Evaluation of Modified Phase Opponency Processing of Noisy Speech

José A. Figueroa Serra, Vikramjit Mitra, Carol Y. Espy-Wilson

MERIT BIEN, University of Maryland, College Park, MD

Abstract

Current state-of-the-art speech recognition systems perform well in clean conditions. However, their performance drops drastically in everyday acoustically noisy environments. To address this issue, we are analyzing the Modified Phase Opponency (MPO) based speech enhancement algorithm which improves the signal-to-noise ratio of the signal while making no assumptions or needing any estimate of the noise. In particular, we compare our MPO-based algorithm against other techniques in a series of quality and intelligibility tests using both normal-hearing and hearing-impaired listeners. The speech signals are corrupted with car noise and speech-shaped noise. The results will show if it is worthwhile for us to explore the use of the MPO in hearing-aid and cochlear-implant devices.

1. Introduction

Everyday environments can be quite noisy and, as a result, pose a significant problem for many technologies. We encounter many different types of noises that can be broadly classified into: (1) Stationary noise (for example the noise created by the motor of an electric-fan), (2) Non-Stationary noise (for example the slamming of a door or the background noise in a restaurant). These noises not only affect the performance of speech recognition systems and speaker recognition systems but also affect the ability of people with hearing impairments to listen and understand speech. Various speech enhancement techniques have been developed to improve the Signal-to-Noise ratio (SNR) of noisy speech signals. In this project, we focus on the evaluation of one such technique called the Modified Phase Opponency (MPO).

MPO detects and retains the harmonic and/or formant regions in speech by assuming that these regions are sufficiently narrowband and in the process attenuates the rest of the background noise. MPO is based on a neural model that is used for detection of tones-in-noise called the Phase Opponency (PO) model [1]. However MPO suffers in two distinct cases (1) when the noise is sufficiently narrow band which results in noise insertions, (2) when two narrow-band formants are close together (as in the case of /r/ when the 2^{nd} and the 3^{rd} formant comes close to one another) so that they appear as wide-band and hence results in speech deletions. To address this Deshmukh et. al. proposed [1] the combination of MPO with the Aperiodic, Periodic, Pitch (APP) detector. The APP detector provides information regarding the periodicity and aperiodicity confidence of each temporal frame. A speech-dominant

region typically has high periodic energy, where as a noise-dominant region has little if any periodic energy. These facts are used in conjunction with the MPO decision to reintroduce deleted wide-band speech segments and remove narrowband noisy regions. The resulting system, which is termed as the MPO-APP addresses the shortfalls of the MPO by reducing insertion and deletion error, unfortunately the perceptual quality of the enhanced speech suffered due to the shadow-effect of the noises at low SNR.

The MPO-APP acts as a switch, by detecting speech dominant regions in a spectro-temporal profile and passing that region 'as-is', as a result of which noise gets passed along with the speech regions which results in the "shadow effect". A pre-processor can be used with the MPO-APP enhancement algorithm that helps to reduce the 'shadow-effect', by reducing the noise level prior to MPO-APP processing. This preprocessing has two-fold effects (a) It improves the SNR of speech signal which will be further processed by MPO-APP; and since it is known that MPO-APP performs better at higher SNR's, hence it ensures better enhancement strategy, (b) it reduces the noise level at the speech dominant regions, hence improves the perceptual quality of the speech. A time-varying generalized Spectral Subtraction (GSS) has been implemented as the preprocessor of the MPO-APP, which assumes that the background noise has a component which is quasi-stationary over a small region [3].

The current research describes our evaluation of the MPO-APP-GSS speech enhancement algorithm using a series of quality and intelligibility tests taken by normal hearing listener and cochlear implant (CI) users (with and without a hearing aid).

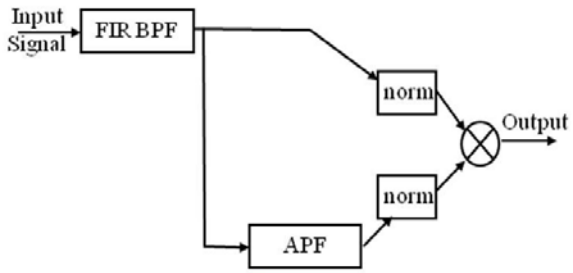


Figure 1: *MPO model*

The speech quality tests will indicate whether the MPO-APP-GSS technique has been able to improve the quality and the ease-of-listening of noise-corrupted speech and how it performs with respect to some of the state-of-the-art enhancement techniques. The Intelligibility test on the other hand aims to evaluate the intelligibility of the enhanced speech as perceived by the listeners.

Figure 1 shows the MPO model. The relative magnitude and phase response of the two independent paths can be controlled independently [1]. The all pass filter (APF) used in one of the paths facilitates the manipulation of the relative phase responses of the two paths without affecting their magnitude responses. The MPO model analyzes an input signal by performing a cross-correlation of the input signal with its phase shifted version as shown in Figure 1. If the input signal is narrow band (almost tone-like), then the cross-correlation of the

signal with it's phase shifted version will be mostly negative. Thus, for most speech sounds which are periodic and, therefore, consist of a combination of tones, we expect the output of the MPO to be negative. On the contrary, if the input signal is wideband or mostly aperiodic, then the cross-correlation of it with its phase shifted version will be mostly positive. Hence by tracking the cross-correlation values, we can detect a narrow band region from a wideband region.

The MPO architecture for speech enhancement is shown in Figure 2; where the analysis and synthesis filterbanks are perfect reconstruction filterbanks. Each of the MPO blocks in the sub-channels is an MPO structure tuned to a different center frequency (CF) and identical in construction to Figure 1. The CFs are spaced every 50 Hz from 100 Hz to just below the maximum frequency [1]. The analysis filterbank splits the input speech into *N* sub-bands and each of those sub-bands are analyzed by the MPO unit. The MPO acts a switch and hence decides which channels to pass and which to attenuate. If the MPO output for a specific channel is mostly negative (below a certain threshold), then it considers the content of that channel as sufficiently narrow-band and hence passes it 'as-is'. On the contrary if the MPO output for a channel is sufficiently positive (above a certain threshold), then it decides the content of that channel to be wideband and hence attenuates the content of that channel. The outputs of the N channels are then processed by the synthesis filterbank to generate the enhanced speech.

One of the salient features of the MPO speech enhancement technique is its noise independent nature. It does not require any noise estimate nor does it require *a priori* knowledge of the noise characteristics to perform the enhancement task

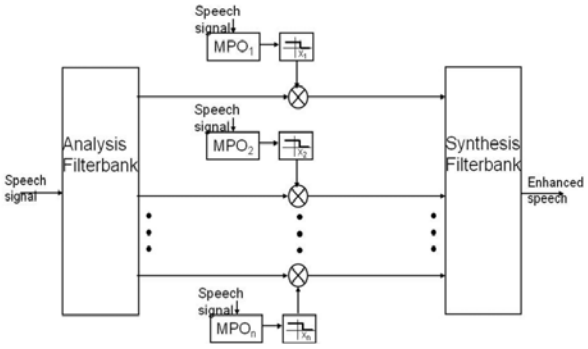


Figure 2: *MPO architecture*

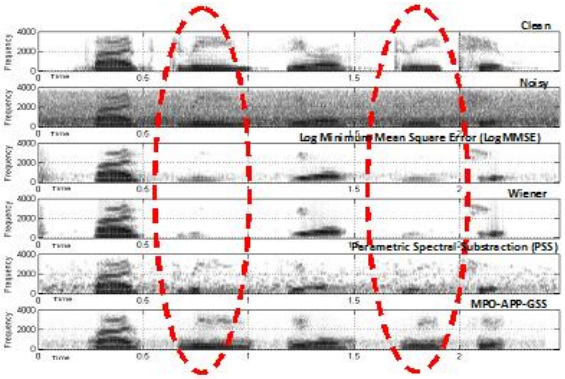


Figure 3: *MPO-APP-GSS Comparison*

A spectrogram of a clean speech signal is show in Figure 3. This signal is corrupted with car noise at 5 dB SNR. The corrupted signal is then processed with four different speech enhancement schemes to remove background noise: Log Minimum Mean Squared Error (LogMMSE), Wiener filter, Parametric Spectral Substraction (PSS) and MPO-APP-GSS. Both the LogMMSE and

Wiener filtering based method delete some of the higher formants reducing the naturalness of the enhanced speech, they are also found to suffer from speech deletions at low SNRs. The major drawback of PSS method is its musical noise addition, which degrades the perceptual quality of the enhance speech significantly. It also suffers from speech deletions. MPO-APP-GSS method retains most of the voiced regions in the speech signal and successfully retains most of the higher formants as well. It can be seen in Figure 3 that the MPO-APP-GSS method is able to retain the second and third formant around 2000 Hz near 0.8s and again near 1.8s while passing very little noise.

2. Methodology

2.a. Databases

The Aurora-2 database was used for the quality tests [4]. This database is a derivative of the TIdigits database re-sampled at 8 kHz and it is composed of three different subsets for testing: subset A, subset B and subset C, each of them with different types of background noise. In the present research, only subset A was used for evaluation. Subset A consists of short recordings consisting of five digits corrupted by four different noise types at seven different SNRs from ∞ to -5 dB. The four different noise types in subset A are: car noise, babble noise, subway noise, and exhibition noise. For the quality tests in this study we used car noise and speech shaped noise. These are referred to as N3 and N5 respectively. The dataset for speech shaped noise was prepared in-house.

The corpus used for the intelligibility test is called the coordinate response measure (CRM) database [2]. The CRM is used to measure speech intelligibility of utterances processed with different enhancement schemes to remove background noise. The speech corpus consists of sentences of the form "Charlie go to (color) (number) now" spoken by eight talkers including four males and four females. Colors in the sentences include: blue, red, green, and white. Numbers include the digits one through eight (1-8). This corpus was originally developed by the Air Force Research Laboratory. For the quality tests in this study, we excluded the sentences that contained the color blue and/or the number seven. Also, all of the sentences used were spoken by a male speaker.

2.b Subjects

Five normal hearing listeners and four CI (with and without a hearing aid) users participated throughout the quality and intelligibility tests. The age range for the normal hearing listeners was 23 to 56 years old. The CI users are of advanced age ranging from 66 to 75 years old. The experience of the CI users with the hearing device varies from 1 year to 10 years.

2.c Testing Protocol

The quality and intelligibility tests were developed with MATLAB in the Speech Communication Laboratory. Subject testing was performed in the Cochlear Implants and Psychophysics Laboratory. To be sure that our normal hearing subjects did in fact have normal hearing, we had to test it. Air conduction pure tone thresholds were

measured using a calibrated audiometer (GSI 10). Testing was performed in a double wall sound proof booth using ER-3A insert earphones. Normal hearing for this study was defined as audiometric thresholds equal to or better than 25 dB HL in both ears from 250-4000 Hz. Participants who did not meet this criteria were excluded from participation in this study.

The quality and intelligibility tests were performed in a sound-proof room that contained a computer with speakers. The room was first calibrated at 70 dB sound pressure level (SPL) using a pure-tone stimulus at 2000 Hz (with equal root mean square (RMS) to the average RMS of the experimental stimulus) with a sound level meter at a length of 1 meter at the approximate location of the listener's head. CI users listened through their own speech processor using their everyday settings. They were instructed to adjust their devices to a comfortable level.

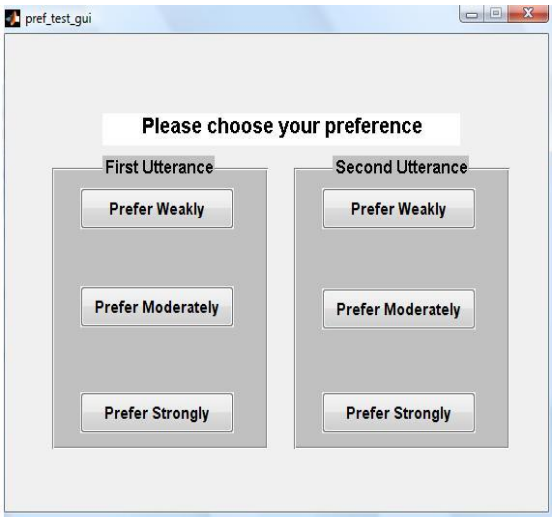


Figure 4: *Quality test GUI*

2.d Experimental Setup for Quality Tests

The quality test consists of paired comparisons. The two instances are processed with different speech enhancement algorithms at different SNRs (-5 dB, 0 dB, 5 dB, 10 dB, and 15 dB). The subject tested had to choose which one of the two sentences he/she prefers and with what degree ranging from: weakly, moderately or strongly in the graphical user interface (GUI) shown in figure 4. If one of them is much better (in terms of ease-of listening, quality of speech etc…), then you would prefer that one "strongly" over the other. On the other hand, if both of them are almost equally good or equally bad, then the subject would prefer one of them "weakly" over the other. In the intermediate case you would prefer one of them "moderately". A subject's decision should be based upon how pleasant the speech sounds, and which one is perceived to be cleaner. The duration of the quality test is about one hour and three minutes (depends on how quickly the user responds). The test is equally divided into four parts with two breaks that range from 1-3 minutes and the halftime break of 5 minutes. The subject is given the option to keep doing the test without any breaks by just hitting "enter" on the keyboard.

After the quality test the subject proceeded to the intelligibility test. The purpose of this test is to know how well the speech enhancement techniques improve the ability of the listeners to understand what they heard.

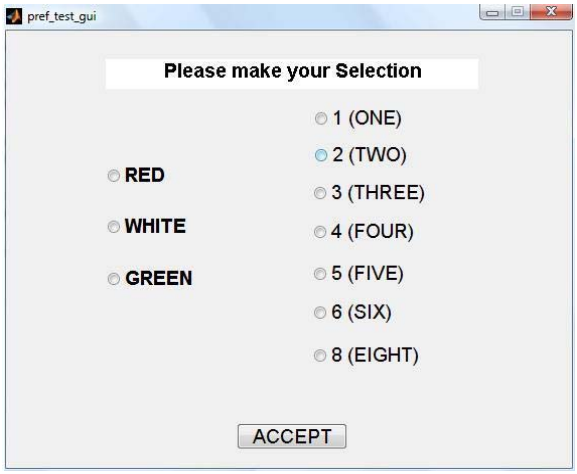


Figure 5: *Intelligibility test GUI*

2.e Experimental Setup for Intelligibility Tests

For the intelligibility test, the sentences were processed with different SNRs (0 dB, 5 dB, and 15 dB). The listeners chose what "color" and "number" they heard using the GUI interface shown in Figure 5. The listener had three breaks of 3 minutes each and they were given the option to continue without the breaks by hitting "enter" on the keyboard. Based on preliminary results obtained from the quality judgments, only two speech enhancement techniques were used in this evaluation: the MPO-APP-GSS and the LogMMSE .

3. Results

The plots shown on the figures 6, 7, 8, 9, 10, and 11 are labeled with abbreviations such as N-1 or CI-1. N-1 means normal hearing listener one while CI-1 means CI user one. For quality test if the bars are on the positive side it means that the subject preferred us over other approaches, on the other hand if they are on the negative side it means they preferred other techniques. Quality test results with car noise added and then processed with different approaches to "clean" the speech signal indicate a preference for us above other techniques most of the time with both normal hearing listeners and CI users (see Figures 6, and 7). Quality test results for speech shaped noise indicate a similar pattern as with car noise results, us being preferred most of the time (See Figures 8 and 9). For the intelligibility test bars reaching one means the subject preferred us weakly, reaching two means moderately and three means strongly over other speech enhancement scheme. Intelligibility results with car and speech shaped noise indicate that CI users found our speech enhancement algorithm more intelligible most of the time over unprocessed noisy speech as the SNR reduces (see Figures 10 and 11). CI-2 also took the intelligibility with a CI device and a hearing aid finding the test more intelligible wearing both devices at the same time (see Figure 12).

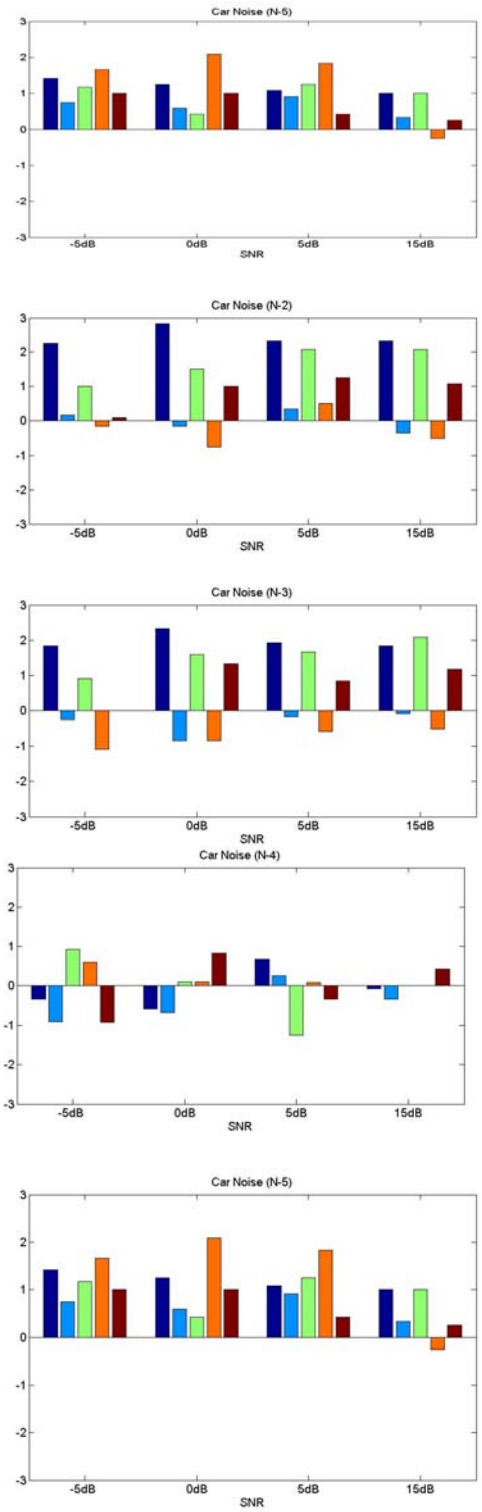


Figure 6: *Quality test results with car noise for normal hearing listeners. Purple: Original unprocessed speech, Blue: LogMMSE, Green: SSB, Orange: Wiener, Brown: MPO-APP. Normal hearing listeners are in numerical order, from 1 to 5.*

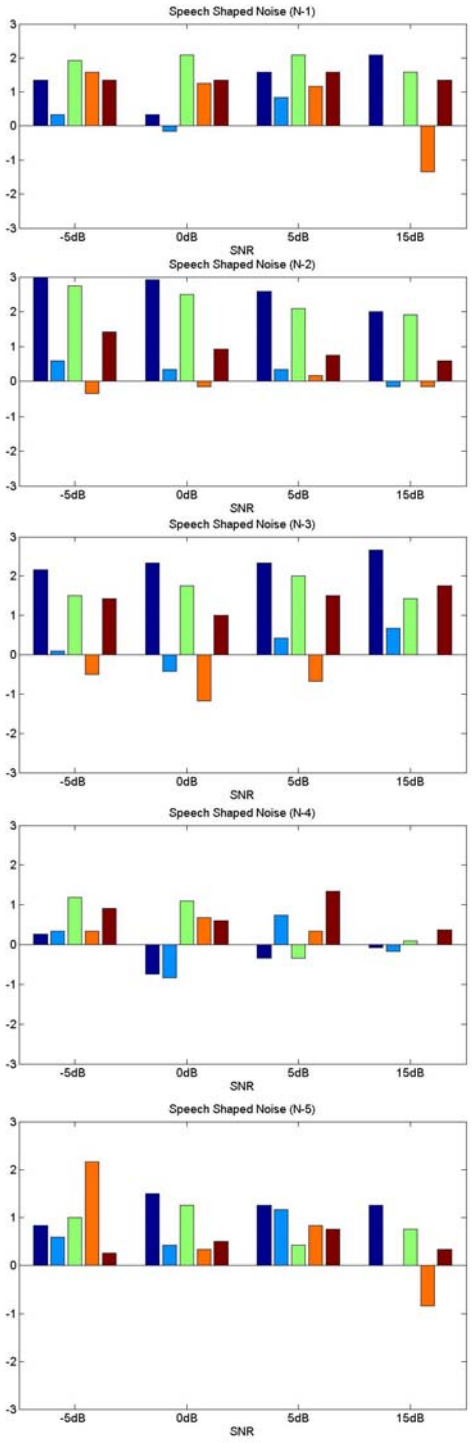
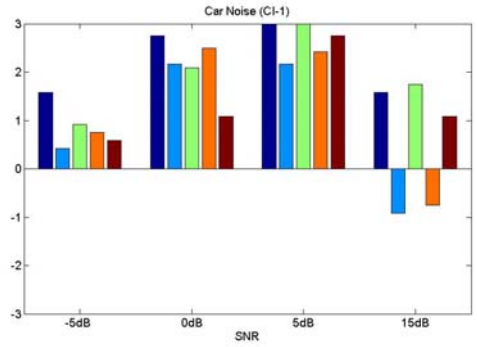
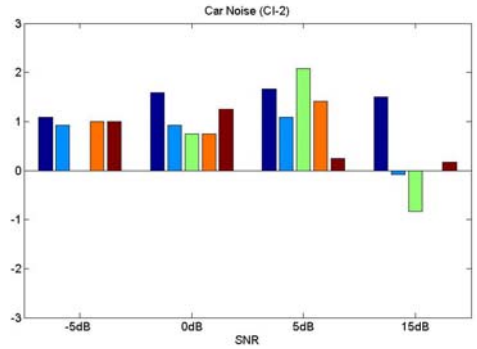
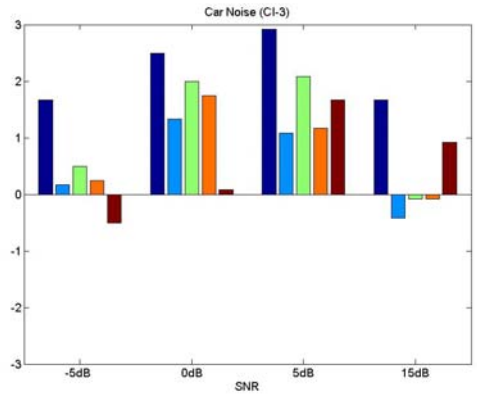


Figure 7: *Quality test results with speech shaped noise for normal hearing listeners. Follows the same legend of figure 6. Normal hearing listeners are in numerical order, from 1 to 5.*







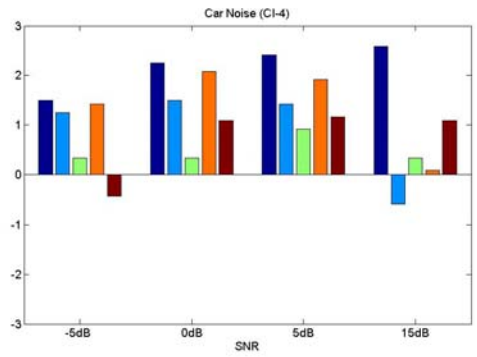
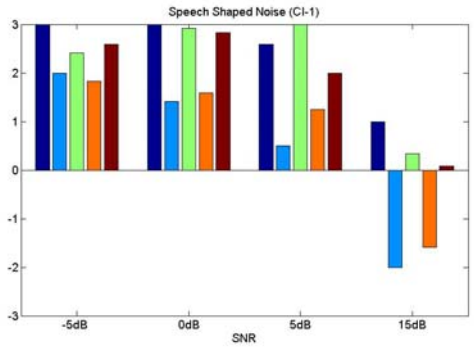
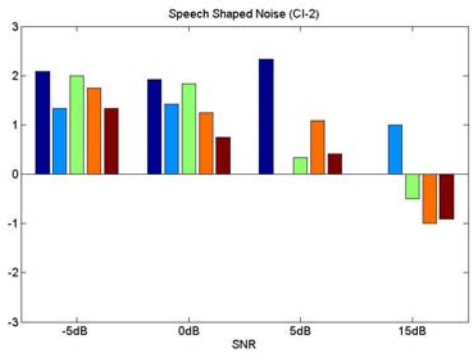
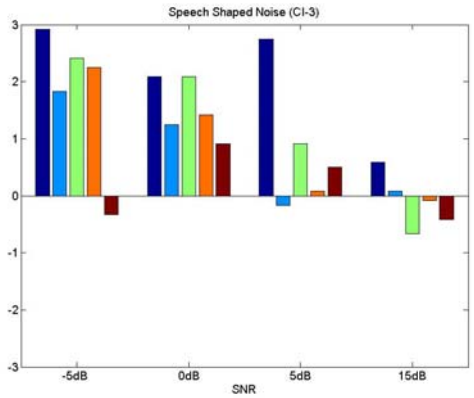


Figure 8: *Quality test results with car noise for CI users. Follows the same legend of figure 6. CI users are in numerical order from 1 to 4.*







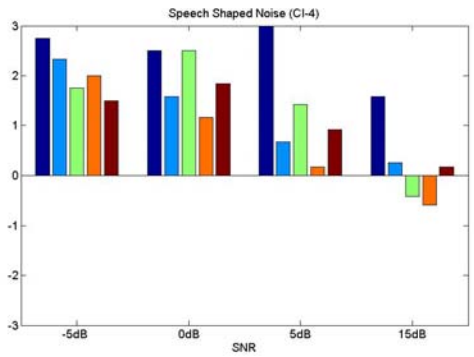


Figure 9: *Quality test results with speech shaped noise for CI users. Follows the same legend of figure 6. CI users are in numerical order from 1 to 4.*

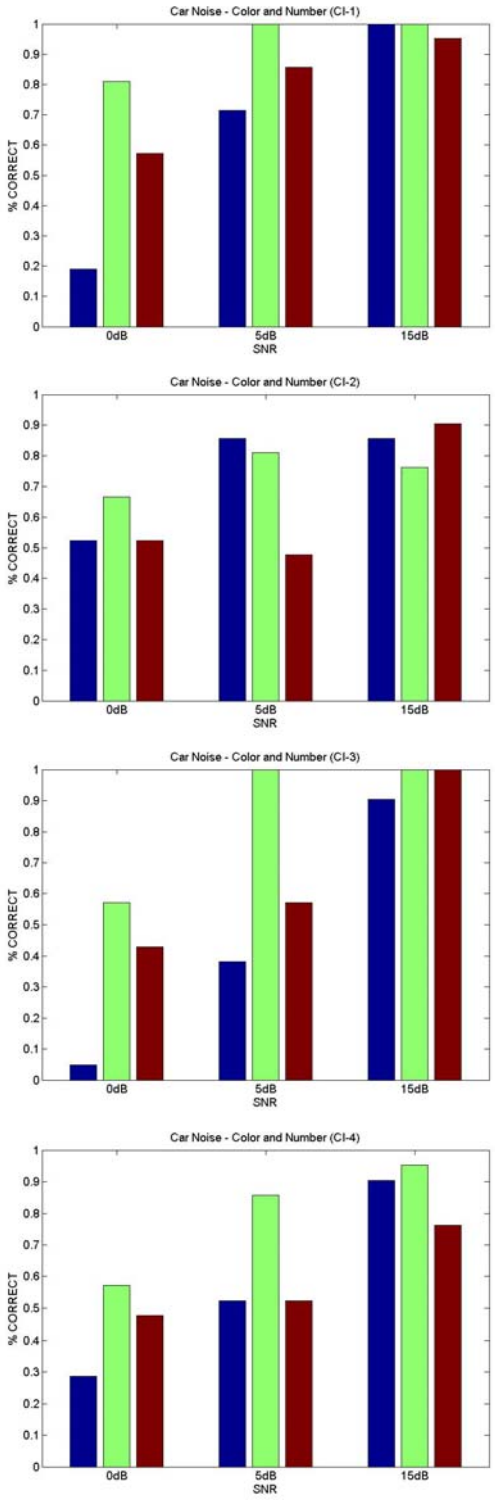


Figure 10: *Intelligibility results with car noise for CI users. Purple: Original unprocessed speech, Green: LogMMSE, and Brown: MPO-APP-GSS. . CI users are in numerical order from 1 to 4.*

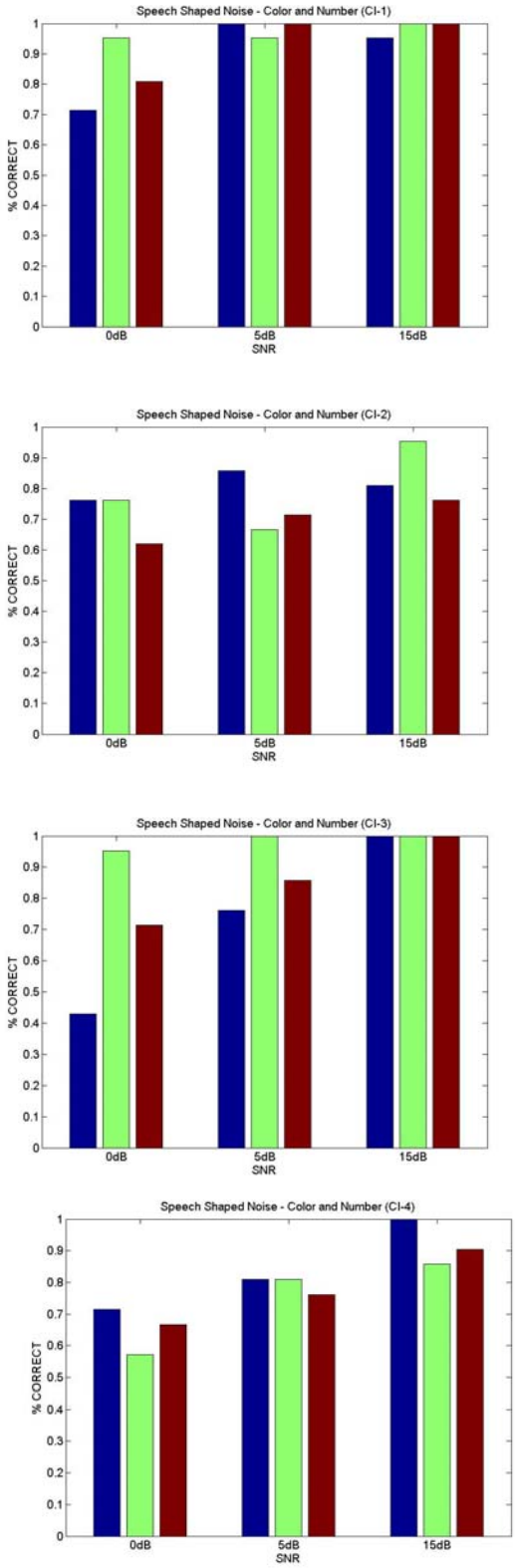


Figure 11: *Intelligibility results with speech shaped noise for CI users. Follows the same legend as figure 10. . CI users are in numerical order from 1 to 4.*

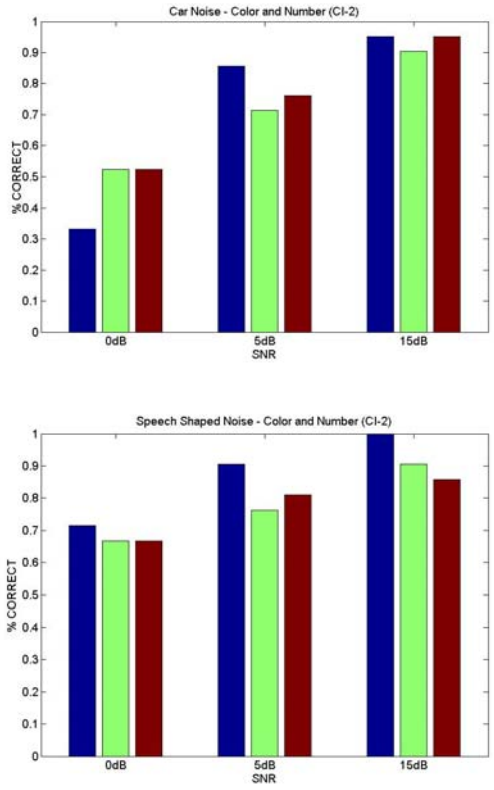


Figure 12: *Intelligibility results with car noise and speech shaped noise for CI-2. CI-2 also took the test with a CI device and a hearing aid. Follows the same legend of figure 10. Follows the same legend as figure 10.*

3. Acknowledgements

We are grateful to the research participants for their support of our work, and to Dr. Monita Chatterjee and Kara Schvartz for advice, providing the laboratory and helping to recruit the subjects used in this study. This study was funded by NSF Grant #IIS0703859, and NIH/NIDCD RO1 DC004786.

4. References

[1] O.D. Deshmukh, C. Espy-Wilson and L.H. Carney, "Speech Enhancement Using The Modified Phase Opponency Model", Journal of Acoustic Society of America, Vol. 121, No. 6, pp 3886-3898, 2007.

[2] Bolia, R., Nelson, W., Ericson, M., and Simpson, B. (2000). “A speech corpus for multitalker communications research,” J. Acoust. Soc. Am. 107,1065-1066.

[3] V. Mitra, C. Espy-Wilson, 'Speech Enhancement for Noise-Robust Speech Recognition', under review for the 156th meeting of the Acoustical Society of America.

[4] Hirsch, H. G., and Pearce, D. (2000). “The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions”, in ISCA ITRW ASR2000 Automatic Speech Recognition: Challenges for the Next Millennium, Paris, France, pp. 18-20.