EVALUATING SENSORINEURAL HEARING LOSS WITH AN AUDITORY NERVE MODEL USING A MEAN STRUCTURAL SIMILARITY MEASURE

Andrew Hines, Naomi Harte

Department of Electronic & Electrical Engineering
Sigmedia Group
Trinity College Dublin, Ireland
email: hinesa@tcd.ie

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. This work seeks to create an objective measure to automate this inspection process and ranks hearing losses based on auditory-nerve discharge patterns. A systematic way of assessing phonemic degradation using the outputs of an auditory nerve model for a range of sensorineural hearing losses would aid in rapid prototyping development of speech-processing algorithms for digital hearing aids. 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. The mean structural 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.

1. INTRODUCTION

This work examines a systematic way of assessing phonemic degradation using the outputs of an auditory nerve (AN) model for a range of sensorineural hearing losses (SNHL). The practical application of this is to allow speech-processing algorithms for hearing aids to be objectively tested in development without human trials. The model used in this study was the cat AN model of Zilany and Bruce [1]. It produces simulated auditory nerve neural spike train outputs at specific characteristic frequencies (CF). The levels of degradation in output due to a 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 power 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. [2]

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.[3] showed that if auditory-nerve discharge patterns in response to sounds as complex as speech can be accurately modelled and predicted, 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 model outputs for progressive hearing losses are displayed in Fig. (1). 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 [4] 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 at higher presentation levels. This paper explores an alternative mean structural similarity measure (MSSIM)[5] and uses it to compare neurograms produced for utterances over a range of SNHL. It was chosen because unlike RMAE, it examines the similarity over patches or windows, capturing changes in features rather than individual point differences. 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 [6]. Section 2 introduces the types of neurograms that were analysed and the MSSIM. Section 3 describes how the measure was assessed. Section 4 presents and discusses important features of the results, with conclusions and future work in Section 5.

2. BACKGROUND

2.1 Auditory Nerve Models

The Zilany & Bruce AN model used in this study[7] was designed with an ultimate goal of predicting human speech recognition performance for both normal hearing and hearing impaired listeners [8]. It builds upon several efforts to develop computational models including [9], [10] and [11]. It matches physiological data over a wider dynamic range than previous auditory models. 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 [12] and optimal phonemic compression schemes [13].

2.2 Neurograms

This study evaluated the effect of SNHL using two types of neurograms- temporal fine structure (TFS) and average discharge or temporal envelope (ENV). Both display the neural response as a function of CF and time. Rosen [2] 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. [14] 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 [15] looked at Mandarin Chinese

 ISSN 2076-1465

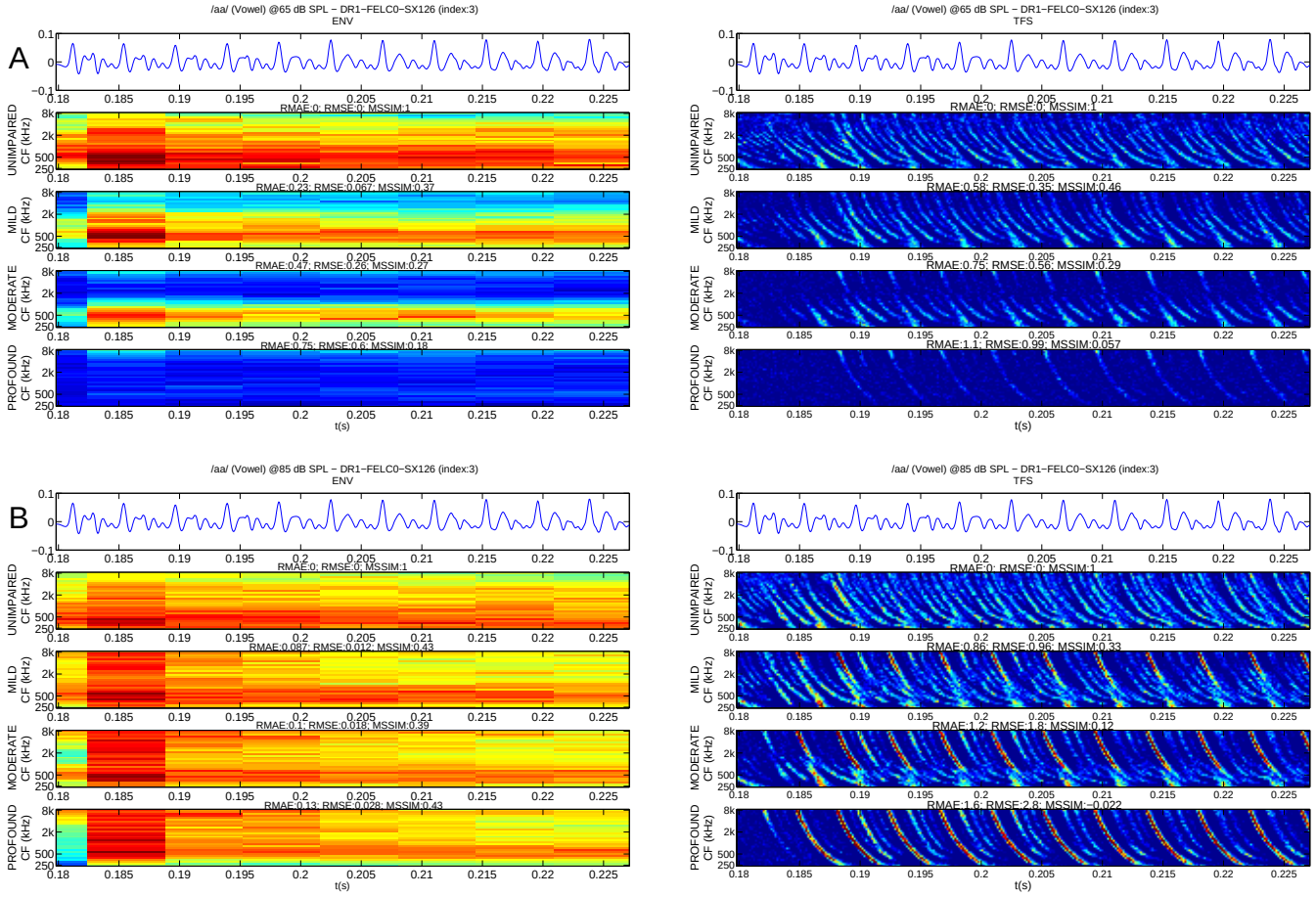

Figure 1: *Sample ENV (left) and TFS (right) neurograms for vowel (/aa/) with progressively degrading hearing loss. Presentation Level 65 dB SPL in (A) and 85 dB SPL in (B). For reference purposes, the top rows in (A) and (B) show the signal, with the time axis shown at a greater resolution in the TFS compared to the ENV. The next row displays the neurograms from a model with unimpaired hearing. The bottom three rows are progressively impaired hearing loss neurograms. The TFS neurograms in (A) show that at lower presentation levels the vowel degrades with progressive 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.*

monosyllables and found that in the majority of trials, identification was based on TFS rather than ENV. Lorenzi et al. [16] 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. [17] compared the amplification schemes of NAL-R and DSL to find single-band gain adjustments to optimize the mean firing rates. They found that in general the optimal lay in the order of +10dB above the prescribed linear gains 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 [18]. It is apparent there is value in analysing both neurograms in optimising hearing aids to restore speech intelligibility to those with SNHL, even though the underlying physiological bases have not been established from a perceptual perspective.

2.3 Mean Structural Similarity Index (MSSIM)

The relative mean absolute error (RMAE) metric was used in [4] to compare neurograms from phonemes presented to unimpaired and impaired ANs. This measure normalised the absolute difference between the unimpaired and impaired representations (x & y) relative to the mean unimpaired representation. The structural similarity index (SSIM) [5] between x and y is defined as a comparison of the original and degraded signal constructed as a function of luminance (l), contrast (c) and structure (s):

$$SSIM(x,y) = f(l(x,y),c(x,y),s(x,y)) \quad (1)$$

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 component also contains constant values (C_n) to avoid instabilities and is weighted with a coefficient > 0 (α, β *and* γ) which can be used to adjust the relative importance of the component. Thus (1) can be expressed as (2). The MSSIM metric has properties similar to RMAE or relative mean squared error (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 property of boundedness $-1 < S(x,y) \leq 1$. See [5] for a full description.

$$SSIM(x,y) = \left(\frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}\right)^{\alpha} \cdot \left(\frac{2\sigma_{xy} + C_2}{\mu_x^2 + \sigma_y^2 + C_2}\right)^{\beta} \cdot \left(\frac{(2\sigma_{xy} + C_3)}{\mu_x^2 + \sigma_y^2 + C_3}\right)^{\gamma} \quad (2)$$

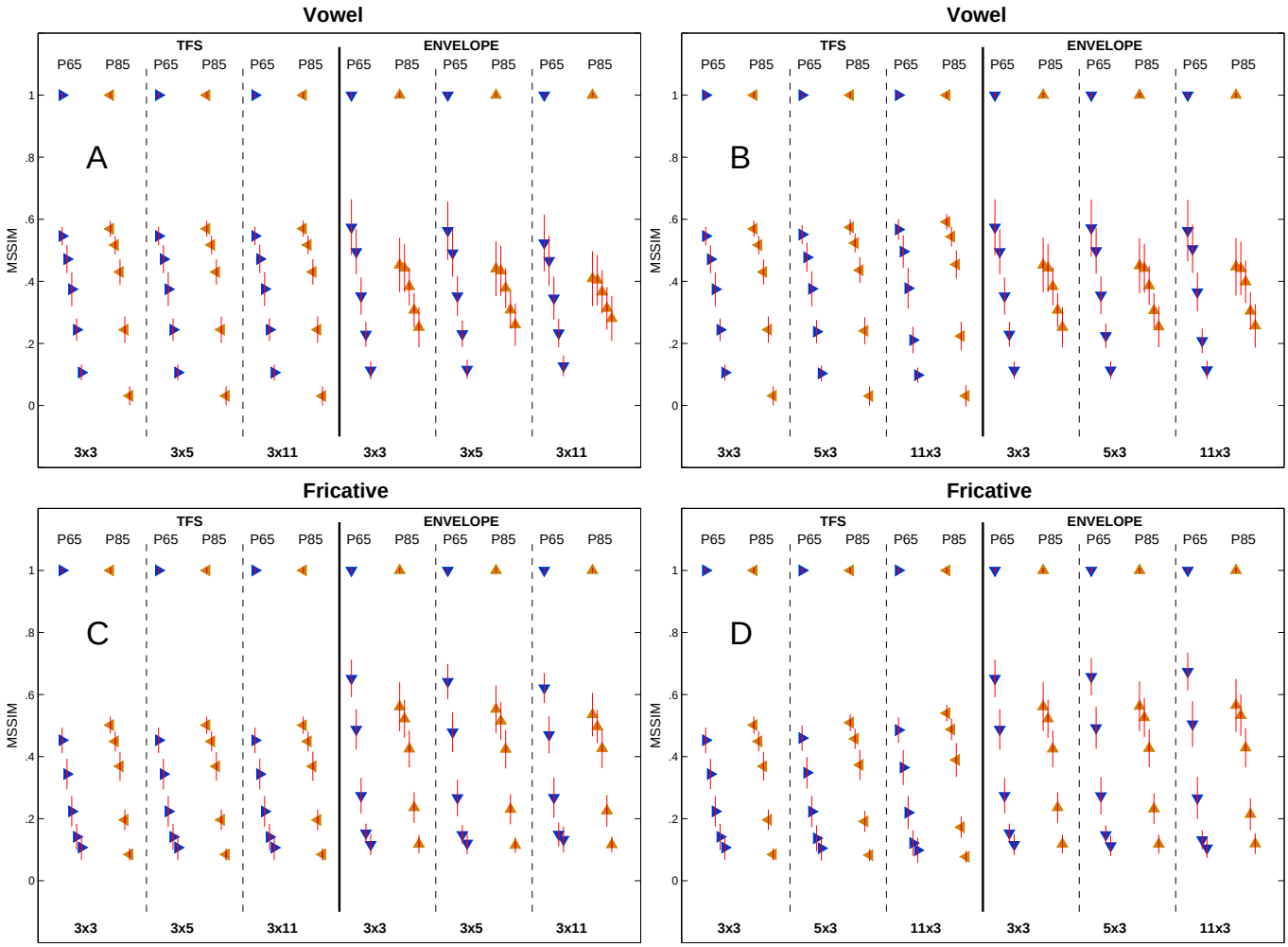

Figure 3: *Data points represent hearing loss levels compared to unimpaired, beginning from MSSIM of 1 for comparison with unimpaired and progressing through FLAT10, FLAT20, MILD, MODERATE and PROFOUND. Along the bottom of each graph 3x3 etc indicate the window size in pixels (as illustrated in Fig.(2)). Each window size is tested for two presentation levels: P65 and P85. Top row vowels, bottom row fricatives. (A+C): varying MSSIM window in time (3 by 3,5, and 11 pixels); (B+D): varying window size across CF bands (3,5 and 11 by 3).*

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". To evaluate the choice of window size and weightings that best suit the proposed application, the following criteria were defined. It should predict correctly 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.)

3. METHOD

Following the methodology in [4], the core TIMIT corpus [19] comprising of 192 sentences and 6854 phoneme utterances was used as the speech waveform source. Each 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). As in [17], an outer ear filter with gains as per [20] was used to pre-filter the speech waveforms to mimic the amplification that occurs prior to the middle and inner ear. Each sentence was then presented at the two presentation levels to the AN Model for an unimpaired hearing profile and hearing losses with flat 10dB, flat 20dB, mild, moderate and profound.

The audiograms used match the samples presented by Dillon[21] to illustrate prescription fitting over a wide range of hearing impairments (Fig.(4)). 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 ability to discriminate between unimpaired and very mild losses in hearing thresholds.

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 [22] and as used for similar AN Model simulations [17], 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

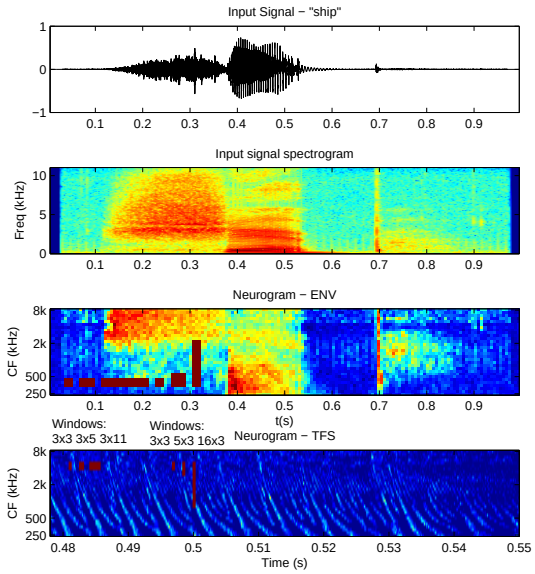

Figure 2: *Illustrative view of window sizes. The input signal and spectrogram are shown for the word "ship" in the top two rows. The ENV neurogram shows the 3 time varied window sizes followed by the 3 frequency varied window sizes. The bottom row shows the TFS neurogram with the same window sizes illustrated. Note that time scale in TFS neurogram is changed (zoomed in on the /i/ vowel)*

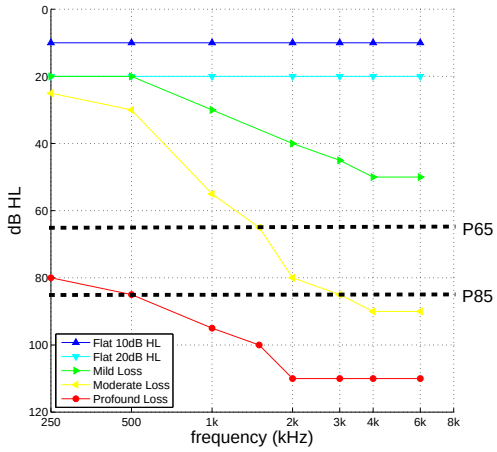

Figure 4: *Hearing Loss Audiograms tested- Flat 10, Flat 20, mild, moderate and profound. To allow comparison with audiogram thresholds the presentation levels of input signals and indicated with dashed lines P65 and P85. Hence the profound loss will cause most information above 500Hz to be sub-threshold.*

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 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 [5] was used varying the window sizing parameter.

Each neurogram was a standard height of 30 pixels (one per CF band) and varied in width with the duration of the phoneme. 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 (stop, affricate, fricative, nasal, SV/glide, vowel) 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).

4. RESULTS & DISCUSSION

The data in Fig.(3) shows results from a subset of the full suite of tests. Figs.(3A) and (3B) present the MSSIM for vowels for both the TFS and ENV neurograms with corresponding fricatives results in Figs.(3C) and (3D). Each shows 3 different NxM windows where N is frequency and M time resolution. For each window size, the MSSIM for both P65(►) and P85(◄) can be seen progressively deteriorating for the Flat10, Flat20, mild, moderate and profound losses. The error bars show one standard devation around the metric.

Fig.(3A) shows the results for progressively longer time samples in the MSSIM window. The TFS is relatively insensitive to increases in the time window. However, the ability to differentiate between SNHL levels reduced in the 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 also limited 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 but the variance starts to increase significantly for larger windows, e.g. 16x3 as shown in (B). The ENV results became more clustered for this larger window size. Space limits the ability to present the results for all phoneme groups. A detailed examination of plots from the other phoneme groups revealed broadly similar behaviour. 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. Fig.3(C) and 3(D) show that the results for fricatives are significantly less sensitive to changes in window size. This is likely due to fricatives having less defined structure than vowels (e.g no formants).

Fig.(5) compares the relative mean absolute error (RMAE) and relative mean squared errors (RMSE) with the 3x3 window MSSIM metric. Again the error bars represent one standard deviation. Note that for RMAE and RMSE the metric is 0 for the equality and increasing, i.e. the reverse to MSSIM. The difficulties with the RMAE metric that were previous reported in [4] and which are illustrated in Fig.(1) are apparent in the TFS behaviour and the MSSIM addresses this. Both measures have less spread at P85 and the use of the MSSIM for ENV is not as beneficial here for vowels although it performed well for other phoneme groups.

5. CONCLUSIONS AND FUTURE WORK

Overall, MSSIM has been demonstrated to have good potential as a metric to quantitatively compare AN outputs via neurograms. A physiological basis of hearing aid desing relies on the restoration of AN outputs of impaired nerves as closely as possible to that of an unimpaired nerve. The ability to quantify this similarity is an important stage of this process. This work shows that as a metric for comparing TFS neurograms, MSSIM is more informative than

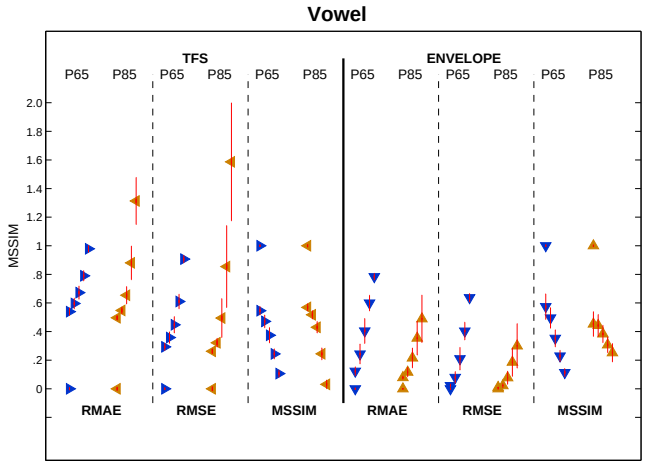

Figure 5: *Comparison of MSSIM with RMAE and RMSE (which are error levels and hence has a 0 data point for unimpaired and increase with hearing loss.)*

RMAE or RMSE. 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. The metric's boundedness and the results for TFS neurograms indicate it is a superior metric to simple RMAE or RMSE. Work is ongoing to correlate these results with data from sensorineural hearing loss listener tests. Ultimately, these results will aim to justify replacing early stage clinical trials with simulated trials.

REFERENCES

[1] M.S.A. Zilany and I.C. Bruce, "Representation of the vowel /E/ in normal and impaired auditory nerve fibers: Model predictions of responses in cats," *J. Acoust. Soc. Am.*, vol. 122, no. 1, pp. 402–417, July 2007.

[2] S. Rosen, "Temporal information in speech: Acoustic, auditory and linguistic aspects," *Philosophical Transactions: Biological Sciences*, vol. 336, no. 1278, pp. 367–373, 1992.

[3] M. B. Sachs, I. C. Bruce, R. L. Miller, and E. D. Young, "Biological basis of hearing-aid design," *Annals of Biomedical Engineering*, vol. 30, pp. 157168, 2002.

[4] A. Hines and N. Harte, "Error metrics for impaired auditory nerve responses of different phoneme groups," in *Interspeech 09*, Brighton, England, 2009, pp. 1119–1122.

[5] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004.

[6] S. Kandadai, J. Hardin, and C.D. Creusere, "Audio quality assessment using the mean structural similarity measure," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 221–224.

[7] M.S.A. Zilany and I.C. Bruce, "Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery," *J. Acoust. Soc. Am.*, vol. 120, no. 3, pp. 1446–1466, Sept 2006.

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

[9] L. Deng and C. D. Geisler, "A composite auditory model for processing speech sounds," *J. Acoust. Soc. Am.*, vol. 82, pp. 2001–2012, 1987.

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

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

[12] F. Dinath and I. C. Bruce, "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*, pp. 1793–1796, 2008.

[13] I.C. Bruce, F. Dinath, and T. J. Zeyl., "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)*, pp. 73–81, 2007.

[14] Z.M. Smith, B. Delgutte, and A.J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, no. 6876, pp. 87–90, 2002, 10.1038/416087a.

[15] L. Xu and B.E. Pfingst, "Relative importance of temporal envelope and fine structure in lexical-tone perception (l)," *The Journal of the Acoustical Society of America*, vol. 114, no. 6, pp. 3024–3027, 2003.

[16] C. Lorenzi, G. Gilbert, and and S. Garnierand B.C.J. Moore H. Carn, "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," *Proceedings of the National Academy of Sciences*, vol. 103, no. 49, pp. 18866–18869, 2006.

[17] I.C. Bruce, F. Dinath, and T. J. Zeyl., "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)*, pp. 73–81, 2007.

[18] M. Heinz and J. Swaminathan, "Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech," *JARO - Journal of the Association for Research in Otolaryngology*, vol. 10, no. 3, pp. 407–423, 2009, 10.1007/s10162-009-0169-8.

[19] U.S. Dept. Commerce DARPA, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus," *NIST Speech Disc 1-1.1*, 1990.

[20] F.M. Wiener and D.A. Ross, "The pressure distribution in the auditory canal in a progressive sound field," *The Journal of the Acoustical Society of America*, vol. 18, no. 2, pp. 401–408, 1946.

[21] H. Dillon, "Hearing Aids," *New York: Thieme Medical Publishers*, 2001.

[22] M.C. Liberman, "Auditory nerve response from cats raised in a low noise chamber," *J. Acoust. Soc. Am.*, vol. 63, pp. 442–455, 1978.