Predictions of Speech Intelligibility with a Model of the Normal and Impaired Auditory-periphery

M. S. Arefeen Zilany (zilany@grads.ece.mcmaster.ca) and **Ian C. Bruce**
Department of Electrical and Computer Engineering, McMaster University, Hamilton, Ontario, Canada

1. Introduction

A fall-off in speech intelligibility at higher-than-normal presentation levels has been observed for listeners with and without hearing loss. Speech intelligibility predictors based on the acoustic signal properties, such as the articulation index and speech transmission index, cannot directly account for the effects of presentation level and hearing impairment. Recently, Elhilali et al. (2003) introduced the spectrotemporal modulation index (STMI), a speech intelligibility predictor based on a model of how the auditory cortex analyzes the joint spectro-temporal modulations present in speech. However, the auditory-periphery model used by Elhilali et al. is very simple and cannot describe many of the nonlinear, level-dependent properties of cochlear processing, nor the effect of hair cell impairment on this processing. In this study, we quantify the effects of speech presentation level and cochlear impairment on speech intelligibility using the STMI with a more physiologically-accurate model of the normal and impaired auditory periphery developed by Zilany and Bruce (2006). This model can accurately represent the auditory-nerve responses to a wide variety of stimuli across a range of characteristic frequencies and intensities spanning the dynamic range of hearing. In addition, outer and inner hair cell impairment can be incorporated. Compared to the experimental word recognition scores, this model-based STMI can qualitatively predict the effect of presentation levels on speech intelligibility for both normal and impaired listeners in a wide variety of conditions.

2. Method

A. Model of the Auditory-periphery

❖ The auditory-periphery model by Zilany and Bruce (2006) addresses level-dependent tuning, two-tone suppression, BF-shift with level.

❖ C1 output dominates at low and moderate levels, and is responsible for the synchrony capture or multi-formant responses seen in vowel responses.

❖ A parallel-path C2 filter has been introduced as a second mode of excitation to the IHC.

❖ C2 output dominates at high levels, and is responsible for the high level effects such as the C1/C2 transition, peak splitting, loss of synchrony capture by a particular formant in vowel responses.

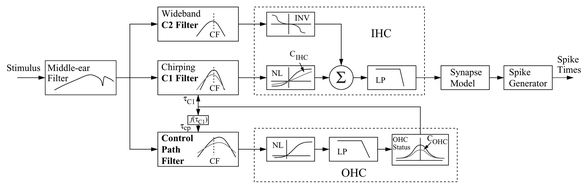

Fig. 1. Schematic diagram of the AN fiber model, reprinted from [2] with permission. The input to the model is an instantaneous pressure waveform of the stimulus in Pascals and the output is the spike times in response to that input. The model has a middle-ear filter, a feed-forward control-path, a signal-path C1 filter and a parallel-path C2 filter, the inner hair-cell (IHC) section followed by the synapse model and the discharge generator. Abbreviations: outer hair cell (OHC), low-pass (LP) filter, static nonlinearity (NL), characteristic frequency (CF), inverting nonlinearity (INV). C_{OHC} and C_{IHC} are scaling constants that indicates OHC and IHC status, respectively.

❖ The output of the model of the auditory-periphery is represented by a time-frequency spectrogram-like output, which is referred to as a "*neurogram*".

❖ Simultaneous outputs (discharge rates averaged over every 8 ms) from 128 AN fibers, CFs ranging from 0.18 to 7.04 kHz spaced logarithmically, make up the neurogram.

❖ The output at each CF represents the average discharge rates of fibers having three different spontaneous rates: 50 (high), 5 (medium) and 0.1 (low) spikes/s.

❖ Consistent with the distribution of spontaneous rates of fibers within an animal, the maximum weight (0.6) goes to high rate fibers, and the weight given to medium and low spontaneous rate fibers is 0.2 each.

❖ In the impaired case, the weights of high spontaneous rate fibers only are scaled down according to the degree of IHC impairment in the cochlea.

B. Model of the Central Auditory System

❖ Analyzes the AN neurogram to estimate the spectral and temporal modulation content.

❖ Implemented by a bank of modulation-selective filters ranging from slow to fast rates (2 to 32 Hz) temporally and narrow to broad (0.25 to 8 cyc/oct) scales spectrally.

C. Spectro-temporal Modulation Index (STMI)

❖ The deviation the model output at the cortical stage has undergone from a template (i.e., the expected response) gives a measure of the STMI.

❖ The template has been chosen as the output of the normal model to the stimulus at 65 dB SPL (conversational speech level) in quiet.

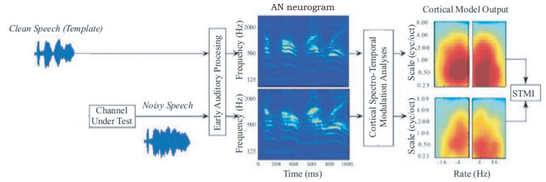

Fig. 2. Schematic showing steps in computing the STMI (taken from Chi et al., 1999, and Elhilali et al., 2003).

❖ After analyzing the two-dimensional (2-D: time and frequency) AN neurogram by the modulation filter banks, the cortical output is a four-dimensional (4-D: time, frequency, rate and scale) complex-valued representation.

❖ Since only temporal and spectral modulations are to be extracted, the cortical output of the model in each case (both for template and test stimulus) has been adjusted by subtracting the model output due to its own base spectrum. The base is a stationary noise with a spectrum identical to that of the long-term spectrum of the stimulus being tested.

❖ Once the cortical output of the test stimulus, N, and the template, T, for that stimulus are computed, the STMI can be calculated as:

$$\text{STMI} = \sqrt{1 - \frac{\|T - N\|^2}{\|T\|^2}}$$

where ||.|| indicates the 2-norm of the corresponding signal.

D. Differences from the study by Elhilai et al.

❖ The AN model employed in this work is a more complete and physiologically-accurate model.

❖ In Elhilali et al., the 4-D cortical output is reduced to 3-D by averaging over the stimulus duration. However, in this study, the 4-D cortical output is used in all cases, as temporal information seems important.

❖ The equation employed to calculate the STMI here is the square root of the expression utilized in Elhilali et al.

❖ A lateral inhibitory network (LIN), between the auditory-periphery and the auditory cortex, was used in Elhilali et al., which is not included in this present work.

❖ Consistent with the physiological and anatomical observations, AN fibers with different spontaneous rates have been considered.

3. Results

A. Effects of Presentation Levels for Listeners with Normal Hearing

1) In Quiet:

❖ Molis and Summers (2003) conducted an experiment on seven normal hearing listeners in quiet, and the task was to identify correct words from 72 lists each having ten low-context sentences, where the sentences were either lowpass- or highpass-filtered.

❖ In this work, a range of lowpass- and highpass-filtered sentences from TIMIT database are applied as the input to the model, and the STMIs are computed. The cut-off frequencies for low- and high-pass filters used here are 1.0 and 2.5 kHz, respectively.

❖ Word recognition scores for highpass-filtered sentences declined more consistently than the decrease in the recognition of the lowpass-filtered sentences.

❖ In our AN model, the lower CFs have relatively less nonlinearity than those at the higher CF fibers, which in turn gives relatively broader tuning at higher CFs at high levels (Robles and Ruggero, 2001). In addition, the loss of synchrony capture by formant 2 (F2) in a vowel response occurs at a lower presentation level for higher CF fibers (Wong et al., 1998; Zilany and Bruce, submitted). These two model properties could explain the observed larger rollover at high levels for highpass-filtered speech materials.

2) In Noisy Conditions:

❖ Dubno et al. (2005) studied the effects of speech and masker level on the recognition of speech for the NU6 monosyllabic words. Broadband (0.165–7.4 kHz) speech was presented in speech-shaped maskers at three speech levels (70, 77 and 84 dB SPL) for each three SNRs (+8, +3 and –2 dB). An additional low level noise was added to produce equivalent masked thresholds for all listeners.

❖ Here we have simulated the same experimental conditions, and subsequently the model-based STMI has been computed for the 40 monosyllabic words from NU6 word lists.

❖ Word recognition declined significantly with increasing level, even when the SNR was held constant. Compared to the experimental word recognition scores reported in Dubno et al. (2005), the results are qualitatively similar.

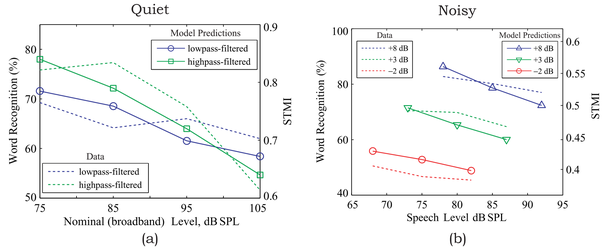

Fig. 3. (a) Mean word recognition performance for normal listeners (dotted lines) from Fig. 1 of Molis and Summers (2003) and STMI (solid lines with symbols) versus presentation level for lowpass- and highpass-filtered sentences.
(b) Mean word recognition performance for normal listeners (dotted lines) from Fig. 4 of Dubno et al. (2005) and STMI (solid lines with symbols) versus presentation level at three SNRs (+8, +3, and -2 dB) for broadband speech from the NU6 word lists.

B. Effects of Presentation Levels for Listeners with Hearing Loss

1) In Unaided Conditions:

❖ Shanks et al. (2002) studied the performance of impaired listeners who were divided into four groups based on their degree and slope of the hearing loss. Three pure tone (0.5, 1 and 2 kHz) averages in dB HL indicated the degree of hearing loss, and the slope was the change in pure tone thresholds between 0.5 and 4 kHz.
Group 1: < 40 dB HL, <10 dB/octave; Group 2: <40 dB HL, >10 dB/octave;
Group 3: > 40 dB HL, <10 dB/octave; Group 4: >40 dB HL, >10 dB/octave.

❖ Speech stimuli (connected speech test, CST) were presented in three speech levels (52, 62 and 74 dB SPL) with three signal to babble (S/B) ratios (+3, 0, and –3 dB).

❖ In this work, we have simulated four example impairments, each representing one of the four groups of impaired listeners.

❖ In the unaided condition, performance of word recognition in multi-talker babble declines slightly with the increased presentation levels for the mild and moderately impaired listeners but increases substantially for severely impaired listeners because of increased audibility.

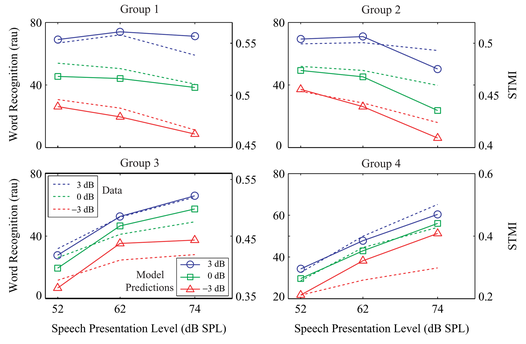

Fig. 4. Mean unaided CST word recognition performance (dotted lines) from Fig. 3 of Shanks et al. (2002) and STMI (solid lines with symbols) versus presentation level for four groups of impaired listeners, for presentation levels (52, 62 and 74 dB SPL) at three S/B ratios (+3, 0, and –3 dB).

2) In Aided Conditions:

❖ Shanks et al. (2002) compared performance for three hearing aid circuits: peak clipping, compression limiting, and wide dynamic range compression. They found that all three hearing aids circuits provided benefit over the unaided conditions.

❖ Here, STMI predictions are computed for the NAL-R (National Acoustic Laboratory) prescription applied to the peak-clipping hearing aid circuit only.

❖ In both experimental and model predictions, performance declines with increasing speech levels for nearly all impaired listeners.

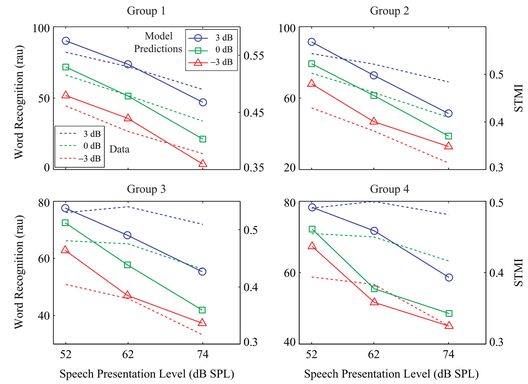

Fig. 5. Mean aided CST word recognition performance (dotted lines) from Fig. 6 of Shanks et al. (2002) collapsed across the three hearing aid circuits and STMI (solid lines with symbols) versus presentation level for four groups of impaired listeners, for presentation levels (52, 62 and 74 dB SPL) at three S/B ratios (+3, 0, and –3 dB).

4. Discussions and Conclusions

The auditory model-based STMI, when implemented with a physiologically-accurate auditory-periphery model, can directly address the effects of presentation level and cochlear impairment on speech intelligibility. In contrast, predictors based on acoustic signal properties need to use ad-hoc methods to account for degradations due to suprathreshold nonlinearities or cochlear impairment. The accuracy in predicting speech intelligibility by this model-based STMI provides strong validation of attempts to design hearing aid algorithms or amplification schemes based on physiological data and models (Sachs et al., 2002; Bruce, 2004; Bondy et al., 2004).

References

[1] M. Elhilali, T. Chi and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," Speech Comm., vol. 41, pp. 331–348, 2003.

[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., vol. 120(3), pp. 1446–1466, 2006.

[3] T. Chi, Y. Gao, M. C. Guyton, P. Ru and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," J. Acoust. Soc. Am., vol. 106(5), pp. 2719–2732, 1999.

[4] M. R. Molis and V. Summers, "Effects of high presentation levels on recognition of low- and high-frequency speech," Acoustics Research Letters Online, vol. 4(4), pp. 124–128, 2003.

[5] L. Robles and M. A. Ruggero, "Mechanics of the mammalian cochlea," Physiol. Rev., vol. 81(3), pp. 1305–1352, 2001.

[6] J. C. Wong, R. L. Miller, B. M. Calhoun, M. B. Sachs and E. D. Young, "Effects of high sound levels on responses to the vowel /ɛ/ in cat auditory nerve," Hear. Res., vol. 123, pp. 61–77, 1998.

[7] M. S. A. Zilany and I. C. Bruce, "Representation of the vowel /ɛ/ in normal and impaired auditory-nerve fibers: model predictions of responses in cats," submitted to J. Acoust. Soc. Am.

[8] J. R. Dubno, A. R. Horwitz and J. B. Ahlstrom, "Word recognition in noise at higher-than-normal levels: decreases in scores and increases in masking," J. Acoust. Soc. Am., vol. 118(2), pp. 914–922, 2005.

[9] J. E. Shanks, R. H. Wilson, V. Larson and D. Williams, "Speech recognition performance of patients with sensorineural hearing loss under unaided and aided conditions using linear and compression hearing aids," Ear & Hearing, vol. 23, pp. 280–290, 2002.

[10] M. B. Sachs, I. C. Bruce, R. L. Miller and E. D. Young, "Biological basis of hearing-aid design," Ann. Biomed. Eng., vol. 30(2), pp. 157–168, 2002.

[11] I. C. Bruce, "Physiological assessment of contrast-enhancing frequency shaping and multiband compression in hearing aids," Physiol. Meas., vol. 25(4), pp. 945–956, 2004.

[12] J. Bondy, S. Becker, I. C. Bruce, L. Trainor and S. Haykin, "A novel signal-processing strategy for hearing-aid design: Neurocompensation," Signal Processing, vol. 84(7), pp. 1239–1253, 2004.