THE RELATIVE ROLES OF TEMPORAL NEURAL CUES IN PREDICTING SPEECH INTELLIGIBILITY

Michael R. Wirtzfeld and Ian C. Bruce

Department of Electrical & Computer Engineering • McMaster University • Hamilton, ON, Canada

(wirtzfmr@mcmaster.ca - ibruce@ieee.org)

Abstract

Speech intelligibility predictors based on the contributions of envelope (ENV) and time fine structure (TFS) neural cues have many potential applications in communications and hearing research. However, establishing robust correlates between subjective speech perception scores and neural cues has not been straightforward. The spectro-temporal modulation index (STMI) [1] is able to predict the effects of presentation level on intelligibility in normal-hearing and hearing-impaired listeners [12] and the effects of hearing aid compression schemes [5]. Despite these results, the STMI metric can not explain speech intelligiblity for "auditory chimaeras" [8] where speech information is primarily in the TFS [4], motivating the inclusion of TFS neural cues in the predictive models.

A speech corpus of 1,750 sentences divided into five chimaera types, subjectively scored by 5 normal hearing listeners, was used. From sentence pairs consisting of an unprocessed NU-6 sentence and one of five respective chimaeric forms, auditory nerve responses [13] were simulated. The resulting characterizations of spectro-temporal variations, called "neurograms", were subsequently processed by the STMI and neural similarity (NSIM) [2] metrics. The latter metric captures TFS cues in addition to ENV related information.

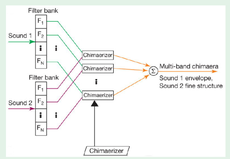
To investigate the ability of these metrics to predict % correct phoneme scores for NU-6 target-word chimaeras, 3 linear regression models were applied: Model 1 uses NSIM ENV and NSIM TFS factors with one interaction term; Model 2 uses STMI with NSIM TFS with one interaction term; and Model 3 considers only the STMI metric to establish a basis of comparison. Model 1 and Model 2 provide better predictive performance with the inclusion of NSIM TFS. However, the combination of STMI and NSIM TFS explained more variation than the NSIM TFS and its native NSIM ENV factor.

It is found that speech intelligibility predictions based solely on ENV neural cues are inadequate. Predictive performance is improved with the inclusion of complementary TFS cues.

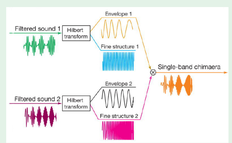
INTRODUCTION

A. Auditory Chimaeras

- Investigation of the relative contributions of envelope (ENV) and temporal fine structure (TFS) cues to speech perception has been ongoing for several years.
- Smith et al. [8] created the idea of "auditory chimaeras" to combine the ENV of one speech signal with the TFS of a second, unrelated signal. By using this idea it was demonstrated that envelope cues are critical for speech perception, while fine structure is necessary for pitch perception and sound localization.



(a) Filterbank Construction



(b) Single Band Construction

Figure 1: *(a.) Perfect-reconstruction filter banks split respective sounds into a complementary set of frequency bands with each matching pair of frequency bands passed to a Hilbert transform based Chimaerizer. (b.) The Hilbert transform factors band-limited signals into their envelope and fine structure components. A single-band chimaera is produced by the product of envelope 1 and fine structure 2 [8].*

- Elhilali et al. [1] established the spectro-temporal modulation index (STMI) that quantifies the degradation in the cortical encoding of spectral and temporal modulations due to noise. This metric depends only on the slow varying envelope and ignores the fast TFS.
- Swaminathan & Heinz [9] have demonstrated the importance of saliant features of both auditory nerve ENV and TFS cues for speech perception in noise. Neural correlates to speech perception have been difficult to establish because of cochlear transformations between TFS and recovered neural ENV [3]. Swaminathan & Heinz found that neural ENV coding was a primary contributor to speech perception, even in noise, while neural TFS contributed in noise but mainly in the presence of neural ENV.

B. Neurograms

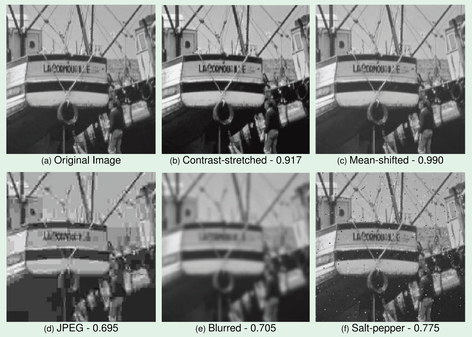
- Examination of ENV and TFS cues is realized by using 2-dimensional images derived from post-stimulus time histograms (PSTHs).
- PSTHs are generated for a given set of basilar membrane characteristic frequencies (CFs) and stacked one on top of the other relative to the time axis. This arrangement is called a "neurogram".
- Adjusting the time resolution of the PSTHs reveals the inherent ENV and TFS information for a given neurogram.

C. Neurogram SIMilarity (NSIM) Metric

- Hines & Harte [2] studied the effect of sensorineural hearing loss on phonemic degradation with ENV and TFS neurograms using a metric based on visual image quality prediction [10].
- Unlike pixel-based metrics, the Neurogram SIMilarity (NSIM) metric uses patches of pixels spanning the image and calculates respective "luminance" (μ_1, μ_2), "contrast" (σ_1, σ_2) and "structure" (σ_{12}) statistics. Weighted contributions (α, β and γ) determine a single patch value.

$$\mathrm{NSIM} = \left(\frac{2\mu_1\mu_2 + C_1}{\mu_1^2 + \mu_2^2 + C_1}\right)^{\alpha} \cdot \left(\frac{2\sigma_1\sigma_2 + C_2}{\sigma_1^2 + \sigma_2^2 + C_2}\right)^{\beta} \cdot \left(\frac{\sigma_{12} + C_3}{\sigma_1\sigma_2 + C_3}\right)^{\gamma} \quad (1)$$

- An overall metric value is found by averaging the NSIM values over time and CF.



(a) Original Image (b) Contrast-stretched - 0.917 (c) Mean-shifted - 0.990

(d) JPEG - 0.695 (e) Blurred - 0.705 (f) Salt-pepper - 0.775

Figure 2: *Illustration of visual image quality predictor from [10]. Each image has a mean squared-error (MSE) of 210. Values shown are the mean structural similarity values.*

METHODS

A. Speech Corpus

- A 1,750 sentence corpus [4] is equally divided across 5 chimaera types.

Speech-ENV+WGN-TFS	Speech-ENV+MN-TFS	
Speech-TFS+WGN-ENV	Speech-TFS+MN-ENV	TFS-only

- Average phonemic perception scores for chimaera-processed CVC NU-6 target-words were obtained from 5 normal hearing subjects [4].
- Sentence pairs, a chimaera and its original, were normalized to 65 dB SPL.

B. Auditory Periphery Model

- The Zilany et al. model [13] is used in this study. This phenomenological model characterizes the auditory pathway from the middle ear to the auditory nerve.
- Real-ear unaided gain is applied using the head-related transfer function described by Wiener and Ross [11].
- A speech stimulus is applied to the model to compute scaled spike-rate PSTH responses for a set of CFs.
- Inner and outer hair cell model parameters C_{IHC} and C_{OHC} are both set to unity to model a normal auditory periphery.

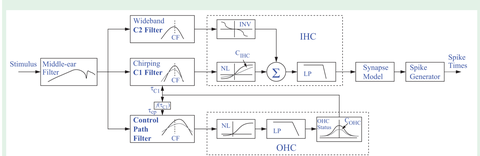


Figure 4: *Zilany et al. Auditory Model [13]*

C. Metrics

C.1 Spectro-temporal Modulation Index (STMI)

- PSTHs are collected for each chimaera-original sentence pair using 128 CFs, logarithmically spaced from 180 to 7,040 Hz.
- A period of silence was appended to the end of the sentence stimulus to account for adaptation behavior of the model.
- Each PSTH response is convolved with a 16-ms rectangular window at 50% overlap to produce mean-rate neurograms.
- The STMI analyzes temporal and spectral modulation content from the mean-rate neurogram. It is based on a cortical model of speech processing investigated by [1].

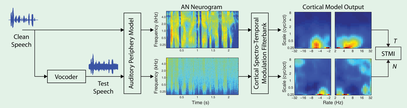


Figure 5: *Spectro-Temporal Modulation Index [4]*

- Two cortical responses are produced: (1) a reference output, *T*, from the original stimulus in quiet and (2) a response to the chimaera stimulus, *N*.

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

- The STMI is a scalar value between 0 and 1. A larger number indicates better predicted speech intelligibility.

C.2 Neurogram SIMiliarity (NSIM)

- PSTHs were collected for each original-chimaera sentence pair at 30 CFs, logarithmically-spaced from 250 Hz to 8 kHz.
- The neural adaptation behavior of the model was accounted for by appending a period of silence at the end of each sentence.
- Neurograms are produced with time bins and convolution with a Hamming window (50% overlap) as follows:

Mean-rate	100 µs	128-sample Hamming
Fine-timing	10 µs	32-sample Hamming

- Weighting parameters (α, β, γ) for Eq. (1) are set to $(1, 0, 1)$, respectively.

RESULTS

A. Perceptual Scores

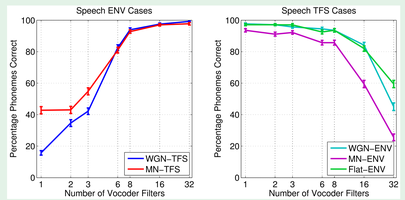


Figure 7: *Subjective Intelligibility Phoneme Scores from [4] (Error-bars: ± 1 SEM)*

B. STMI Predictions

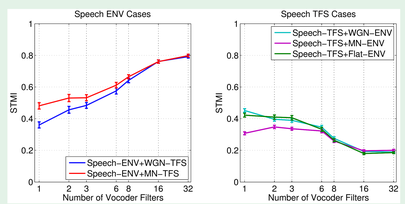


Figure 8: *STMI Predictions (Error-bars: ± 1 SEM)*

- Fig. 9 illustrates the relationship between subjective phoneme scoring and the STMI prediction. For the Speech-ENV+WGN-TFS and Speech-ENV+MN-TFS chimaeras, the STMI metric is more correlated to the subjective evaluation of intelligibility. The relationship between the STMI metric and the Speech-TFS chimaeras is less correlated.

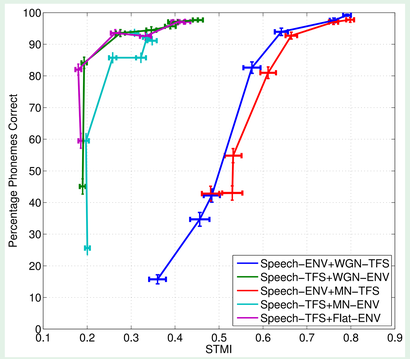


Figure 9: *Phoneme Scoring Versus STMI (Error-bars: ± 1 SEM)*

C. NSIM Predictions

- Figures 10 and 11 illustrate the relationship between subjective phoneme scoring and the NSIM ENV and TFS metrics, respectively. Like the STMI metric, the NSIM ENV is correlated with the phoneme scoring of the Speech-ENV chimaeras.

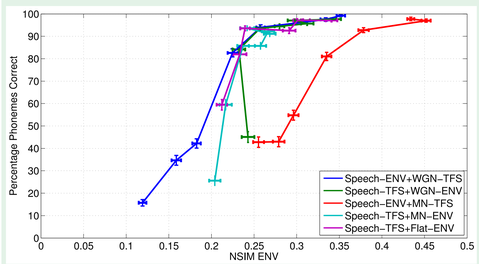


Figure 10: *Phoneme Scoring Versus NSIM ENV (Error-bars: ± 1 SEM)*

- As shown in Fig. 11, the NSIM TFS metric is able to segregate the Speech-TFS and Speech-ENV chimaeras. However, it does not establish a clear correlation with the phoneme scoring.

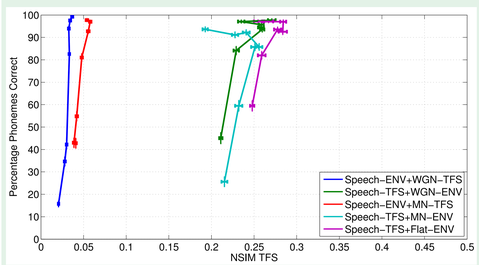


Figure 11: *Phoneme Scoring Versus NSIM TFS (Error-bars: ± 1 SEM)*

D. Modeling

- Table 1 summarizes the results of fitting different multiple-regression models to the NSIM ENV, NSIM TFS and STMI metrics along with their interactions. Regression coefficients are given with p-values in parenthesis. Fig. 12 illustrates the performance of the different regression models in predicting the RAU transformed percent correct phoneme scores.

Table 1: *Regression Model Parameter Summary*

Model 1		Model 2		Model 3	
ENV	258.2 (0.0002)	STMI	231.6 (< 0.0001)	STMI	46.4 (0.075)
TFS	-2.5 (0.986)	TFS	483.1 (< 0.0001)		
ENV × TFS	397.8 (0.443)	STMI × TFS	-347.2 (0.09)		
Adj. R^2	0.65	Adj. R^2	0.83	Adj. R^2	0.09
p-value	< 0.0001	p-value	< 0.0001	p-value	0.075

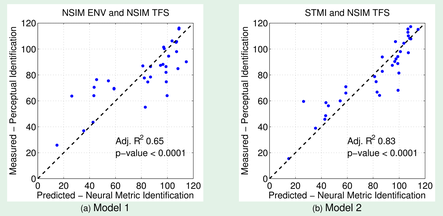


Figure 12: *First-order Linear Regression Model Predictions*

CONCLUSIONS

- Preliminary modeling indicates the NSIM TFS metric is complementary to both the NSIM ENV and STMI metrics.
- We are continuing to investigate several issues encountered with neurogram scaling, regularization terms in the NSIM metric, and phase sensitivity of the NSIM TFS.
- Other linear and nonlinear regression models will be explored for fusion of the NSIM TFS and STMI values.

ACKNOWLEDGMENTS

- This work is supported by NSERC Discovery Grant 261736.

REFERENCES

[1] Elhilali, M., et al., "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Communication*. Volume 41, Issue 2-3, pgs. 331-348, October 2003

[2] Hines, A. and Harte, N., "Speech Intelligibility Prediction using a Neurogram Similarity Index Measure," *Speech Communication*. Volume 54, Issue 2, pgs. 306-320, February 2012

[3] Ibrahim, R. A. and Bruce, I. C., "Effects of peripheral tuning on the auditory nerve's representation of speech envelope and temporal fine structure cues," in *The Neurophysiological Bases of Auditory Perception*. Editors: E. A. Lopez-Poveda et al., Springer, NY, pgs. 429-438, 2010

[4] Ibrahim, R. A. and Bruce, I. C., "The interaction of envelope and temporal fine structure cues in speech perception," in *Abstracts of the 34th ARO Midwinter Research Meeting*, 2011

[5] Leung, B. and Bruce, I. C., "Physiological assessment of nonlinear hearing aid amplification schemes," at *International Hearing Aid Research Conference (IHCON 2008)*, Lake Tahoe, CA, August 2008

[6] Lorenzi, C., et al., "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," *PNAS*. Volume 103, No. 49, pgs. 18866-18869, 2006

[7] Shera, C. A., et al., "Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements," *PNAS*. Volume 99, No. 5, pgs. 3318-3323, 2002

[8] Smith, Z. M., et al., "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*. Volume 416, No. 6876, pgs. 87-90, 2002

[9] Swaminathan, J. and Heinz, M. G., "Psychophysiological Analyses Demonstrate the Importance of Neural Envelope Code for Speech Perception in Noise," *The Journal of Neuroscience*. Volume 32, Issue 5, pgs. 1747-1756, 2012

[10] Wang, Z., et al., "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*. Volume 13, Issue 4, pgs. 600-612, 2004

[11] Wiener, F. M. and Ross, D. A., "The pressure distribution in the auditory canal in a progressive sound field," *Journal of the Acoustical Society of America*. Volume 18, No. 2, pgs. 401-408, October 1946

[12] Zilany, M. S. A. and Bruce, I. C., "Predictions of Speech Intelligibility with a Model of the Normal and Impaired Auditory-periphery," in *Proceedings of 3rd International IEEE EMBS Conference on Neural Engineering*. IEEE. Piscataway, NJ, pgs. 481-485, 2007

[13] Zilany, M. S. A., Bruce, I. C., Nelson, P. C., and Carney, L. H. (2009), "A phenomenological model of the synapse between the inner hair cell and auditory nerve: Long-term adaptation with power-law dynamics," *Journal of the Acoustical Society of America*. 126(5), 2390-2412