Spike-Time Coding and Auditory-Nerve Degeneration Best Explain Speech Intelligibility in Noise for Normal and Near-Normal Low-Frequency Hearing

Ian C. Bruce[1] (ibruce@ieee.org), Agnès C. Léger[2], Michael R. Wirtzfeld[1], Brian C. J. Moore[3], and Christian Lorenzi[4]

[1]McMaster University, Hamilton, Canada; [2]University of Manchester, UK; [3]University of Cambridge, UK; [4]École Normale Supérieure, Paris, France

ABSTRACT

Léger et al. (2012) measured the intelligibility of speech that was lowpass filtered at 1.5 kHz in background noise for a group of hearing-impaired (HI) listeners who had normal or near-normal hearing below 1.5 kHz. Compared to a control group of normal hearing (NH) listeners, the HI listeners displayed an overall deficit in speech understanding. However, the improvement in intelligibility obtained by introducing temporal or spectral dips into the masking noise (referred to as "masking release") was similar for the NH and HI groups. It was not possible to explain the patterns of masking release exhibited by the two groups using the extended speech intelligibility index (ESII). Also, the ESII only allows for ad hoc implementation of hearing impairment. This motivated the use of a neural-based intelligibility predictor, to see what forms of neural speech coding can explain masking release and what types of cochlear pathology best describe the suprathreshold deficit of the HI listeners while preserving masking release.

The auditory-periphery model of Zilany et al. (2009, 2014) was used to obtain predictions of auditory nerve (AN) responses to the stimuli used by Léger et al. (2012). The effects of outer hair cell (OHC) impairment, inner hair cell (IHC) impairment, and degeneration of AN fibers were studied. Two different neural-based predictors were investigated: the Spectro-Temporal Modulation Index (STMI; Elhilali et al., 2003) and the Neurogram SIMilarity index (NSIM; Hines and Harte, 2010, 2012). Two versions of each of these metrics were assessed, one that depends only on the mean-rate AN representation and another that depends on the all-information AN representation (i.e., includes spike-timing information).

The mean-rate versions of the STMI and NSIM did not accurately predict the patterns of masking release seen in the human data. The all-information version of the STMI gave somewhat improved predictions of the NH data but over-predicted the effects of impairment for the HI group. In contrast, the all-information NSIM gave accurate predictions of the data for both groups. The best predictions of the deficits in overall intelligibility for the HI group were obtained with mixed OHC/IHC impairment and some degradation of AN fibers.

These results strongly suggest that spike-time coding of speech is required to explain masking release and that some suprathreshold deficits in intelligibility may be caused by AN degradation.

I INTRODUCTION

A long-standing question in auditory research has been whether speech features are represented by spike-timing (Young and Sachs, 1979) or mean-rate (Sachs and Young, 1979) cues in the auditory nerve (AN) response. Spike-timing cues are generally more robust in background noise (Sachs et al., 1983), but their necessity cannot be determined without quantitative predictions of speech intelligibility data. In addition, both forms of neural coding could be disrupted by cochlear pathology, but the exact deficits are likely to be dependent on the pattern of cochlear impairment. In this study, we used a computational model of the AN to explore these issues by means of direct predictions of a set of speech perception data from Léger et al. (2012).

A Experimental Design of Léger et al. (2012)

The study of Léger et al. (2012) used four sets of 48 Vowel-Consonant-Vowel (VCV) stimuli in a consonant identification task. Experiment I measured the perception of Low-Frequency Speech, that is, speech lowpass (LP) filtered at 1.5 kHz. The mean audiograms of the normal hearing (NH) and hearing impaired (HI) groups are shown in the left panel of Fig. 1. All the HI subjects have clinically normal or near-normal thresholds below 1.5 kHz. In addition to outer hair cell (OHC) and inner hair cell (IHC) impairment producing these audiograms, these participants may also suffer from degradation of their low-spont AN fibers in cases of normal audiograms (Kujawa and Liberman, 2009) or of both their high-spont and low-spont AN fibers in cases of threshold shift (Liberman and Dodds, 1984)—see the right panel of Fig. 1.

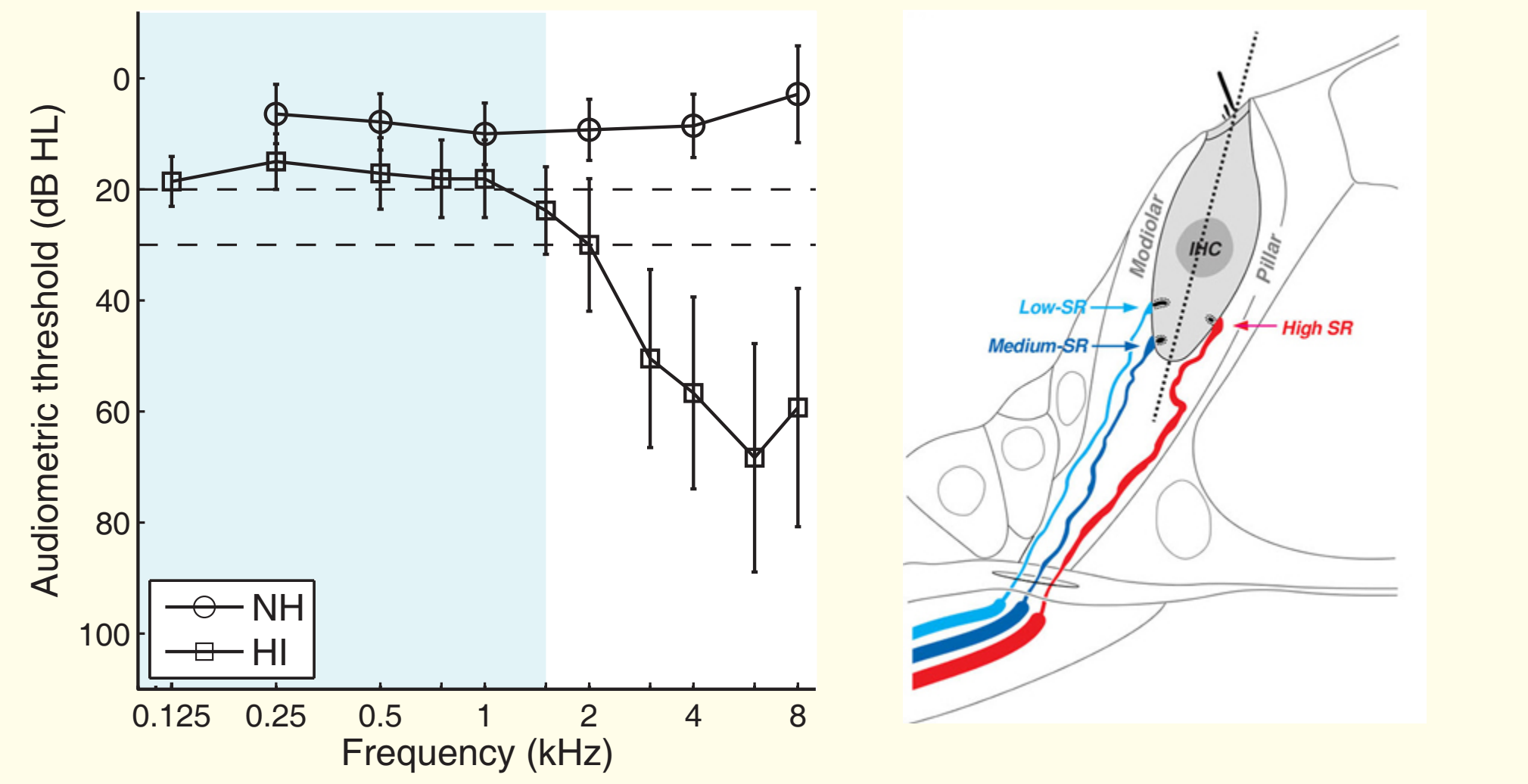

Figure 1: ***What type of cochlear damage is present for near-normal audiograms?*** **Left: Audiograms for the normal hearing (NH) and hearing impaired (HI) listener groups.** The curves show mean air conduction thresholds of the test ears, with error bars indicating ± 1 SD. Horizontal dashed lines show the limits of normal (≤ 20 dB HL) and near-normal (≤ 30 dB HL) audiometric thresholds. The shaded area shows the frequency limit of the stimuli used in the experiment (i.e., < 1.5 kHz). Adapted from Léger et al. (2012) © Acoustical Society of America. **Right: Innervation of inner hair cells (IHCs) by low spont rate (LSR), medium spont rate (MSR) and high spont rate (HSR) auditory nerve fibers (ANFs).** Reprinted from Bharadwaj et al. (2014) © 2014 Bharadwaj, Verhulst, Shaheen, Liberman and Shinn-Cunningham.

Speech identification was measured in quiet and using six different masker noises presented at three different signal-to-noise ratios (-6, -3 and 0 dB SNR):

- Notionally steady (unmodulated) speech-shaped noise.
- Temporally-modulated noises (8-Hz square wave, 100% modulation depth) at a:
  - 50% duty-cycle (DC), or
  - 25% duty-cycle.
- Spectrally-modulated noises created by passing noise through an array of gammatone filters, each with a bandwidth of 1 ERB_N (Glasberg and Moore, 1990), and setting to zero the output of:
  - one filter out of every two ($1\text{ERB}_\text{N}/2$),
  - two adjacent filters out of every four ($2\text{ERB}_\text{N}/4$), or
  - three adjacent filters out of every four ($3\text{ERB}_\text{N}/4$).

B Perceptual Data from Léger et al. (2012)

As shown in Fig. 2, the HI group had an overall deficit in speech perception compared to the NH group, even though audibility differences were controlled for (by testing HI listeners in frequency regions of clinically near-normal audibility and amplifying speech for listeners with thresholds > 30 dB HL). Furthermore, the degree of temporal and spectral masking release was the same across the two groups.

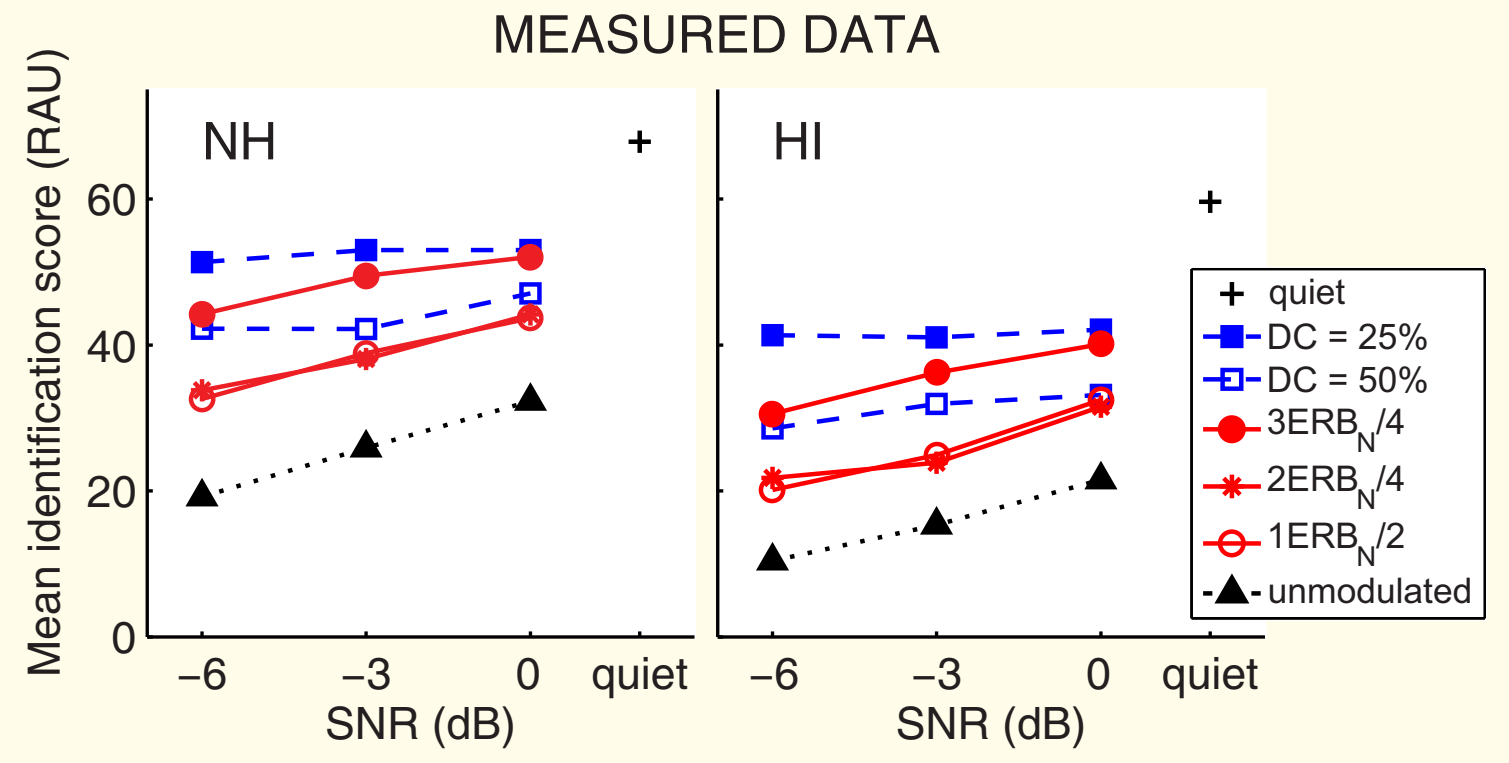

Figure 2: Mean consonant identification scores for VCVs in rationalized arcsine units (RAU) for the NH and HI listener groups, plotted as a function of SNR (in dB). Conditions were: in quiet (crosses), in temporally-modulated noise (squares), in spectrally-modulated noise (circles and asterisks), and in unmodulated noise (triangles). Adapted from Léger et al. (2012) © Acoustical Society of America.

II SPEECH INTELLIGIBILITY PREDICTORS

A Extended Speech Intelligibility Index (ESII)

The extended speech intelligibility index (ESII) of Rhebergen and Versfeld (2005) and Rhebergen et al. (2006) is an acoustic-based metric. Predictions are based on averaging the time-varying SNR. Hearing loss and suprathreshold deficits are implemented in an *ad hoc* fashion.

B Spectro-temporal Modulation Index (STMI)

Elhilali et al. (2003) developed the spectro-temporal modulation index (STMI) to quantify the fidelity of the cortical representation of spectral and temporal modulations. AN fiber PSTHs were generated for 128 characteristic frequencies (CFs), logarithmically spaced from 180 to 7,040 Hz. Each PSTH response was convolved with a 16-ms rectangular window at 50% overlap to produce mean-rate neurograms. As shown in Fig. 3, two cortical responses are produced: (1) a reference output, *T*, from the original stimulus in quiet and (2) a response to the processed stimulus, *N*.

$$\text{STMI} = 1 - \frac{\|T-N\|^2}{\|T\|^2} \quad (1)$$

The STMI is a scalar value between 0 and 1, with a larger number indicating better predicted speech intelligibility. Without the lateral inhibitory network (LIN) between the auditory periphery and cortical models, the STMI is sensitive only to mean-rate cues. With the introduction of the LIN, some spike-timing cues in the AN are converted to mean-rate cues.

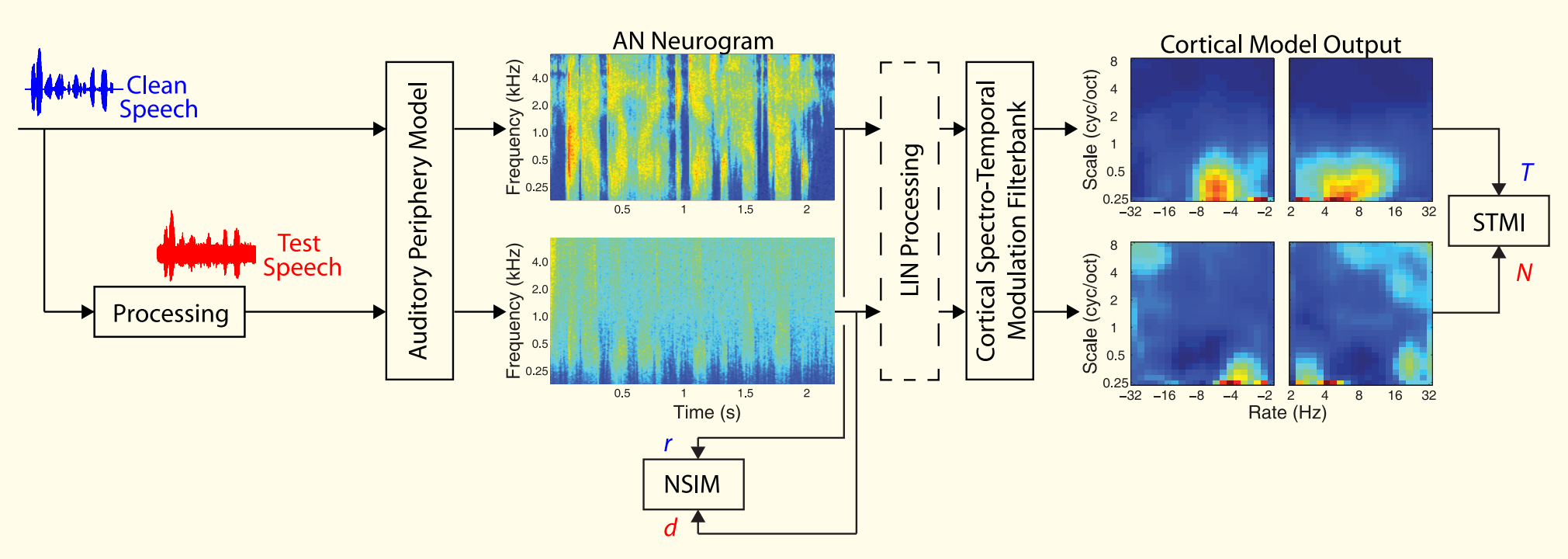

Figure 3: Schematic of calculation of the Neurogram SIMilarity (NSIM) metric based on the reference r and degraded d auditory nerve (AN) neurograms and the Spectro-Temporal Modulation Index (STMI) based on the processing of the AN neurograms by a bank of cortical spectro-temporal modulation filters producing template T and "noisy" N auditory cortex outputs. In some simulations the AN neurograms are passed through a lateral inhibitory network (LIN) before the cortical modulation filterbank. In this study, the *Processing* to produce the *Test Speech* includes the LP filtering of the speech and the different background noise conditions. In addition, the *Test Speech* is passed through a version of the *Auditory Periphery Model* of Zilany et al. (2014, 2009) that incorporates any hair cell impairment and/or AN fiber degeneration when computing the degraded AN neurogram d and the noisy cortical output N. Human cochlear tuning from Ibrahim and Bruce (2010) was used in all cases.

C Neurogram SIMiliarity (NSIM) metric

Hines and Harte (2012, 2010) developed the Neurogram SIMilarity (NSIM) metric based on a visual image quality metric. In this study, AN fiber PSTHs were collected for CFs at the 21 frequencies used in the ESII. *Mean-rate (MR) neurograms* were produced with 100-µs time bins and convolution with a 128-sample Hamming window (50% overlap), while *all-information (AI) neurograms* were produced with 10-µs time bins and convolution with a 32-sample Hamming window (50% overlap). "Luminance" (μ_r, μ_d), "contrast" (σ_r, σ_d) and "structure" (σ_{rd}) statistics were calculated for 3×3 patches of the neurogram, and contributions were weighted (α, β and γ) to determine a single patch NSIM value according to:

$$\text{NSIM} = \left(\frac{2\mu_r\mu_d + C_1}{\mu_r^2 + \mu_d^2 + C_1}\right)^{\alpha} \cdot \left(\frac{2\sigma_r\sigma_d + C_2}{\sigma_r^2 + \sigma_d^2 + C_2}\right)^{\beta} \cdot \left(\frac{\sigma_{rd} + C_3}{\sigma_r\sigma_d + C_3}\right)^{\gamma} \quad (2)$$

Weighting parameters $(\alpha,\ \beta,\ \gamma)$ were set to $(1,\ 0,\ 1)$, respectively. A set of regularization parameters C_1, C_2 and C_3 was found that gave greatly improved predictions compared to the parameters used by Hines and Harte (2012, 2010). An overall metric value was found by averaging the NSIM values over time and CF.

III RESULTS

Prediction results were obtained for the ESII, the STMI with (w/) and without (w/o) LIN processing, the all-information (AI) NSIM, and the mean-rate (MR) NSIM.

Regression analysis was conducted first with the data for the NH and HI groups fitted separately and then with the data for the groups combined. The former allows for the possibility that there may be some group differences not accounted for by cochlear impairment, while the latter requires that all differences between the NH and HI group be captured by the AN model.

For all the intelligibility predictors, the mean audiograms of the NH and HI groups were used. For the STMI, a mixed OHC/IHC impairment with no AN fiber degeneration was investigated. For the NSIM metrics, the effects partial low-spont fiber (LSF) loss, partial high-spont fiber (HSF) and LSF loss, OHC impairment alone and IHC impairment alone were also investigated.

A Prediction Results Summary

Table I: Regression Analysis Results

Model	NH Group Adj. R^2	HI Group Adj. R^2	Comb. Groups Adj. R^2
ESII	0.60	0.70	0.55
STMI w/o LIN, mixed HC imp., all ANFs	0.74	0.78	0.58
STMI w/ LIN, mixed HC impairment, all ANFs	0.86	0.93	0.71
AI NSIM, mixed HC impairment, all ANFs	0.91	0.97	0.80
AI NSIM, mixed HC imp., LSF loss	0.91	0.97	0.80
AI NSIM, mixed HC imp., HSF & LSF loss	<u>0.93</u>	<u>0.98</u>	<u>0.86</u>
AI NSIM, OHC impairment, all ANFs	0.91	0.97	0.79
AI NSIM, IHC impairment, all ANFs	0.92	<u>0.98</u>	0.81
MR NSIM, mixed HC impairment, all ANFs	0.78	0.96	0.28
MR NSIM, mixed HC imp., LSF loss	0.77	0.96	0.28
MR NSIM, mixed HC imp., HSF & LSF loss	0.76	0.58	0.69
MR NSIM, OHC impairment, all ANFs	0.77	0.95	0.33
MR NSIM, IHC impairment, all ANFs	0.80	0.96	0.16

Abbreviations: ESII = Extended Speech Intelligibility Index; STMI w/o LIN = Spectro-Temporal Modulation Index without lateral inhibitory network; STMI w/ LIN = Spectro-Temporal Modulation Index with lateral inhibitory network; AI NSIM = all-information Neurogram SIMilarity index; HC = hair cell; ANFs = auditory nerve fibers; LSF = low-spont fiber; HSF = high-spont fiber; OHC = outer hair cell; IHC = inner hair cell; MR NSIM = mean-rate Neurogram SIMilarity index. p values < 0.01 for all fits. Values for the models providing the best fits are underlined.

B Groupwise Regression Predictions

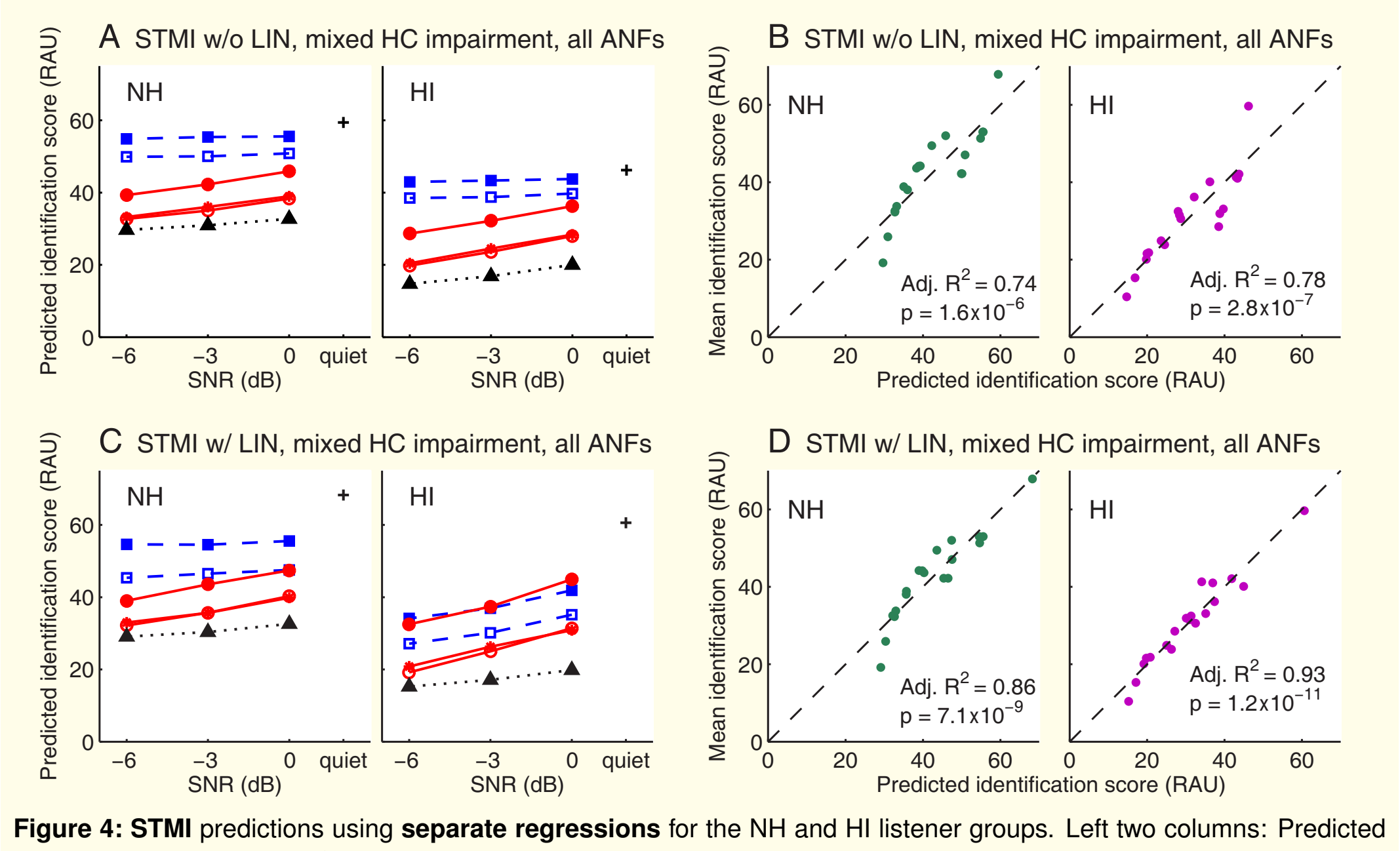

Figure 4: STMI predictions using **separate regressions** for the NH and HI listener groups. Left two columns: Predicted mean RAU scores versus SNR. The plotting convention is the same as for Fig. 2. Right two columns: Measured mean RAU scores versus predicted mean RAU scores. Dots show individual values for each of the 19 processing conditions. **A & B: STMI without a LIN**. **C & D: STMI with a LIN**.

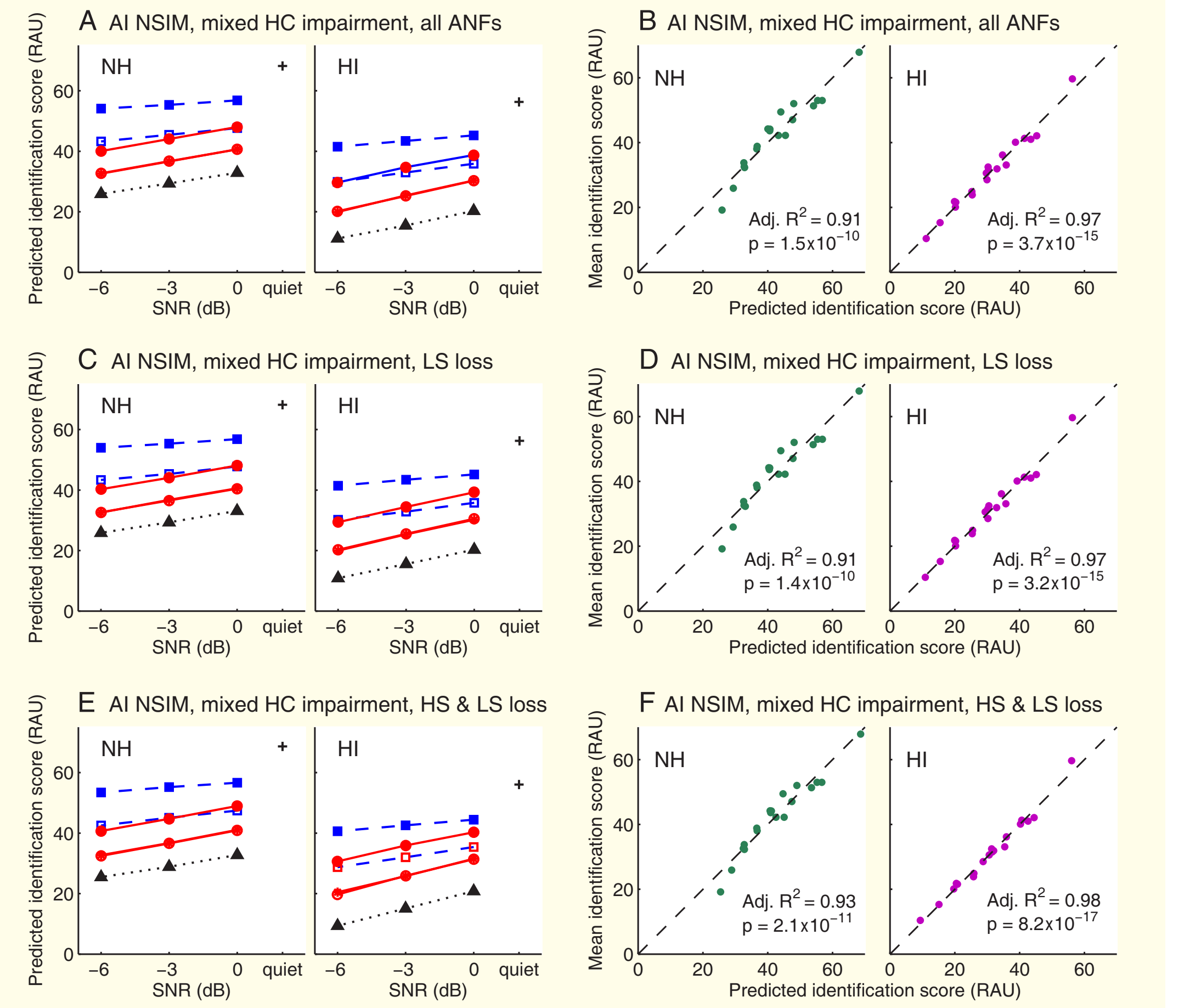

Figure 5: All-information (AI) NSIM and mixed hair-cell (HC) impairment predictions using **separate regressions** for the NH and HI listener groups. **A & B: All ANFs intact**, i.e., without deafferentation. **C & D: Some loss of low-spont (LS), high-threshold ANFs**. **E & F: Some loss of high-spont (HS), low-threshold ANFs, in addition to loss of LS ANFs**.

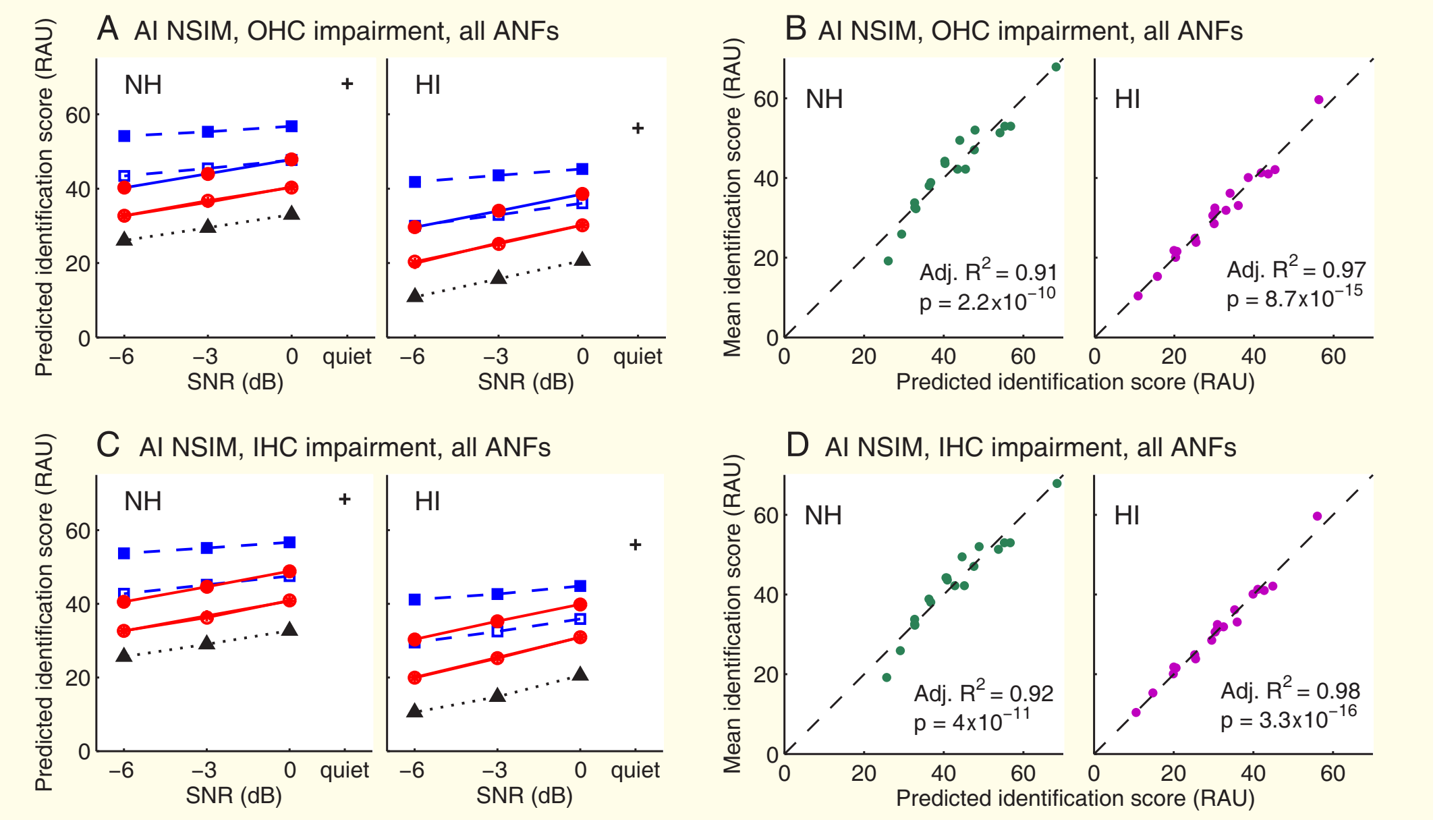

Figure 6: AI NSIM and either outer-hair-cell (OHC) or inner-hair-cell (IHC) impairment alone predictions using **separate regressions** for the NH and HI listener groups. Each row is for a different auditory model predictor. **A & B: OHC impairment alone** giving the audiometric threshold shifts and no deafferentation. **C & D: IHC impairment alone** giving the audiometric threshold shifts and no deafferentation.

C Regression Predictions for Combined Groups

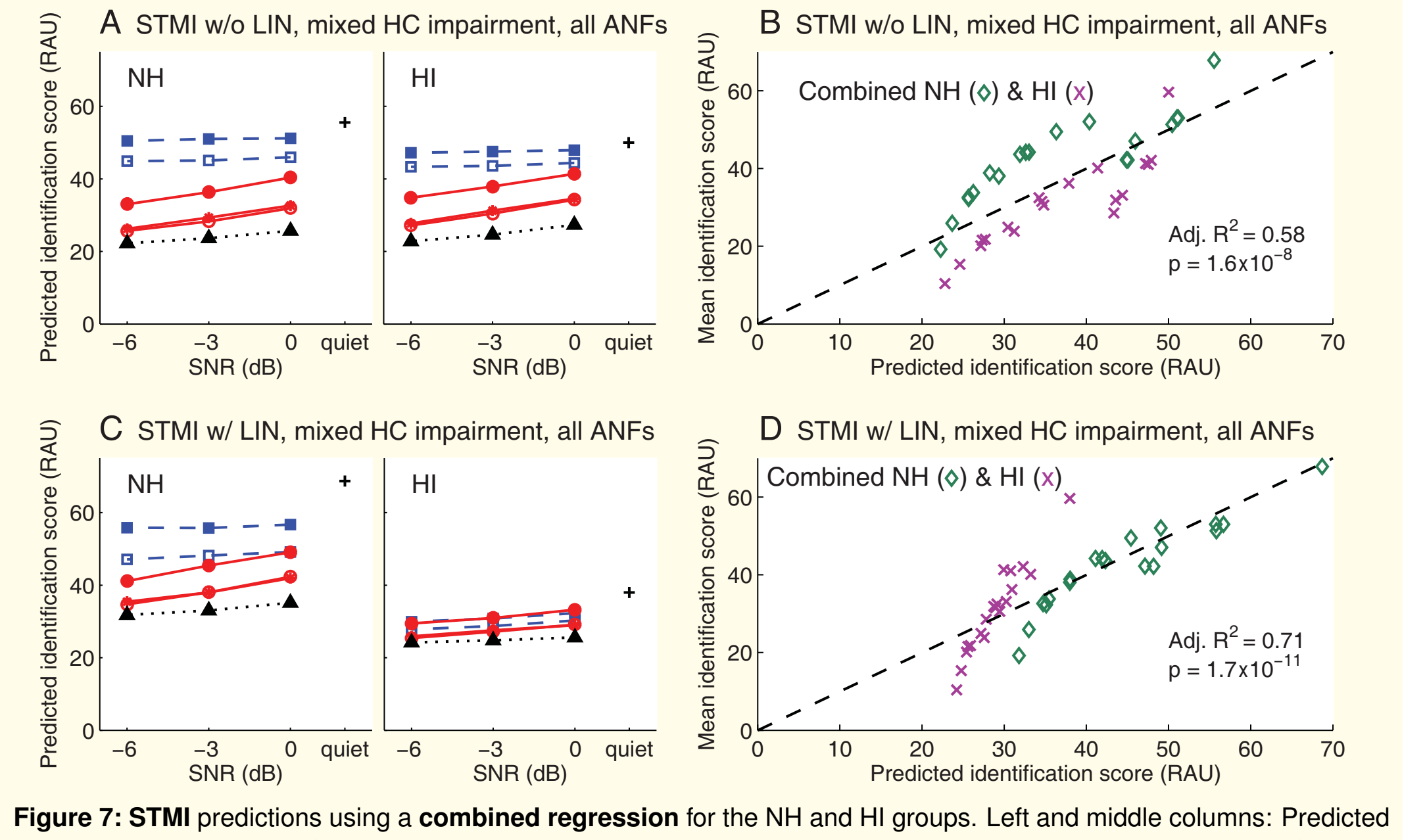

Figure 7: STMI predictions using a **combined regression** for the NH and HI groups. Left and middle columns: Predicted mean RAU scores versus SNR. Right column: Measured mean RAU scores versus predicted mean RAU scores. Symbols show individual values for each of the 19 processing conditions. **A & B: STMI without a LIN**. **C & D: STMI with a LIN**.

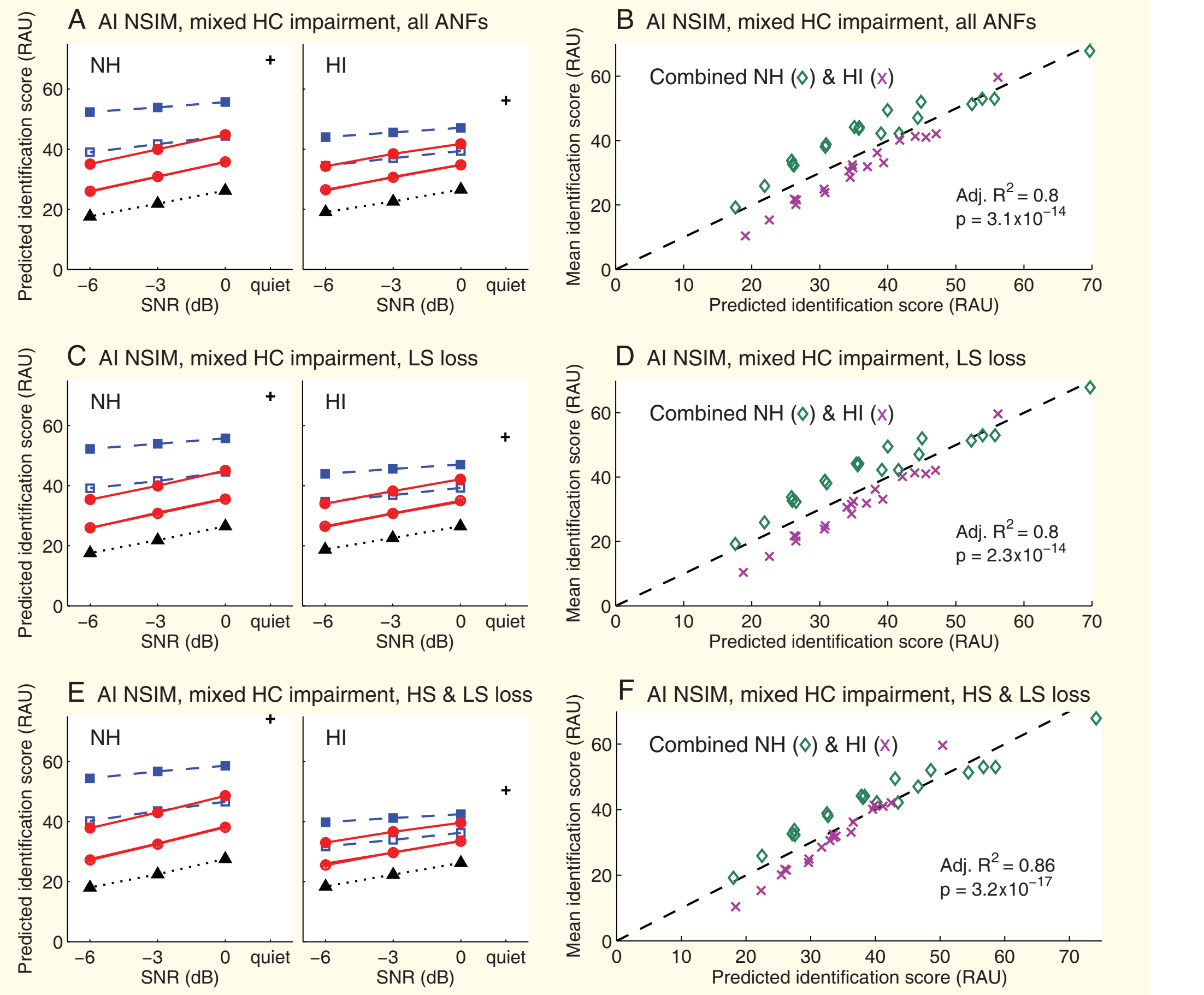

Figure 8: AI NSIM and mixed hair-cell HC impairment predictions using a **combined regression** for the NH and HI groups. Plotting convention as for Fig. 7 **A & B: All ANFs intact**, i.e., without deafferentation. **C & D: Some loss of low-spont (LS), high-threshold ANFs**. **E & F: Some loss of high-spont (HS), low-threshold ANFs, in addition to loss of LS ANFs**.

IV CONCLUSIONS

- The predictions of the all-information NSIM metric were superior to those of the other metrics, suggesting that spike-timing cues in the AN representation of speech are crucial in explaining the masking release data of Léger et al. (2012).
- Adding a LIN in the STMI processing, which makes the STMI somewhat sensitive to spike time cues, did improve the predictions substantially, consistent with the theory proposed by Shamma and Lorenzi (2013), but the predictions were still inferior to those of the AI NSIM.
- Partial loss of both high-spont and low-spont AN fibers and mixed OHC/IHC impairment, consistent with Liberman and Dodds (1984), best explained the degraded speech perception of the HI group.

V ACKNOWLEDGMENTS

Supported by: Erasmus Mundus ACN Training Network, NSERC Discovery Grant 261736, the Royal Society (International Joint Project, 2009R3), 7FP-SME222291 Dual-Pro program, MRC (UK), and CIFRE Grant from ANRT and Neurelec.

REFERENCES

Bharadwaj, H. M., Verhulst, S., Shaheen, L., Liberman, M. C., and Shinn-Cunningham, B. G. (2014). "Cochlear neuropathy and the coding of supra-threshold sound," Front. Syst. Neurosci. **8**, 26.

Elhilali, M., Chi, T., and Shamma, S. A. (2003). "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Comm. **41**, 331–348.

Glasberg, B. R. and Moore, B. C. J. (1990). "Derivation of auditory filter shapes from notched-noise data," Hear. Res. **47**, 103–138.

Hines, A. and Harte, N. (2010). "Speech intelligibility from image processing," Speech Comm. **52**, 736 – 752.

Hines, A. and Harte, N. (2012). "Speech intelligibility prediction using a neurogram similarity index measure," Speech Comm. **54**, 306 – 320.

Ibrahim, R. A. and Bruce, I. C. (2010). "Effects of peripheral tuning on the auditory nerve's representation of speech envelope and temporal fine structure cues," in *The Neurophysiological Bases of Auditory Perception*, edited by E. A. Lopez-Poveda, A. R. Palmer, and R. Meddis (Springer, New York), Chap. 40, pp. 429–438.

Kujawa, S. G. and Liberman, M. C. (2009). "Adding insult to injury: cochlear nerve degeneration after "temporary" noise-induced hearing loss," J. Neurosci. **29**, 14077–14085.

Léger, A. C., Moore, B. C. J., and Lorenzi, C. (2012). "Temporal and spectral masking release in low- and mid-frequency regions for normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **131**, 1502–1514.

Liberman, M. C. and Dodds, L. W. (1984). "Single-neuron labeling and chronic cochlear pathology. II. Stereocilia damage and alterations of spontaneous discharge rates," Hear. Res. **16**, 43–53.

Rhebergen, K. S. and Versfeld, N. J. (2005). "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," J. Acoust. Soc. Am. **117**, 2181–2192.

Rhebergen, K. S., Versfeld, N. J., and Dreschler, W. A. (2006). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," J. Acoust. Soc. Am. **120**, 3988–3997.

Sachs, M. B., Voigt, H. F., and Young, E. D. (1983). "Auditory nerve representation of vowels in background noise," J. Neurophysiol. **50**, 27–45.

Sachs, M. B. and Young, E. D. (1979). "Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate," J. Acoust. Soc. Am. **66**, 470–479.

Shamma, S. and Lorenzi, C. (2013). "On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system," J. Acoust. Soc. Am. **133**, 2818–2833.

Young, E. D. and Sachs, M. B. (1979). "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," J. Acoust. Soc. Am. **66**, 1381–1403.

Zilany, M. S. A., Bruce, I. C., and Carney, L. H. (2014). "Updated parameters and expanded simulation options for a model of the auditory periphery," J. Acoust. Soc. Am. **135**, 283–286.

Zilany, M. S. A., Bruce, I. C., Nelson, P. C., and 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," J. Acoust. Soc. Am. **126**, 2390–2412.

ARO 2015 Midwinter Meeting - Baltimore, MD - February 21–25