Proceedings of the 3rd International
IEEE EMBS Conference on Neural Engineering
Kohala Coast, Hawaii, USA, May 2-5, 2007

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

Muhammad S. A. Zilany[†] and Ian C. Bruce[*]
Department of Electrical & Computer Engineering, McMaster University
Hamilton, ON, Canada

Abstract— **A fall-off in speech intelligibility at higher-than-normal presentation levels has been observed for listeners with and without hearing loss [1]–[5]. 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. [6] introduced the spectro-temporal 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 [7]. 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 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.**

I. INTRODUCTION

The deterioration of speech recognition at high signal and noise levels has been observed for both normal hearing and hearing impaired listeners [1]–[4]. This declination in recognition has been referred to as the rollover effect, and the magnitude depends on several factors: speech and noise levels, noise shape, normal quiet thresholds, and the amount and extent of impairment. For normal hearing subjects listening in quiet, recognition of key words in sentences exhibits more rollover for highpass-filtered sentences than for lowpass-filtered sentences [4], [8]. The greater rollover observed at high-frequencies is consistent with the physiological and psychoacoustic data suggesting that cochlear processing shows more level dependence in basal regions tuned to high frequencies than in the apical, low frequency regions [9]. In noisy conditions, performance of normal hearing listeners declines substantially above conversational level with increasing speech levels at a fixed signal to noise ratio (SNR). The effect is larger for nonsense or monosyllabic words and when speech is presented in a masker with a spectrum that matches the spectrum of the speech [1], [2]. Lowpass- and highpass-filtered word recognition in noise also declines with increasing speech levels at high levels, although highpass-filtered words show relatively more decline in recognition [3]. From audibility estimates based on the articulation index, the deterioration in intelligibility can be modeled as a relative increase in effective masking level for low frequency speech, but the decline in recognition for high frequency speech can not be explained by the masking growth alone [3].

For listeners with hearing loss, speech recognition in noisy conditions is affected in the same way as for normal hearing listeners when differences in audibility are considered [1]. Shanks et al. [5] found that in unaided conditions, speech intelligibility decreases with increasing speech levels for listeners with relatively mild hearing loss but *increases* for listeners having severe to profound hearing loss. However, in aided conditions, they found that word recognition declines at higher than normal levels for almost all listeners.

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. In the speech intelligibility index (SII), a frequency-independent level distortion factor (LDF) has been introduced to take into account the declination in speech intelligibility when the overall speech level exceeds 73 dB SPL. However, it has been shown that the rollover effect at high levels is frequency dependent [4], [8]. Recently, Elhilali et al. [6] introduced the spectro-temporal modulation index (STMI), a speech intelligibility predictor that employs an auditory model to analyze the effects of noise, reverberations, and other distortions on the joint spectro-temporal modulations present in speech. This model-based STMI can account for some difficult and nonlinear distortions of speech such as phase-jitter and phase shifts, which other estimators can not address. However, the auditory-periphery model used by Elhilali et al. is very simple and lacks many of the nonlinear, level-dependent phenomena of the cochlear processing such as two-tone suppression, level-dependent tuning, and adaptation properties which are considered critical for the task in hand. Additionally, this model does not incorporate the effects of hair cell impairment on the processing, and thus is not able to predict the effects of impairment on speech intelligibility.

The goal of our study is to quantify the effects of speech presentation level and cochlear impairment on speech intelligibility using the STMI with a more physiologically-accurate

[†] Email: zilany@grads.ece.mcmaster.ca
[*] Email: ibruce@ieee.org

1-4244-0792-3/07/$20.00©2007 IEEE.

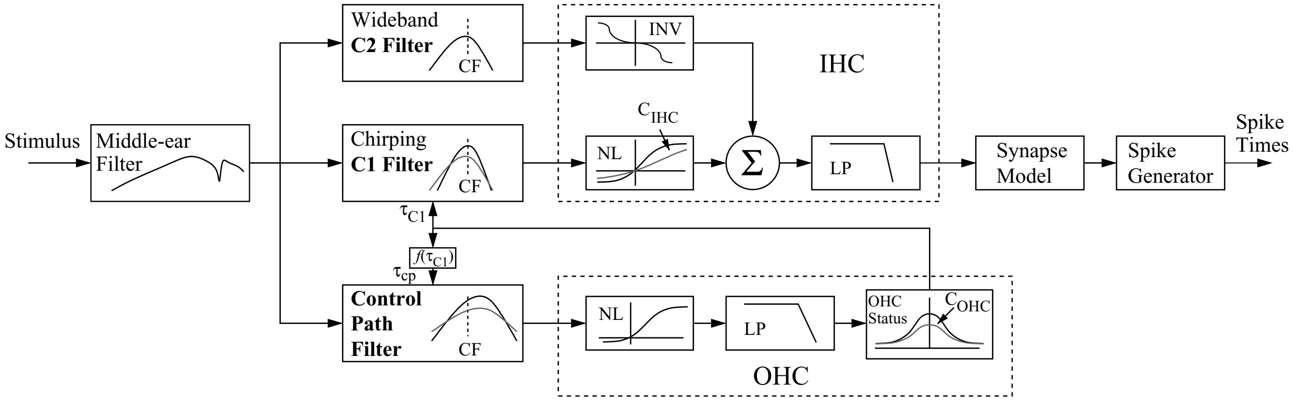

Fig. 1. Schematic diagram of the AN fiber model, reprinted from [7] with permission. The input to the model is an instantaneous pressure waveform of the stimulus in Pascals and the output is the spike times in response to that input. The model has a middle-ear filter, a feed-forward control-path, a signal-path 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.

model of the normal and impaired auditory periphery [7]. This model features a number of important phenomena seen in auditory-nerve (AN) fiber responses at high presentation levels, such as the elevation, broadening and frequency-shift in tuning, the component-1/component-2 (C1/C2) transition, and peak splitting [7]. These effects may be important in predicting the effects of presentation level on speech intelligibility. In addition, outer and inner hair cell impairment can be adjusted to simulate different degree of hearing loss, which enables this model-based STMI to predict the effects of impairment on speech intelligibility. In this paper, effects of presentation levels on speech intelligibility have been predicted for a wide variety of experimental conditions (such as in quiet or in background noise, broadband or filtered stimuli, unaided or aided conditions) and for listeners with different degree of hearing loss.

II. METHOD

A. Model of the Auditory-periphery

The auditory-periphery model has been developed by Zilany and Bruce [7], and is capable of generating realistic response properties of the AN fibers in cats across a wide range of characteristic frequencies (CFs) and intensities spanning the dynamic range of hearing. The schematic diagram of the model is illustrated in Fig. 1. Each section of the model provides a phenomenological description of the major functional components of the auditory-periphery, from the middle ear (ME) to the auditory nerve.

The first section models the filtering properties of the ME, which affects the relative levels of the components of the input wide-band stimuli, and hence plays an important role in simulating responses to multi-component stimuli such as vowels. The input to the ME is the instantaneous pressure waveform of the stimulus in Pa (pascal) sampled at 500 kHz. The ME filter is followed by a signal-path C1 filter which sets up the baseline tuning for the AN fibers. A feed-forward control path regulates the gain and bandwidth of the C1 filter to account for several level-dependent properties in the cochlea. The C1 filter has been designed in such a way that it can address a range of realistic response properties of the cochlea. A parallel-path C2 filter has been introduced as a second mode of excitation to the IHC and is critical for simulating the transition region effects at high levels. The C2 filter is linear, static, and is the same as the C1 filter at high-levels in the normal cochlea or the completely OHC-impaired version in the damaged cochlea, i.e., the tuning of the C2 filter is same as the broadest possible tuning of the C1 filter. The two filters (C1 and C2) are followed by two separate transduction functions, referred to as the C1 and C2 transduction functions. The summed output of the two transduction functions is then passed through a seventh-order IHC low-pass filter with a cut-off frequency of 3.8 kHz that describes the fall-off in pure tone synchrony with CF above 1 kHz. The IHC output drives the IHC-AN synapse which provides the instantaneous synaptic release rate as output. Finally the AN discharge times are produced in the model by a renewal process that includes refractory effects. More details of the model can be found in [7], and model code is available from the authors on request.

To compute the STMI, 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 to be analyzed by the central auditory system. 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. It is to be noted that 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. Consistent with the physiological observation, the number of low and medium rate fibers remains almost unaltered in the impaired case.

B. Model of the Central Auditory System

This stage analyzes the AN neurogram to separate different features and cues associated with different sound percepts. More specifically, this stage estimates the spectral and temporal modulation content of the AN neurogram. There is physiological and psychoacoustical evidence that the auditory system, particularly at the level of primary auditory cortex (AI), analyzes the dynamic acoustic spectrum of the stimulus (extracted at its earlier stages) by employing arrays of so-called spectro-temporal response fields (STRFs). In fact, each STRF acts as a modulation-selective filter of its input neurogram, and thus summarizes the way a cell responds to a stimulus. To calculate the STMI, this stage has been 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 STMI is a measure of speech integrity as viewed by a model of the auditory system. In other words, 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.

1) Computing the STMI: The AN fibers outputs for 128 CFs spaced logarithmically in the tonotopically organized cochlea are analyzed by the bank of modulation filters. The temporal rates of the filters range from 2 to 32 cyc/sec (Hz), and the scales are in the range from 0.25 to 8 cyc/oct, which covers the range of perceptually important spectro-temporal modulations available in speech for human. 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. The details of the implementation issues are available in [6].

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 computed as:

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

where $\|\cdot\|$ indicates the 2-norm of the corresponding signal.

2) Differences from the study by Elhilai et al. [6]: Although this work has been based on the study by Elhilali et al. [6], there exists some differences in the method leading to the calculation of the STMI. First, the AN model employed in this paper is a more complete and physiologically-accurate model, meaning that it has incorporated almost all of the nonlinearities seen in the AN fiber responses. Thus, the effects of presentation level on speech intelligibility could easily be investigated with this model, whereas the model of the early auditory processing in [6] cannot explain any effects of level or cochlear impairment. Second, in Elhilali et al. [6], 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. Third, the equation employed to calculate the STMI here is the square root of the expression used in [6]. Fourth, a lateral inhibitory network (LIN), between the auditory-periphery and the auditory cortex, was used in [6], which is not included in this present work. Fifth, consistent with the physiological and anatomical observations, AN fibers with different spontaneous rates have been considered.

III. Results

This section presents the effects of presentation level and impairment in the cochlea on the prediction of speech intelligibility using the model-based STMI approach. Several experimental conditions have been considered: normal listeners in quiet and noisy conditions, and impaired listeners in unaided and aided conditions. For comparison, word recognition scores found experimentally have been shown along with the STMI predictions.

A. Effects of Presentation Levels for Listeners with Normal Hearing

1) In Quiet: Word recognition scores for normal listeners in a quiet condition show larger and more consistent decreases (i.e., greater rollover) at high levels for highpass-filtered speech than for lowpass-filtered speech [4]. Molis and Summers [4] 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. To avoid ceiling effects, adaptive tracking was used to determine the lowpass and highpass cut-off frequencies that resulted in around 70%-correct keyword recognition. The broadband speech levels were varied from 75 to 105 dB SPL in steps of 10 dB SPL. They observed that word recognition scores for highpass-filtered sentences declined more consistently than the decrease in the recognition of the lowpass-filtered sentences. Similar behavior has been observed by Studebaker and Sherbocoe [8] in the performance-intensity function for narrow-band stimuli. As the presentation level increases, high-frequency speech shows greater rollover than low and mid-frequency speech. The speech test stimuli in this case were digitized copies of recordings of the NU6 monosyllabic words test (Auditec, St. Louis). So, the frequency dependent rollover can be generalized from single words to sentences and from narrow-band to highpass- and lowpass-filtered stimuli.

In this paper, 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. Figure 2 shows the averaged STMI for both highpass- and lowpass-filtered sentences along with the experimental word recognition scores in percent correct. Compared to the word recognition scores reported in [4], the results are qualitatively similar.

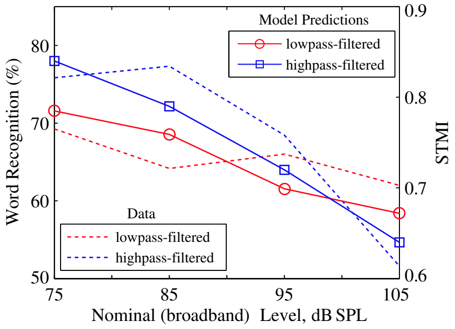

Fig. 2. Averaged word recognition performance for normal listeners (dotted lines) from Fig. 1 of [4] and STMI (solid lines with symbols) versus presentation level for lowpass- and highpass-filtered sentences.

The greater rollover observed at high-frequencies is consistent with the cochlear processing that shows more level dependence in basal regions tuned to high frequencies than in the apical, low frequency regions [9]. 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. 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 [10]. These two model properties could explain the observed larger rollover at high levels for highpass-filtered speech materials.

2) In Noisy Conditions: Speech recognition in background noise decreases at presentation levels above conversational levels, even when the SNR is held constant. The decline is greater when speech is presented in a masker with a spectrum that matches the spectrum of the speech [1]–[3]. Dubno et al. [2], [3] 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. Word recognition declined significantly with increasing level, even when the SNR was held constant. From the audibility estimates based on the articulation index, this decrease was attributed to the nonlinear growth of masking that effectively reduces the SNR at high speech-shaped masker levels. Masked pure tone thresholds measured in the speech-shaped maskers increased linearly with increasing masker level at lower frequencies but nonlinearly at higher frequencies, consistent with the nonlinear growth of upward spread of masking that followed the peaks in the spectrum of the speech-shaped masker.

In this paper, 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. The averaged STMI for three SNRs at three different speech levels are shown along with the experimental word recognition scores in Fig 3. Compared to the experimental word recognition scores reported in [2], the results are qualitatively similar.

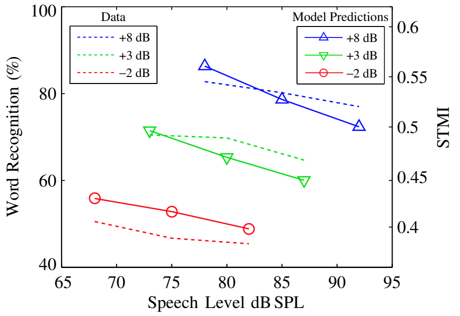

Fig. 3. Averaged word recognition performance for normal listeners (dotted lines) from Fig. 4 of [2] and STMI (solid lines with symbols) versus presentation level at three SNRs (+8, +3, and −2 dB) for broadband speech from the NU6 word lists.

B. Effects of Presentation Levels for Listeners with Hearing Loss

Shanks et al. [5] studied the performance of word recognition for listeners with different degree and slope of 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. Groups 1 and 2 had pure tone threshold averages <40 dB HL, and groups 3 and 4 had >40 dB HL. Groups 1 and 3 had slopes >10 dB/octave, and groups 2 and 4 had >10 dB/octave. Performance was evaluated on the connected speech test (CST, Hearing Aid Research Laboratory at the University of Memphis) for three hearing aid circuits: peak clipping, compression limiting, and wide dynamic range compression. Speech stimuli were presented at three presentation levels (52, 62 and 74 dB SPL), and performance was evaluated at three signal to babble (S/B) ratios (3, 0 and −3 dB). The reference signal to babble ratio (0 dB) is defined as the babble level that resulted in 50% performance in word recognition, and is assumed constant for all subsequent experiments.

In this work, we have simulated four example impairments each representing the degree and slope of hearing loss from one of the four groups. To compute the STMI, similar experimental conditions have been applied.

1) Unaided Condition: Shanks et al. [5] showed that in the unaided condition, performance of word recognition in multi-talker babble declined slightly with increased presentation levels for the mild and moderately impaired listeners (groups 1 and 2), but increased substantially for severely impaired listeners (groups 3 and 4). Fig. 4 shows the mean unaided CST word recognition performance versus presentation level for four groups of impaired listeners. Dotted lines indicate experimental performance in rationalized arcsine units (raus), and the model-based STMI are shown by the solid lines with symbols. It is clear that the model-based STMI are qualitatively same as the experimental observations. For listeners with mild and moderate hearing losses, STMI decreases slightly with increasing presentation level, whereas for severely impaired listeners, STMI increases sharply as the audibility for these listeners becomes substantial at higher levels.

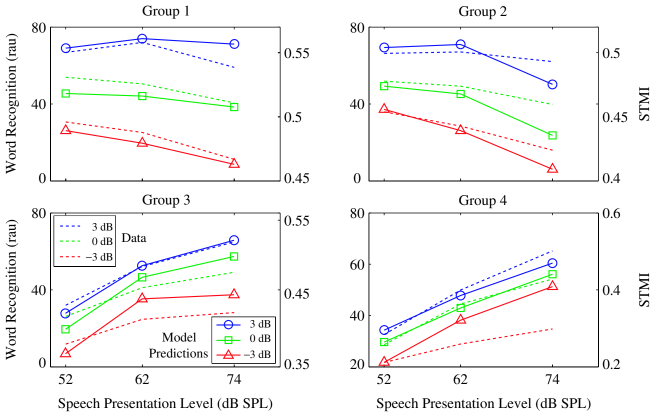

Fig. 4. Mean unaided CST word recognition scores (dotted lines) in rationalized arcsine units (raus) from Fig. 3 of [5] and STMI (solid lines with symbols) as a function of presentation level for four groups of listeners with different degree and slope of hearing loss. CST stimuli were presented at three levels (52, 62 and 74 dB SPL) and in background multi-talker babble corresponding to three S/B ratios (3, 0 and −3 dB).

2) Aided Condition: In the aided condition, Shanks et al. [5] compared the performance for three hearing aid circuits: peak clipping, compression limiting, and wide dynamic range compression. They observed that all three hearing aid circuits provided benefit over the unaided condition in both quiet and background noise. However, the performance of all circuits is similar and is thus averaged over the three hearing aid circuits. Dotted lines in Fig. 5 show the mean aided CST word recognition performance as a function of presentation level for the four groups of impaired listeners. It has been found that in aided conditions, speech recognition performance declined with increasing speech levels for nearly all impaired listeners.

We have simulated the same experimental conditions, and the model-based STMI is computed for the NAL-R (National Acoustic Laboratory Revised) prescription applied to the peak-clipping hearing aid circuit for four representative groups of impaired listeners. Model predictions are shown by the solid lines with symbols in Fig. 5. Compared to the experimental observations, model-based STMI also declines in all cases with increasing presentation levels. Furthermore, the aided STMI in all four groups improves over the unaided conditions at the lowest presentation level (52 dB SPL in this case), consistent with the data.

IV. 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 [11].

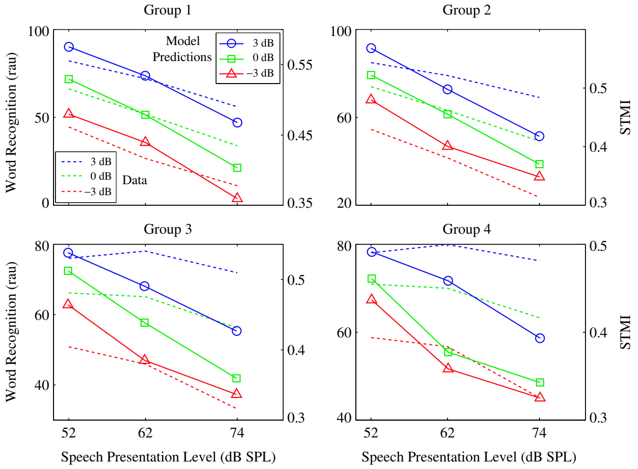

Fig. 5. Mean aided CST word recognition scores (dotted lines) in raus collapsed across the three hearing aid circuits (from Fig. 6 of [5]) and STMI (solid lines with symbols) as a function of presentation level for four groups of listeners with different degree and slope of hearing loss. CST stimuli were presented at three levels (52, 62 and 74 dB SPL) and in background multi-talker babble corresponding to three S/B ratios (3, 0 and −3 dB). The STMI is computed for the NAL-R prescription applied to the peak-clipping hearing aid circuit only.

V. ACKNOWLEDGMENTS

This research was supported by NSERC Discovery Grant 261736 and the Barber-Gennum Chair Endowment.

REFERENCES

[1] G. A. Studebaker, R. L. Sherbecoe, D. M. McDaniel and C. A. Gwaltney, “Monosyllabic word recognition at higher-than-normal speech and noise levels,” *J. Acoust. Soc. Am.*, vol. 105(4), pp. 2431–2444, 1999.

[2] 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.

[3] J. R. Dubno, A. R. Horwitz and J. B. Ahlstrom, “Recognition of filtered words in noise at higher-than-normal levels: decreases in scores with and without increases in masking,” *J. Acoust. Soc. Am.*, vol. 118(2), pp. 923–933, 2005.

[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] 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.

[6] 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.

[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(3), pp. 1446–1466, 2006.

[8] G. A. Studebaker and R. L. Sherbecoe, “Intensity-importance functions for bandlimited monosyllabic words,” *J. Acoust. Soc. Am.*, vol. 111(3), pp. 1422–1436, 2002.

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

[10] 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 /ε/ in cat auditory nerve,” *Hear. Res.*, vol. 123, pp. 61–77, 1998.

[11] 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.