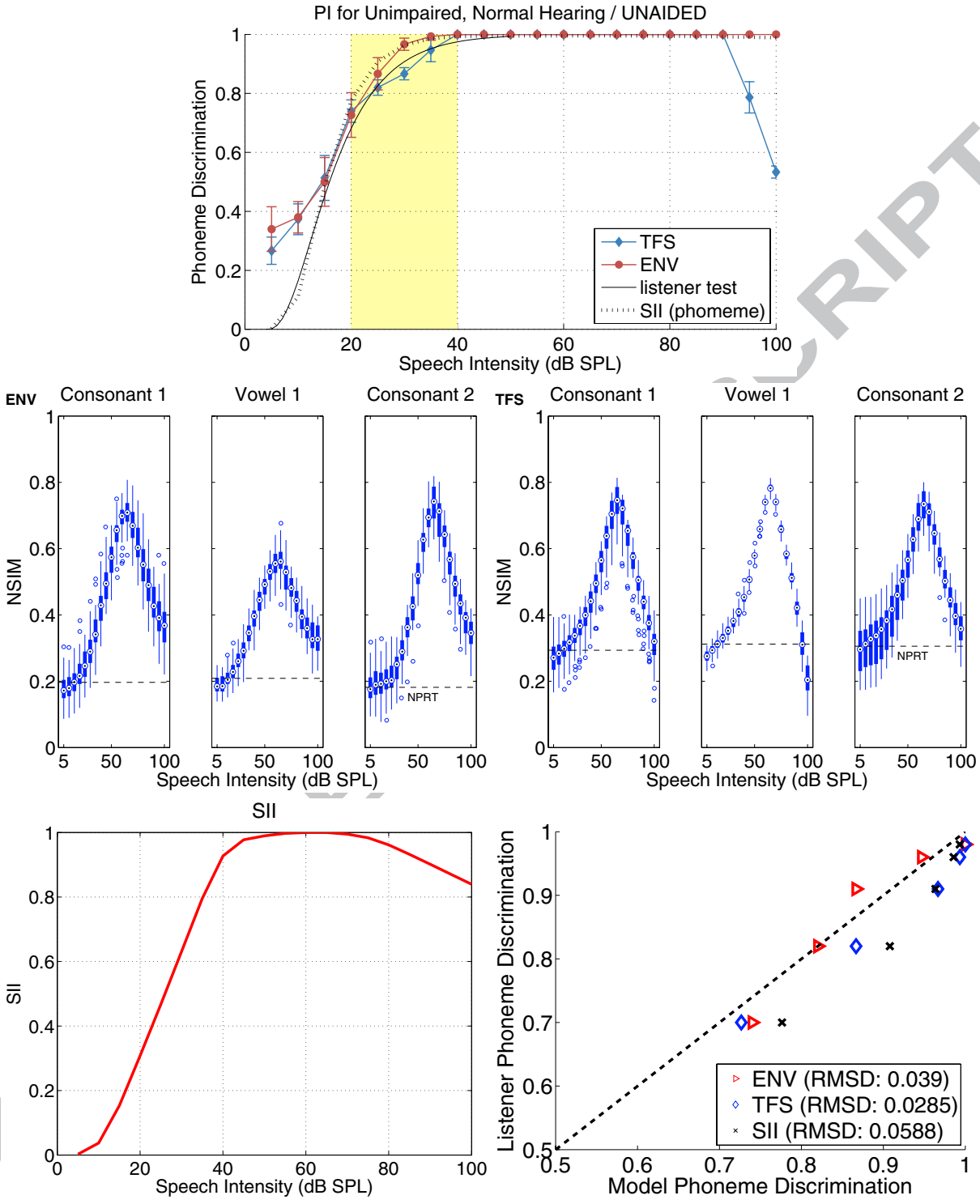

Fig. 10: Top: Simulated performance intensity functions for NSIM evaluation of ENV and TFS neurograms in quiet with SII phoneme discrimination predition plotted for comparison. Second Row: NSIM scores plotted per phoneme position with NPRT level at 15 dB SPL. Third Row: SII plot and real versus modelled data linear correlation and RMSD.

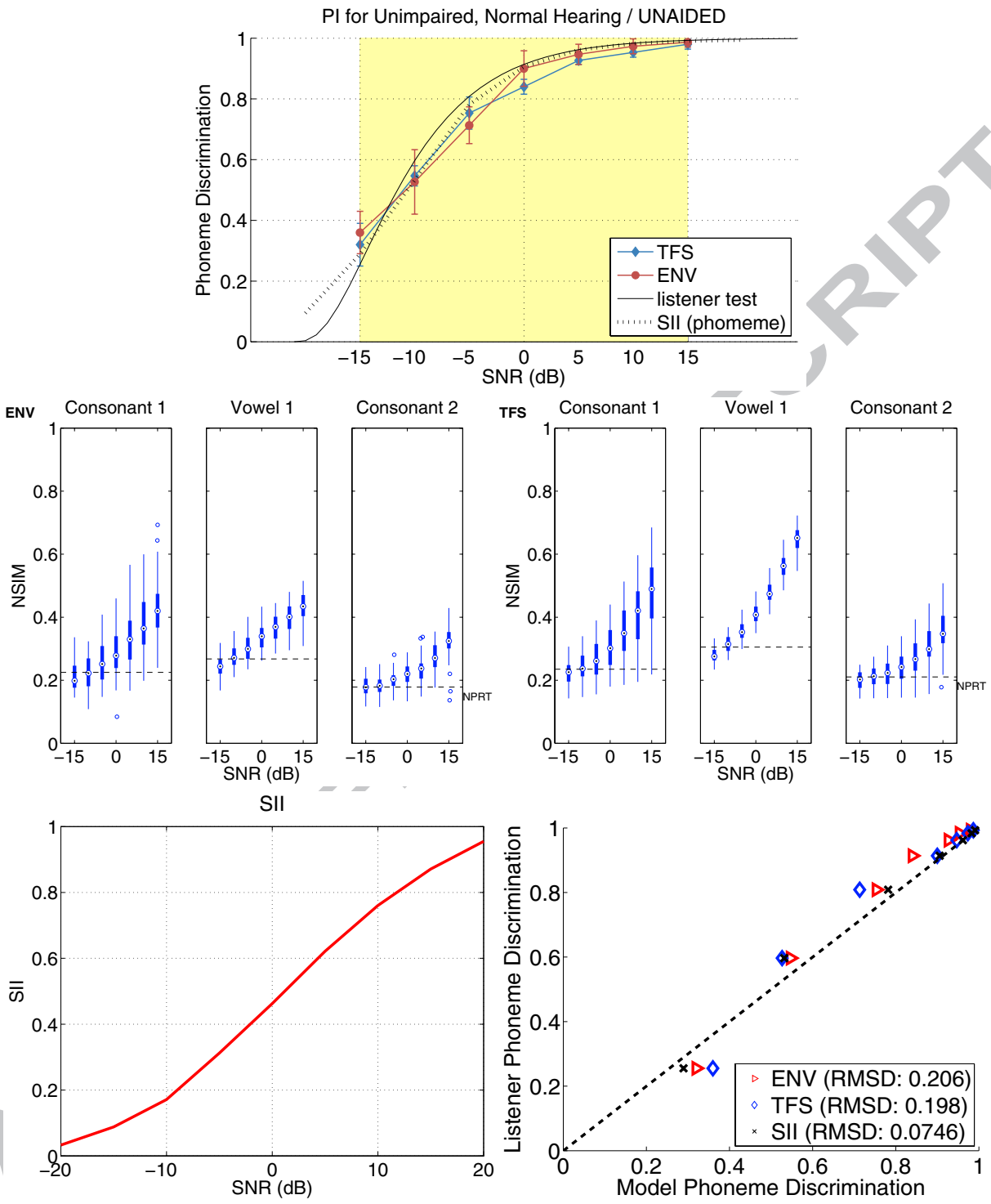

Fig. 11: Top: Simulated performance intensity functions for NSIM evaluation of ENV and TFS neurograms in 55 dB SPL steady state noise with SII phoneme discrimination prediction plotted for comparison. Second Row: NSIM scores plotted per phoneme position with NPRT level at -11 dB SPL. Third Row: SII plot and real versus modelled data linear correlation and RMSD.

Highlights

> We predict Speech intelligibility by modelling tests for normal hearing listeners. > A neurogram similarity index quantifies intelligibility from computational auditory model outputs. > Simulated performance intensity function results are of comparable accuracy to real listener tests. > It represents an important step in validating the use of auditory nerve models to predict speech intelligibility.