2010

Speech Intelligibility from Image Processing

Andrew Hines
Technological University Dublin, andrew.hines@tudublin.ie

Naomi Harte
University of Dublin, Trinity College

Recommended Citation

Hines, A. & Harte, N. (2010) Speech Intelligibility from Image Processing,*Speech Communication,* iss. 9, 736-752 pp. doi:10.1016/j.specom.2010.04.006

Speech Intelligibility from Image Processing

Andrew Hines, Naomi Harte

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

Abstract

Hearing loss research has traditionally been based on perceptual criteria, speech intelligibility and threshold levels. The development of computational models of the auditory-periphery has allowed experimentation via simulation to provide quantitative, repeatable results at a more granular level than would be practical with clinical research on human subjects. The responses of the model used in this study have been previously shown to be consistent with a wide range of physiological data from both normal and impaired ears for stimuli presentation levels spanning the dynamic range of hearing.

The model output can be assessed by examination of the spectro-temporal output visualised as neurograms. The effect of sensorineural hearing loss (SNHL) on phonemic structure was evaluated in this study using two types of neurograms: temporal fine structure (TFS) and average discharge rate or temporal envelope. A new systematic way of assessing phonemic degradation is proposed using the outputs of an auditory nerve model for a range of SNHLs. The mean structured similarity index (MSSIM) is an objective measure originally developed to assess perceptual image quality. The measure is adapted here for use in measuring the phonemic degradation in neurograms derived from impaired auditory nerve outputs. A full evaluation of the choice of parameters for the metric is presented using a large amount of natural human speech.

The metric's boundedness and the results for TFS neurograms indicate it is a superior metric to standard point to point metrics of relative mean absolute error and relative mean squared error. MSSIM as an indicative score of intelligibility is also promising, with results similar to those of the standard Speech Intelligibility Index metric.

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

Key words:
auditory periphery model, hearing aids, sensorineural hearing loss,
structural similarity, MSSIM, Speech Intelligibility

1. Introduction

Hearing loss research has traditionally been based on perceptual criteria, speech intelligibility and threshold levels. The development of computational models of the auditory-periphery has allowed experimentation via simulation to provide quantitative, repeatable results at a more granular level than would be practical with clinical research on human subjects.

Several models have been proposed, integrating physiological data and theories from a large number of studies of the cochlea. The model used in this paper is the cat auditory nerve (AN) model of Zilany and Bruce (2007). The code for the model is shared by the authors and the model responses have been shown to be consistent with a wide range of physiological data from both normal and impaired ears for stimuli presentation levels spanning the dynamic range of hearing (Zilany and Bruce, 2006). It produces simulated auditory nerve neural spike train outputs at specific characteristic frequencies (CF). The levels of degradation in output due to a sensorineural hearing loss (SNHL) configured in the model can be assessed by examination of the spectro-temporal output visualised as neurograms. Two distinct types of neurograms are considered important in describing speech signals: a temporal envelope (ENV) measurement; and a temporal fine structure (TFS). The first averages the poststimulus time histogram (PSTH) intensity at each CF over a number of time bins while the latter preserves fine timing structure of the auditory nerve spikes. They are both seen as useful for cues to speech intelligibility (Rosen, 1992).

This work examines a systematic way of assessing phonemic degradation using the outputs of an auditory nerve (AN) model for a range of SNHLs. The practical application of this is to allow speech-processing algorithms for hearing aids to be objectively tested in early stage development without having to resort to extensive human trials. The proposed strategy is to design hearing aids by looking to restore normal patterns of auditory nerve activity rather than focusing on human perception of sounds. Sachs et al. (2002) showed that auditory-nerve discharge patterns in response to sounds as complex as speach can be accurately modelled and predicted that this knowledge

could be used to test new strategies for hearing-aid signal processing. They demonstated examples of auditory-nerve representations of vowels in normal and noise-damaged ears and discussed from a subjective visual inspection how the impaired representations differ from the normal. Comparable examples are displayed in Figs. (5&6). This work seeks to create an objective measure to automate this inspection process and ranks hearing losses based on auditory-nerve discharge patterns.

Previous work (Hines and Harte, 2009) showed that a relative mean absolute error metric (RMAE) that compared the neurogram outputs of phonemes for impaired AN models relative to the output for an unimpaired model "hearing" the same input, was not fully reflecting the complexity of TFS effects - particularly in vowels. This paper explores the use of an alternative mean structural similarity measure (MSSIM)(Wang et al., 2004) and uses it to compare neurograms produced for utterances over a range of SNHL. MSSIM is a statistical metric popular in image processing that was originally developed to estimate the reconstruction quality of compressed images. It has also been shown to have potential in audio quality assessment to compare and optimise audio compression algorithms (Kandadai et al., 2008).

Speech intelligibility is a method for computing a physical measure that is highly correlated with the intelligibility of speech as evaluated by speech perception tests given a group of talkers and listeners. The Speech Intelligibility Index (SII) has been standardised by ANSI (1997). While SII is calculated from acoustical measurements of speech and noise this work looks at computing intelligibility through the measurement of simulated auditory nerve output.

Section 2 introduces the computational modelling of the auditory periphery and how their outputs can produce neurograms. It also introduces the structured similarity measure used in this study and other speech intelligibility measures. Section 3 describes the speech corpus used and the methodology employed to assess the measure using the computational model and progressively degrading SNHLs. Section 4 presents and discusses important features of the results, with conclusions and future work presented in Section 5.

2. Background

2.1. Auditory Periphery Model

A phenomenological based AN model matches its responses to experimental results measured for physiological tests. To date, no model claims to fully implement all the current knowledge of physiological characteristics, specifically: fibre types, dynamic range, adaptation, synchronisation, frequency selectivity, level-dependent rate and phase responses, suppression, and distortion (Lopez-Poveda et al., 2005).

The auditory nerve (AN) model used in this study was designed with an ultimate goal of predicting human speech recognition performance for both normal hearing and hearing impaired listeners (Zilany, 2007). It builds upon several efforts to develop computational models including Deng and Geisler (1987), Zhang et al. (2001) and Bruce et al. (2003). The Deng and Geisler (1987) design sought to account for "synchrony capture" but was unable to deal with longer duration signals due to round-off errors accumulating. It sought to model both suppression and adaptation but not two-tone suppression or basilar membrane (BM) compression. The Zhang et al. (2001) model featured non-linear tuning with compression. Two tone suppression was handled through a broad control path with respect to the signal path. Compression (level dependant gain) was also implemented. The signal path was implemented with a fourth order gammatone filter. The design of Bruce et al. (2003) modelled both normal and impaired auditory peripheries. It looked at aspect of the damage within the periphery such as inner hair cells (IHC) and outer hair cells (OHC) damage and the effects on tuning versus compression. Two-tone rate supression and basilar membrane compression were supported. A middle ear filter was added.

The Zilany and Bruce (2006) model builds upon the previous designs and matched to physiological data over a wider dynamic range than previous auditory models. This was achieved by providing two modes of basilar membrane excitation to the IHC rather than one. The gammatone filter was replaced by a tenth order chirp filter. The model responses are consistent with a wide range of physiological data from both normal and impaired ears for stimuli presented at levels spanning the dynamic range of hearing. It has recently been used to conduct studies into hearing aid gain prescriptions (Dinath and Bruce, 2008) and optimal phonemic compression schemes (Bruce et al., 2007a).

A schematic diagram of the current model is available in Fig.(1) of Zilany and Bruce (2006) which illustrates how model responses matched physiological data over a wider dynamic range than previous models by providing two modes of basilar membrane excitation to the inner hair cell rather than one.

The model is composed of several modules each providing a phenomenological emulation of a particular function of the auditory periphery. First, the stimulus is passed through a filter mimicking the middle ear. The output is then passed to a control path and a signal path. The control path handles the wideband BM filter, followed by modules for non-linearity and low pass filtering by the OHC. The control path feeds back into itself and into the signal path to the time-varying narrowband filter. This filter is designed to simulate the travelling wave delay caused by the BM. The signal is then passed through the non-linear and low pass filters simulating IHCs. A synapse model and spike generator follow allowing for spontaneous and driven activity, adaptation, spike generation and refractoriness in the AN. The model allows hair cell constants C_{IHC} and C_{OHC} to be configured which control the IHC and OHC scaling factors and allow SNHL hearing thresholds to be simulated.

The AN model takes speech waveforms which are used to derive an AN spike train for a fibre with a specific characteristic frequency (CF). By simulating the model over a range of CF it is possible to capture the AN response to speech input in time and frequency. This allows neurogram outputs to be generated. These are similar to spectrograms, except displaying the neural response as a function of CF and time.

Two neurogram representations are produced from the AN model output: a spike timing neurogram (fine timing over several microseconds); and an average discharge rate (time resolution averaged over several milliseconds). The neurograms allow comparative evaluation of the performance of unimpaired versus impaired auditory nerves.

2.2. Neurograms

The effect of SNHL was evaluated in this study using two types of neurograms: temporal fine structure (TFS) and average discharge rate or temporal envelope (ENV). Both display the neural response as a function of CF and time. Rosen (1992) breaks the temporal features of speech into three primary groups: envelope (2-50 Hz), periodicity (50-500 Hz) and TFS (600 Hz and 10kHz). The envelope's relative amplitude and duration are cues and translate to manner of articulation, voicing, vowel identity and prosody of

speech. Periodicity is information on whether the signal is primarily periodic or aperiodic, e.g. whether the signal is a nasal or a stop phoneme. TFS is the small variation that occurs between periods of a periodic signal or for short periods in an aperiodic sound and contains information useful to sound identification such as vowel formants.

Smith et al. (2002) looked at the relative importance of ENV and TFS in speech and music perception finding that recognition of English speech was dominated by the envelope while melody recognition used the TFS. Xu and Pfingst (2003) investigated Mandarin Chinese monosyllables and found that in the majority of trials, identification was based on TFS rather than ENV. In a general sense, these findings show that while ENV is important for understand speech tokens (not only of English), TFS is important for F0 variations and is important in the speech intelligibility tone languages. Lorenzi et al. (2006) showed hearing impaired listeners had a reduced ability to process the TFS of sounds which plays an important role in speech intelligibility especially when background sounds are present, suggesting that the ability to use TFS may be critical for "listening in the background dips." They concluded that TFS stimuli may be useful in evaluating impaired hearing and in guiding the design of hearing aids. Work by Bruce et al. (2007b) compared the amplification schemes of NAR-R (National Acoustics Laboratories of Australia, Revised) and DSL (Desired Sensation Level) to find an optimal single-band gain adjustment, finding that the optimal lay in the order of +10dB for envelope evaluations but -10dB to optimise with respect to TFS. The relationship between the acoustic and neural envelope and TFS was examined by Heinz and Swaminathan (2009). Even though the underlying physiological bases has not been established from a perceptual perspective, current research indicates that there is value in analysing both ENV and TFS neurograms. While ENV is seen as more important for spoken English, the importance of TFS to melody, Mandarin Chinese, and English in noise suggests measuring both ENV and TFS restoration when looking to optimise hearing aids to increase speech intelligibility to those with SNHL.

2.3. Mean Structural Similarity Index (MSSIM)

The relative mean absolute error (RMAE) metric was used in previous work by the authors to compare neurograms from phonemes presented to unimpaired and impaired ANs (Hines and Harte, 2009). As MAE is a multiplicative scale, it is comparatively meaningless without normalisation. Thus

for a given unimpaired representation $x(i,j)$, defined on the integer time-frequency grid and an impaired representation $y(i,j)$, the RMAE, calculated relative to the mean unimpaired representation is given by

$$RMAE = \frac{\sum |x(i,j) - y(i,j)|}{\sum |x(i,j)|} \tag{1}$$

For comparative purposes, a relative mean squared error (RMSE) can be calculated in a similar fashion as:

$$RMSE = \sqrt{\frac{\sum |x(i,j) - y(i,j)|^2}{\sum |x(i,j)|^2}} \tag{2}$$

The structured similarity index (SSIM), was proposed by Wang et al. (2004) as an objective method for assessing perceptual image quality. It is a full-reference metric, i.e. it is measured against a known, error free original image. The metric seeks to use the degradation of structural information as a component of its measurement under the assumption that human perception is adapted to structural feature extraction within images. It was found to be superior to MSE for image quality comparison and better at reflecting the overall similarity of two pictures in terms of appearance rather a than simple mathematical point-to-point difference. SSIM is defined as a comparison of the original and degraded signal, x and y, constructed as a function of luminance (l), contrast (c) and structure (s) with the (i,j) grid dropped for clarity:

$$SSIM(x,y) = f(l(x,y), c(x,y), s(x,y)) \tag{3}$$

Luminance, $l(x,y)$, looks at a comparison of the mean (μ) values across the two signals. The contrast, $c(x,y)$ is a variance measure, constructed in a similar manner to the luminance but using the relative standard deviations (σ) of the two signals. The structure is measured as a inner product of two N-dimensional unit norm vectors, equivalent to the correlation coefficient between the original x and y. Each factor is weighted with a coefficient > 0 which can be used to adjust the relative importance of the component, allowing the right hand side of (3) to be expressed as (4). The SSIM metric has properties similar to RMAE or RMSE, as it provides symmetry, $S(x,y) = S(y,x)$, identity $S(x,y) = 1$ if, and only if, $x = y$. However, in addition, it satisfies a desirable property of boundedness $-1 < S(x,y) \leq 1$. See Wang et al. (2004) for a full description.

$$SSIM = (\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1})^\alpha.(\frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2})^\beta.(\frac{(2\sigma_{xy} + C_3)}{\sigma_x\sigma_y + C_3})^\gamma \quad (4)$$

The SSIM metric is applied locally over a window rather than globally, as when comparing images the human observer can only perceive a local area in the image at high resolution at one time instance. The MSSIM is the mean of the SSIM calculated at each comparative point. The choice of window size used by the SSIM for image processing is related to how a person perceives an image, or "how closely they look". The authors suggest values suitable for image comparison. The MSSIM is used in this work to compare neurograms from an impaired AN to that of an unimpaired AN neurogram, e.g. Figs.(5, 6).

To evaluate the choice of window size and weightings that best suit the proposed application, the following criteria were defined. It should correctly predict the order of hearing losses i.e. the metric should deteriorate with increased hearing loss. Secondly it should minimise variance between error metrics for a given phoneme type, given a fixed presentation level and hearing loss. Thirdly, the chosen parameters should make sense in terms of the physiological and signal processing boundaries on the system. (e.g. the choice of window size makes sense in terms of allowing different types of phonemes to be measured by being short enough in the time axis to allow a measurement but long enough to take into account the structural points of interest on longer phonemes.)

Wang et al. (2004) point out that as it is a symmetric measure it can be thought of as a similarity measure for comparing any two signals, not just images. Kandadai et al. (2008) assessed audio quality, both temporally, using short and fixed time-domain frames, and spectro-temporally, using a decomposed non-redundant, time-frequency map. They compared results with human listener tests and found a best fit with weightings towards contrast (variance) and structure rather than the luminance (mean) component, particularly for their time-frequency comparisons.

2.4. Speech Intelligibility

Quantitive prediction of the intelligibility of speech as judged by a human listener is a critical metric in the evaluation of many audio systems from telephone channels through to hearing aids. A number of metrics have been developed to measure speech intelligibility, including static CF mea-

sures (AI/SII), temporal measures (STI), and measures taking account of the physiological effects of the auditory periphery (STMI and NAI).

The Articulation Index (AI) was developed as the result of work carried out in Bell Labs over a number of decades. It was first described by French and Steinberg (1947) and subsequently incorporated into the standard which is now entitled ANSI S3.5-1997 (R2007), "Methods for the Calculation of the Speech Intelligibility Index" (SII) (ANSI, 1997). Additions to AI mean that SII now allows for hearing thresholds, self masking of the speech signal for closely spaced frequency bands and upward spread of masking as well as high presentation level distortions.

The AI measure is described as a range from 0 to 1 or a percentage, where 1 represents perfect information transmission through the channel. As summarised by Steeneken and Houtgast (1980), the computing the AI consists of 3 steps: calculation of the effective signal-to-noise ratio (SNR) within a number of frequency bands; a linear transformation of the effective SNR to an octave-band-specific contribution to the AI; a weighed mean of the contributions of all relevant octave bands. The original definition of AI summed over twenty equally spaced, contiguous frequency bands the equal 5% contribtions , W_i.

$$AI = \frac{1}{20} \sum_{i=1}^{20} W_i \tag{5}$$

Steeneken and Houtgast (1980) proposed an alternative, temporal metric called the Speech-Transmission Index (STI) which was essentially an extension of the AI concept that handled distortion in the time domaim using an underlying Modulation Transfer Function (MTF) concept for the transmission channel.

Elhilali et al. (2003) presented a Spectro-Temporal Modulation Index (STMI) for assessment of speech intelligibility. Their primary motivation was employing an auditory model to allow the analysis of joint spectro-temporal modulations in speech to assess the effects of noise, reverberations and other distortions. STMI was shown to be sensitive to non-linear distortions to which simpler measures, like STI, were not sensitive.

The Neural Articulation Index (NAI), developed by Bondy et al. (2004) estimates speech intelligibility from the instantaneous neural spike rate over time, produced when a signal is processed by an auditory neural model. The NAI uses band weightings and compared favourably with intelligibility pre-

dictions of STI. The authors point out that while NAI is more computationally complex than STI, it can be used for hearing impairment intelligibility applications where AI and STI are only able to account for threshold shifts in hearing loss, not sensorineural supra-threshold degradations. This was examined by Schijndel et al. (2001) who found that for SNHL listeners, detection thresholds for distortions in spectral information where significantly higher than for normal hearing listeners while thresholds in intensity and temporal information distortion thresholds were not significantly different.

Ultimately, the goal is to use MSSIM as a metric of phonemic degradation to quantify loss of speech intelligibility in simulated AN responses for particular SNHL.

3. METHOD

3.1. Test Corpus

The TIMIT corpus of read speech was selected as the speech waveform source (DARPA, 1990). The TIMIT test data has a core portion containing 24 speakers, 2 male and 1 female from each of the 8 American dialect regions. Each speaker reads a different set of SX sentences. The SX sentences are phonetically-compact sentences designed to provide a good coverage of pairs of phones, while the SI sentences are phonetically-diverse. Thus the core test material contains 192 sentences, 5 SX and 3 SI for each speaker, each having a distinct text prompt. The core test set maintains a consistent ratio of phoneme occurrences as the larger "full test set" (2340 sentences). The speech provided by TIMIT is sampled at 16 kHz.

TIMIT classifies fifty seven distinct phoneme types and groups them into 6 phoneme groups (Table. 1) and 1 group of "others" (e.g. pauses). There are 6854 phoneme utterances in the core test set and the number of occurrence of each group is given in Table. 1. The TIMIT corpus of sentences contains phoneme timings for each sentence. These were used in the experiments presented here to analyse neurograms at a phonetic level.

3.2. Audiograms and Presentation Levels

The audiograms used match the samples presented by Dillon (2001) to illustrate prescription fitting over a wide range of hearing impairments. The hearing loss profiles selected were mild, moderate and profound. Two flat hearing losses 10 and 20 dB HL were also included in testing to investigate the

Phoneme Group	Number in core test set	Phonemes
Stops	1989	b d g p t k dx q tcl bcl dcl pcl kcl gcl
Affricates	82	jh ch
Fricatives	969	s sh z zh f th v dh
Nasals	641	m n ng em en eng nx
SV/Glides	832	l r w y hh hv el
Vowels	2341	iy ih eh ey ae aa aw ay ah ao oy ow uh uw ux er ax ix axr ax-h

Table 1: TIMIT phoneme groups. (Stop closures annotated with cl, e.g. tcl)

ability to discriminate between unimpaired and very mild losses in hearing thresholds.

For comparative analysis of responses, it was necessary to create and store AN responses for each of the 192 test sentences. The original TIMIT sentence was resampled to the stimulated minimum sample rate for the AN Model (100kHz) and scaled to 2 presentation levels 65 and 85 dB SPL (denoted P65/P85) representing normal and shouted speech. The head related transfer function (HRTF) from Wiener and Ross (1946) of the human head was used to pre-filter the speech waveforms to mimic the amplification that occurs prior to the middle and inner ear. This technique has been used in other physiological and simulation studies (Zilany and Bruce, 2007).

The response of the AN to acoustic stimuli was quantified with neurogram images. 30 CFs were used, spaced logarithmically between 250 and 8000 Hz. The neural response at each CF was created from the responses of 50 simulated AN fibres. In accordance with Liberman (1978) and as used for similar AN Model simulations (Bruce et al., 2007b), 60% of the fibres were chosen to be high spontaneous rate (>18 spikes/s), 20% medium (0.5 to 18 spikes/s), and 20% low (<0.5 spikes/s). Two neurogram representations were created for analysis, one by maintaining a small time bin size (10μs) for analysing the TFS and another with a larger bin size (100μs) for the ENV. The TFS and ENV responses were smoothed by convolving them with 50% overlap, 128 and 32 sample Hamming window respectively.

The phoneme timing information from TIMIT was used to extract the neurogram information on a per phoneme basis at P65 and P85. This yielded

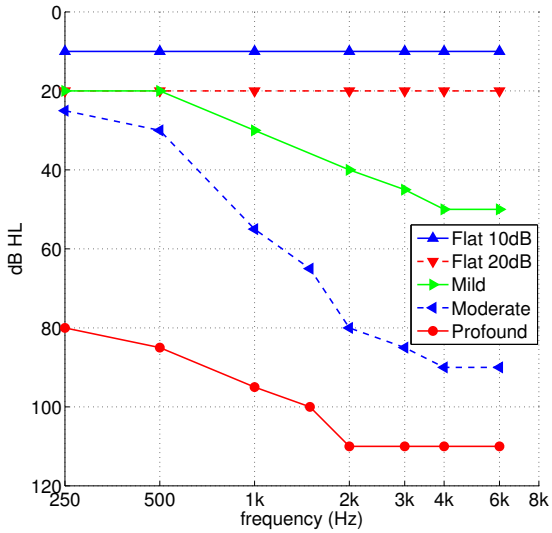

Figure 1: *Audiograms of sample hearing losses tested*

a pair of neurograms for each phoneme utterance representing the original, distortion free reference TFS and ENV images from the unimpaired AN model, and pairs of progressively deteriorating images. The MSSIM measure was calculated between the unimpaired reference image and each of the impaired images. The basic metric described in Wang et al. (2004) was used varying the window sizing parameter. A modified version of Wang's published SSIM code for MATLAB (The MathWorks, Natick, MA) was used to allow variations on α, β and γ weightings.

Treating a neurogram as a picture, each neurogram was a standard height of 30 pixels (one per CF band) and varied in width with the duration of the phoneme. Due to the natural variation in duration of phonemes, the length varied considerably in the region of 3-30 pixels for ENV neurograms and from 100-1200 pixels for TFS neurograms. To assess the impact of these parameters, the MSSIM was calculated across the full data set and an average MSSIM and standard deviation were calculated and aggregated by phoneme group, as per Table.(1), for each hearing loss. The window size was assessed by altering its size in CF from 3 to 30 and then in time coverage from 3 to 11 as illustrated in Fig.(2). The weights α, β & γ were investigated, using the weightings proposed for audio in Kandadai et al. (2008), specifically, $\alpha = 0$,

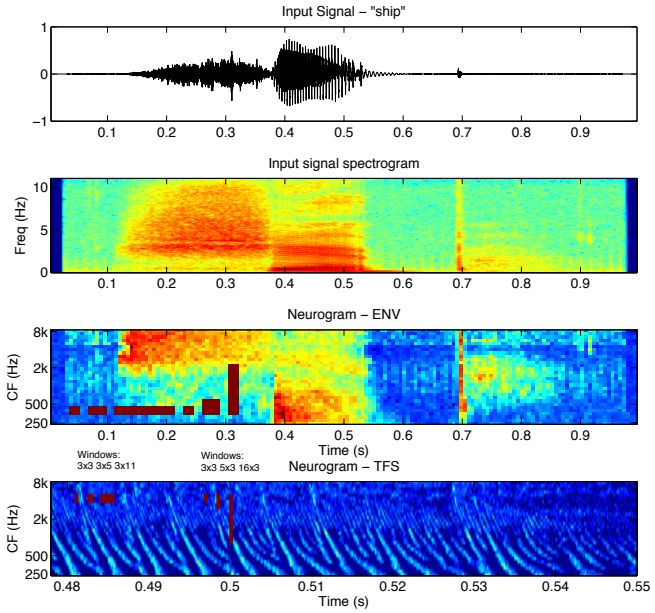

Figure 2: *Illustrative view of window sizes reported on a TFS vowel neurogram. Note that time scale in TFS neurogram is changed (zoomed in on vowel). The neurograms display the sound over logarithmically scaled CF bands in the y-axis against time in the x-axis. The colour represents the intensity of stimulus.*

$\beta = 0.8$ & $\gamma = 0.2$.

4. RESULTS & DISCUSSION

4.1. MSSIM Window Size

The data in Fig.(3) shows results from a subset of the full suite of tests for vowels and fricative phoneme groups. The figure is split into six panels with the left-hand column showing vowels and the right-hand column showing fricatives. Each panel in rows (A) and (B) present 3 different NxM windows where N is frequency and M time resolution. The top row, (A), shows windows with CF fixed and time varying. The MSSIM at any data point represents the similarity between the unimpaired and impaired neurograms for a phoneme group with a particular MSSIM window size. The middle row, (B), shows results with time fixed and CF window size varying. Each panel shows results for both the TFS and ENV neurograms. For each window size, the MSSIM for both TFS(P65:▶;P85:◀) and ENV(P65:▲;P85:▼) can

be seen progressively deteriorating for the hearing loss: flat 10, flat 20, mild, moderate and profound loss. The error bars show one standard devation around the metric as an indication of spread.

Fig.3(A) shows the results for progressively longer time samples in the MSSIM window. The TFS is relatively insensitive to increases in the time window in both vowels and fricatives. However, the ability to differentiate between SNHL levels reduced in the vowel ENV results as they clustered over a smaller range as the time window expanded. This can be seen in moving from 3x3 to 3x11 in (A). The choice of ENV window size was further influenced by the number of samples in the neurogram as for some phonemes, stops in particular, may only be 3 pixels wide.

The effect of including progressively more CF bands is shown in Fig.3(B). The MSSIM is stable for frequency windows of 3-5 pixels for the TFS for both vowels and fricatives as shown in (B) but the ability to distinguish between moderate and profound losses in fricatives diminished for the larger 11x3 window size. The ENV results became marginally more clustered in both vowels and fricatives as the number of CF bands in the window size increased. Results for the other phoneme groups are presented in Appendix A. A detailed examination of plots from the other phoneme groups revealed broadly similar behaviour to changes in window size. This led to the overall conclusion that a suitable window size is 3-5 pixels wide for comparing both the TFS and ENV neurograms. Intuitively this makes sense insofar as the resolution of both has been determined in the choice of window size used to construct the neurograms. In frequency, the MSSIM is looking at information in just 1 or 2 CF bands around the 'ideal' band and the time resolution is $\pm 20\mu s$ for TFS and $\pm 200\mu s$ for ENV. Overall, it is interesting to note the significant drop between unimpaired and Flat 10 and the noticeable difference between Flat 10 and Flat 20, demonstrating the ability of the metric to reflect even small changes in the AN response.

4.2. MSSIM Weighting

Fig.(3)(C) shows the MSSIM for vowels and fricatives with a fixed 3x3 window where luminance, contrast and structure weightings, α, β & γ from (3), were varied. W1 is the unweighted MSSIM with $\alpha = \beta = \gamma = 1$. W2 shows the results with the optimal time-frequency audio weightings as found by Kandadai et al. (2008). Their results found that a zero weighting for luminance (α) and dominance of contrast (β) over structure (γ) provided the best correlation with listener tests. W3 shows an alternate weighting to

W2 keeping $\alpha = 0$ but switching the dominance to structure rather than contrast.

Altering the α, β and γ weightings resulted in the variance increasing for the TFS results (3(C)). However it also shifted the scale by reducing the error difference between unimpaired and the flat 10 loss. The ENV results clustered over a smaller range for the alternative W2 and W3 weightings which can be seen both vowels and fricatives. It is clear that the weighting are important and correlation of the results from this study with listener tests is required to find an optimal weighting balance for neurogram assessment.

4.3. Comparison of MSSIM to RMAE/RMSE

Fig.(4) compares the 3x3 unweighted MSSIM measure to RMAE and RMSE noting that for RMAE and RMSE the metric is 0 for the equality and increasing, i.e. the reverse to MSSIM. The error bars again show one standard deviation around the metric. As observed in prior work, RMAE has difficulties in accurately capturing the degradation occurring in some phonemes TFS behaviour (Hines and Harte, 2009). This caused a re-evaluation of the RMAE and RMSE error metrics for TFS comparisons. The RMAE metric has been expressed as a fraction of the normal unimpaired response's average power, presuming that with a degradation of the AN response, less information will be present and hence the impaired neurogram will be lower in power than the unimpaired neurogram. While this is true overall, examination of fine timing of vowels shows that the choice of error measure may cause unexpected results particularly at high presentation levels. The situation can arise where due to the phenomena of spread of synchrony (which generally occurs above 80 dB SPL), AN fibres start to show synchrony to other stimulus frequency components with fibres responding to stimulus at lower frequencies than their own characteristic frequency(CF) (Wong et al., 1998).

4.4. Effect of Hearing Loss on Neurograms

Figs.(5, 6) show sample ENV and TFS neurograms at P65 and P85 presentation levels for unimpaired and progressively impaired hearing losses. The fricative example, Fig.(5), illustrates that the intensity diminishes as the hearing loss increases: from a neurogram persepective, there is less information in the plot. The vowel example, Fig.(6), illustrates a different behaviour. The TFS neurogram for the unimpaired model shows a strong periodic response pattern in the low frequency range. It is information rich

with fine timing information and has speckled power gradient. The moderate loss neurogram shows similar periodic information in the lower frequencies but has lost much of the fine timing response in between. In the higher frequencies the low power information has been lost and the onset of synchrony spread is apparent. Finally for the profound loss, it can be seen that most of the lower frequency and fine timing data has been lost. Phase locking has occurred along with a spread of synchrony, with the phase locking to the formant frequency and erroneous power spreading across higher frequency bands. The MSSIM addresses this and captures the degradation in a bounded metric, with an range of -1 to +1, limiting phonemic group comparisons within a common range. The results in Fig.(4) demonstrate the wide variation in vowels for RMAE and RMSE, which occurs because the spread of synchrony is not as pronounced in every instance as it is in the illustrated case. The variation in MSSIM is much smaller as it appears to classify the profound losses with moderate or severe synchrony spread as a similarly poor result.

Examining the ENV examples illustrates that for fricatives the all three metrics capture the loss if activity within the progressively degrading neurograms at both P65 and P85 (Fig.(5)). At P65, the vowel degraded in a similar manner to the fricative. At P85, the spreading and phase locking has kept the ENV neurogram's average discharge rate up.

Fig.(7 & 8) show MSSIM results for all phoneme groups. A spider plot representation has been used to allow trends to be clearly seen. Each plot shows the MSSIM for the 6 phoneme groups with the different coloured rings depicting hearing loss (from blue flat 10 to red profound). The scale has been reversed, going from 1 in the axis centre out to 0 to allow for visual comparison to RMAE and RMSE. The RMAE and MSE results go from 0 and are unbounded, hence the scales have been set to display all results. The MSSIM performance was consistent across phoneme groups, presentation levels and neurogram resolution (ENV/TFS). For MSSIM, there is good delineation of each HL level. For P85, the ENV shows almost no difference between flat 10 and flat 20 for vowels and SV/glides. The problems highlighted in Fig.(4) are also illustrated in the spider plots where MAE displays vowel errors for TFS neurograms much larger than the errors in other phoneme groups. Vowels and SV/Glides RMAE displayed similar RMAE errors and this behaviour was compounded in the RMSE results.

4.5. Comparison to NAI

The NAI evaluates spectro-temporal outputs, looking at bands over time. It is a phenomenological metric based on empirical data, and like STI it uses band weightings and a redundancy factor across bands. In contrast MSSIM is a full-reference comparative metric, looking at the spectro-temporal information and does not rely on prior knowledge of which frequency bands are important to calculate speech intelligibility. The choice of component weighting, window size, and neurogram resolutions (i.e. number of CF bands tested; using ENV and TFS) are critical factors in configuring MSSIM for this application, but it does not introduce prior knowledge of the importance of one CF band over another for the intelligibility of a particular phoneme.

4.6. Limitations of MSSIM

While MSSIM is a more promising metric of phonemic degradation than either RMAE or RMSE, it is worth commenting on some of its limitations. Computationally, it is more expensive than RMAE. The full reference nature of the metric means that it will not handle even small timing mismatches, limiting its potential use utterances of the same word. Practically, this means it is not suitable for comparing different utterances of the same phoneme even by the same speaker. There is an alternate version, CW-SSIM (Wang and Simoncelli, 2005) that uses complex wavelets to handle offsets and roations in pictures, however this is significantly more computationally intensive and has not been tested in this study.

4.7. Towards a single AN fidelity metric

This study sought to investigate the suitability of an MSSIM based metric for quantifying SNHL degradations through neurogram comparisons. This was done for ENV and TFS neurograms and their effectiveness at distinguishing losses for progressively deteriorating audiograms was measured and evaluated for different phoneme groups. Ultimately, a single, weighted measure that can compare auditory nerve outputs yielding a single comparative metric is desirable.

Steeneken and Houtgast (2002) found that CF frequency weightings do not vary significantly for SNR or gender, but other studies found that the test speech material used resulted in different frequency weightings depending on whether the tests used nonsense words, phonetically balanced words or connected discourse. The results presented in this paper are measures at a phoneme group level. Fig. 9 shows the SII as calculated using various

nonsense syllable tests where most English phonemes occur equally often (as specified in Table B.2 (ANSI, 1997)). By equally weighting and combining the results by phoneme group into a single metric, the comparable plots for TFS and ENV neurogram can be seen in Fig.(10) for MSSIM and RMAE. The first two plots show the ENV, TFS followed by a combined ENV/TFS plot where the mean of the ENV and TFS value is plotted. Comparing the SII to the combined MSSIM, the main difference is the large drop from unimpaired to Flat 10.

It can also be seen that the higher presentation level has a lower SII score for mild hearing losses. This is caused by a phenomena known as the rollover effect (Jerger and Jerger, 1971; Studebaker et al., 1999) because over a range of increasing presentation levels the intelligibility score reaches a maximum and then declines as the level continues to increase. This characteristic appears to have been captured by MSSIM in the ENV neurogram but not by RMAE: Fig.(10) shows flat 10 with lower scores for P85 than P65 in MSSIM but not in RMAE.

5. CONCLUSIONS AND FUTURE WORK

As a metric for comparing TFS neurograms, MSSIM is more informative than RMAE or RMSE. The measure has fulfilled the original criteria set down for a useful metric. It has correctly predicted the order of hearing losses i.e. the metric deteriorates with increased hearing loss showing how different phoneme groups degrade with SNHL. Secondly it has low variance for a phoneme class, given a fixed presentation level and hearing loss. Thirdly, the established parameters for the window size make sense in terms of the physiological and signal processing boundaries on the system.

The choice of window size was significant in the ENV neurograms but the TFS results were not as sensitive to the size of window. A window size of up to 5 pixels was optimal for both neurograms. Further experimentation is required to establish whether alternative weightings will be beneficial for this application. The metric's boundedness and the results for TFS neurograms indicate it is a superior metric to simple RMAE or RMSE.

The use of MSSIM as an indicative score of intelligibility is promising, despite the absence of listener tests. The AN responses are taken from a model based on sound physiological data and the model has been demonstrated as capable of capturing a range of responses of hearing, both impaired and

unimpaired (Zilany and Bruce, 2006). Correlation of these results with listener tests is required to further demonstrate the ability of MSSIM to capture phonemic degradation. The goal is to take hearing aid design a step closer to removing the necessity for extensive listener tests in early stage algorithm design by substituting the use of a computational AN model and suitable speech intelligibility metric.

Appendix A. Full Result Set

Results for the analysis of MSSIM for vowels and fricatives were presented in the results section in Fig.(3). The overall results at the optimal window size for all phoneme groups where summarised in spider plots in Figs.(7 & 8).

Analysis of the performance of MSSIM for other phoneme groups are included here for completeness. Fig.(A.11) shows affricates and nasal phoneme groups; Fig.(A.12) shows stops and SV/glides phoneme groups.

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.

Bruce, I., Dinath, F., Zeyl., T. J., 2007a. 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), 73–81.

Bruce, I., Dinath, F., Zeyl., T. J., 2007b. 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), 73–81.

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.

DARPA, U. D. C., 1990. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. NIST Speech Disc 1-1.1.

Deng, L., Geisler, C. D., 1987. A composite auditory model for processing speech sounds. J. Acoust. Soc. Am. 82, 2001–2012.

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

Dinath, F., Bruce, I. C., 2008. Hearing aid gain prescriptions balance restoration of auditory nerve mean-rate and spike-timing representations of speech. Proceedings of 30th International IEEE Engineering in Medicine and Biology Conference, IEEE, Piscataway, NJ, 1793–1796.

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.

French, N. R., Steinberg, J. C., 1947. Factors governing the intelligibility of speech sounds. The Journal of the Acoustical Society of America 19 (1), 90–119.

Heinz, M., Swaminathan, J., 2009. Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech. JARO - Journal of the Association for Research in Otolaryngology 10 (3), 407–423, 10.1007/s10162-009-0169-8.

Hines, A., Harte, N., 2009. Error metrics for impaired auditory nerve responses of different phoneme groups. In: Interspeech. Brighton, pp. 1119–1122.

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

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.

Liberman, M., 1978. Auditory nerve response from cats raised in a low noise chamber. J. Acoust. Soc. Am. 63, 442–455.

Lopez-Poveda, E. A., Manuel, S. M., Dexter, R. F. I., 2005. Spectral processing by the peripheral auditory system: Facts and models. In: International Review of Neurobiology. Vol. Volume 70. Academic Press, pp. 7–48.

Lorenzi, C., Gilbert, G., H. Carn, a. S. G. B. M., 2006. Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. Proceedings of the National Academy of Sciences 103 (49), 18866–18869.

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., and Young, E. D., 2002. Biological basis of hearing-aid design. Annals of Biomedical Engineering, 30, 157-168.

Schijndel, N. H. v., Houtgast, T., Festen, J. M., 2001. Effects of degradation of intensity, time, or frequency content on speech intelligibility for normal-hearing and hearing-impaired listeners. The Journal of the Acoustical Society of America 110 (1), 529–542.

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

Steeneken, H. J. M., Houtgast, T., 1980. A physical method for measuring speech-transmission quality. The Journal of the Acoustical Society of America 67 (1), 318–326, sTI.

Steeneken, H. J. M., Houtgast, T., 2002. Phoneme-group specific octave-band weights in predicting speech intelligibility. Speech Communication 38 (3-4), 399–411.

Studebaker, G. A., Sherbecoe, R. L., McDaniel, D. M., Gwaltney, C. A., 1999. Monosyllabic word recognition at higher-than-normal speech and noise levels. The Journal of the Acoustical Society of America 105 (4), 2431–2444.

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.

Wang, Z., Simoncelli, E. P., 2005. Translation insensitive image similarity in complex wavelet domain. In: Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on. Vol. 2. pp. 573–576.

Wiener, F., Ross, D., 1946. The pressure distribution in the auditory canal in a progressive sound field. The Journal of the Acoustical Society of America 18 (2), 401–408.

Wong, J. C., Miller, R. L., Calhoun, B. M., Sachs, M. B., Young, E. D., 1998. Effects of high sound levels on responses to the vowel /[var epsilon]/ in cat auditory nerve. Hearing Research 123 (1-2), 61–77, doi: DOI: 10.1016/S0378-5955(98)00098-7.

Xu, L., Pfingst, B., 2003. Relative importance of temporal envelope and fine structure in lexical-tone perception (l). The Journal of the Acoustical Society of America 114 (6), 3024–3027.

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., Bruce, I., July 2007. Representation of the vowel /E/ in normal and impaired auditory nerve fibers: Model predictions of responses in cats. J. Acoust. Soc. Am. 122 (1), 402–417.

Zilany, M. S. A., 2007. Modeling the neural representation of speech in normal hearing and hearing impaired listeners. PhD Thesis, McMaster University, Hamilton, ON.

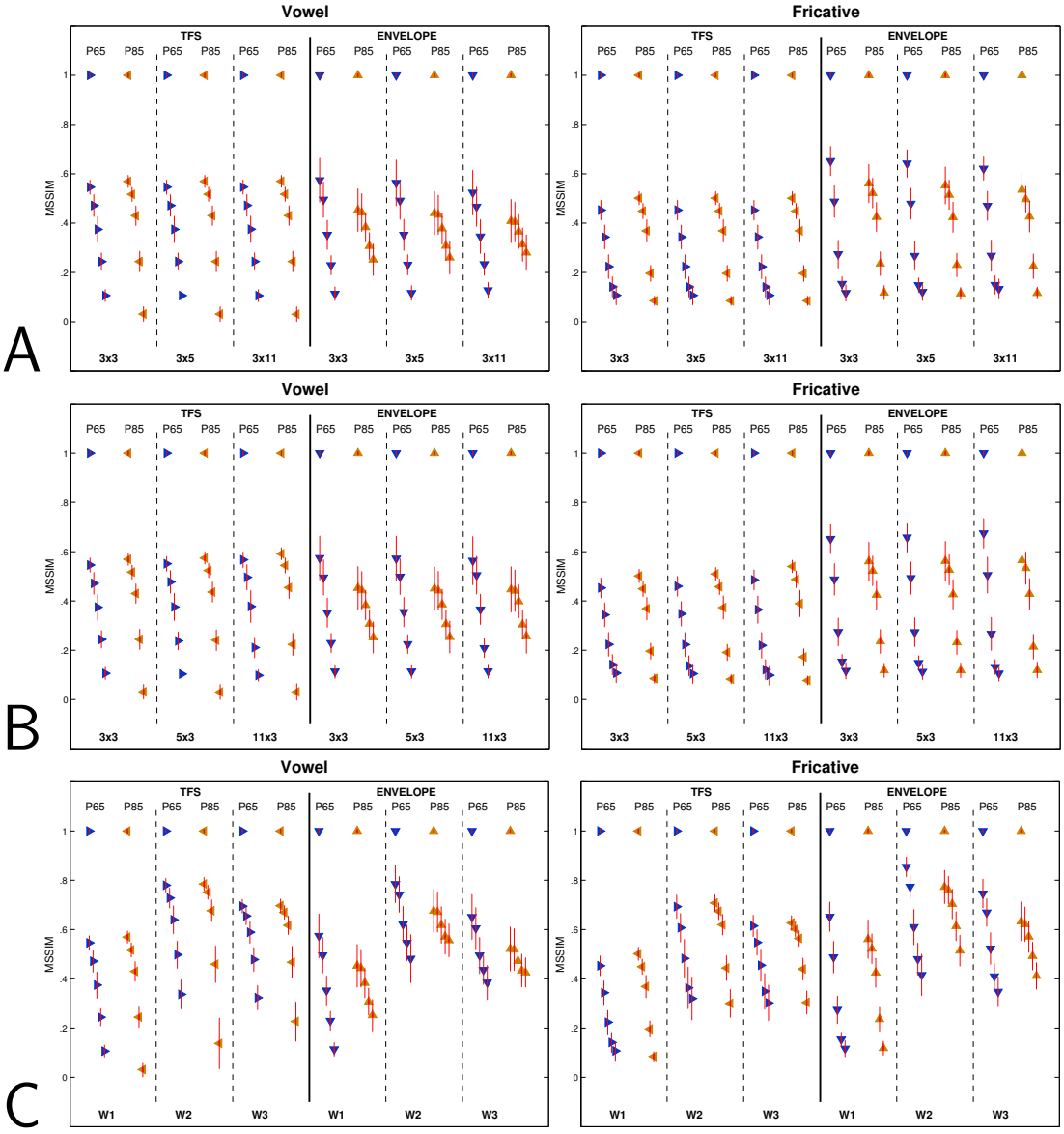

Figure 3: *Left: Vowels; Right: Fricatives. Data points represent hearing loss levels compared to unimpaired, beginning from MSSIM of 1 for comparison with unimpaired and progressing through FLAT10, FLAT20, MILD, MODERATE and PROFOUND. Top Row (A): varying MSSIM window in time; Middle Row (B): varying MSSIM window in CF; Bottom Row (C): Varying MSSIM weighting* $(\alpha, \beta, \gamma) W1 = (1, 1, 1) W2 = (0, 0.8, 0.2) W3 = (0, 0.2, 0.8)$*, window size fixed at 3x3;*

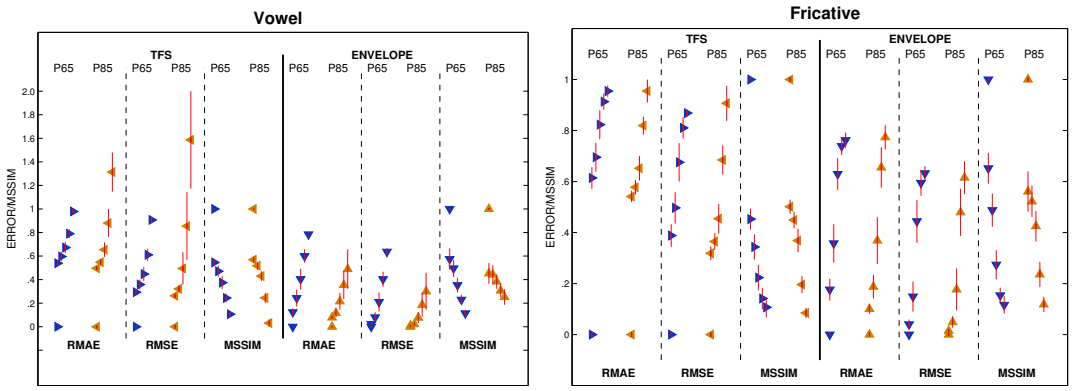

Figure 4: *Left: Vowels; Right: Fricatives. Comparison of MSSIM with RMAE and RMSE (which are error levels and hence have a 0 data point for unimpaired and increase with hearing loss, i.e. read MSSIM top to bottom and RMAE/RMSE bottom to top.)*

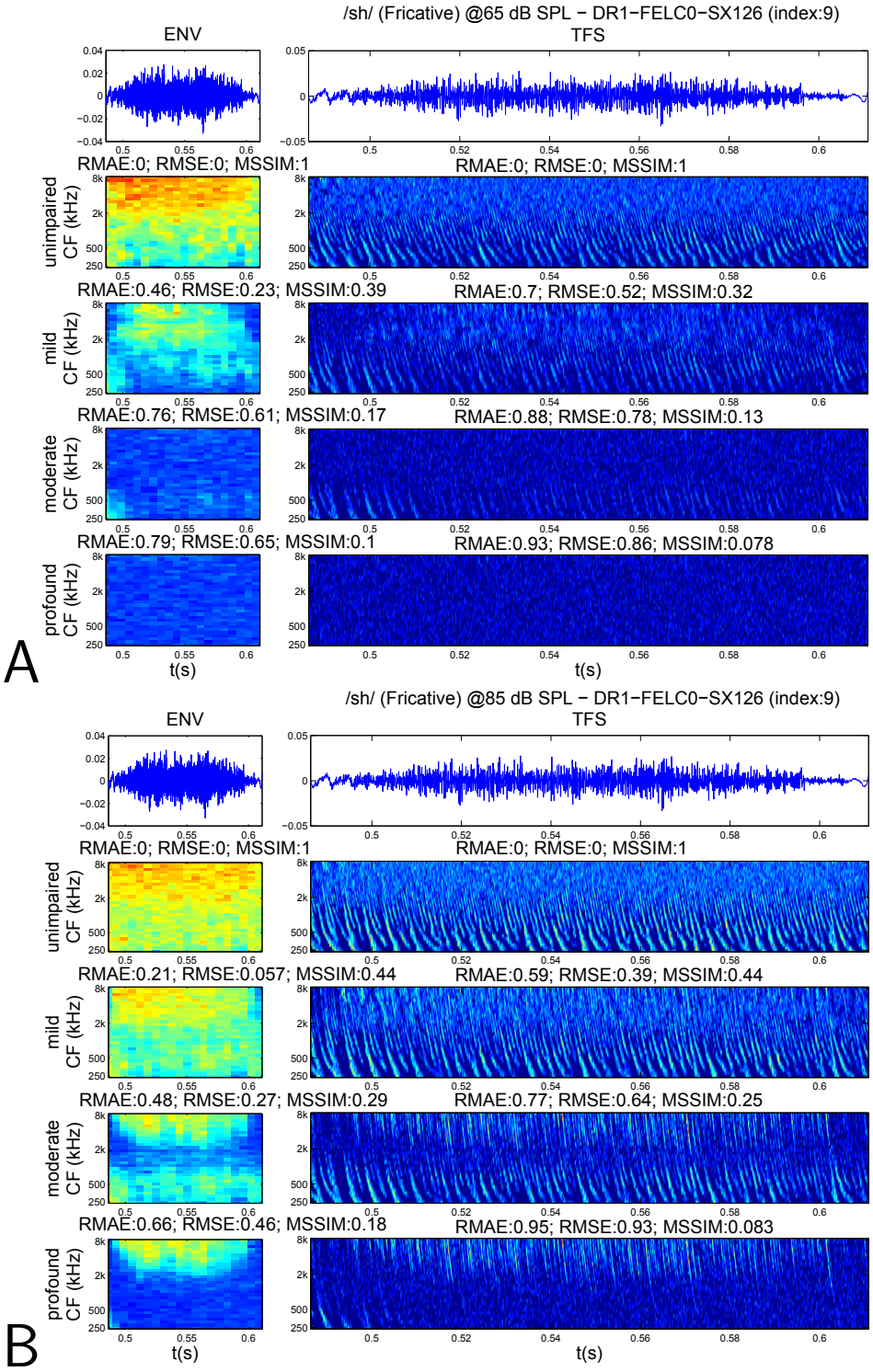

Figure 5: Sample ENV (left) and TFS (right) neurograms for fricative /sh/ with progressively degrading hearing loss. Presentation Level 65 dB SPL in (A) and 85 dB SPL in (B). For reference purposes, the top rows in (A) and (B) show the signal, with the time axis shown at a greater resolution in the TFS compared to the ENV. The next row displays the neurograms from an model with unimpaired hearing. The bottom three rows are progressively impaired hearing loss neurograms. It can be seen that the amount of information in contained in the neurogram diminishes rapidly with hearing loss in (A), as would be expected by examining the audiogram thresholds Fig.(1) for the tested hearing losses at

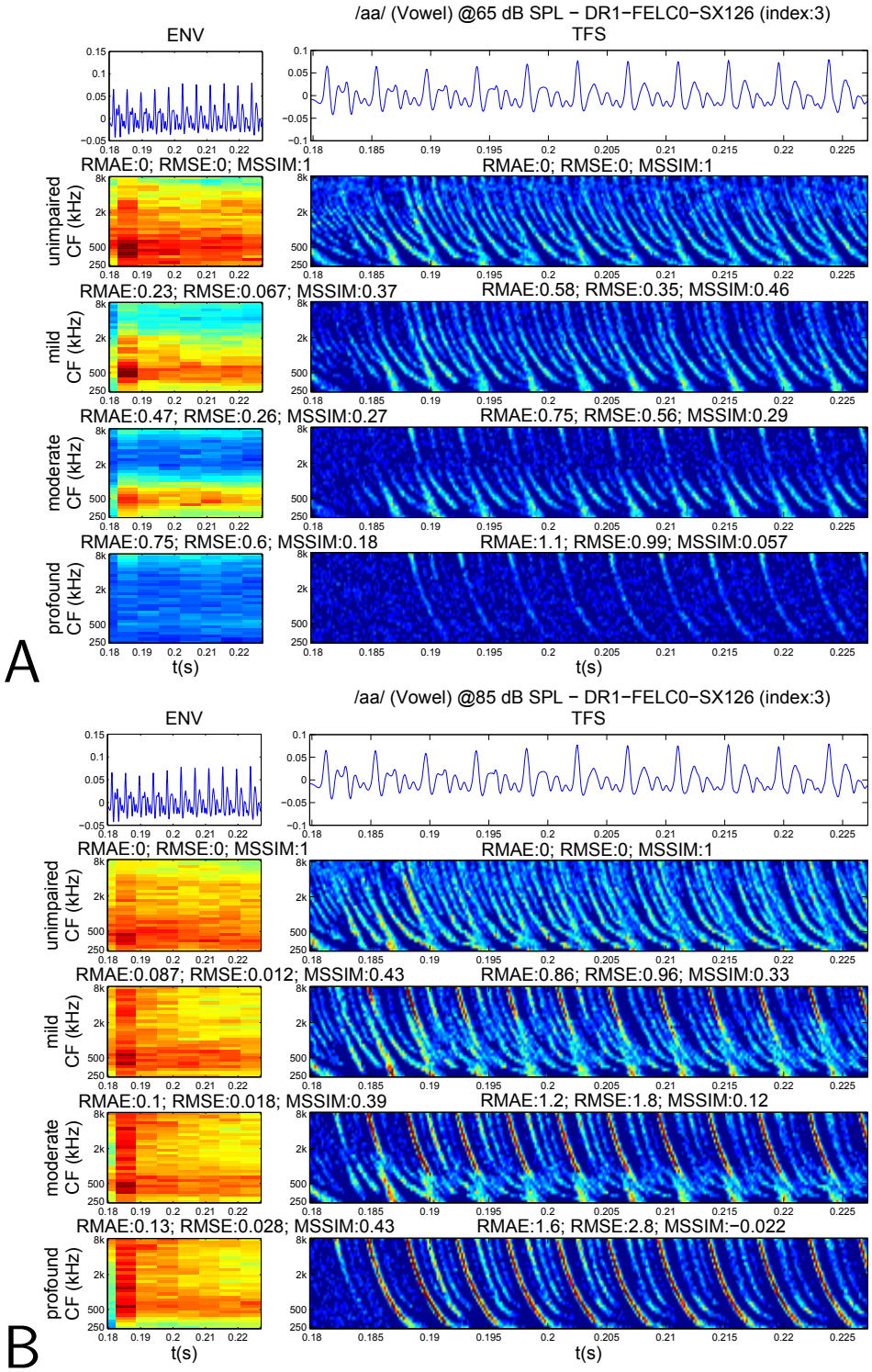

Figure 6: Corresponding samples to Fig.(5) ENV and TFS neurograms for a vowel (/aa/) with progressively degrading hearing loss. The TFS neurograms in (A) show that at lower presentation levels the vowel degrades with progressive hearing loss of fine timing information. In (B), it can be seen that at 85 dB SPL not only is information being lost, phase locking and a spread of synchrony across CF bands is causing the addition of erroneous information with progressive hearing loss.

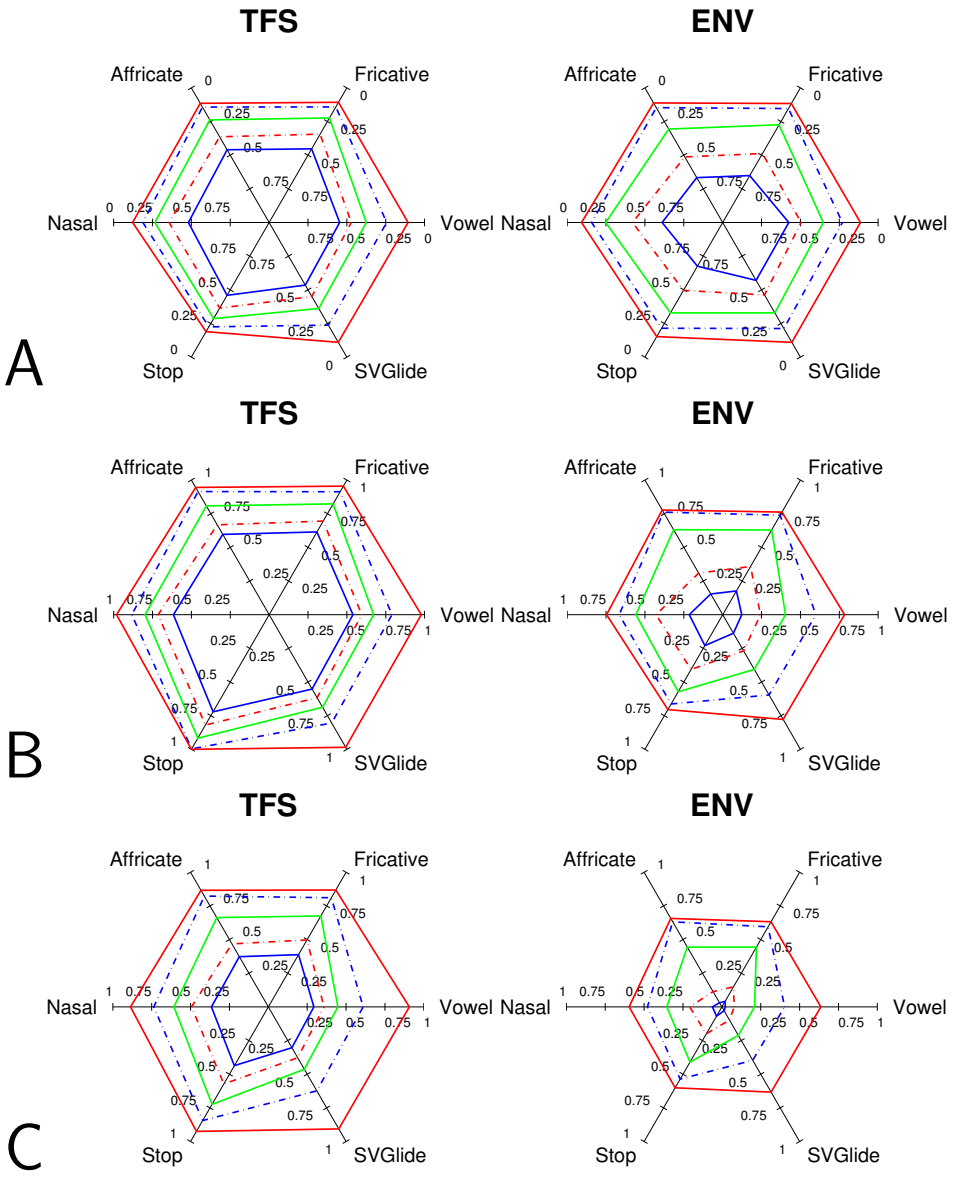

Figure 7: Results for all phoneme groups at 65 dB SPL. Coloured lines represent audiograms (blue to red: flat 10 to profound). (A): MSSIM. Scaled inverted (1 to 0) to allow trend comparison with RMAE and RMSE; (B): Mean Absolute Error (RMAE); (C): Mean Squared Error (RMSE)

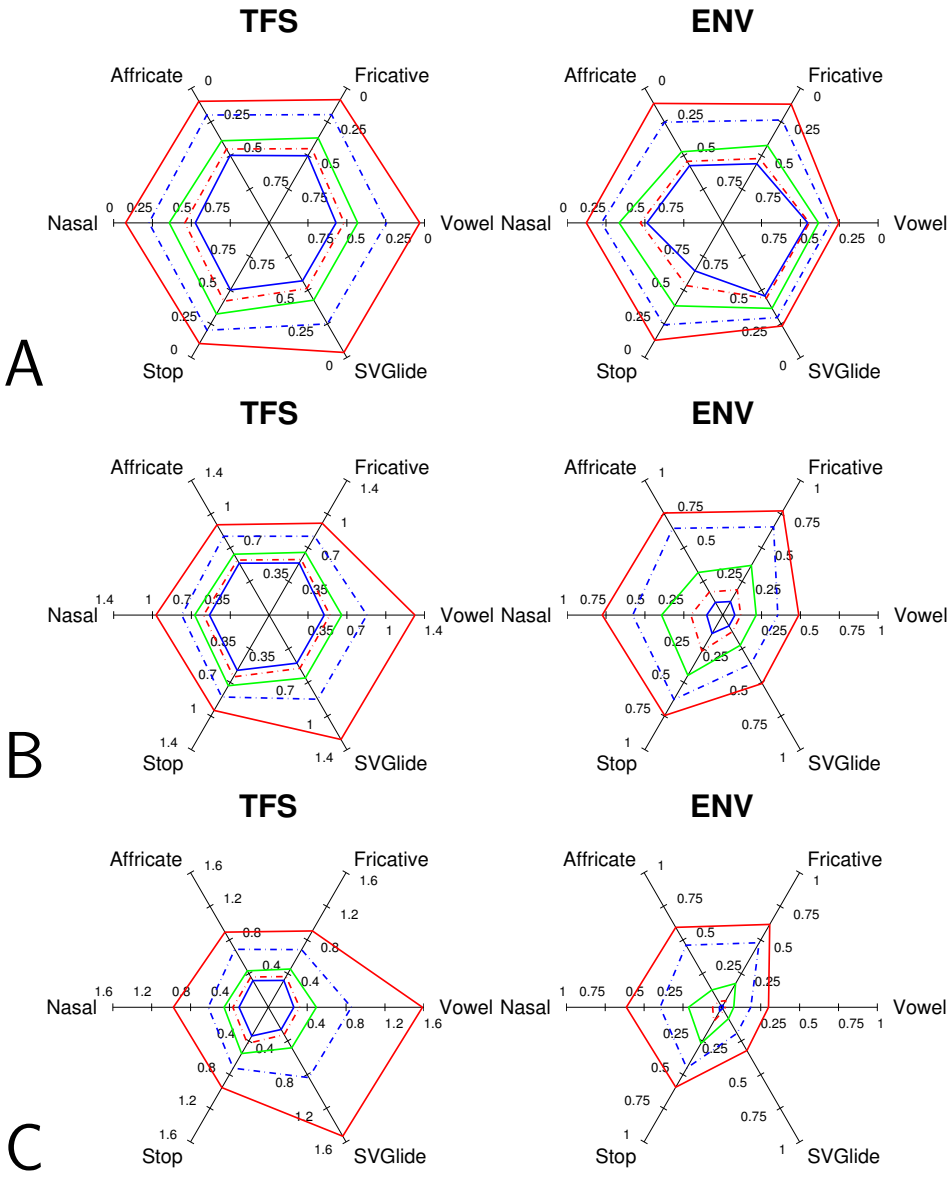

Figure 8: Results for all phoneme groups at 85 dB SPL. Coloured lines represent audiograms (blue to red: flat 10 to profound). (A): MSSIM. Scaled inverted (1 to 0) to allow trend comparison with RMAE and RMSE; (B): Relative Mean Absolute Error (RMAE). Range > 1 for Vowel TFS at P85 ; (C): Relative Mean Squared Error (RMSE). Range $>$ 1.5 for Vowel TFS at P85

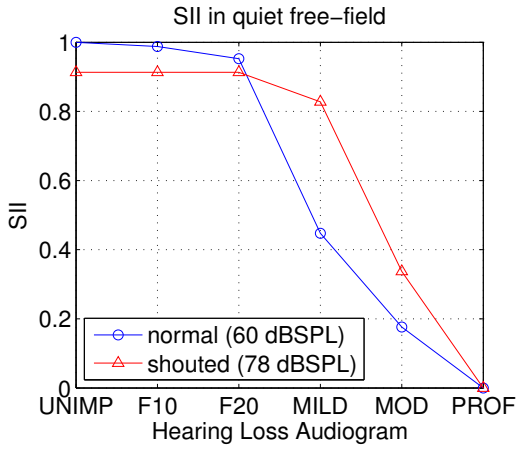

Figure 9: SII as calculated using various nonsense syllable tests where most English phonemes occur equally often (as specified in Table B.2 (ANSI, 1997))

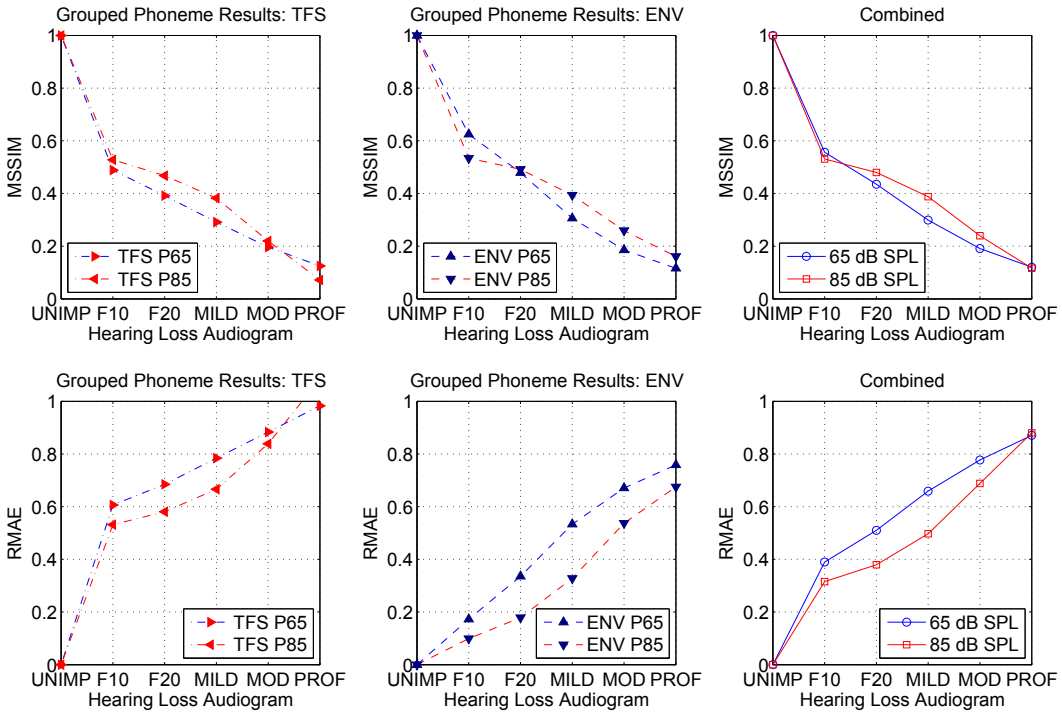

Figure 10: Above: MSSIM and below: RMAE. Mean TFS, ENV, and combined metrics for all phoneme groups, equally weighted

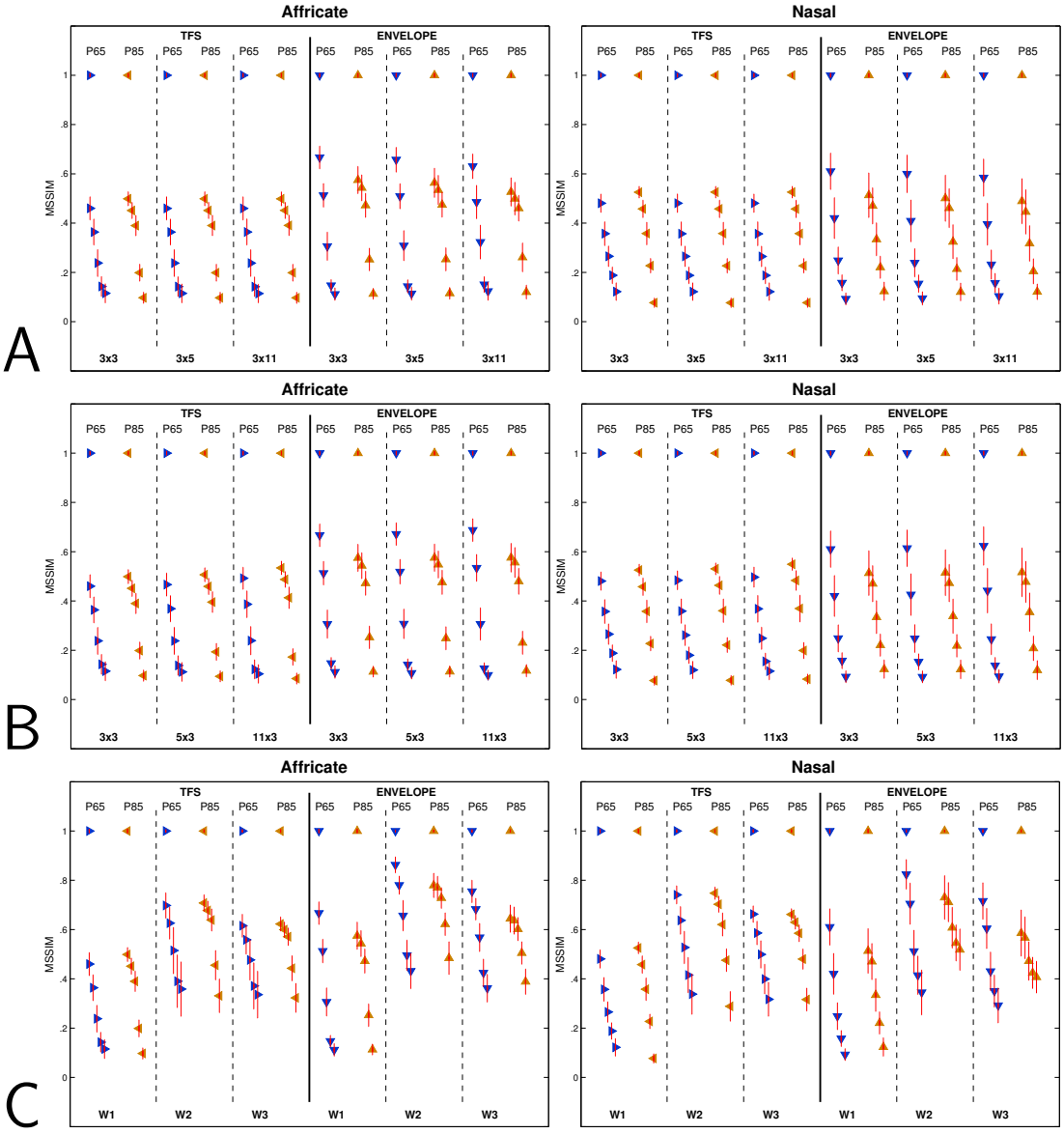

Figure A.11: *Left: Affricate; Right: Nasal. Data points represent hearing loss levels compared to unimpaired, beginning from MSSIM of 1 for comparison with unimpaired and progressing through FLAT10, FLAT20, MILD, MODERATE and PROFOUND. Top Row (A): varying MSSIM window in time; Middle Row (B): varying MSSIM window in CF; Bottom Row (C): Varying MSSIM weighting* $(\alpha, \beta, \gamma) W1 = (1, 1, 1) W2 = (0, 0.8, 0.2) W3 = (0, 0.2, 0.8)$*, window size fixed at 3x3;*

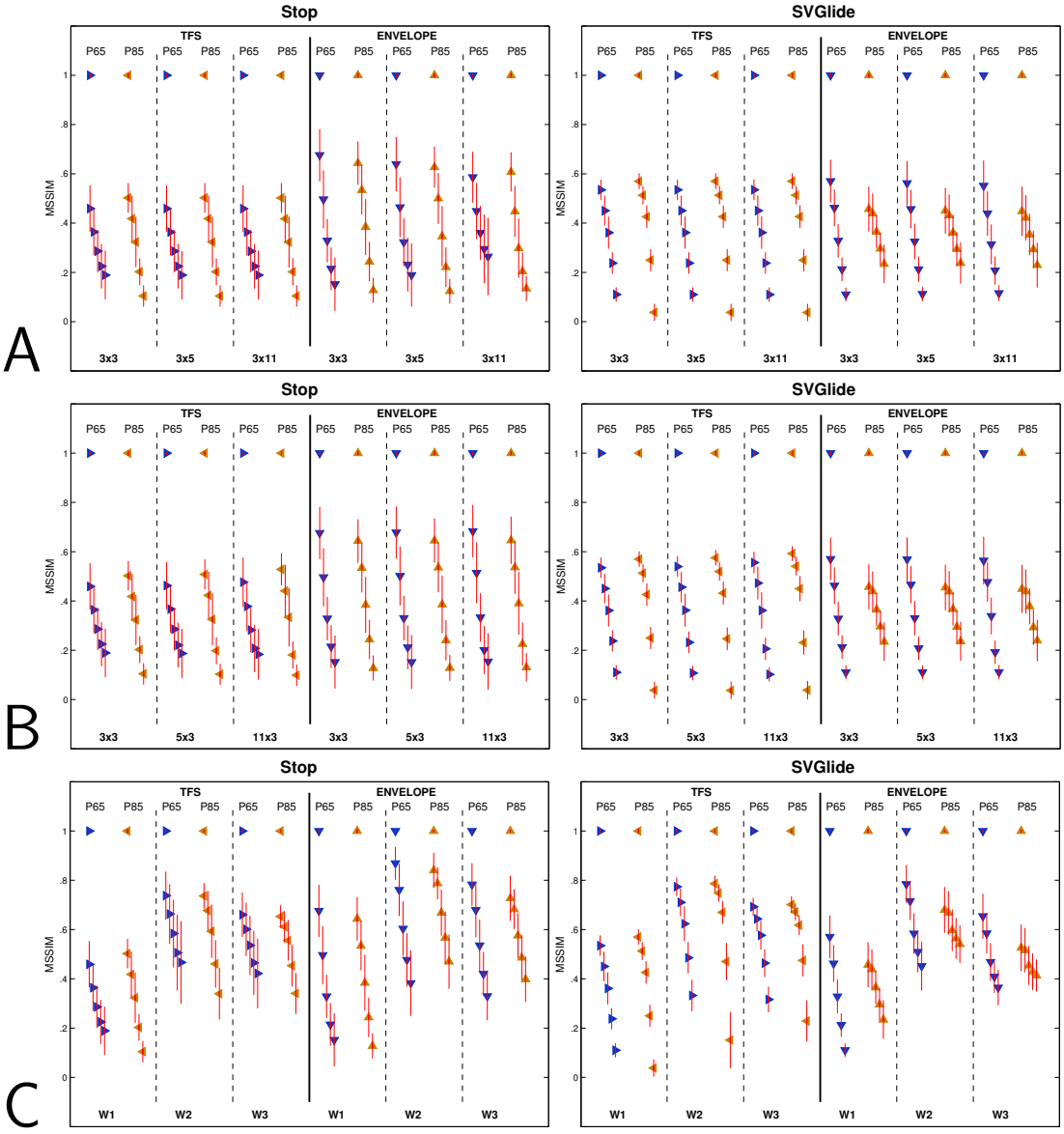

Figure A.12: *Left: Stop; Right: SV/Glide. Data points represent hearing loss levels compared to unimpaired, beginning from MSSIM of 1 for comparison with unimpaired and progressing through FLAT10, FLAT20, MILD, MODERATE and PROFOUND. Top Row (A): varying MSSIM window in time; Middle Row (B): varying MSSIM window in CF; Bottom Row (C): Varying MSSIM weighting* $(\alpha, \beta, \gamma) W1 = (1, 1, 1) W2 = (0, 0.8, 0.2) W3 = (0, 0.2, 0.8)$*, window size fixed at 3x3;*