THE NEUROGRAM MATCHING SIMILARITY INDEX (NMSI) FOR THE ASSESSMENT OF SIMILARITIES AMONG NEUROGRAMS

Michael Drews[⋆] *Michele Nicoletti*[⋆] *Werner Hemmert*[⋆] *Stefano Rini*[†]

[⋆] Institute for Medical Engineering
[†] Institute for Communications Engineering
Technische Universität München, Germany

ABSTRACT

In this paper a new similarity index for neurograms is proposed. This index is inspired by the Needleman-Wunsch algorithm which determines the minimum number of operations to transform a vector into another in terms of insertions, deletions and substitutions. The Needleman-Wunsch algorithm can be extended to the two dimensional case and the number of transformations required to change a matrix into another is used to define a measure of similarity. This similarity measure is applied to neurograms and optimized to perform prediction of speech intelligibility in noise. Word recognition scores for for speech samples in noise are evaluated using the proposed similarity index, showing a clear improvement in speech intelligibility estimation with respect to other neurogram similarity metrics in the literature. The proposed similarity index is not restricted to a certain time resolution and could serve to evaluate neurogram similarity with respect to temporal fine structure in future.

Index Terms— neurogram, similarity measure, edit distance, speech intelligibility.

1. INTRODUCTION

The development of sophisticated computer models of the human inner ear [1, 2, 3] have made it possible to efficiently and precisely simulate the discharge patterns of multiple Auditory Nerve Fibers (ANFs) as a response to auditory stimuli.

Despite of these advances, much is still to be understood to how the auditory information is coded and represented in those discharge patterns.

A common tool to represent the neural response of the auditory nerve are "neurograms", matrices, which show the intensity of neural spiking of multiple ANFs as a function of time and cochlear location. As the inner ear decomposes signals according to their frequency components, each cochlear location has a corresponding characteristic frequency. Therefore, neurograms look similar as short-term spectrograms, with the difference that the signal phase is coded in spike times, at least in the low-frequency range up to 1-3 kHz. Neurograms are useful to characterize the effects of sensorineural hearing loss [4] and of electrical stimulation in cochlear implant patients [5]. Neurograms of the auditory nerve are also interesting because they carry all the information of a sound available to the central nervous system.

In this investigation we add noise to speech signals and quantify the degradation of the neurograms. We then compare this degradation to the speech understanding of human listeners. With this procedure we hope to answer the following questions: How robust is speech coded in neurograms? How does speech coding in noise degrade? How does human speech understanding correlate with the degrading neurograms?

Some similarity indexes have already been proposed in the literature:

•**Articulation Index [6] (AI) and Neural Articulation Index[7] (NAI):** The AI evaluates speech intelligibility purely as a sum of the Signal-to-Noise Ratio (SNR) in twenty frequency bands of the speech sample. The NAI is a variation of the AI obtained as a weighted sum of the SNR in seven frequency bands of the neurogram.

•**Neurogram Similarity Index Measure(NSIM) [8]:** Neurograms can be regarded as images and the similarity among them assessed using image processing techniques. The Structural Similarity Index Measure (SSIM) was developed by Wang et al. [9] to evaluate JPEG compression quality by assessing image similarity between compressed and uncompressed images in terms of three indexes: intensity, variance and cross-correlation. The NSIM is obtained by applying the SSIM to two neurograms.

• **Spectro-Temporal Modulation Index [10]:** Neurograms are convolved with the Spectro-Temporal Response Field which has the form of a spectro-temporal Gabor function [11] to produce a four dimensional function in time, frequency, rate and scale. This response is averaged over the frequency to obtain a scale-rate plot. The Spectro-Temporal Modulation Index is obtained as the average over scale and rate of the time correlation of the scale-rate plot.

Supported by within the Munich Bernstein Center for Computational Neuroscience by the German Federal Ministry of Education and Research (reference number 01GQ1004B).

In this paper we define a similarity measure across neurograms which is based on the number of changes required to transform one neurogram into another in terms of insertions, deletions and substitutions. We argue that this measure of neuronal similarity captures the perceptual distance between two sound samples and outperforms other similarity indexes proposed in literature. In this paper we focus on predicting the performance of human listeners to discriminate Consonant-Vocal-Consonant (CVC) samples.

2. BACKGROUND

2.1. Inner Ear Model

We generated neurograms with a modified inner ear model developed by the Carney group [1], in which we have tuned the middle ear filters to human-like hearing thresholds [12]. This model generates random spike trains of multiple ANFs at different center frequencies. It includes a model of synaptic adaptation and generates realistic responses of three different fiber types: High- (HSR), Medium- (MSR) and Low- (LSR) Spontaneous Rate fibers. Deviating from the physiological composition of the auditory nerve, where HSR fibers are the dominating population [13], we consider a composition of ANFs as 10 % for HSR, 20 % for MSR and 70 % for LSR fibers. This is consistent with the theory that neural coding may be disproportionately based on the enhanced dynamic range of LSR fibers in noisy environments as suggested in [14].

2.2. Speech Samples

The speech samples considered in our simulations were CVC words, such as *"ship"*, from the Arthur Boothroyd (AB) word list [15] used as a standardized listener test for English native speakers. In particular, we considered different SNRs of the speech samples in pink noise; the noise starts 0.2 ms before each word's onset, which is required for the adaptation of the auditory model. This initial stimulation was subsequently removed from the responses. Multiple neurograms were obtained for each speech sample and at SNRs from -15 to 20 dB. Minor adjustments were performed to obtain neurograms with a time duration of 0.8 ms.

2.3. Neurogram Pre-Procession

The auditory model generates ANF spike trains at a time resolution of 10 μs and for 100 different Center Frequencies (CFs), Greenwood-spaced between 80 Hz and 20 kHz. The model output was used to obtain the *rate-place code* (see [16], [17], [18]), in this paper we used the average discharge rate of each ANF in 10 ms time bins. The rate-place code captures only the information contained in the envelope of the sound signal, nevertheless experiments with chimaeric sounds reveal that for simple word recognition tasks as few as four frequency bands, providing only envelope information, are sufficient for good performance ($> 85\%$ word recognition [19]). Finally the rate-place code was further downsampled to 23 frequency bands using Root-Raised-Cosine (RRC) sampling windows with 50% overlap. However, as it will be clear in the next section, the proposed similarity index is not restricted to these particular assumptions on the time and frequency sampling. The reason for this sampling window is to obtain a representation which is sufficiently smooth and to reduce aliasing. In the frequency domain we utilize a RRC window of unitary power while in the time domain the RRC window is normalized as to obtain the same mean firing rate before and after sampling, thus preserving the rate-place code.

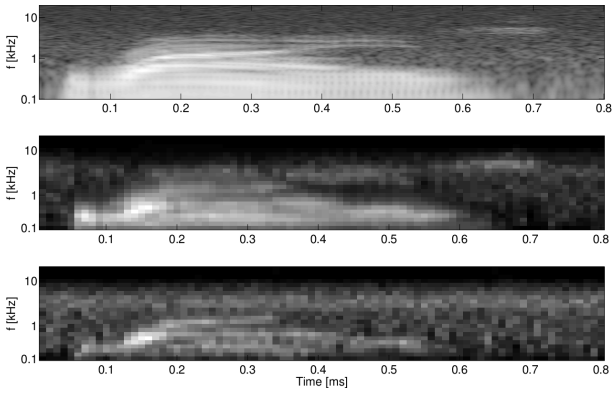

Fig. 1. Sample signal. *Top*: Spectrogram of the word "wise" (at +20 dB SNR). *Second row*: Pre-processed neurogram of the signal at +20 dB SNR. *Third row*: Pre-processed neurogram at 0 dB SNR.

Figure 1 shows the spectrogram for a given CVC sound (top panel) and the corresponding neurograms after time and frequency averaging for +20 dB SNR (mid panel) and 0 dB SNR (lower panel).

3. NEUROGRAM MATCHING TRANSFORMATION INDEX

3.1. The Two Dimensional Levenshtein (2DL) Algorithm

In this section we introduce the Neurogram Matching Similarity Index (NMSI) which is obtained from the approximate calculation of the two dimensional Levenshtein [20] distance among neurograms. The Levenshtein distance, also known as edit distance, is a distance measure for strings of characters defined as the minimum cost required to change one string into another in terms of cost of inserting, deleting and substituting a single character. The Levenshtein distance can be efficiently evaluated using dynamic programming using the Needleman-Wunsch algorithm [21] - also known as Sellers algorithm [22] and "optimal string matching" algorithm - which

has emerged independently in computer science, speech processing and bio-informatics.

We focus on a two dimensional extension of the Needleman-Wunsch algorithm, the 2DL algorithm, which was first proposed independently by Moore in [23] and by Tanaka and Kikuchi in [24]. In comparing two neurograms, we utilize a simple variation of the 2DL algorithm where the costs of substituting an element depend linearly on the absolute value of the difference between the elements. We introduce three cost parameters q_t, q_f and q_a for time-, frequency- and amplitude shift of a spectro-temporal element in a neurogram. Insertion and deletion cost of an element is set to 1.

We won't restate the algorithm here but rather refer the interested reader to [25] and present instead an example.

Example Consider the problem of calculating the edit distance between the two matrices

$$M_1 = \begin{bmatrix} 1 & 2 & 3 \\ 5 & 8 & 4 \end{bmatrix}, \quad M_2 = \begin{bmatrix} 1 & 1 \\ 5 & 9 \\ 3 & 4 \end{bmatrix} \qquad (1)$$

when the cost of each transformation is one. The 2DL algorithm provides the the following sequence of operations to transformation M_1 to M_2
1) delete the third column, $[3\ 4]^T$,
2) substitute the element $(1, 2)$ from 2 to 1,
3) substitute the element $(2, 2)$ from 8 to 9,
4) insert a third row, $[3\ 4]$.
in which case the two matrices are at a distance of 6. This example is also illustrated in Fig. 2: the elements in the red frames are deleted, the blue ones inserted and the green ones are amplitude-shifted.

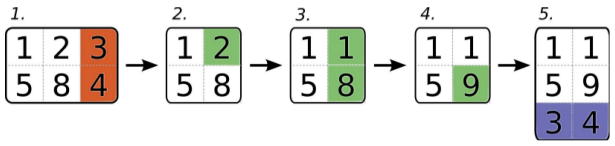

Fig. 2. The 2DL algorithm applied to the matrices M_1 and M_2 in (1).

3.2. Predicting Speech Recognition Performance in Noise

A simulated listener test was carried out replacing the human listener by the auditory model and expressing speech recognition in terms of neurogram distance using the NMSI.
We utilized the similarity index from Sec. 3.1 to predict speech intelligibility as to mimic human performance. We optimized the cost of time-, frequency- and amplitude-shift.

Speech samples were taken from the phonetically balanced CVC word list by Arthur Boothroyd [15]. We used the AB word list which contained 12 lists of 10 phonemically balanced CVC words (120 words). Word recognition performance of human test subjects in steady-state noise was simulated by keeping the noise level constant and increasing the speech level. Noise was spectrally matched to the speech and held at a constant level of 40 dB$_{SPL}$ (SPL: sound pressure level, relative to 20 μPa). Speech signals were added with SNRs from -15 dB to +20 dB in steps of 5 dB. The Speech Reception Threshold (SRT), defined as the signal level at which human test subjects can understand 50% of spoken words, was reached at an SNR of -11.5 dB [26].

The intuitive reason why we expect the NMSI to be able to predict speech intelligibility is that the 2DL algorithm provides a measure on how much the speech signal is impaired by the added noise. However, as the human auditory system is able to understand speech in noise even at negative SNRs, we found it more effective to calculate the distance of speech in noise to the condition of noise only. This measure is very sensitive to detect speech information in noise.

4. RESULTS

Figure 3 shows how the NMSI behaves as a function of SNR. NMSI was calculated for 120 words from the CVC word list. The distance was calculated for every word relative to 10 reference neurograms at an SNR of -21 dB, where the noise dominated. With increasing speech level, the neurograms significantly deviate from the reference condition. To convert the NMSI into recognition rates, we apply a trick ([27]): With a threshold at the median NMSI at the SRT (-11.5 dB, green dashed line in Fig. 3) – per definition – 50% of the NMSI values will be above this threshold at this point, which replicates the case for normal hearing subjects. With increasing SNR, a larger number of NMSIs calculated for different words will be above this threshold, predicting a higher recognition score. For SNRs higher than 10 dB the NMSI predicts already perfect recognition. The resulting function is shown in figure 4 (dashed red line) and replicates human performance (solid blue line) better than the NSIM proposed by Hines et al. [8] (dashed green line). Please note that the NSIM is intended to replicate only recognition rates higher than 50%, as for lower SNRs the NSIM saturates. To match the curve for lower SNRs would require to set the reference neurogram to clean speech. Then, noisy speech would have large NSIM values and a similar thresholding approach would lead to low recognition rates.

We now look into how the NMSI was optimized to predict human speech perception in noise [26]. This was achieved by finding appropriate values for the three cost parameters time-shift q_t, frequency-shift q_f and amplitude-shift q_a. The penalty function was set to the Mean Square Error (MSE) between the NMSI prediction and the listener data from [26] at all simulated SNRs:

$$MSE = \sum_{SNR\,=\,-10dB}^{+20dB} |NMSI(SNR) - data(SNR)|^2$$

The cost of a substitution should always be less than 2. Otherwise it would be cheaper to simply delete and add an element which has a cost of 1 for both cases. Therefore cost of time-, channel- and amplitude-shift also should not be bigger than 2. Fig. 5 shows the MSE value for $q_f = 1.6$ and varying q_t and q_a in that region. The best minimum MSE is found at approximately $(q_t, q_f, q_a) = (2, 1.6, 1)$.

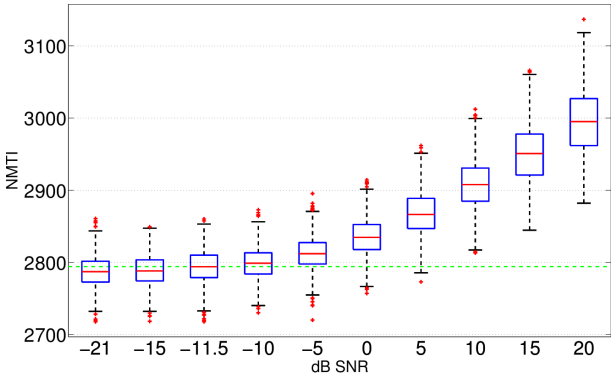

Fig. 3. NMSI of all neurograms at their respective SNR to the reference neurograms at -21 dB SNR. Blue boxes mark the 25% and 75% interquartiles. Red horizontal lines indicate the mean NMSI. Whisker length is 1.5 times the interquartile range. Green dashed line shows mean NPRT.

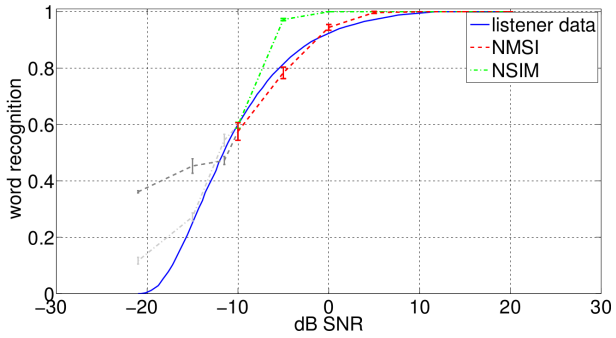

Fig. 4. Simulated speech recognition from the data from figure 3. *Red dashed* line shows NMSI results, *Green dash-dotted* line shows results for NSIM. *Blue solid* line show listener data from [26]. Data are generated by using optimal parameters as described in section 4. Grey parts of the curves are not considered in optimization process.

5. CONCLUSION

In this correspondence we propose a new similarity index for neurograms which we term Neurogram Matching Similarity Index (NMSI). This index is derived from the edit distance between matrices and is computed using an extension of the Needleman-Wunsch algorithm. The NMSI is obtained as the total cost to transform one neurogram into another using

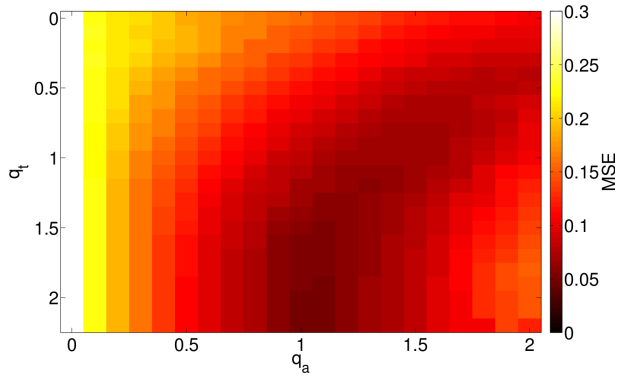

Fig. 5. MSE between SPIF and PIF for changing parameters q_t and q_a. Cost of channel-shift $q_f = 1.6$ is fixed. Minimum lies around the point $(q_t, q_f, q_a) = (2, 1.6, 1)$.

elementwise time-shifts, frequency-shifts, amplitude-shifts, deletions and insertions for the case in which a cost is assigned to each such operation. In this investigation the cost of each operation was optimized to predict speech recognition performance in noise and it was shown that the NMSI outperforms other similarity indexes proposed in the literature. Given that the underlying algorithms would support much higher temporal resolutions than we used in this investigation (10 ms), we predict a high potential of the NMSI to evaluate also the temporal finestructure present in auditory neurograms.

6. RELATION TO PRIOR WORK

Hines et al. [8] proposed a Neurogram Similarity Index (NSIM), which is based on image processing techniques. Other approaches are given in [7, 6, 10]. To our knowledge the 2DL algorithm [23, 24] proposed in this paper as basis for a new Neurogram Matching Similarity Index (NMSI) was applied to auditory neurograms for the first time here.

Acknowledgements

The authors are thankful to Professor Gerhard Kramer for the insightful discussions.

7. REFERENCES

[1] P. C. Nelson M. S. A. Zilany, I. C. Bruce and L. H. Carney, "A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics.," *The Journal of the Acoustical Society of America*, vol. 126, no. 5, pp. 2390–2412, Nov. 2009.

[2] G.J. Brown, R.T. Ferry, and R. Meddis, "A computer model of auditory efferent suppression: Implications for

the recognition of speech in noise," *The Journal of the Acoustical Society of America*, vol. 127, pp. 943, 2010.

[3] M. Holmberg, D. Gelbart, and W. Hemmert, "Speech encoding in a model of peripheral auditory processing: Quantitative assessment by means of automatic speech recognition," *Speech Communication*, vol. 49, no. 12, pp. 917–932, 2007.

[4] 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, no. 2, pp. 157–168, 2002.

[5] N.Y.S. Kiang, D.K. Eddington, and B. Delgutte, "Fundamental considerations in designing auditory implants," *Acta Oto-Laryngologica*, vol. 87, no. 3-6, pp. 204–218, 1979.

[6] N.R. French and J.C. Steinberg, "Factors governing the intelligibility of speech sounds," *The Journal of the Acoustical Society of America*, vol. 19, no. 1, pp. 90–119, 1947.

[7] J. Bondy, I.C. Bruce, S. Becker, and S. Haykin, "Predicting speech intelligibility from a population of neurons," *Advances in Neural Information Processing Systems*, vol. 16, 2003.

[8] A. Hines and N. Harte, "Speech intelligibility from image processing," *Speech Communication*, vol. 52, no. 9, pp. 736–752, 2010.

[9] 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.

[10] M. Elhilali, T. Chi, and S.A. Shamma, "A spectro-temporal modulation index (stmi) for assessment of speech intelligibility," *Speech Communication*, vol. 41, no. 2, pp. 331–348, 2003.

[11] T. Chi, Y. Gao, M.C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 106, pp. 2719, 1999.

[12] E. Terhardt, "Calculating virtual pitch," *Hearing Research*, vol. 1, no. 2, pp. 155–182, 1979.

[13] M.C. Liberman, "Physiology of cochlear efferent and afferent neurons: direct comparisons in the same animal," *Hearing Research*, vol. 34, no. 2, pp. 179–191, 1988.

[14] L.A.J. Reiss, R. Ramachandran, and B.J. May, "Effects of signal level and background noise on spectral representations in the auditory nerve of the domestic cat," *JARO-Journal of the Association for Research in Otolaryngology*, vol. 12, no. 1, pp. 71–88, 2011.

[15] A. Boothroyd, "Developments in speech audiometry," *British Journal of Audiology*, vol. 2, no. 1, pp. 3–10, 1968.

[16] M.B. Sachs and E.D. Young, "Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate," *The Journal of the Acoustical Society of America*, vol. 66, no. 2, pp. 470–479, 1979.

[17] B. Delgutte and N. Kiang, "Speech coding in the auditory nerve: (iii). voiceless fricative consonants," *The Journal of the Acoustical Society of America*, vol. 75, no. 3, pp. 887–896, 1984.

[18] E.D. Young, "Neural representation of spectral and temporal information in speech," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 363, no. 1493, pp. 923–945, 2008.

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

[20] V. Levenshtein, "Binary coors capable or correcting deletions, insertions, and reversals," in *Soviet Physics-Doklady*, 1966, vol. 10.

[21] S.B. Needleman, C.D. Wunsch, et al., "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[22] P.H. Sellers, "On the theory and computation of evolutionary distances," *SIAM Journal on Applied Mathematics*, pp. 787–793, 1974.

[23] R.K. Moore, "A dynamic programming algorithm for the distance between two finite areas," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, , no. 1, pp. 86–88, 1979.

[24] E. Tanaka and Y. Kikuchi, "A metric between pictures," *Trans. IEICE*, vol. 63, pp. 1018–1025, 1980.

[25] M.S. Waterman, "Dynamic programming algorithms for picture comparison," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 129–134, 1985.

[26] A. Boothroyd, "The performance/intensity function: an underused resource.," *Ear Hear*, vol. 29, no. 4, pp. 479–491, Aug 2008.

[27] A. Hines and N. Harte, "Simulated performance intensity functions," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, 2011, pp. 7139–7142.