2009

Error Metrics for Impaired Auditory Nerve Responses of Different Phoneme Groups

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

Naomi Harte
University of Dublin, Trinity College

Recommended Citation

Hines, A. & Harte, N. (2009). Error Metrics for Impaired Auditory Nerve Responses of Different Phoneme Groups. *INTERSPEECH*, Brighton, United Kingdom, 6-10 September. doi:10.21427/m56b-yx68

Error metrics for impaired auditory nerve responses of different phoneme groups

Andrew Hines[1]*, Naomi Harte*

Department of Electronic & Electrical Engineering
Sigmedia Group
Trinity College Dublin
Ireland
hinesa@tcd.ie[1]

Abstract

An auditory nerve model allows faster investigation of new signal processing algorithms for hearing aids. This paper presents a study of the degradation of auditory nerve (AN) responses at a phonetic level for a range of sensorineural hearing losses and flat audiograms. The AN model of Zilany & Bruce was used to compute responses to a diverse set of phoneme rich sentences from the TIMIT database. The characteristics of both the average discharge rate and spike timing of the responses are discussed. The experiments demonstrate that a mean absolute error metric provides a useful measure of average discharge rates but a more complex measure is required to capture spike timing response errors.

Index Terms: auditory periphery model, hearing aids, sensorineural hearing loss, phonemic degradation

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 [1]. 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[2].

The goal of this study was to further analyse the degradation of AN responses at a phoneme level for a range of sensorineural hearing losses, by using the neural representations of speech provided by the model rather than perceptual feedback. By presenting a phonetically rich selection of sentences to the AN model, the differences between an unimpaired ear model and three progressively impaired ear models were examined. Prior work [3] examined 3 progressively impaired models over 3 presentation levels. Analysis of the results pointed towards further research to examine the choice of metric. An examination of additional audiograms was undertaken to further investigate saturation point boundary conditions.

This analysis serves a number of objectives, primarily by providing a better understanding of the rate of phonemic presentation degradation the the auditory nerve based on hearing loss profile and presentation level variables. By quantifying the degradation with a suitable measure of phoneme fidelity, it may provide the basis for design of new hearing aid algorithms based on optimal phonemic response restoration.

Section II introduces the chosen computational model and hearing loss profiles to be examined. Section III presents the methodology employed in gathering the results. Section IV presents and analyses the results. Further work is then proposed based on the results presented.

2. Background

2.1. Model

This study used the cat auditory nerve (AN) model developed and validated against physiological data by Zilany and Bruce [2]. The ultimate goal of the model is to predict human speech recognition performance for both normal hearing and hearing impaired listeners [4]. It has recently been used to conduct studies into hearing aid gain prescriptions [5] and optimal phonemic compression schemes[6].

The Zilany and Bruce AN model builds upon several efforts to develop computational models including Deng and Geisler [7], Zhang et al.[8] and Bruce et al.[9]. A schematic diagram of the model is available in Fig. 1 of Zilany and Bruce [2], 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 AN model takes speech waveforms, resampled at 100kHz with instantaneous pressures in units of Pascal. These are used to derive an AN spike train for a fibre with a specific characteristic frequency (CF). Running the model at a range of CFs 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.

The current version of the AN model was enhanced to address higher presentation levels[4]. The model can now handle the shift in best frequency (BF) that occurs and is partially responsible for the loss of synchrony capture at high levels.

2.2. Audiograms

Four audiograms representing hearing loss profiles (Fig. 1) were selected to represent a mild, moderate, severe and profound hearing loss. These were the same as those used in prior work[3] with the addition of a severe loss. In addition, flat audiograms at 10, 20, 85 and 120 dB HL, i.e. with a constant dB loss across all frequency ranges, were examined to help understand the extremities of the error ranges. These audiograms are denoted FLAT10, FLAT20, FLAT85 and FLAT120 in the results section.

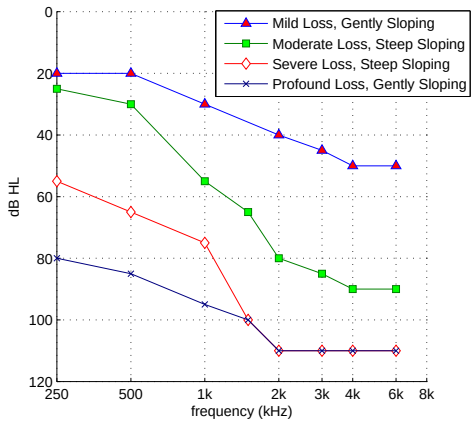

Figure 1: *Hearing Loss Audiograms*

3. Method

3.1. Collection of Data

The TIMIT corpus of read speech[10] was selected as the speech waveform source. The TIMIT test data has a core portion containing 24 speakers, 2 male and 1 female from each of the 8 dialect regions. This core test set of 192 sentences maintains a consistent ratio of phoneme occurences as the larger "full test set" (2340 sentences). TIMIT classifies 57 distinct phoneme types and groups them into 6 phoneme groups - stops, affricates, fricatives, nasals, sv/glides and vowels.

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 (P65/P85). A head related transfer function (HRTF) [11] of the human head was used to pre-filter the speech waverforms to mimic the amplification that occurs prior to the middle and inner ear. This technique has been used in other physiological and simulation studies [4]. Each sentence was presented at the two presentation levels to the AN Model for an unimpaired hearing profile and for each hearing loss profile.

3.2. Analysis of neural responses

The response of the AN to acoustic stimuli was quantified by the creation of "neurograms". As previously stated, these display the neural response as a function of CF and time. 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 [12] and as used for similar AN Model simulations [6][5], 60% of the fibers 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) which retained granular spike timing information and another with a larger bin size (312.5μs) which gave a moving average discharge rate.

3.3. Aggregating Phoneme Error Data

The phoneme timing information from TIMIT was used to extract the neurogram information on a per phoneme basis. For each phoneme occurrence, a mean absolute error was calculated between the unimpaired average discharge rate neurogram output and the three impaired models' neurograms. The mean absolute error for a phoneme was divided by the mean of the unimpaired neurogram for that phoneme, to normalise the error with respect to the phoneme sample's input pressure. In effect, the error is then expressed as a fraction of the normal response for the phoneme. This allows for comparisons at different presentation levels and across phoneme types.

This process was repeated using the spike timing neurograms to give two error metrics per phoneme at each hearing loss and presentation level. The errors per phoneme occurrence were collected to find a mean error per phoneme type. These were then sorted into their respective phoneme groupings to find a group mean error.

4. Results

Fig. 2 shows the average discharge errors and spike timing errors for five hearing loss profiles at the two presentation levels. The model exhibited saturation in error losses for the severe, profound, FLAT85 and FLAT120 hearing losses with minimal (<0.01) difference in error levels for phoneme groups. Hence, for clarity, only the FLAT120 value is shown in Fig. 2. Both the average discharge errors and spike timing errors clearly demonstrate the increased degradation of the AN response across all phoneme groups with increasing severity of hearing loss.

Examining the average discharge errors, both the FLAT10 and FLAT20 losses exhibit improvements of approximately 50% moving from P65 to P85 across all phoneme groups. For the mild and moderate losses the improvements are more pronounced for vowel and SV/glide groupings. This can be attributed to the the importance of the low frequency information for the reconstruction of the F1 and F2 formants for vowel-like sounds in the auditory nerve. The stops, affricate and fricatives have significant energy in the higher frequency bands (greater than 2kHz). This is reflected in the drop in errors between P65 and P85 for these groups where the mild loss benefits more than the moderate loss which has a steep drop off in hearing loss at higher frequencies.

The spike timing errors preserve the temporal fine structure, i.e. the rapid oscillations with a rate close to the centre frequency of the band. As would be expected, the mean absolute errors for the spike timing are higher than their corresponding average discharge errors. This is true across all phoneme groups. For the mild loss the P85 presentation level resulted in a more significant error reduction for the stops, affricates and fricatives than for the nasals, glides and vowels. This difference was less pronounced for the moderate loss and reversed in the FLAT120 loss where the error in the nasals, glides and vowels significantly exceed those of the stops, affricates and fricatives.

This caused a reevaluation of the chosen error metric for spike timing comparisons. The metric has been expressed as

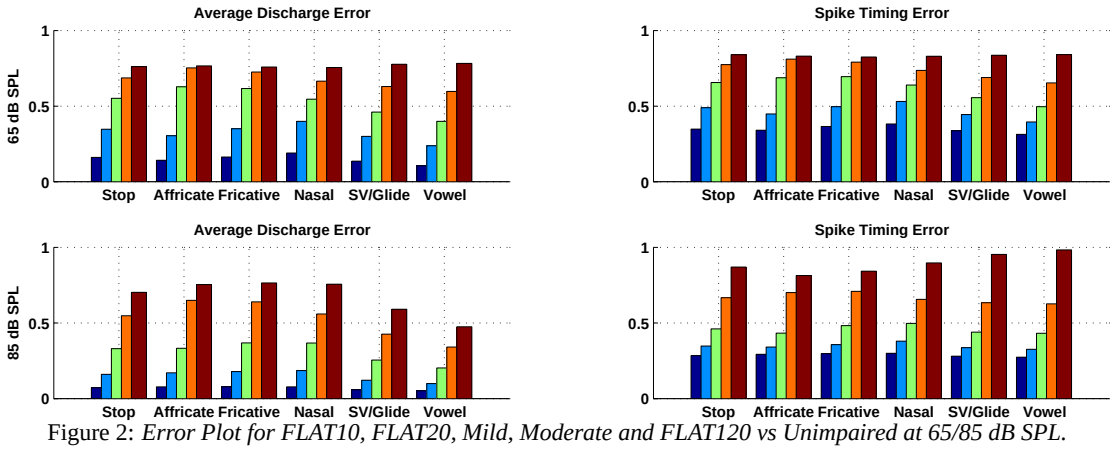

Figure 2: *Error Plot for FLAT10, FLAT20, Mild, Moderate and FLAT120 vs Unimpaired at 65/85 dB SPL.*

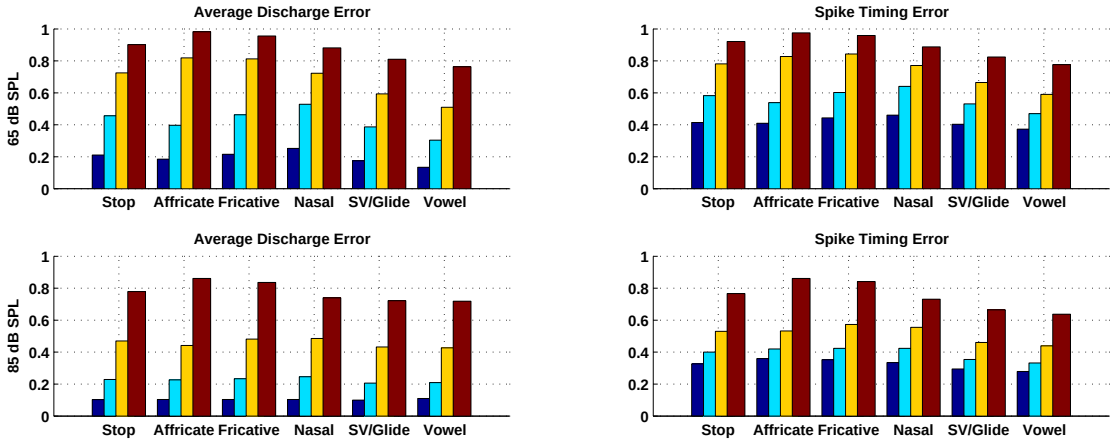

Figure 3: *Error Plot for FLAT10, FLAT20, Mild, Moderate vs Unimpaired at 65/85 dB SPL normalised against saturated levels.*

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) [13].

The sample spike timing neurograms for the vowel /eh/ presented to three models at P85 illustrate this phenomenon (Fig. 4). The spike timing 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 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 FLAT120 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.

4.1. Correlation

A 2-D cross correlation is a useful metric to compare two-dimensional data such as images. This motivated the calculation of the 2-D cross-correlation between the unimpaired and impaired spike timing neurograms for the vowel sounds. As the two neurograms being compared are always time aligned, any mismatches occurring are as a result of model impairments rather than synchronisation issues. The diagram in Fig. 5 represents the peak in the 2-D correlation across frequency dimension for the three illustrated neurograms. The auto-correlation and mild loss correlation illustrate the periodic nature of the spike timing neurogram. The loss of fine timing information is beginning to be apparent in the overall reduction in correlation. The collapse in the fine timing data is clear in the dramatic reduction in correlation for the FLAT120 loss. The sharp peaks apparent from phase locking to formants F1, F2, F3 for the unimpaired and mild loss have been flattened. This confirms that capturing the complicated information interaction within spike timing neurograms with a single measure, such as the mean absolute error or peak correlation, remains a challenge.

4.2. Normalised Results

Despite the fact that the maximum mean absolute error levels rise between P65 and P85 for the spike timing errors, they do continue to saturate, with the four super-threshold audiograms tested (FLAT85, severe, profound and FLAT120) all saturating

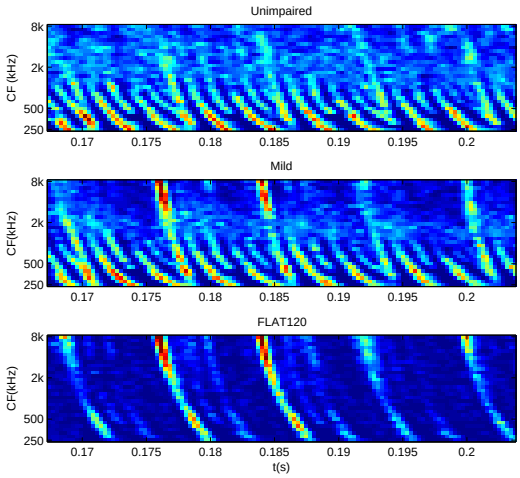

Figure 4: *Spike Timing Neurograms for vowel /eh/ presented at 85 dB SPL to three AN Models*

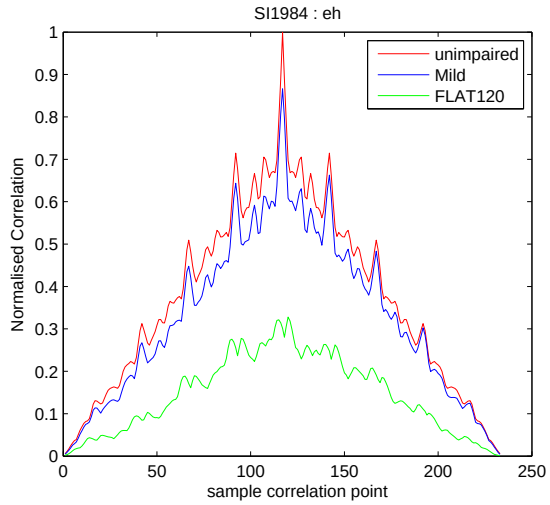

Figure 5: *Cross correlation of Spike Timing Neurogram for vowel /eh/ presented at 85 dB SPL to three AN Models*

to the same level. Hence, these saturation levels can be taken to be a maximum error value for the phoneme group (with the minimum still being 0 for comparison of unimpaired with itself). Normalising the errors by dividing each error by the saturation level for the phoneme group at each presentation level allows the results in Fig. 3 to be presented.

The results in Fig. 3 present more intuitive comparison between the performance of the mild and moderate hearing losses at each presentation level. Examining average discharge errors, the mild hearing loss has vowels performing significantly better than fricatives at P65. At P85 all phoneme groups are performing at a similar level, indicating the higher presentation level has helped the performance of stops, affricates and fricatives significantly more than the nasals, glides and vowels.

5. Conclusions and Future Work

This study used a number of hearing loss profiles to examine the error levels in auditory nerve representation of phonemic groups for real speech. The results showed that a mean absolute error metric provides a good comparative measure for average discharge rate neurograms due to the longer time windows used in capturing them. However the fine timing information in the spike timing neurograms combined with the response of the AN model at high presentation levels mean a more rigorous analysis is required to produce meaningful error metrics. This is a vital step in moving towards using the AN Model to assess the performance of different speech processing algorithms for hearing aids at a phonetic level. Work is ongoing to develop a fidelity measure related to a structural similarity (SSIM) index to accurately capture a meaningful spike timing error metric. Future study using existing human perceptual data would further validate the AN model's predictions.

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.*, 122(1):402–417, July 2007.

[2] 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.*, 120(3):1446–1466, Sept 2006.

[3] A. J. Hines and N. Harte. Measurement of phonemic degradation in sensorineural hearing loss using a computational model of the auditory periphery. *Irish Signals and Systems Conference, 2009. IET (To appear)*, 2009.

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

[5] 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*, pages 1793–1796, 2008.

[6] 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)*, pages 73–81, 2007.

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

[8] 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.*, 109:648–670, 2001.

[9] 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.*, 113:369–388, 2003.

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

[11] Francis M. Wiener and Douglas A. Ross. The pressure distribution in the auditory canal in a progressive sound field. *The Journal of the Acoustical Society of America*, 18(2):401–408, 1946.

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

[13] Jeff C. Wong, Roger L. Miller, Barbara M. Calhoun, Murray B. Sachs, and Eric D. Young. Effects of high sound levels on responses to the vowel /[var epsilon]/ in cat auditory nerve. *Hearing Research*, 123(1-2):61 – 77, 1998.