Hearing Aid Research

Sponsor
National Institutes of Health Grants R01 DC00117, R01 DC007152, AFOSR Contract No. FA9550-05-C-0032.

Academic and Research Staff
Professor Louis D. Braida, Dr. Joseph Desloge, Dr. Raymond Goldsworthy, Dr. Karen L. Payton, Dr. Charlotte M. Reed

Visiting Scientists and Research Affiliates
Dr. Paul Duchnowski, Dr. Oded Ghitza, Dr. Kenneth W. Grant, Professor Ying-Yee Kong, Professor Jean C. Krause, Dr. Peninah S. Rosengard

Graduate Students
David Messing, Michael Naber, Mona Shretha,

Undergraduate Students
Ian Smith

Technical and Support Staff
Loraine Delhorne, Denise Rosetti

Our long-term goal is to develop improved hearing aids for people suffering from sensorineural hearing impairments and cochlear implants for the deaf. Our efforts are focused on problems resulting from inadequate knowledge of the effects of various transformations of speech signals on speech reception by impaired listeners, specifically on the fundamental limitations on the improvements in speech reception that can be achieved by processing speech. Our aims are

To develop and evaluate analytical models that can predict the effects of a variety of alterations of the speech signal on intelligibility.

To evaluate the effects of style of speech articulation and variability in speech production on speech reception by hearing impaired listeners.

To assess the relative contributions of various functional characteristics of hearing impairments to reduced speech-reception capacity.

To develop and evaluate signal processing techniques that hold promise for increasing the effectiveness of hearing aids.

Studies and Results

I-A. Role of Reduced Audibility

This research is concerned with analyzing the factors responsible for poor speech reception by listeners with hearing impairments, and with developing techniques for overcoming these degradations. To the extent the research is successful, it will help determine design goals for improved wearable hearing aids, establish new criteria and techniques for aural rehabilitation, and contribute to improved understand of both residual auditory function and speech perception.

We have made substantial progress over the past year toward our research goal of developing a greater understanding the role of reduced audibility in the speech-reception performance of listeners with moderate-to-profound degrees of hearing loss.

Progress over the past year is reported in three major areas: (1) the role of audibility in understanding the speech and psychoacoustic abilities of listeners with cochlear hearing impairment; (2) the effects of age on auditory gap-duration discrimination; and (3) a critical review of the literature on the role of audibility in the temporal and intensive abilities of listeners with cochlear hearing loss.

I-B-1. Role of Audibility in Speech and Psychoacoustic Performance of Listeners with Cochlear Hearing Loss

This research is aimed at providing direct tests of the role of audibility in explaining the speech-reception and psychoacoustic performance of listeners with moderate-to-profound hearing impairments. Individual hearing losses are simulated in age-matched normal-hearing listeners using two techniques to produce the desired shifts in hearing threshold and the effects of loudness recruitment present in cochlear loss: addition of a spectrally-shaped masking noise and multi-band expansion. Stimuli presented to hearing-impaired and simulated-loss listeners are thus equated in stimulus level specified in both dB SPL and dB SL. Any differences in performance observed on speech-reception tests between hearing-impaired and normal-hearing listeners can then be ascribed to supra-threshold deficits associated with hearing impairment. A battery of psychoacoustic measurements is employed to determine the source of any such suprathreshold components to speech-reception performance.

During the past year, we have made progress in implementing the experimental protocols for speech and psychoacoustic testing and in initiating data collection on hearing-impaired listeners and age-matched normal-hearing listeners with simulated hearing loss.

Development of Experimental Protocols. Experiments are controlled by a Dell 1.0 GHz Pentium III PC equipped with a high-quality 24-bit LynxOne sound card (Lynx Studios). The primary experimental engine used to generate and adaptively modify the experimental stimuli is the Alternative Forced Choice (AFC) program suite for Mathworks MATLAB (developed at the University of Oldenburg, Germany). Seven basic experimental protocols have been developed for data collection, including measurements of speech reception, spectral and temporal resolution, cross-modality processing ability, and cognitive ability.

Specifically, we are collecting data on the following tasks:

(i) Speech-Reception Thresholds (SRT) for HINT sentences (Nilsson et al., 1994) in steady-state or temporally-interrupted background noise at two overall levels (65 and 80 dB SPL) and three types of speech processing [unprocessed, linear-gain hearing aid with NAL-PR prescription (Byrne and Dillon, 1986), and amplitude-compression hearing aid (Goldstein et al., 2003)] ;

(ii) Pure-tone threshold detection at frequencies of 250, 500, 1000, 2000, 4000, and 8000 Hz and durations of 10 and 500 ms in quiet and in background noise;

(iii) Notched-noise masking, which tests the detection of 200-msec probe tones (250, 500, 1000, 2000, and 4000 Hz) in a 220-ms notched-noise simultaneous masker consisting of two bands of noise (one located above and the other below the probe signal, each with bandwidth of 0.25 times the frequency of the probe tone) selected to create notch widths ranging from 0 to 0.8 times the frequency of the probe tone;

(iv) Tone-on-tone forward masking, examining the detection of a 10-msec probe (500, 1000, 2000, and 4000 Hz) using 110-msec maskers (at frequencies of 0.55, 1.0 and 1.15 times the probe frequency) and five values of delay time between offset of masker and onset of probe (0, 10, 20, 60, and 100 ms);

(v) Temporal-modulation detection in a 500-msec broadband noise at each of ten modulation frequencies (ranging from 2 to 1024 Hz);

(vi) Temporal gap detection for auditory and tactual stimuli using 250 and 400 Hz leading and trailing markers with a nominal duration of 100 ms and a reference gap of 6.4 ms; and

(vii) A Reading-Span test (Ronnberg et al.,1999) which examines the subject's ability to retain the final word of orally-read sentences in lists of nonsense and sensible sentences ranging in set size from 2 to 5.

Subjects. We have completed data collection on five hearing-impaired subjects with a variety of audiometric configurations and severity of loss and with an age range of 24 to 69 years. The hearing loss of each of these hearing-impaired subjects is then simulated in a group of three normal-hearing listeners who are matched in age to the impaired listener (plus or minus five years). Thus far, we have collected full sets of data on 11 of the 15 planned simulated-loss listeners for the first five hearing-impaired subjects.

Data Analysis and Results. For each of the experimental paradigms described above, data-analysis software has been prepared that allows us to examine the performance of individual subjects and to compare performance between a given hearing-impaired listener and the simulated-loss subjects. Using quantitative methods of comparison on each task, we have begun to relate patterns of performance on the speech-reception task to basic psychoacoustic and cognitive abilities.

I-B-2. Effects of Age on Auditory Gap-Discrimination Ability

Based on previous results indicating a deterioration in temporal-processing ability with age (e.g., Lister et al., 2002), we examined the effect of age on the ability of listeners with clinically normal hearing to discriminate the duration of gaps in third-octave bands of noise (center frequencies of 500, 1000, and 2000 Hz) for two baseline values of gap duration (0 and 250 ms) and for both same-frequency and frequency-disparate leading and trailing markers. Gap-duration difference limens were measured in three groups of listeners with mean age of 28, 48, and 62 years, respectively. Unlike previously reported results, our study did not reveal a clear effect of age on the gap-duration discrimination ability for either of the two baseline gap durations or for any of the leading and trailing marker combinations. We plan to extend this research by increasing both the age range of the subjects and the number of subjects within each age category.

I-B-3. Review of Past Research on the Role of Audibility in Predicting Effects of Hearing Impairment

Although supra-threshold effects of hearing impairment are widely believed to be related to the decreased resolution on psychoacoustic tasks and the poorer speech-reception abilities of hearing-impaired listeners, the role of reduced audibility itself in explaining the consequences of hearing loss is as yet not completely understood. During the past year, we have completed a critical review of research on temporal resolution in listeners with hearing impairment. Within the temporal domain, audibility effects appear to play a large role in certain tasks (such as gap detection and tone detection in modulated noise) but not in others (including temporal integration). A manuscript reviewing the role of audibility on intensity perception is also in preparation.

1-C. Significance

Our research is concerned with analyzing the factors responsible for poor speech reception by listeners with hearing impairments, and with developing techniques for overcoming these degradations. To the extent the research is successful, it will help determine design goals for

improved wearable hearing aids, establish new criteria and techniques for aural rehabilitation, and contribute to improved understand of both residual auditory function and speech perception.

I-D Plans for the Coming Year

Area 1: Role of Audibility in Speech Reception. We will continue this research through: collection of data on an additional 5 listeners with hearing impairment and 15 age-matched normal-hearing controls with simulated hearing loss; analysis of results including quantitative comparisons between hearing-impaired and simulated-loss subjects and correlation of speech performance with psychoacoustic abilities; and preparation of manuscripts for publication.

Area 2: Integration Across Frequency. We will process CVC nonsense syllables using linear filtering and noise for use in consonant-identification experiments examining performance in each of two filtered bands separately and in combination. Testing will be conducted on four normal-hearing listeners, two listeners with hearing impairment and six normal-hearing listeners with simulated hearing loss.

II-A. Models of Speech Intelligibility

This section discusses our research on cochlear implants as well as our research concerned with developing a model of speech intelligibility.

II-B Cochlear Implant Research

In this research project, we are collecting performance measures from cochlear implant recipients on both psychoacoustic and speech measures. For the psychoacoustic measures, subjects listen to a number of sounds and are asked to make a decision as to which sound is higher in pitch (or which sound is louder). We have tested nine adult subjects so far, and will begin testing of children in May. The psychoacoustic tests we are currently conducting consist of 4 measures: pure tone pitch discrimination, pure tone loudness discrimination, tone complex pitch discrimination, and a musical interval test. Each of these tests is administered using a two alternative forced choice procedure. The procedure adaptively converges to a level where subjects correctly discriminate 70% of the sounds correctly. The tests are given in a sound proof booth using a computer and custom developed software. Subjects are tested on speech using a vowel and a consonant database. The vowel database consists of 12 vowels spoken by 5 different males and females for a total of 120 tokens. The consonant database consists of 20 consonants spoken by 5 different males and females for a total of 200 tokens. Subjects are presented with a screen with either the 12 vowels or 20 consonants and listen to the tokens presented in random order. They are asked to use respond to what they hear. The tests have been developed so that noise can be adaptively controlled in the background as well.

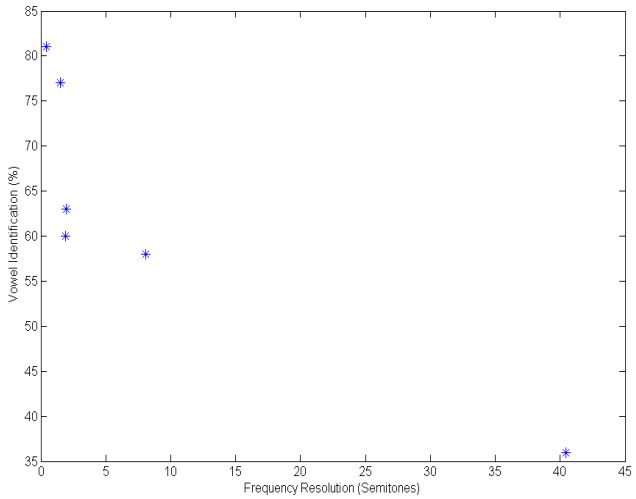


Figure II-B-1: Average reception scores on the vowel test versus average frequency resolution scores on the pure tone test.

Our initial data collected supports the hypothesis that psychoacoustic and speech perception are related. The subject's exhibiting the highest speech reception scores also exhibited the best pitch perception. Fig. II-B-1 shows the average reception scores on the vowel test versus the average frequency resolution on the pitch test. While only six subjects have been tested, it is clear that the one poor performer of the group on the vowel test also had exceptional difficulty in pure tone frequency discrimination.

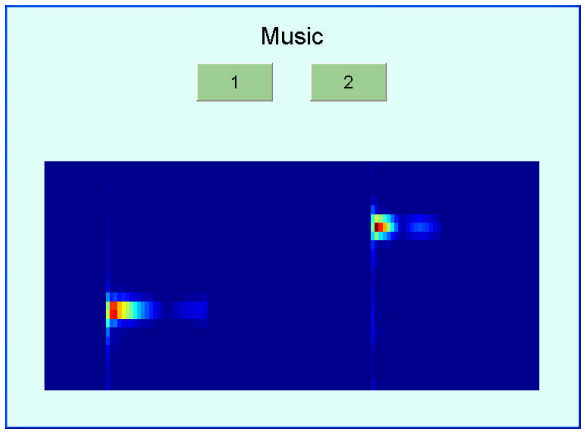


Figure II-B-2: A feedback tool for listening to the differences between musical notes.

We have begun a series of software modifications that will allow subjects to practice the psychoacoustic tasks on a personal computer at home. We have developed a protocol designed to evaluate the potential benefits of specific training on pitch and loudness differences. The software modification allows the user to have a visual cue associated with the sound that they hear. An example is given in Fig. II-B-2. The hearing impaired individual would sit and listen to two sounds, either pure tones or musical instruments, and either the pitch or the loudness would be adaptively changed as in the test. But the visual feedback would reinforce the pitch or

loudness difference, and hopefully allow the impaired individual to improve performance over time. It is a simple tool, but also provides tools for monitoring progress over time. The training tool can also be used for speech practice, but our initial protocol will investigate training using simple sound stimuli and use the speech measures for independent assessment.

II-C STI Based Models of Speech Intelligibility

We focused primarily on evaluating the short-term metrics we developed last year.

Fig. II-C-1 below shows the asymptotic performance of the 3 metrics evaluated as a function of window length for speech in stationary speech-shaped noise at 0dB SNR. Note that the Envelope Regression method is identical to the theoretical method (based on SNR in 7 octave bands) for speech in noise and both averages effectively reach the long-term value for window lengths greater than 4 or 5 seconds. All three metrics behave differently when the window length is reduced below 1/3 s.

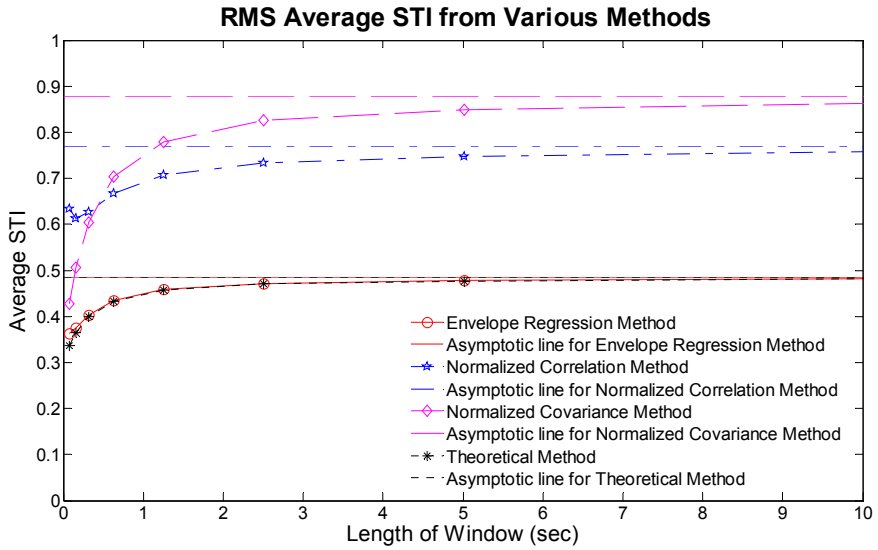


Figure II-C-1. Average STI vs window length for the three short-time metrics and the theoretical method; all compared with their long-term asymptotes for 0dB SNR speech in stationary speech-shaped noise.

In addition, regression analyses were performed on each metric for long (~2.5 s), moderate (1/3 s) and short (~80 ms) window lengths. Figure 2 depicts those plots for the Envelope Regression method which had the best fit to the theoretical method over all the conditions tested.

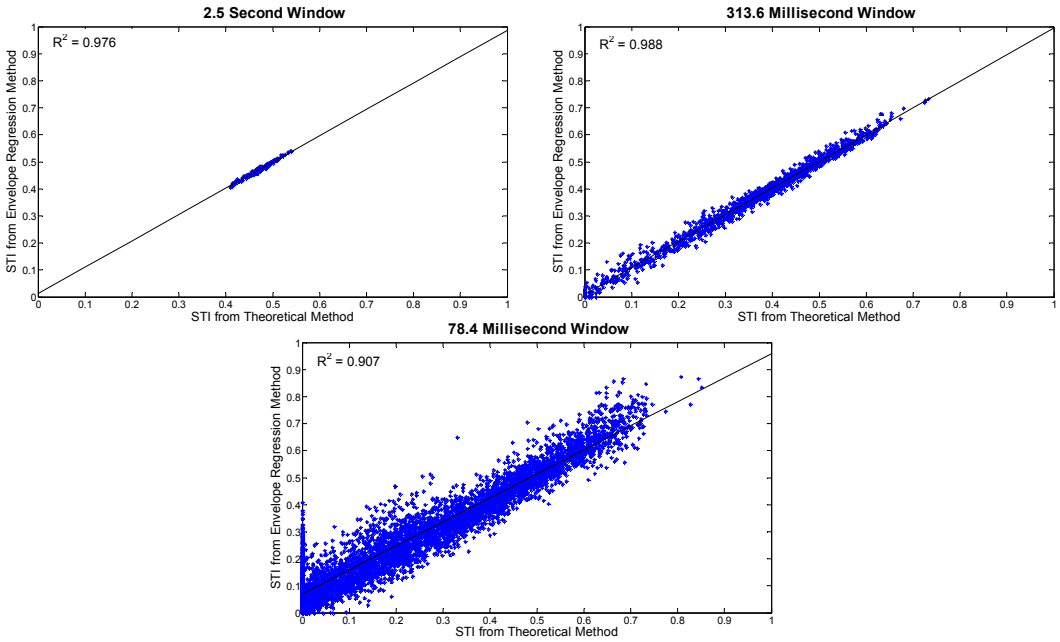
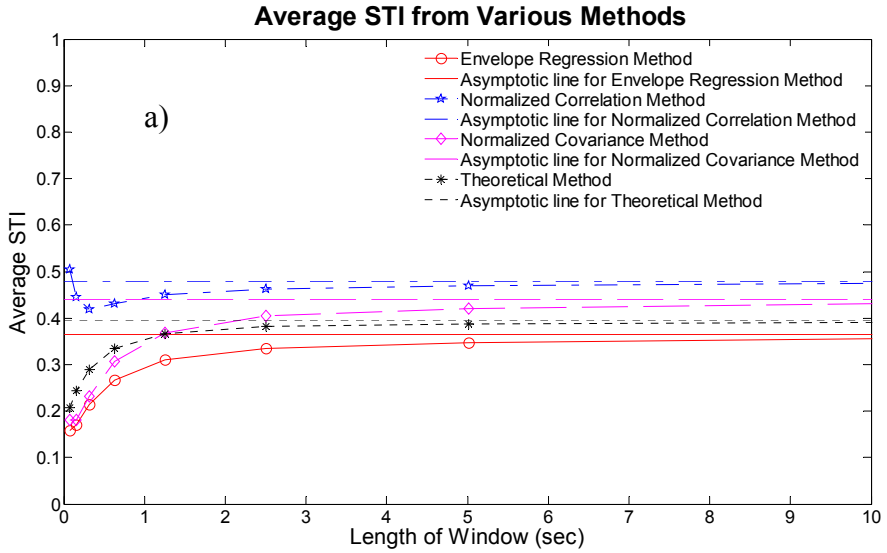


Figure II-C-2. Regression analysis for STI computed using the theoretical method as compared with the STI from the Envelope Regression method for the speech in stationary noise condition. In each case, the R^2, goodness of fit, term is greater than 0.9 and greater than 0.97 for the two longer windows.

The metrics were also evaluated for asymptotic behavior in speech-shaped noise plus reverberation and in the presence of multi-talker babble. Below, in Fig. II-C-3, are the asymptotic convergence plots for those two conditions, noise plus reverberation in the upper plot and multi-talker babble in the lower plot.



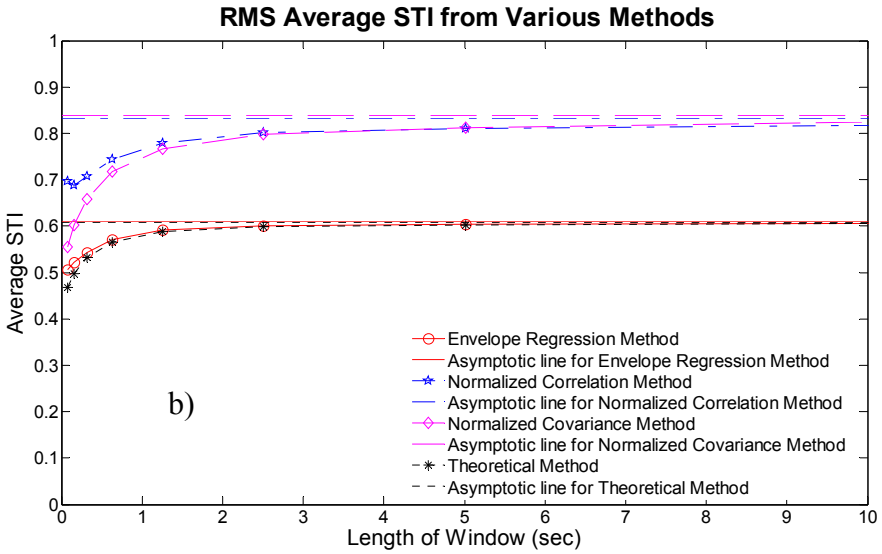


Figure II-C-3. Asymptotic performance of the three metrics in 0dB SNR stationary noise plus reverberation (top plot) and in 0dB SNR multi-talker babble (bottom plot) as compared to the theoretical method. Also plotted are the long-term asymptotes for each metric and the theoretical method.

It is interesting to note that while the ER method most closely matches the theoretical STI for all conditions in Fig. II-C-3, it deviates the most in the noise plus reverberation condition. This could be due, in part, to the fact that the theoretical method uses the formula provided by Houtgast et al. 1980 to predict the impact of reverberation, effectively assuming an exponential decay and no distinct echoes while the metrics are based on the actual speech in the simulated reverberant environment in which there were a few initial echoes.

The linear regression analysis for the ER method in the noise plus reverberation is shown below in Fig. II-C-4 for the shorter two window lengths: 1/3 sec and ~80 ms. Note that the R^2 goodness of fit for the longer window is now 0.785 and the R^2 for the shorter window is only 0.345.

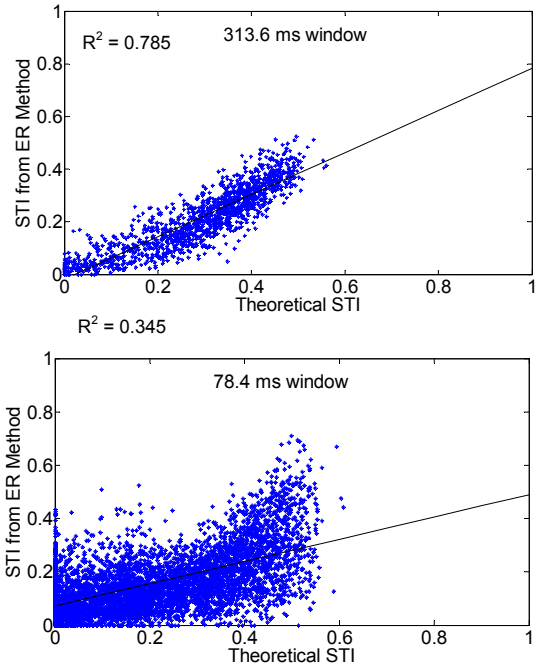


Figure II-C-4. Regression analysis for 0dB noise plus reverberation (RT_{60} = 0.6s) for two window lengths: 1/3 sec (upper plot) and ~80 ms (lower plot).

After this thorough analysis on a single-talker database, the new short-time metrics were then tested on two new talkers, speaking both conversationally and using “fast-clear” speech, speech which is articulated clearly but spoken at approximately normal speaking rates [Krause, 2001]. Both speaking styles were mixed with stationary speech-shaped noise at 01.4 dB SNR. Dr. Krause provided subject scores at both the sentence and word levels for the test materials. One important feature to point out is that, while the subject scores and STI values overlap for the two speaking styles, on average the fast clear sentences were both more intelligible and had higher STI values than the conversational sentences.

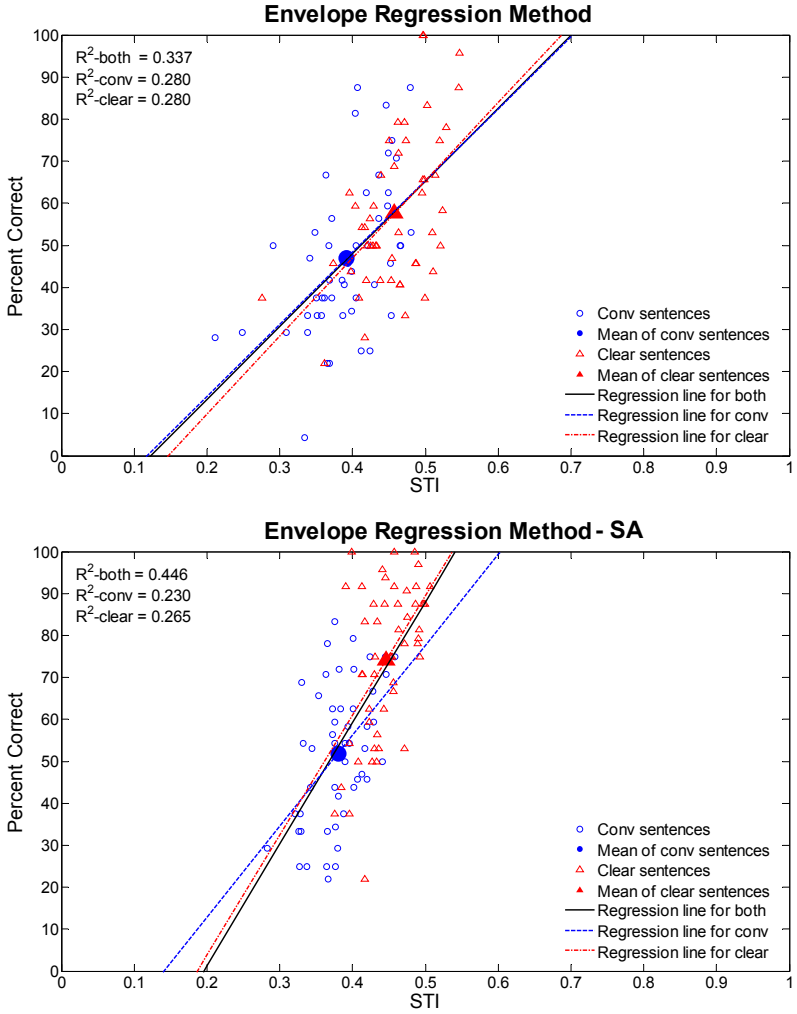


Figure II-C-5. Plots of STI using the ER method vs subject percent correct scores on sentences for the two speaking styles. Red symbols and lines correspond to clear sentences and blue symbols and lines correspond to conversational sentences. The upper plot shows the data for talker RG and the lower plot depicts the data for talker SA. Large symbols indicate means for all sentences of a given speaking style and talker.

The data also were examined for change in STI and percent correct going from conversational to clear. The results are shown in Fig. 6 below for the ER method. The vast majority of the sentences were more intelligible in the fast clear speaking style and they had higher STI values (large symbols indicate means of all sentences).

Figure II-C-6. Change in percent correct as a function of change in STI using the ER method for talkers RG (blue symbols) and SA (red symbols). The large symbols are the means across all sentences for each talker.

The data were also analyzed at the word level. Fig. II-C-7 shows how much each word in the initial and final positions change in STI and intelligibility going from conversational to fast/clear for each talker.

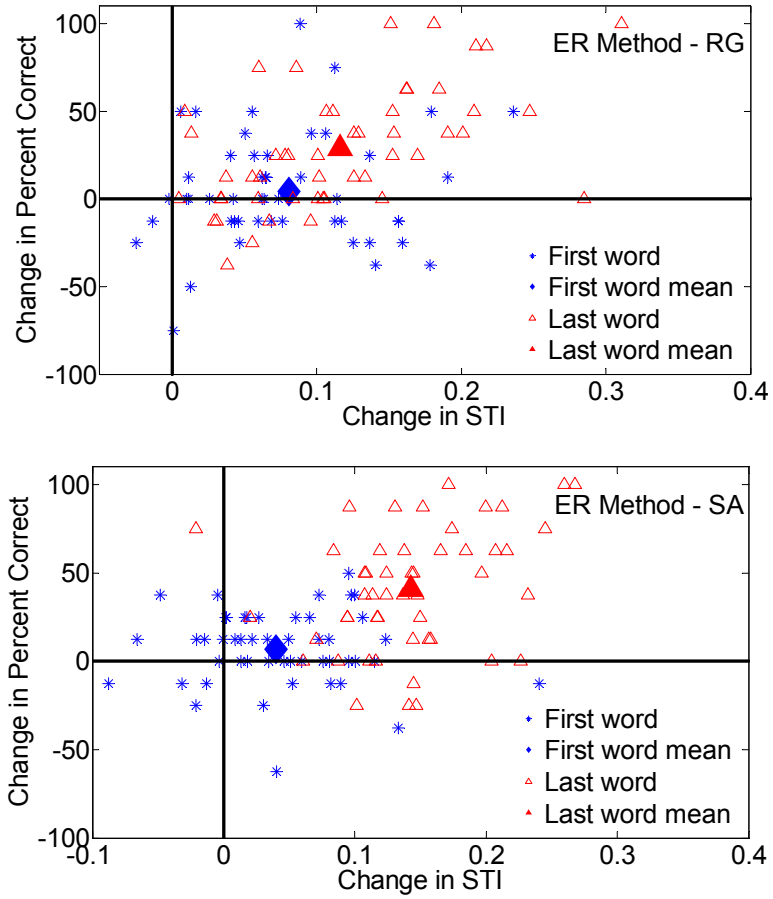


Figure II-C-7. Change in percent correct vs change in STI from conversational to fast/clear speech for words in the initial position (blue symbols) and final positions (red symbols). The upper plot is for talker RG and the lower plot is for talker SA. The large filled symbols indicate the means for each word position.

The most important thing to note about these results is the greater change in both STI and percent correct for the last words in the sentence over the first words in the sentence. This result is consistent with observations that one of the reasons conversational speech is less intelligible than either clear or fast/clear is that the voice level of the talker decreases toward the end of the sentence.

III-A. Application of Cortical Processing Theory to Acoustical Analysis

The goal of this research is to develop a machine which will use state-of-the-art non-linear peripheral auditory models (PAM) connected to a perceptually inspired model of template matching to (1) predict phonetic confusions made by normally-hearing listeners, and (2) predict intelligibility of distorted speech generated by passing naturally spoken speech through realistic communication systems.

Success in this project will contribute to and have significance for the following:

1) Revising models of auditory periphery by including the role of the descending pathway in making the cochlear response to speech sounds robust to degradation in acoustic conditions.

2) Establishing models of template-matching in the context of human perception of degraded speech. These models will provide guidance to physiological studies of cortical processing.

3) Enabling diagnostic assessment of speech intelligibility by using closed-loop models of the auditory periphery integrated with perception-based template matching.

4) Improving the performance of automatic speech recognition systems in acoustically adverse conditions.

We have developed a model of auditory speech processing capable of predicting consonant confusions by normal hearing listeners. In particular, we develop a phenomenological model of the Medial Olivocochlear efferent pathway. We then use this model to predict human error patterns of initial consonants in consonant-vowel-consonant words. In the process we demonstrate its potential for speech identification in noise. Our results produced performance that was robust to varying levels of additive noise and which mimicked human performance in discrimination of synthetic speech.

III-A-1. Introduction

Medial olivocochlear (MOC) efferent activity is believed to regulate the cochlear operating point depending on background acoustic stimulation, resulting in robust human performance in perceiving speech in a noisy background (e.g., Kiang et al., 1987). By reducing outer hair cell (OHC) motility and changing OHC shape, MOC stimulation increases basilar membrane stiffness, and in turn inhibits inner hair cell (IHC) response in the presence of noise. This paper develops this picture into a closed-loop model of the peripheral auditory system, a model that adaptively adjusts its cochlear operating point. Specifically we develop our model to attempt to match and predict human confusions of initial consonants in speech-shaped additive Gaussian noise.

Our long-term goal is to formulate a template-matching operation, with perception-related rules of integration over time and frequency at its core, in the context of human perception of degraded speech, but in this paper we concentrate on separating the backend development from the front-end. Our approach is to minimize the influence of cognitive and memory factors while preserving the complex acoustic cues that differentiate initial diphones. Hence we tune the parameters of the peripheral auditory model in as much isolation as possible by reducing the effect of the

backend system. Once the basic signal processing front-end of our model is tuned, we can then freeze the front-end and develop the back-end pattern recognition template matching system.

II-A-2 Peripheral Auditory Model (PAM)

We have developed a closed-loop model of the auditory periphery that was inspired by current evidence about the possible role of the efferent system in regulating the operating point of the cochlea. This, in turn, results in an auditory nerve (AN) representation that is less sensitive to changes in environmental conditions. In implementing the cochlear model we use a bank of overlapping cochlear channels uniformly distributed along the ERB scale, four channels per ERB. Each cochlear channel comprises a nonlinear filter and a model of the IHC (half-wave rectification followed by a low-pass filter, representing the reduction of synchrony with CF). The dynamic range of the simulated IHC response is restricted – both below and above – to a dynamic range window (DRW), representing the observed dynamic range at the AN level.

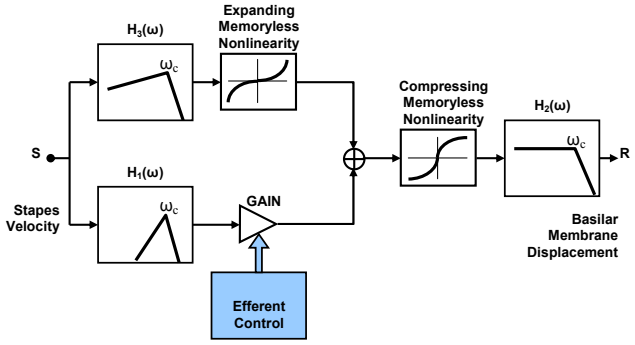


Figure III-A-1. MBPNL filterbank. A parameter GAIN controls the gain of the tip of the basilar membrane tuning curves. To best mimic psychophysical tuning curves of a healthy cochlea in quiet, the tip gain is set to GAIN =40dB (Goldstein, 1990)

The filter (Fig. III-A-I) is Goldstein's model of nonlinear cochlear mechanics (MBPNL; Goldstein, 1990). This model operates in the time domain and changes its gain and bandwidth with changes in the input intensity, in accordance with observed physiological and psychophysical behavior. The lower path (H1/H2) is a compressive nonlinear filter that represents the sensitive, narrowband compressive nonlinearity at the tip of the basilar membrane tuning curves. The upper path (H3/H2) is a linear filter (the expanding function preceded by its inverse compressive function results in a unitary transformation) that represents the insensitive, broadband linear tail response of basilar-membrane tuning curves. The gain parameter (GAIN) controls the gain of the tip of the basilar membrane tuning curves, and is used to model the inhibitory efferent-induced response in the presence of noise. For the open-loop (ie without adaptive feedback) MBPNL model GAIN is set to 40dB, to best mimic psychophysical tuning curves of a healthy cochlea in quiet.

As for the efferent-inspired part of the model we mimic the effect of the Medial Olivocochlear efferent path (MOC). Morphologically (e.g. Guinan, 1996), MOC neurons project to different places along the cochlea partition in a tonotopical manner, making synapse connections to the outer hair cells and, hence, affecting the mechanical properties of the cochlea (e.g. increase in basilar membrane stiffness). Therefore, we introduce a frequency dependent feedback mechanism which controls the tip-gain (G) of each MBPNL channel according to the intensity level of the sustained noise in that frequency channel.

The "iso-input" frequency response of an MBPNL filter at CF of 3600Hz with various tip gain settings is shown in Fig/ III-A-2. For an input signal $s(t) = A\sin(2\pi f_o t)$, with A and f_o fixed, the

MBPNL behaves as a linear system with a fixed "operating point" on the expanding and compressive nonlinear curves, determined by A. For a given A, a discrete "chirp" signal was presented to the MBPNL, with a slowly changing frequency. Changes in f_o occurred only after the system reached steady-state, for a proper gain measurement. The frequency response for the open-loop MBPNL model is shown at the upper-left corner (i.e. for GAIN = 40dB). Fig. III-A-2 shows the iso-input frequency response of the system for different values of input SPL level. As the input level increases the output gain drops and the bandwidth increases, in accordance with physiological and psychophysical behavior (Glasberg and Moore, 1990). As the gain increases, the distance between the maximum and minimum peaks, corresponding to inputs of 40 dBSPL and 120 dBSPL in Figure 2, increases.

In our closed-loop model (ie with feedback), the tip GAIN parameter is adjusted based on the efferent response, which in turn is calculated based on the amount of noise present. For our experiments, we limited the range that GAIN can vary, based on biological observations.

This adjustment of the GAIN parameter has several consequences. Besides making the energy of the noise at the output of each filter more consistent, it also affects the properties of each filter. The general effect is that loud noises reduce the non-linear amplification of small amplitude sounds while weak noises maintain the larger amplification of small amplitude sounds. Hence the overall effect of the efferent system in our model is to amplify small amplitude components of the speech stimulus by an amount that depends on the noise level. This point is illustrated in more detail in Fig. III-A-2. In this figure, the upper-left panel represents the nominal response (i.e. in quiet), with GAIN set to 40dB. In this quiet condition, weaker amplitude sounds such as the 40dB SPL sound are amplified greatly (in this case roughly 20 dB more) relative to louder sounds such as the 120 dB SPL stimulus. By increasing the efferent response in noise, we reduce the GAIN and the MBPNL response to weaker stimuli such as the 40 dB SPL tone (and background noise), as shown in the lower right pane of Fig. III-A-2 where the GAIN parameter is set to 10dB. Hence for high energy tone stimuli the MBPNL response is hardly affected, while the response for low energy stimuli (e.g. 40 or 60dB SPL signals) is reduced by some 30dB in the presence of noise.

Fig. III-A-3 shows – in terms of a spectrogram – simulated IHC responses to speech in noise for two conditions (SPL levels of 50dB or 70dB, and SNR of 10dB), for an open-loop system with a linear cochlear model (left-hand side) and for the closed-loop system (right-hand side). Due to the nature of the noise-responsive feedback, the closed- loop system spectrograms fluctuate with changes in background noise considerably less than are spectrograms produced by the open-loop model. This property is desirable for stabilizing the performance of the template-match operation with varying noise conditions, as reflected in the quantitative evaluation reported next.

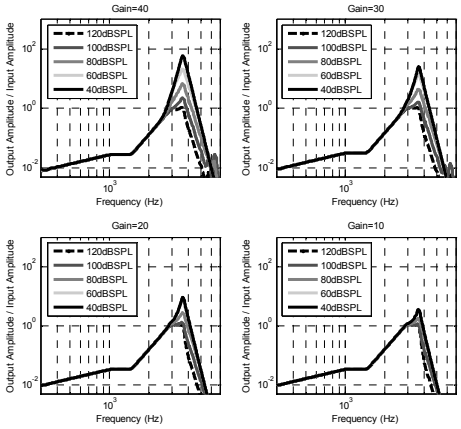


Fig. III-A-2. MBPNL frequency responses Iso-input frequency responses of an MBPNL filter (at CF of 3641Hz) for different values of GAIN parameter. From Upper-left, clockwise: GAIN =40, 30, 20 and 10dB. Upper-left corner (Gain=40dB) is for healthy cochlea in quiet (Goldstein, 1990). Input sinusoids are varied from 40 to 120 dB SPL.

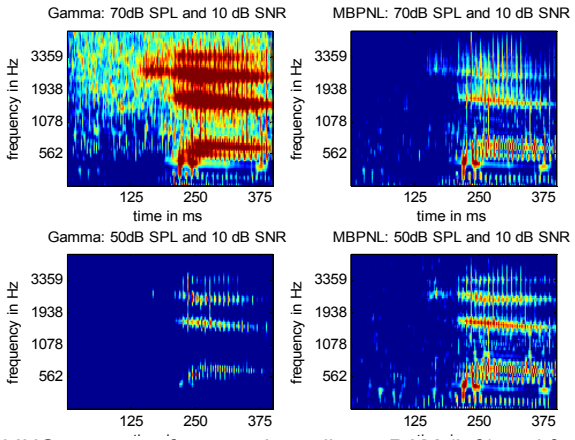


Fig. III-A-3 Simulated IHC response for open-loop, linear PAM (left) and for closed-loop PAM (right).

III-A-3. Quantitative evaluation – template-matching

Our long-term objective is to predict consonant confusions made by normally-hearing listeners, listening to degraded speech. Our prediction engine comprises the efferent-inspired peripheral auditory model followed by a template matching operation (using a distance measure between template and test tokens). Our focus in this worfk is to find the parameters of the first stage with a minimal interference of the second.

Ideally, to eliminate unwanted interaction between stages, errors due to template matching should be reduced to zero. In reality we could only try to minimize interaction by taking the following three steps: (1) we use the simplest possible psychophysical task in the context of speech perception, namely a binary discrimination test. In particular, we use Voiers' DRT (1983) which presents the subject with a two alternative forced choice between two alternative CVC words that differ in their initial consonants. Such task minimizes the influence of cognitive and memory factors while maintaining the complex acoustic cues that differentiate initial diphones (recall the central role of diphones in speech perception, e.g. Ghitza, 1993); (2) we use the DRT paradigm with synthetic speech stimuli. An acoustic realization of the DRT word-pairs was synthesized so that the target values for the formants of the vowel in a word-pair are identical, restricting stimulus differences to the initial diphones; and (3) we use a "frozen speech" methodology (e.g. Hant and Alwan, 2003): the same acoustic speech token is used for training and for testing, so that testing tokens differ from training tokens only by the acoustic distortion.

For these studies the amount of noise allowed over the lower bound of the DRW was set to 2dB, 6dB, or 10dB, with different combinations of level per frequency band. The frequency bands examined were divided roughly according to the first formant, second formant, and third formant regions for clean speech. Specifically, the first frequency band had channels with center frequency of 266 Hz to 844Hz; the second frequency band had channels with center frequency of 875 Hz to 2359 Hz; and the final frequency band examined had channels with center frequency of 2422Hz to 5141Hz.

A Chi-squared metric with a significance level of 95% based on contingency table analysis of data (Zar, 1999) was used to evaluate how closely machine performance matched that of humans, and to tune the front-end auditory model parameters. The settings that yielded the best match to human in the Chi-squared sense were a DRW lower bound of 65dB, with noise allowed per frequency band according to table 1, with stretching, and with a 10-ms window.

Frequency Band CF	Noise Above DRW Lower Bound
266-844 Hz	10 dB
875-2359 Hz	6 dB
2422-5141 Hz	6 dB

Table III-A-1. Noise allowed above the lower bound of the DRW per frequency band for the system with the best match to human.

Cumulative Chi-squared analysis per DRT dimension using a template token at 60dBSPL and 10dBSNR are shown in figures 4 and 5; for these tests the noise level and SNR or the test token was varied between SPL=70, 60, and 50dB and SNR=0, 5, and 10dB. These results suggest that the acoustic dimensions of voicing minus and sustention minus were significantly different from human for the majority of the conditions tested. When examining figure 4, the negative bars for the voicing minus and sustention minus categories imply that the machine is performing better than humans. The reason for this better machine performance is unknown; however it could be due to the simple pixel by pixel MSE computation of our backend. For the voicing category, timing differences between voiced and unvoiced sounds due to voiced onset times could make discrimination easier for the machine model and hence bias results. For the sustension category, continuants (such as /f/) which belong to the ST+ category tend to occur in initial consonants that are much more gradual and spread over time while obstruents (such as /p/) which belong to the ST- category are much more abrupt and compact over time. It is possible that these timing differences are over-emphasized by the nature of our simple MSE backend comparison on time-aligned speech, hence biasing performance in favor of the machine for these two categories.

All other DRT acoustic categories have cues that are less dependent on timing differences. Machine performance over these categories also matched humans much better with a few exceptions. The graveness plus category significantly differs for the 60dBSPL x 5dBSNR condition, and the graveness minus category significantly differs for the 50dBSPL x 10dBSNR condition.

Despite the differences for a few acoustic categories and for a few presentation conditions, the average Chi-squared metric of 2.37 suggests that on average, machine performance was close to human (and certainly within the Chi-squared significance level of 3.84).

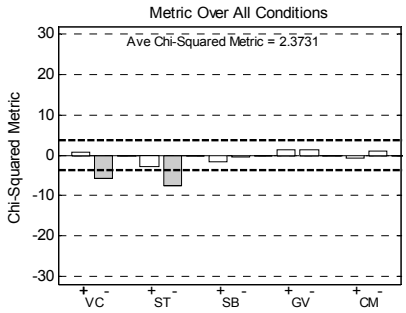


Figure III-A-4. Overall Chi-squared results for the system that yielded the best match to humans. Negative bars indicate human errors exceed that of machine. Positive bars indicate machine errors exceed that of humans. The absolute value of each bar is the Chi-squared value for that acoustic dimension. The performance on voicing-minus and sustension-minus categories is much better than that of human and significantly contributes to the overall Chi-squared metric. Grey bars indicate differences between machine and human that were statistically significant according to the Chi-squared test.

III-A-5. Discussion

We have described a model of the signal processing of the human auditory periphery and demonstrated how several of the non-linear operations taking place in the biological system can be used to develop a system that improves our capability to predict human performance in noise. One of the key non-linear interactions of our system that is regulated by efferent-inspired feedback control is that of the MPBNL gain versus the lower bound of the DRW. Besides affecting filter shapes in response of noise, this interaction aids in making the output more consistent. In part, this is due to the normalizing effect that efferent control has on the output: it makes outputs fall into the DRW of interest and be consistent across input levels. However it also yields a performance gain across SNR levels that traditional linear processing does not provide. At low noise levels, the gain is high, making the filters more responsive to small amplitude signals. This in turn amplifies small amplitude sounds such as some transients in consonants, which may be very useful for speech recognition in environments with low levels of noise. At high noise levels, the gain is low, making the filters much less responsive to small amplitude signals. Hence smaller short-time noise transients are attenuated and effectively masked below our DRW rate window of interest. At these higher noise levels, this noise masking effect allows the higher SNR regions of the speech signal to emerge from the noise background. This effectively yields an "unmasking" of sounds in noisy backgrounds, similar to the affect Ferry (2007) describes in his work and that of Dolan and Nuttal (1988), and Kawase et al. (1993)

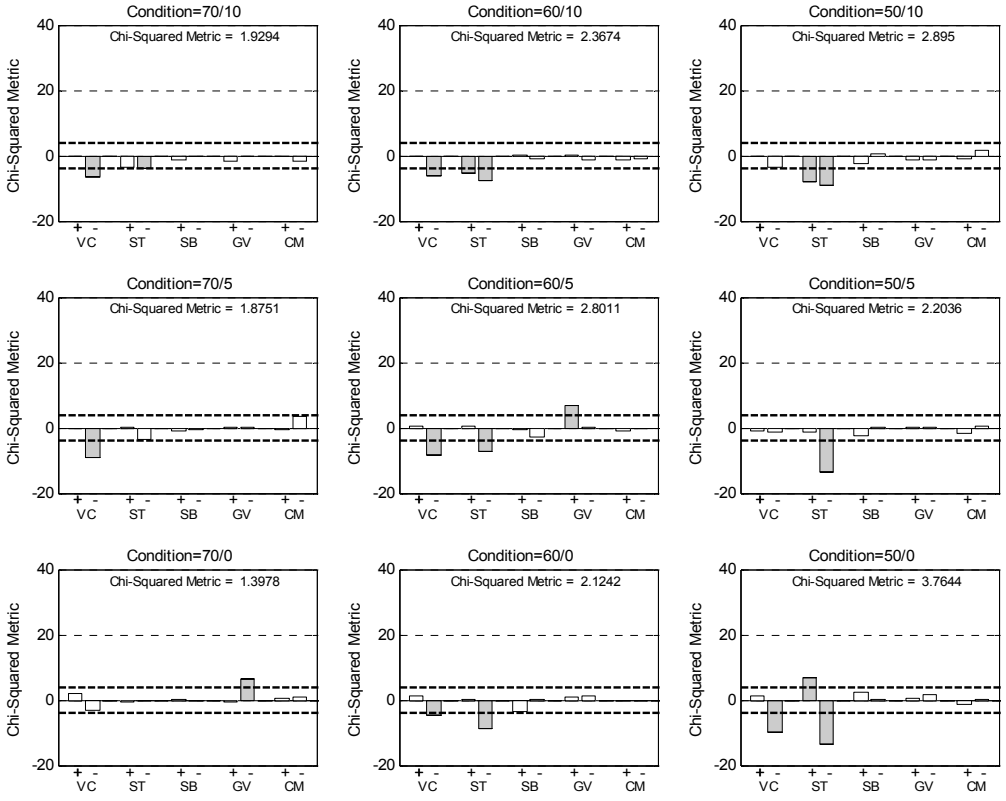


Figure III-A-5. Detailed Chi-squared metric results computed separately for each noise condition for the system that yielded the best match to humans. The noise condition is specificed in each panel by the SPL/SNR levels. The machine performance on a few acoustic dimensions, especially voicing-minus and sustension-minus, is significantly better than human performance. Overall the Chi-squared metrics here indicate that this system was a much better match than any other we had evaluated.

IV-A Novel Means of Controlling Psychoacoustic Experiments

We developed a novel means of controlling psychoacoustic experiments. Soundgen (Naber, 2008) is a web services based sound generation system. Sounds are produced by a dedicated server running Linux, MATLAB, Apache, and PHP. The sounds created by Soundgen are combinations of monaural or binaural tones and filtered noises. The characteristics of the tones and noises are passed to the Soundgen web service via a JSON object sent over HTTP. When the generation is complete, the web service replies with another JSON object containing the URL of the generated sound file. Accompanying the Soundgen web service is a small library, easing the service's use in JavaScript. This library allows JavaScript programmers to call a function, which triggers a callback function that executes when the request has been processed by Soundgen. The library and web service allow users to create portable psychoacoustic experiments as web-applications. The server can handle requests from multiple computers that are equipped only with a modern web browser and headphones or loudspeakers.

Publications

Chapters in Books

Ghitza, O., Messing, D., Delhorne, L., Braida, L., Bruckert, E., and M. M. Sondhi (2007). Towards predicting consonant confusions of degraded speech. In: Hearing – from basic research to applications (Eds.) B. Kollmeier, G. Klump, V. Hohmann, U. Langemann, M. Mauermann, S. Uppenkamp and J. Verhey, Springer-Verlag.

Meeting Papers

Presented

K. L. Payton, M. Shrestha (2008) "Analysis of short-time Speech Transmission Index algorithms," Proc. Acoustics'08, June 29-July 4, 2008, Paris, France

K. L. Payton, M. Shrestha (2008) "Evaluation of short-time speech-based intelligibility metrics," Proc. Int. Commission on Biol. Effects Noise 2008, July 21-25, 2008, Foxwoods, CT.

Theses

M. R. Naber, (2008). "Soundgen: A web services based sound generation system for the psychoacoustics laboratory," M. Eng Thesis, Dept .of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

References:

Byrne, D. and Dillon, H. (1986). "The National Acoustics Laboratory new procedure for selecting the gain and frequency response of a hearing aid," Ear and Hearing 7, 257-265.

Dolan, D. F., and Nuttal, A. L. (1988). "Masked cochlear whole-nerve response intensity functions altered by electrical stimulation of the crossed olivocochlear bundle," J. Acoust. Soc. Am. 83: 1081–1086.

Ferry, R., and Meddis, R. (2007). "A computer model of medial efferent suppression in the mammalian auditory system." J. Acoust. Soc. Am. 122(6): 3519–3526.

Goldstein, J. L. (1990). Modeling rapid waveform compression on the basilar membrane as a multiple-bandpass-nonlinearity filtering, Hearing Research, 49, 39-60.

Goldstein, J.L., Oz, M., Gilchrist, P.M., and Valente, M. (2003). Signal processing strategies and clinical outcomes for gain and waveform compression in hearing aids. Proc. 37th Asilomar Conf. on Signals, Systems, and Computers, 391-398 (IEEE Pub. ISBN 0-7803-8104-1.

Goldsworthy, R. and J. Greenberg, (2004). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," J. Acoust Soc Am. 116(6), 3679-3689.

Holube, I. and B. Kollmeier (1996). "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model." J. Acoust. Soc. Am. 100(3): 1703-1716.

T. Houtgast, H. J. M. Steeneken, and R. Plomp (1980), "Predicting Speech Intelligibility in Rooms from the Modulation Transfer Function I. General Room Acoustics," Acustica, 46, 60-72.

Kawase, T., Delgutte, B., and Liberman, M. C. (1993). "Antimasking effects of the olivocochlear reflex. II. Enhancement of auditory-nerve response to masked tones." Journal of Neurophysiology, Vol 70, Issue 6 2519-2532.

Kiang, N. Y. S., Guinan, J. J., Liberman, M. C., Brown, M. C., and Eddington, D. K. (1987). Feedback control mechanisms of the auditory periphery: implication for cochlear implants. In Banfai, P., editor, International Cochlear Implant Symposium. Duren,West Germany.

Koch, R. (1992). Gehörgerechte Schallanalyse zur Vorhersage und Verbesserung der Sprachverständlichkeit (Auditory sound analysis for the prediction and improvement of speech intelligibility), Universität Göttingen.

Krause, J. C. (2001), "Properties of Naturally Produced Clear Speech at Normal Rates and Implications for Intelligibility Enhancement," Ph.D., Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA.

Lister, J., Besing, J., and Koehnke, J. (2002). "Effects of age and frequency disparity on gap discrimination,: J. Acoust. Soc. Am., 111, 2793-2800.

Ludvigsen, C., C. Elberling, G. Keidser, T. Poulsen, (1990). "Prediction of intelligibility of non-linearly processed speech." Acta Otolaryngol Suppl 469: 190-195.

Nilsson, M., Soli, S.D., and Sullivan, J.A. (1994). "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am., 95, 1085-1099.

Payton, K. L., L. D. Braida, S. Chen, P. Rosengard, R. Goldsworthy, (2002). Chapter 11. Computing the STI using speech as the probe stimulus. in Past Present and Future of the Speech Transmission Index. S. J. van Wijngaarden Ed. TNO Human Factors, The Netherlands: 125-138.

Ronnberg, J., Andersson, J., Samuelsson, S., Soderfeldt, B., Lyxell, B., and Risberg, J. (1999). "A speechreading expert: The case of MM," J. Speech Hearing Lang. Res., 42, 5-20.