EURASIP Journal on Applied Signal Processing 2005:18, 2938–2953

An Auditory-Masking-Threshold-Based Noise Suppression Algorithm GMMSE-AMT[ERB] for Listeners with Sensorineural Hearing Loss

Ajay Natarajan

Robust Speech Processing Group, Center for Spoken Language Research, and Department of Electrical and Computer Engineering, University of Colorado at Boulder, Boulder, CO 80309-0309, USA
Email: nataraja@cslr.colorado.edu

John H. L. Hansen

Center for Robust Speech Systems, Department of Electrical Engineering, Erik Jonsson School of Engineering & Computer Science, and Callier Center (Speech and Hearing), School of Behavioral and Brain Sciences, University of Texas at Dallas, Richardson, TX 75083, USA
Email: john.hansen@utdallas.edu

Kathryn Hoberg Arehart

Department of Speech, Language and Hearing Sciences, University of Colorado at Boulder, 2501 Kittredge Loop Road, UCB 409, Boulder, CO 80309-0409, USA
Email: arehart@colorado.edu

Jessica Rossi-Katz

Department of Speech, Language and Hearing Sciences, University of Colorado at Boulder, 2501 Kittredge Loop Road, UCB 409, Boulder, CO 80309-0409, USA
Email: rossija@colorado.edu

Received 6 May 2004; Revised 21 December 2004

This study describes a new noise suppression scheme for hearing aid applications based on the auditory masking threshold (AMT) in conjunction with a modified generalized minimum mean square error estimator (GMMSE) for individual subjects with hearing loss. The representation of cochlear frequency resolution is achieved in terms of auditory filter equivalent rectangular bandwidths (ERBs). Estimation of AMT and spreading functions for masking are implemented in two ways: with normal auditory thresholds and normal auditory filter bandwidths (GMMSE-AMT[ERB]-NH) and with elevated thresholds and broader auditory filters characteristic of cochlear hearing loss (GMMSE-AMT[ERB]-HI). Evaluation is performed using speech corpora with objective quality measures (segmental SNR, Itakura-Saito), along with formal listener evaluations of speech quality rating and intelligibility. While no measurable changes in intelligibility occurred, evaluations showed quality improvement with both algorithm implementations. However, the customized formulation based on individual hearing losses was similar in performance to the formulation based on the normal auditory system.

Keywords and phrases: normal hearing, hearing impaired, auditory masking threshold, equivalent rectangular bandwidth, generalized minimum mean square estimation.

1. INTRODUCTION

Individuals with sensorineural hearing loss have more difficulty understanding speech compared to those with normal hearing. This effect is compounded in diverse environments that may contain time varying cues/signals or multiple competing speakers. This increased difficulty in understanding speech in noise is due to (a) reduced audibility of speech sounds in listeners with elevated auditory thresholds, and (b) suprathreshold processing deficits characteristic of sensorineural hearing loss. Hearing aids incorporate different strategies to compensate for reduced audibility and for suprathreshold processing deficits. These strategies include frequency-dependent amplification, compression, and directional microphones. Hearing aids based on digital signal processing may also include algorithms for feedback cancellation and active noise reduction. Spectral subtraction is one possible noise reduction algorithm for hearing aid applications

because of its simplicity and low computational requirements. In general, noise reduction circuits employing spectral subtraction use mathematical criteria based on the estimated speech-to-noise ratio. One of the primary objectives in speech enhancement is to achieve a balance between pure noise suppression and the musical noise-like artifacts that may be introduced by the processing techniques. Most noise suppression methods are based on a signal-plus-noise model, and mathematical criteria (such as signal-to-noise ratio) are used to evaluate their performance. In an effort to achieve a better balance between audible musical artifacts and noise suppression, a number of previous studies in speech enhancement have considered incorporating aspects of the human auditory system including masking [1, 2, 3, 4, 5, 6]. In an earlier study, Tsoukalas et al. [1] used a spectral subtraction technique based on aspects of the auditory process. Their method considers an enhancement approach that uses the auditory masking threshold (AMT) [7] in conjunction with a version of spectral subtraction. The AMT in their implementation was calculated in four steps: (1) obtain energies in speech critical band (CB) frequency analysis, (2) convolve a spreading function [8] with the CB spectrum to obtain a masking spread threshold, (3) compute an offset term for masking spread thresholds that takes into account signal tonality, and (4) normalize/compare and account for absolute auditory thresholds. This speech enhancement method is referred to as the TMK algorithm in the present study.

Based on the work in [1], Arehart et al. [9], implemented a version of the TMK algorithm and evaluated its effectiveness in improving speech-perception in noise for both normal-hearing and hearing-impaired listeners. This implementation is referred to as the auditory masking threshold-noise suppression (AMT-NS) scheme in the present study. The AMT-NS algorithm yielded better quality ratings and better intelligibility scores in both normal-hearing and hearing-impaired listeners in some but not all of the test conditions. Their implementation of the TMK scheme employed speech and noise sampled at 8 kHz, while the original TMK [1] used 16 kHz samples of speech and noise. Also, the level of intelligibility improvement reported in [1] was significantly higher than those demonstrated in [9] when using an 8 kHz sample rate version of the enhancement method.

The TMK and the AMT-NS algorithms are based on masking properties of the normal auditory system, with its theoretical underpinnings based on MPEG-4 audio coding [7]. Alternate processing strategies that specifically consider hearing aid applications and the effects of sensorineural hearing loss may optimize the AMT-NS approach to speech enhancement for hearing-impaired listeners. The present study describes a new noise suppression scheme. Referred to here as GMMSE-AMT[ERB], this new scheme includes two primary modification of previous formulations.

The first change is that the new algorithm includes a modification of the suppression structure. Specifically, it is implemented using the modified generalized minimum mean square error (GMMSE) estimators which provide improvement over traditional spectral subtraction estimators [10, 11]. The suppression structure has also been modified so that tonality is not included. Preliminary evaluations in our laboratory indicated that listeners preferred algorithm formulations with tonality disabled. Furthermore, inclusion of tonality would introduce additional complexity to the algorithm formulation, which would impact the ability for real-time implementation in digital hearing aid applications. Finally, the assumptions of the tonality offset, originally formulated for use in MPEG-4 audio coding applications, are primarily related to the harmonic structure of music or audio. While there is some justification in using tonality offset with voiced signals due to the harmonic structure present in formant regions, some assumptions regarding tonality may not be appropriate for hearing aid applications. Therefore, we do not include a tonality offset in the formulation presented here.

The second primary modification is that the new algorithm establishes a framework for customization of the AMT estimation to individual subjects with hearing loss. To accommodate this framework, the algorithm requires estimation of normal frequency resolution as well as the degraded frequency resolution characteristic of cochlear hearing loss. Therefore, the frequency resolution of the cochlea is represented in the algorithm with an auditory filter bank using equivalent rectangular bandwidths (ERBs) [8]. While related to the critical band scale, the ERB scale is used in the algorithm formulation because present-day experimental studies estimating degraded frequency resolution in listeners with sensorineural hearing loss have used the ERB scale and not the critical band scale (e.g., [12, 13, 14]). The estimation of the AMT and of the spreading functions for masking are implemented in two ways: with normal auditory thresholds and normal auditory filter bandwidths (GMMSE-AMT[ERB-NH]) and with the elevated thresholds and broader auditory filters characteristic of cochlear hearing loss (GMMSE-AMT[ERB-HI]).

Section 2 of this paper presents details of the algorithm derivation including the modified structure and framework for customization of the AMT based on individual listener profiles. Section 3 presents evaluation of both GMMSE-AMT[ERB-NH] and GMMSE-AMT[ERB-HI] implementations. GMMSE-AMT[ERB-NH] is evaluated over several speech corpora, using detailed objective quality tests based on segmental SNR and the Itakura-Saito objective quality measures. Formal listener evaluations with normal and hearing impaired subjects of speech quality rating and intelligibility are also used to test performance for both the NH and HI formulations.

2. GMMSE-AMT[ERB] ALGORITHM FORMULATION

The flowchart of the proposed algorithm is presented in Figure 1. The algorithm can be partitioned into three phases that include: (1) enrollment (GMMSE spectral estimation), (2) AMT threshold estimation, and (3) noise suppression. For normal-hearing listeners, only the GMMSE-AMT[ERB-NH] is implemented. For hearing-impaired listeners, both

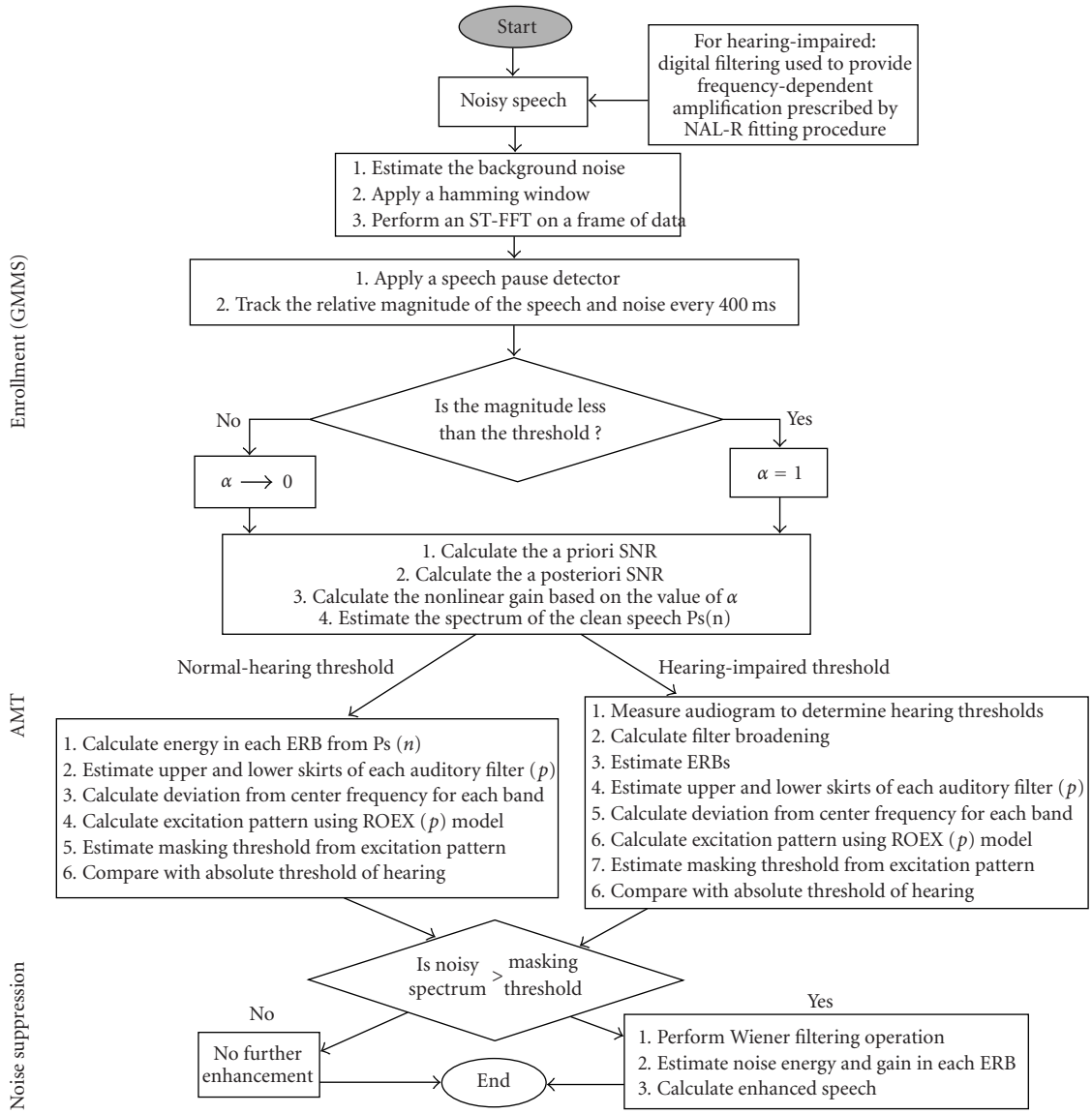

Figure 1: Flowchart of the GMMSE-AMT[ERB] enhancement algorithm.

the GMMSE-AMT[ERB-NH] and GMMSE-AMT[ERB-HI] versions are implemented and customized for individual hearing-impaired listeners by including frequency-dependent amplification approximating the linear gain prescribed by the NAL-R hearing aid fitting procedure [15]. GMMSE-AMT[ERB-HI] is further customized for each individual hearing-impaired listener by considering individual hearing losses in the AMT estimation (i.e., broader auditory filters and elevated thresholds).

2.1. Enrollment: GMMSE spectral estimation

The first processing step is to obtain an estimate of the clean speech power spectrum through a modified generalized minimum mean square estimation algorithm that is needed to calculate the AMT. The original speech signal $x(n)$ is assumed to be degraded by an additive uncorrelated noise source $d(n)$, resulting in the noisy speech signal,

$$y(n) = x(n) + d(n). \tag{1}$$

Under this assumed model, one can obtain a generalized family of MMSE speech spectral estimators as [10, 11]

$$\hat{X}_p = \left(E\{X_p^{\alpha} \mid Y_p\}\right)^{1/\alpha}, \tag{2}$$

where X_p is the power spectrum of the clean speech, and Y_p is the power spectrum of the noisy speech (both of which are real quantities). This MMSE estimator attempts to strike a balance between the *a priori* information and the noisy data information (in this case the *a posteriori* SNR $\gamma - 1$). One of the main advantages of the MMSE amplitude estimator is that it results in colorless residual noise in the enhanced speech [16]. We note that substitution of $\alpha = 0.5$ into (1)

gives the traditional Ephraim-Malah [17] amplitude estimator, and $\alpha = 1$ gives the MMSE power spectral estimator. For MMSE, if the real and imaginary parts of the Fourier coefficients of the clean speech and noise power spectra are modeled as independent zero mean Gaussian random variables with variances $\sigma_x^2(\omega,i)/2$ and $\sigma_d^2(\omega,i)/2$, respectively, and $\alpha = 0.5$, the MMSE estimate of $X(\omega,i)$ is given by [17],[1]

$$\hat{X}(\omega,i) = \left[\Gamma(1.5)\left(\frac{v(\omega,i)}{\gamma(\omega,i)^2}\right)^{0.5}\Phi(-0.5:1:-v(\omega,i))\right]Y(\omega,i), \tag{3}$$

where $\Gamma[\cdot]$ is the Gamma function, and $\Phi(a,b;z)$ is the confluent hypergeometric series (see (4)) defined in [18], and is dependent on the *a priori* SNR and *a posteriori* SNR,

$$\Phi(a:b:z) = 1 + \frac{a}{b}\frac{z}{1!} + \frac{a(a+1)}{b(b+1)}\frac{z^2}{2!} + \frac{a(a+1)(a+2)}{b(b+1)(b+2)}\frac{z^3}{3!} + \cdots, \tag{4}$$

with

$$v(\omega,i) = \frac{\xi(\omega,i)}{1+\xi(\omega,i)}\gamma(\omega,i), \tag{5}$$

where $\xi(\omega,i)$ and $\gamma(\omega,i)$ are defined as

$$\xi(\omega,i) = \frac{\sigma_x^2(\omega,i)}{\sigma_d^2(\omega,i)}, \qquad \gamma(\omega,i) = \frac{|Y(\omega,i)|^2}{\sigma_d^2(\omega,i)}, \tag{6}$$

where $\xi(\omega,i)$ is the *a priori* SNR and $\gamma(\omega,i) - 1$ is the *a posteriori* SNR as a function of frequency ω and frame index i. The definitions in (6) suggest a general representation of the terms $\xi(\omega,i)$ and $\gamma(\omega,i)$, where $\xi(\omega,i)$ is the SNR using the clean speech X, and $\gamma(\omega,i)$ being the ratio of the noisy speech spectrum of $Y(\omega,i)$ to the background noise spectrum assuming that the noise is statistically white. While $\gamma(\omega,i)$ can be obtained from an accurate estimate of the background noise, a decision-directed approach is used to estimate $\xi(\omega,i)$. The estimate for $\xi(\omega,i)$ is given by [17]

$$\xi(\omega,i) = (1-\beta)P[\gamma(\omega,i-1)-1] + \beta\frac{|\hat{X}(\omega,i-1)|^2}{\sigma_d^2(\omega,i-1)}, \tag{7}$$

where β is chosen to be between 0 and 1, and $P[x] = x$ for $x \geq 0$, and $P[x] = 0$ for $x < 0$.

It can be shown that a small value of α (e.g., $\lim_{\alpha\to 0}$) is suitable for noise suppression that improves the segmental SNR [11]. A larger value of α (e.g., $\lim_{\alpha\to 1}$) reduces the amount of musical processing artifacts and speech distortion (note that this balance is illustrated in *Enrollment* phase in Figure 1). This suggests a benefit from a method that dynamically changes the value of α, rather than restricting the processing to a single value. Using a speech/pause detection algorithm, one can dynamically change the value of α. In the noisy signal, if a pause is encountered, the value of α is dynamically adjusted (i.e., $\alpha \to 0$), and in regions where speech is present, the value α is set to 1.

The voice activity detector (VAD) algorithm [19] used to dynamically adjust α is described below. Let P_{dk} be the power spectrum of the distortion/noise for the kth ERB frequency subband, and $\hat{P}_{xk}$ be the estimated power spectrum of the clean speech signal for the kth ERB frequency subband. The values of P_{dk} and $\hat{P}_{xk}$ are obtained from the following relations:

$$\begin{aligned} P_{dk}[n] &= \eta P_{dk}[n-1] + \frac{1-\eta}{1-\kappa}\left(\hat{P}_{xk}[n] - \kappa\hat{P}_{xk}[n-1]\right), \\ \hat{P}_{dk}[n] &= \mu\hat{P}_{dk}[n-1] + (1-\mu)\left(|\hat{X}_k[n]|^2\right), \end{aligned} \tag{8}$$

where $\mu = 0.7$, $\kappa = 0.998$, and $\eta = 0.45$. These values are used for our implementation with an analysis (FFT) frame size of 128 samples, with a skip rate of 64 samples (i.e., overlap of 50% between adjacent analysis windows) using an 8 kHz sample rate. These values were determined to be reasonable for the noise types considered through a pilot experiment, and kept fixed for all processing in the present study. The speech pause detector algorithm is applied as follows:

$$NX_{\text{rel}\,k}[n] = \frac{NX_k[n] - NX_{\min k}[n]}{NX_{\max k}[n] - NX_{\min k}[n]}, \tag{9}$$

where $NX[n] = P_{dk}[n]/\hat{P}_{xk}[n]$. The term $NX_{\text{rel}\,k}[n]$ is the relative ratio of the noise energy to the signal-plus-noise energy for each subband [19]. The values of $NX_{\min k}[n]$ and $NX_{\max k}[n]$ represent the minimum and maximum ratios, and are calculated looking back across the previous 400 milliseconds portion of the speech signal. The value of the power spectrum of the distortion in subband k, P_{dk}, is modified if $NX_k[n]$ is less than a predetermined threshold. We then apply a nonlinear gain term, based on the value of α from the GMMSE algorithm, the *a priori* SNR and the *a posteriori* SNR, to the noisy power spectrum to obtain the estimate of the clean power spectrum.

2.2. AMT threshold estimation

Having presented the GMMSE enhancement scheme and voice activity detector, we now shift to the auditory masking threshold estimation scheme. It is important to note that the use of an AMT is not by itself a speech enhancement process, since it essentially allows the enhancement method to balance noise suppression versus potential processing artifacts. The use of the AMT is of particular interest for hearing-impaired individuals since, in theory, one would expect that the AMT would be shifted for such individuals and allow for a different level of either background noise or processing artifacts in the processed signal.

The steps for calculating the AMT (as shown in Figure 1) in the present algorithm are as follows:

[1]Note that for (3), we use $\hat{X}(\omega,i)$ to represent the spectral estimate of the clean speech which is $\hat{X}_p$ in (2). This was done to be consistent with the notation in [11, 17].

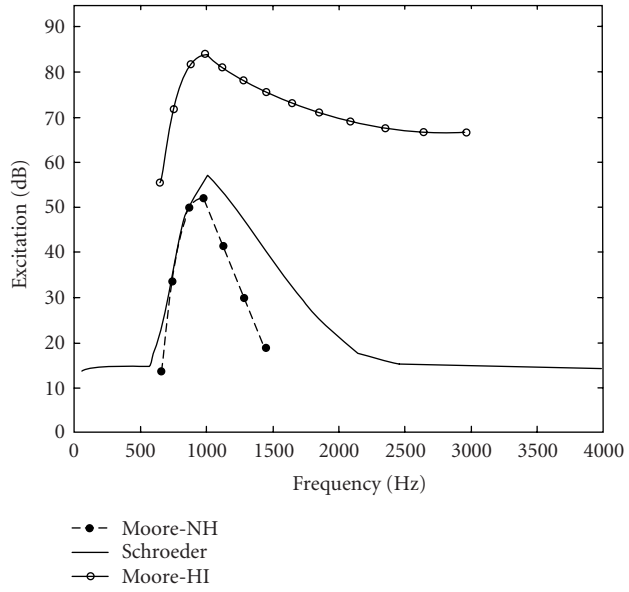

Figure 2: Comparison of excitation patterns estimated for a normal-hearing individual from the Schroeder model, and both normal and hearing impaired individuals using the ROEX auditory filter model (labeled as Moore-HI and Moore-NH from [8]).

(a) determine the auditory filter bandwidth in normal and impaired ears,
(b) calculate the total energy in each auditory filter (ERB),
(c) compute the excitation pattern based on the auditory filter characteristics,
(d) compare the excitation pattern with the absolute threshold of hearing.

The auditory filters are represented using their equivalent rectangular bandwidth [12]. For normal-hearing (NH) individuals, the hearing thresholds across all frequencies are assumed to be 0 dB HL. The hearing thresholds in quiet for hearing-impaired (HI) individuals are obtained from audiometric testing. The ERB values for a normal-hearing individual over the whole frequency range are described by the following equation [12]:

$$\text{ERB} = 24.7(4.37F + 1), \tag{10}$$

where ERB is in Hz, and F is the center frequency in kHz. For the hearing-impaired individual, the ERB is equal to $24.7(4.37F + 1)^{*}B$, where B ($B > 1$) is the frequency broadening term which is described below. The total threshold for HI listeners is a combination of threshold loss due to outer and inner hair cell damage.

The broadening of the auditory filters due to hearing loss can be described by [13, 14]

$$B = (10)^{0.01757(\text{HL}_{\text{ohc}}-22)\cdot([1-(f_c-1)^2]/3.09)} \tag{11}$$

up to a frequency of 1 kHz, and

$$B = (10)^{0.01757(\text{HL}_{\text{ohc}}-22)} \tag{12}$$

for higher frequencies, where f_c is the center frequency in kHz, and HL_{ohc} is the amount of hearing loss due to outer hair cell damage. Eighty percent of the total threshold loss is assumed to be due to loss of outer hair cell function, with the auditory filter bandwidth at 2000 Hz corresponding to filters that are approximately 2.7 times the bandwidth of normal auditory filters (Moore and Glasberg [14]). The constant 0.01757 is chosen so that B has a value of 3.8 when $\text{HL}_{\text{ohc}} = 55$ dB, which the model assumes is the maximum value of broadening due to outer hair cell loss below 2000 Hz. For NH individuals, the value of B is set to 1. Thus, the total number of estimated ERB filters in the frequency partition will be smaller for impaired ears. Once the filter shapes are defined, the signal power in each critical subband is calculated as X_{ERB}. The excitation pattern is derived from the output of the auditory filters as a function of their center frequency. Specifically, the excitation pattern is calculated by summing up the power of each signal component with the filter weighting function that is given by the ROEX(p) model, which is described in [8], as

$$W(g) = (1 + pg)\exp(-pg), \tag{13}$$

where W is the filter shape. We note that the signal power for calculating the excitation pattern must be recalculated to match the audiometric testing results. The correction thresholds for this recalculation are obtained from the TDH-39 headphones for both the normal and impaired ear.

The normalized distance of the signal component from the center frequency f_c of the filter involved is described as

$$g = \left(\frac{|f - f_c|}{f_c}\right). \tag{14}$$

The parameter p in (13) describes both the bandwidth and slope of the skirts of the auditory filter and can be used to derived p_l and p_u, which, respectively, describe the sharpness of the lower and upper sides of the ERB-based bandpass filters. The lower frequency skirt p_l of the auditory filter becomes less sharp with increasing level. Here, p_l varies with broadening and level as

$$p_l(x) = p_{l(51)} - \left(0.35 - 0.35\left(\frac{B-1}{3}\right)\right)\left(\frac{p_{l(51)}}{p_{l(51,1k)}}\right)(X_{\text{ERB}} - 51), \tag{15}$$

where $p_{l(51)}$ is the value of the skirt p for an equivalent noise level of 51 dB/ERB, and $p_{l(51,1k)}$ is the value of $p_l(x)$ at 1 kHz for a noise level of 51 dB/ERB. X_{ERB} is the signal power in each critical subband which can also be stated as the equivalent input power in dB/ERB. The upper frequency skirt, p_u, of the auditory filter does not vary largely with level and can be described as

$$p_u = \frac{4^{*} f_c}{24.7(4.37F + 1)}. \tag{16}$$

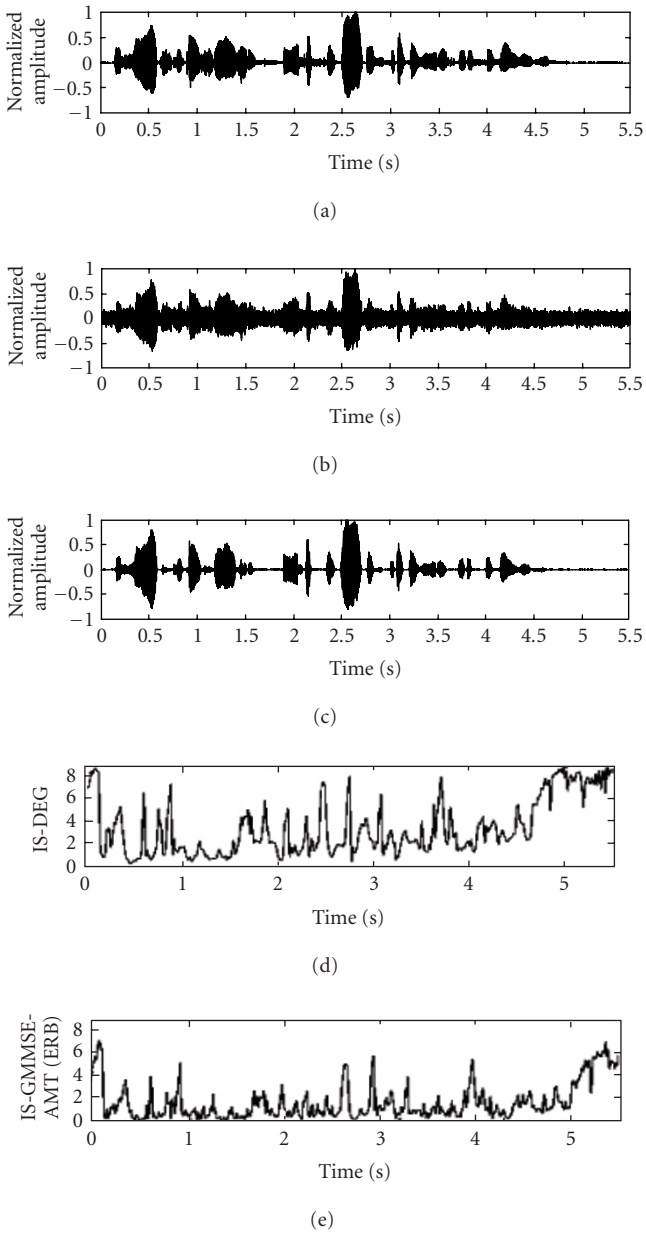

Figure 3: Time domain plots of (a) clean, (b) degraded with FLN noise at 5 dB SNR, and (c) GMMSE-AMT[ERB] enhanced speech "In wage negotiations the industry bargains as a unit with the single union," and IS objective measure versus time for (d) degraded and (e) enhanced speech signals. Average IS measures for degraded and enhanced are 3.23 and 1.8.

Figure 2 compares the excitation pattern based on Schroeder's spreading function and the masking in the ROEX (rounded exponential) model [12]. The excitation pattern does not vary with the level for the critical bands (CB) in the Schroeder model [20]. The excitation pattern for the impaired ear is consistent with broader filter shapes characteristic of sensorineural hearing loss. The excitation pattern is compared with the absolute threshold of hearing and the AMT is set as the greater of the two.

2.3. Scaling issues

Auditory filter shape is dependent on stimulus level [12, 13]. Therefore it is necessary to scale the signal appropriately to represent the actual playback level in dB SPL. This is achieved in the following way.

(a) The output level of the speech waveform is set to 60 dB (SPL) for normal hearing subject and 90 dB (SPL) for individuals with hearing loss.

(b) The maximum dB value of the signal is identified after performing a frame-based FFT analysis of the signal.

(c) A scaling factor is chosen to convert the power spectrum of the signal in dB to a dB (SPL) scale such that the maximum dB (SPL) is limited to 60 dB (SPL) for normal hearing and 90 dB (SPL) for hearing-impaired individuals.

2.4. Audible noise suppression

In our formulation, we use a window frame of the noisy speech $Y_w(i,k)$ and clean speech $X_w(i,k)$ frequency responses in the following power spectral representations (in a manner similar to [1]):

$$X_p(i,k) = \left[\,|X_w(i,k)|^2\right], \qquad Y_p(i,k) = \left[\,|Y_w(i,k)|^2\right]. \tag{17}$$

The noisy speech spectrum is compared with the AMT as calculated in the previous section. The clean speech spectrum is estimated using a nonlinear gain function that is derived using a nonlinear filtering operation for the ith frame and kth subband as shown below [1]:

$$X_p(i,k) = \left[\frac{Y_p(i,k)}{a_b(i) + Y_p(i,k)}\right] Y_p(i,k), \quad \text{with subband } b = k, \tag{18}$$

where the parameter $a_b(i)$ is given by

$$a_b(i) = D_{pb} + \frac{D_{pb}^2}{T_b(i,k)}, \tag{19}$$

where D_{pb} is the mean noise power spectrum of the noise in ERB subband b, and T_b is the masking threshold in the same subband. We can see from (19) that if the noise level approaches the masked threshold $T_b(i,k)$, then the value of $a_b(i)$ approaches $2D_{pb}$, and therefore the suppression in (18) is always greater than the traditional Wiener filter solution (i.e., the Wiener filter solution would have $a_b(i) = D_{pb}$, so $a_b(i) = 2D_{pb}$ will produce a greater suppression value as a function of frequency). If the noise spectrum is below this threshold, no further enhancement processing is performed (as illustrated in Figure 1). The enhanced signal is renormalized[2] and converted back to the time domain.

[2]The renormalization here is essentially converting from power to magnitude spectrum, transforming the frequency domain signal back to the time domain, tracking the maximum and minimum of the waveform to avoid clipping, and finally scaling the input and output signal by a fixed ratio determined by the ratio of their maxima.

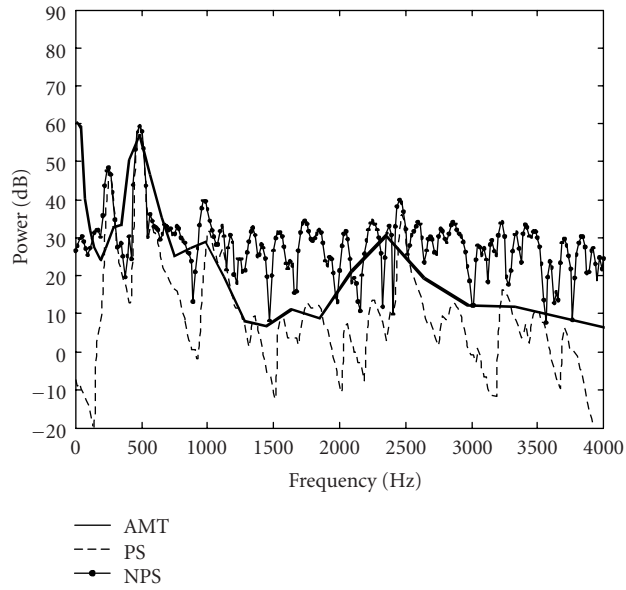

Figure 4: Plot of (i) AMT (solid line), (ii) noisy power spectrum (NPS: solid line with a dot), and (iii) clean power spectrum (PS: dashed line) of the voiced vowel /EY/ for the GMMSE-AMT[ERB-NH] scheme implemented for an individual with normal thresholds.

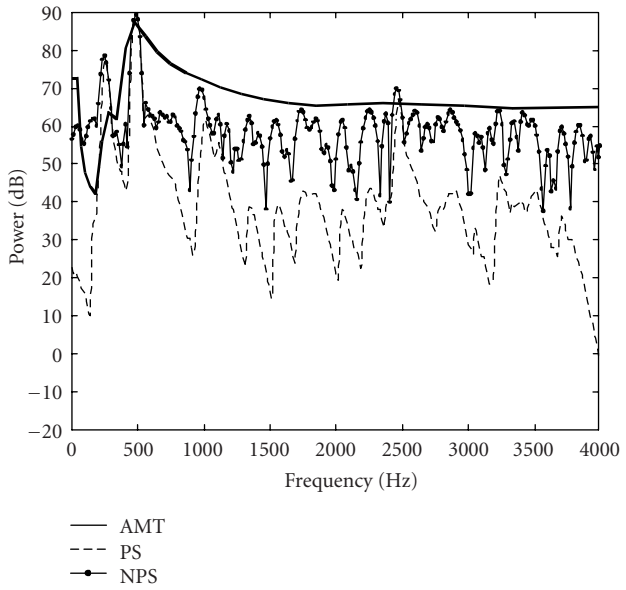

Figure 5: Plots of (i) AMT (solid line), (ii) noisy power spectrum (NPS: solid line with a dot), and (iii) clean power spectrum (PS: dashed line) of the voiced vowel /EY/ for the GMMSE-AMT[ERB-HI] scheme implemented for a typical hearing-impaired listener.

3. EVALUATION

In this section, a detailed performance evaluation is presented for the formulated GMMSE-AMT[ERB] algorithm in the form of objective speech quality results as well as results from subjective speech quality and intelligibility tests. The objective quality of the enhanced speech is assessed in terms of segmental SNRs (SegSNR) as well as the Itakura-Saito (IS) objective speech quality measure [21] for the GMMSE-AMT[ERB-NH] implementation. These measures are explained below in detail. Finally, detailed subjective speech quality tests using a quality rating scale and intelligibility tests using the nonsense syllable test (NST) are presented for individuals with and without hearing loss to assess the performance of the GMMSE-AMT[ERB-NH] and GMMSE-AMT[ERB-HI] algorithm implementations.

For our evaluation, we considered two types of noise with different frequency and temporal structure: (i) stationary flat communications channel noise (FLN), and (ii) large crowd noise from within an open room (LCR). These noise sources have previously been used for speech enhancement and robust speech recognition evaluations [22]. The FLN noise represents a broadband noise source that is quite stationary. The LCR noise is slowly varying and primarily low frequency, where high-frequency (4 kHz) content is approximately 10 dB lower than that seen in the low-frequency region.

3.1. Temporal and spectral plots

Figure 3 shows time waveforms of (i) clean speech, (ii) speech degraded with background FLN noise at 5 dB SNR, and (iii) speech enhanced using the present algorithm (GMMSE-AMT[ERB-NH]) for a single sentence to illustrate detailed processing performance. The processed sentence, "In wage negotiations the industry bargains as a unit with the single union," is taken from the TIMIT speech corpus, and is approximately 5.5 seconds in duration and sampled at an 8 kHz sample rate. Figure 3 also shows the IS objective speech quality measures for the same sentence, (iv) degraded with the FLN 5 dB noise, and (v) enhanced with GMMSE-AMT[ERB-NH]. From this figure, one can observe noticeable noise suppression performed by the GMMSE-AMT[ERB-NH] scheme. The cumulative area under the IS curves in the bottom two panels represents the total amount of distortion as estimated with the IS measure. The enhanced sentence IS plot (v) shows noticeably less distortion than the degraded sentence across the phoneme sequence. This single sentence result has therefore confirmed that the proposed enhancement method provides noise suppression and quality improvement, which is in proportion to the level and type of distortion. We consider a more extensive set of speech enhancement evaluations using objective speech quality measures (overall and within each phoneme) and subjective speech quality measures in the next section. Before considering this, we will briefly consider an example comparison of the AMT used in the GMMSE-AMT[ERB] enhancement scheme.

Figures 4 and 5 show the spectral plots of (1) the noisy speech power spectrum, (2) the clean speech power spectrum, and (3) the audible masked threshold (AMT) for the vowel /EY/ for the GMMSE-AMT[ERB-NH] and GMMSE-AMT[ERB-HI] implementations. Any portion of the power spectrum of the noisy speech that falls below the AMT is

assumed to be inaudible and therefore will not be suppressed. Comparing the AMT for the voiced speech in the NH and HI schemes, one can see that for the HI scheme (Figure 5) there would be far less suppression than in the NH scheme (Figure 4). Because of the pronounced effect of masking in HI individuals, more signal components are masked. On average, noise suppression is performed approximately 80% of the time for the NH scheme and about 40% of the time for HI scheme if we consider each ERB-based filter band and time-based analysis frame. Next, we consider objective measures of processed speech quality over a larger speech corpus.

3.2. Objective quality measures

The performance of an enhancement algorithm can be assessed in two ways: (a) employing objective speech quality measures and/or (b) subjective listener tests, which have as their goal to quantify the improvement/distortion that a human listener would perceive. Two of the most widely used objective quality measures are the segmental SNR (SegSNR) and the Itakura-Saito (IS) distance measure [21, 22]. In normal-hearing listeners, the SegSNR and IS measures have been benchmarked against subjective speech quality measures such as the diagnostic acceptability measure (DAM). The correlation between DAM and IS is 0.59 and between DAM and SegSNR is 0.77. These values are based on a variety of distortions including additive noise, communication distortions, nonlinear distortions, and vocoder distortions [21].

We note that the research performed on objective speech quality measures have focused almost exclusively on measures for predicting speech quality for voice coding applications ([21], [23, Chapter 9]). However, these objective measures have been used extensively to assess the performance of speech enhancement and noise suppression schemes as well. An important issue to note is that for the present study, we employ an AMT. In many objective measures, such as SegSNR, overall speech signal energy and noise signal energy are used within a frame-by-frame basis. Since the purpose of the AMT is to balance noise suppression versus processing artifacts, the AMT is in effect disabling the noise suppression scheme in regions, where further noise suppression would, only introduce, audible processing artifacts. Therefore, for measures such as SegSNR, methods which did not employ an AMT would, in theory, always be selected over those with an AMT since more noise power is left behind (even if that noise is not audible). As such, it would be appropriate to consider a direct comparison of speech enhancement methods that either (i) process noisy speech without an AMT or (ii) employ an AMT, but do not compare between methods that have AMT engaged and disabled. For this reason, we do not report objective measures within our enhancement methods for engaged/disabled AMT processing.

For a broad objective quality evaluation, the 192-sentence core test set in the TIMIT database, with both male and female speakers, was degraded with both stationary (FLN) and nonstationary (LCR) additive noise sources. The noise levels were set at 0 dB and 5 dB SNR. Overall average objective quality measures for the entire 192-sentence TIMIT core set, spoken by both male and female speakers, are presented in Table 1. There are approximately 67 000 speech frames and 8000 silence frames in each test. These results are indicative of the algorithm performance for large speech corpus.

Table 1: Comparison of the objective quality measures across different noise SNRs for the degraded and enhanced GMMSE-AMT[ERB-NH] speech corpus. (SegSNR is in dB, so larger is better; IS measure reflects distortion, so closer to 0 is better.)

	SegSNR		IS	
Noise	DEG	ENH	DEG	ENH
FLN 0 dB	−4.95	−1.63	4.23	2.45
FLN 5 dB	−2.09	0.87	3.35	1.90
FLN 8 dB	−0.62	2.38	2.95	1.64
LCR 0 dB	−4.41	−1.73	3.03	2.16
LCR 5 dB	−1.85	0.59	2.38	1.63
LCR 8 dB	−0.06	2.08	2.01	1.40

Table 2: Comparison of the overall objective quality measures for the degraded speech corpus at 0 dB SNR and the speech corpus enhanced with the TMK algorithm and with the GMMSE-AMT[ERB-NH] speech corpus. (SegSNR is in dB, so larger is better; IS measure reflects distortion, so closer to 0 is better.)

	FLN: Flat comm. noise		LCR: Large crowd noise	
Algorithm	SegSNR	IS	SegSNR	IS
Degraded	−4.95	4.23	−4.41	3.03
TMK	−1.56	2.46	−1.66	2.54
GMMSE-AMT[ERB-NH]	−1.63	2.45	−1.73	2.16

The objective quality results of speech degraded with FLN and LCR noise with different SNRs (0 dB, 5 dB, 8 dB) and enhanced with GMMSE-AMT[ERB-NH] are presented in Table 1 (note that each entry represents an average over 192-TIMIT-sentences). There is a measurable improvement in SegSNRs for both noise types at all SNR levels. There is also a corresponding level of improvement in the IS measure for the enhanced speech over the degraded speech for all conditions (this is especially true for noise types at 5 dB SNR).

Next, we consider performance of the proposed enhancement method with respect to TMK. In Table 2, we present the average SegSNR and IS objective speech quality measures for the 192-TIMIT-sentence test set for FLN and LCR noise distortions at 0 dB SNR. Both noise level (SegSNR) and speech quality (IS) are significantly impacted by both noise sources. Using the TMK algorithm, we performed enhancement for all 192-sentences, and measurable improvement is seen. Since FLN noise is closer to white Gaussian noise, the level of improvement in IS is slightly larger than for the LCR noise, which has multiple speakers in a crowd setting and

is more time varying.[3] The results from Table 2 confirm a similar level of noise suppression, as represented in SegSNR measure, between GMMSE-AMT[ERB-NH] and TMK algorithms. For quality improvement, the performance is comparable for FLN, and GMMSE-AMT[ERB-NH] is slightly better than TMK for LCR. Having considered overall performance, we now wish to examine where in the acoustic phoneme space TMK versus GMMSE-AMT[ERB-NH] shows improvement. In Table 3, we summarize individual IS objective measure performance for each phoneme from the 192 TIMIT sentence test set. The original degraded speech at an SNR of 5 dB with FLN noise is shown under "DEG," and corresponding IS measures for the TMK and proposed enhancement method (labeled as ERB). There are 76 876 frames of speech processed in each case. From this table, we see that GMMSE-AMT[ERB-NH] provides a consistently higher level of quality for nasals, vowels, diphthongs, and semi-vowels. Fricatives and stops resulted in similar level of performance for both enhancement methods. The only class which showed a slight loss for GMMSE-AMT[ERB-NH] was for the silence class (a reduction in IS of 0.15 when going to GMMSE-AMT[ERB-NH] from TMK).

3.3. Listener evaluations

In this section, we describe the procedures used to evaluate the effectiveness of the GMMSE-AMT[ERB-NH] scheme in normal-hearing listeners and the GMMSE-AMT[ERB-NH] and GMMSE-AMT[ERB-HI] schemes in hearing-impaired listeners. Our current evaluation uses a sampling rate of 8000 Hz, which was motivated by our earlier studies on speech enhancement for telephone/telecommunication applications [24], as well as limited computational resources for hearing aid systems.

3.3.1. Listeners

Six listeners with normal hearing and ten listeners with hearing loss participated in this study. Listeners with normal hearing had thresholds of 20 dB HL (ANSI, 1989) or better at octave frequencies from 250–8000 Hz, inclusive. Listeners with hearing loss demonstrated test results consistent with sensorineural pathology: normal tympanometry; absence of otoacoustic emissions in regions of threshold loss and absence of an air-bone gap exceeding 10 dB at any frequency. Listeners with hearing loss had a mild-to-severe hearing loss. All listeners were tested monaurally. Table 4 provides a summary of the characteristics of the listeners with hearing loss, including the audiometric thresholds of the test ear. The test ear of the hearing-impaired listeners was chosen based on the ear with a threshold configuration, allowing the best digital filter design for linear amplification (see below). Listeners were tested individually in a double-walled sound booth. Daily test sessions typically lasted one hour but did not extend beyond two hours. Listeners were compensated 8 USD/hour for their participation.

3.3.2. Stimuli

Speech materials. Two different sets of speech stimuli were used in this study. Speech quality was assessed using 256 sentences from the hearing-in-noise test [25]. Speech intelligibility was assessed using 102 syllables from the CUNY nonsense syllable test [26]. The speech stimuli were digitized at an 8 kHz sampling rate and stored on a Pentium IV computer.

Noise conditions. Speech stimuli were degraded with large crowd room noise (LCR) and flat channel noise (FLN) at overall SNRs of 0 dB and +5 dB.

Signal processing. Digitized speech was degraded with sample noise files with appropriate scaling to generate each SNR. This set of "degraded" signals was then processed by the GMMSE-AMT(ERB) scheme to generate the set of "enhanced" speech signals. In all enhancement processing, the noise spectrum was estimated during an initial portion of silence/noise prior to speech activity, and this estimate was kept constant across the syllable (NST material) or sentence (HINT). The GMMSE-AMT(ERB) scheme was applied in two ways. The first approach GMMSE-AMT(ERB-NH) used thresholds and auditory filter bandwidths characteristic of a normally functioning auditory system. Both listener groups were evaluated with the GMMSE-AMT(ERB-NH) approach. Implemented only for the hearing-impaired listener group, the second approach GMMSE-AMT(ERB-HI) used thresholds and auditory filter bandwidths characteristic of sensorineural hearing loss. Customized for each individual hearing-impaired listener, the GMMSE-AMT(ERB-HI) implementation adjusted the spread-of-masking functions based on individual thresholds and auditory filter bandwidths [14].

Table 5 provides a summary of the stimulus conditions. Quality and intelligibility were measured in a total of eight conditions for the normal-hearing group (2 noise types with 2 SNRs for 2 processing conditions) and a total of 12 conditions for the hearing-impaired group (2 noise types with 2 SNRs for 3 processing conditions).

3.3.3. Equipment

For listener presentation, the digitally stored stimuli went through a digital-to-analog converter (TDT AP2, DD1), a 4000 Hz anti-aliasing filter (TDT FT3), an attenuator (TDT PA4), and a headphone buffer (TDT HB6). Finally, the stimuli were presented monaurally to the test ear of each listener through a TDH-49 earphone.

3.3.4. Presentation level

All stimuli were presented to normal-hearing listeners at an equalized RMS level of 60 dB SPL. Because listeners with hearing loss were not wearing hearing aids, the preprocessed stimuli were frequency-shaped through digital filtering to simulate amplification. Thus, the stimuli presented to the hearing-impaired subjects through headphones was an amplified version of the signal presented to the normal-hearing

[3] We note that the study by Hansen and Arslan [24] does compare stationarity of FLN, LCR, and other noise sources in the context of speech enhancement and robust speech recognition in noise.

Table 3: A comparison of individual phoneme Itakura-Saito objective speech quality measures for the 192-TIMIT-sentence test set for FLN noise at 5 dB SNR (labeled DEG), TMK processed, and GMMSE-AMT[ERB-NH] (labeled ERB) enhancement algorithms. Here, #Fr refers to the number of frames for each individual phoneme with a total of 76 870 frames in the test set.

Objective speech quality across American phonemes										
Ph.		DEG	TMK	ERB	#Fr	Ph.		DEG	TMK	ERB	#Fr
Consonants-nasals						Consonant-unvoiced stops					
/m/	me	3.508	1.886	1.743	1645	/p/	pan	2.447	1.270	1.302	796
/n/	no	3.847	2.141	1.932	2270	/t/	tan	1.914	0.925	0.909	1114
/ng/	sing	3.955	2.230	1.967	402	/k/	key	2.293	1.204	1.190	1132
/nx/	many	1.605	0.867	0.823	141	Consonant-voiced stops					
/em/	problem	3.612	2.573	1.944	37	/b/	be	2.012	1.071	0.995	304
/en/	traction	3.984	2.309	1.994	283	/d/	dawn	2.228	1.142	1.059	375
/eng/	greasing	3.243	1.366	1.264	6	/g/	give	2.399	1.177	1.187	255
Consonant-unvoiced fricatives						Consonant-closure stops					
/s/	sip	2.819	1.341	1.299	4892	/tcl/	it pays	5.519	3.486	3.489	1732
/th/	thing	4.042	2.146	2.116	392	/kcl/	pockets	5.847	3.737	3.789	1583
/f/	fan	3.248	1.503	1.508	1825	/bcl/	to buy	6.255	4.127	3.929	972
/sh/	show	1.772	0.924	0.820	1109	/dcl/	sandwich	5.226	3.392	3.215	1212
Consonant-voiced fricatives						/gcl/	iguanas	5.577	3.616	3.467	527
/z/	zip	3.232	1.596	1.446	2036	/pcl/	accomplish	6.593	4.291	4.282	1247
/zh/	garage	1.960	0.920	0.910	115	Consonant-glottal stop flap					
/dh/	that	2.807	1.494	1.421	630	/q/	_allow	3.017	1.695	1.627	898
/v/	van	3.378	1.749	1.596	741	/dx/	put_in	1.699	0.924	0.862	327
Consonants-affricates						Consonant-unvoiced whisper					
/jh/	joke	2.354	1.270	1.164	357	/hh/	had	2.761	1.451	1.398	414
/ch/	chop	2.263	1.092	1.111	477	Consonant-voiced whisper					
						/hv/	you have	2.148	1.208	1.280	275
Vowels-front						Diphthongs					
/ih/	hid	1.433	0.812	0.755	2070	/ay/	hide	1.046	0.633	0.588	1818
/eh/	head	1.225	0.695	0.657	2265	/oy/	coin	1.712	0.896	0.769	396
/ae/	had	0.996	0.575	0.557	1940	/ey/	pain	1.161	0.664	0.602	2064
/ux/	to buy	1.999	1.005	0.952	603	/ow/	code	2.072	1.083	1.045	1540
Vowels-mid						/aw/	pout	1.267	0.791	0.697	696
/aa/	odd	1.507	0.865	0.758	2227	/iy/	new	1.712	0.983	0.835	2841
/er/	earth	2.146	1.186	1.110	1582	Semivowel-liquids					
/ah/	up	1.556	0.870	0.828	1524	/r/	ran	2.279	1.245	1.206	2071
/ao/	all	2.105	1.167	1.004	1622	/l/	lawn	2.397	1.361	1.258	1895
Vowels-back						/el/	chemicals	3.194	1.809	1.693	702
/uw/	boot	2.466	1.378	1.349	313	Semivowels-glides					
/uh/	foot	1.972	1.119	1.077	295	/w/	wet	3.095	1.715	1.619	1179
Vowel-front schwa						/y/	you	1.743	0.987	0.890	390
/ix/	heed	2.508	1.332	1.249	2527	Silence					
Vowel-back schwa						/#/	extended	7.479	4.739	4.882	9716
/ax/	a ton	2.627	1.448	1.387	1119	/pau/	pause	6.000	3.599	3.589	1158
Vowel-retroflexed schwa						/epi/	epenthetic	4.881	2.729	2.621	253
/axr/	after	2.877	1.617	1.605	1488						
Vowel-voiceless schwa						Overall		3.345	1.950	1.904	76870
/ax-h/	sub	3.846	2.230	1.890	55	Overall-/#/		2.747	1.547	1.473	67154

Table 4: Age (yrs), test ear (left/right), and audiometric thresholds (in dB HL) of listeners with hearing loss. ("na" means threshold measurements were not available.)

HI listener	Age	Test ear	250 Hz	500 Hz	1000 Hz	2000 Hz	4000 Hz	6000 Hz	8000 Hz
1	65	R	30	55	60	65	55	50	70
2	40	R	25	35	50	60	75	70	55
3	44	R	30	25	35	55	105	na	90
4	23	R	35	35	50	60	50	na	35
5	70	L	85	80	65	60	50	45	55
6	80	L	25	10	15	45	70	70	70
7	25	R	5	5	15	40	40	55	50
8	59	R	5	5	5	50	70	50	40
9	56	L	5	15	35	70	90	85	90
10	47	L	15	15	25	30	45	55	60

Table 5: Conditions in which subjective speech intelligibility and quality were evaluated for the group of normal-hearing listeners (NH) and the group of hearing-impaired listeners (HI).

Group	Normal-hearing (NH)	Hearing-impaired (HI)
Noise type	(1) Flat channel noise (FCN) (2) Large crowd noise (LCR)	(1) Flat channel noise (FCN) (2) Large crowd noise (LCR)
SNR	(1) 0 dB (2) 5 dB	(1) 0 dB (2) 5 dB
Processing conditions	(1) Degraded (2) GMMSE-AMT[ERB-NH]	(1) Degraded (2) GMMSE-AMT[ERB-NH] (3) GMMSE-AMT[ERB-HI]

subjects, with the amplification approximating the linear gain prescribed by the NAL-R fitting procedure [15].

3.3.5. Speech quality ratings

The categorical rating scales used for the quality ratings are the same as those used by Neuman et al. [27] and are similar to those developed by Gabrielson et al. [28]. A 10-point rating scale was used to obtain ratings on five different stimulus attributes: clarity, pleasantness, background noise, loudness, and overall impression, with a rating of "0" being worst and a rating of "10" being best. Listeners used a written response form containing the five quality scales to record their ratings. For each condition, participants listened to a block of 30 of the 256 HINT sentences and then used the 10-point scales to rate the quality of the speech for each of the five attributes. The starting sentence for each block of 30 sentences was randomly selected such that on one block of trials the subject would listen to sentences 45 through 75, on the next block sentences 125 through 155 and so forth. A set of quality ratings consisted of ratings on each of the five attributes in each of the eight conditions. The order of the conditions in each set was randomized. Three sets of quality ratings were obtained. Each set took about 40 minutes to complete.

3.3.6. Intelligibility

Nonsense syllable test. The nonsense syllable test (NST) [26, 29] is a closed-set test in which a listener hears a nonsense syllable and then chooses between seven and nine response alternatives. The test consists of 102 syllables contained in 11 subtests, each of which contains between seven and nine syllables. The subtests differ in terms of voicing and position of consonants as well as the vowel. The order of presentation of the 102 nonsense syllables was randomized on each block of trials. The intelligibility session for each listener included one 102-syllable list in each condition, with the order of the conditions randomized within the set. The overall measure of performance is the percentage of correctly identified nonsense syllables.

3.4. Results

3.4.1. Speech quality ratings

Speech quality ratings for each attribute were first averaged for the three trials for each listener. Ratings were then averaged across listeners in each group. Average ratings for the five attributes of quality for the normal-hearing listeners and hearing-impaired listeners are shown in Figure 6. A separate repeated measures analysis of variance (ANOVA) was done for each quality attribute for each of the listener groups. Listener groups were considered separately because the number of processing conditions differed between the two groups. The results of these statistical analyses are shown in Table 6. Enhancement with the GMMSE-AMT[ERB] technique resulted in significant benefit in quality ratings on

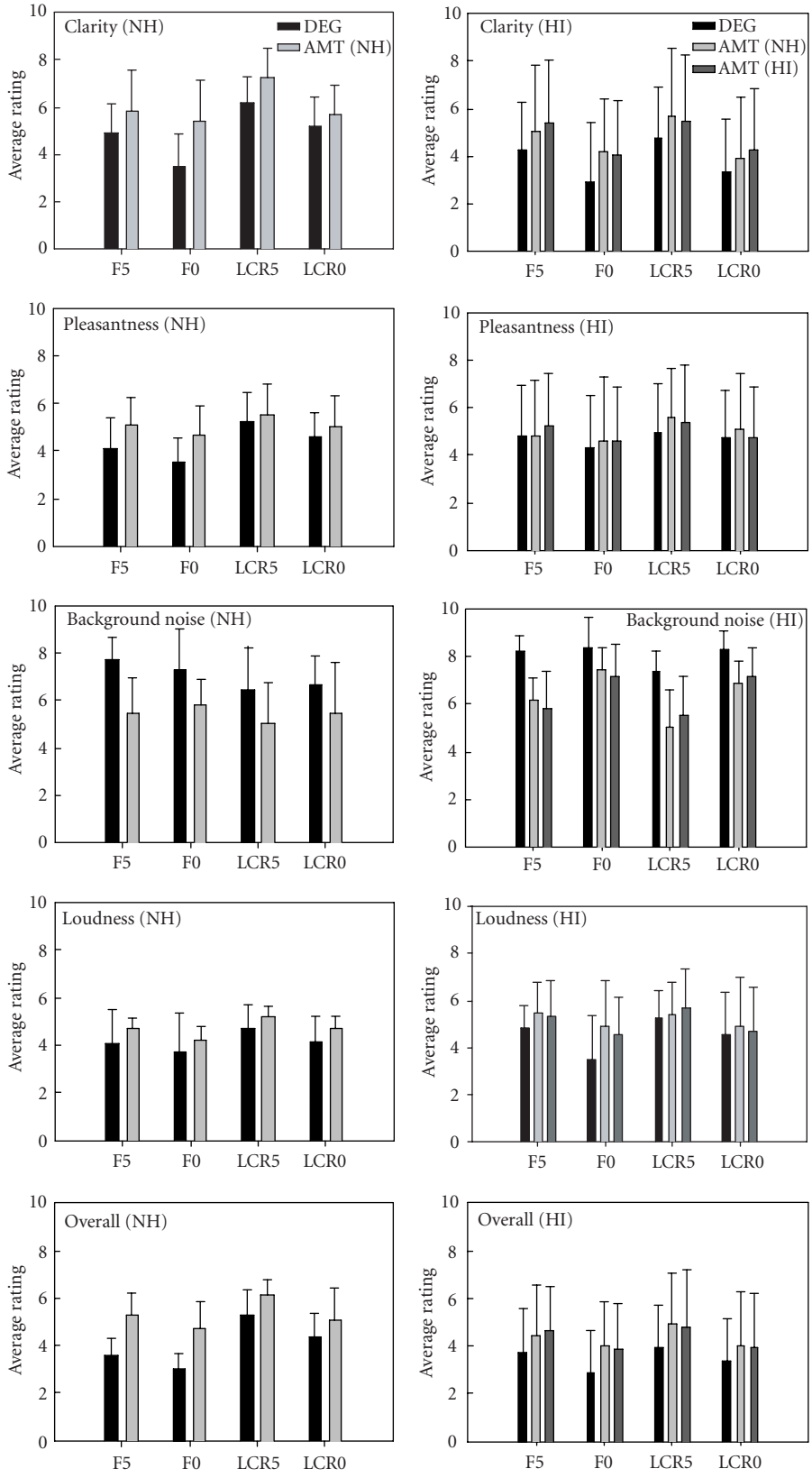

Figure 6: Average ratings of the normal-hearing listeners (left column) and the hearing-impaired listeners (right column) for the five attributes of quality (clarity, pleasantness, background noise, loudness, and overall impression) for degraded (DEG) and enhanced (AMT(NH) and AMT(HI)) speech conditions.

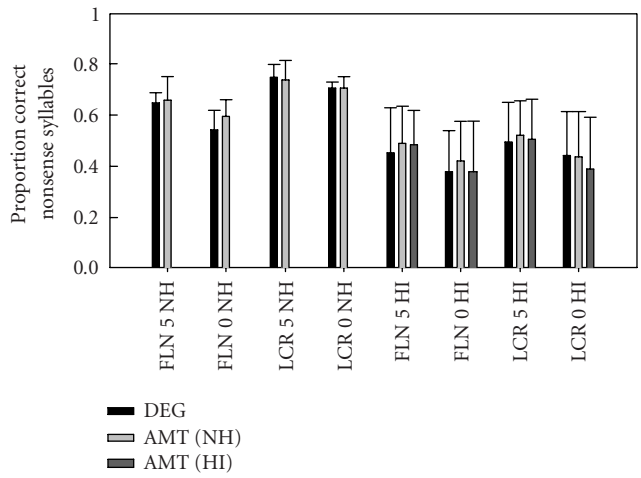

Figure 7: Intelligibility percent-correct scores on the nonsense syllable test scores for normal-hearing listeners (left) and for hearing-impaired listeners (right) for degraded (DEG) and enhanced (AMT(NH) and AMT(HI)) speech conditions.

several attributes in both subject groups. In normal-hearing listeners, enhancement resulted in significantly less noisy ratings, better clarity ratings, and better overall quality ratings. In hearing-impaired listeners, enhancement resulted in significantly better clarity ratings, significantly less noisy ratings, and significantly better overall quality ratings. In the hearing-impaired group, loudness ratings increased slightly (albeit significantly) in the enhanced conditions. Increasing SNR had a significant effect on four of the five rating scales in each listener group (NH: ratings of clarity, pleasantness, loudness, and overall quality; HI: ratings of clarity, background noise, loudness, and overall quality). Overall variability was greater in the HI group versus the NH group. In the normal-hearing group, noise type was a significant factor in quality ratings: LCR was consistently rated more favorably compared to FLN. In both listener groups, the (processing $\times$ SNR) interaction was significant for the background noise scale: stimuli enhanced with GMMSE-AMT[ERB] showed significantly larger changes (decreases) in ratings of noisiness in the 5 dB SNR condition.

3.4.2. Intelligibility: NST

Figure 7 shows NST scores (in proportion correct) for degraded and enhanced conditions for both normal-hearing listeners (left) and hearing-impaired listeners (right). The NST percent-correct scores were first subjected to an arcsin transform [30] and then submitted to repeated measures ANOVAs. The ANOVA results are shown in Table 7. NST scores were better (20% on average) and less variable in the normal-hearing listeners than in the hearing-impaired listeners. In the normal-hearing group, the main effects of noise and SNR were significant: intelligibility scores were better in the +5 dB SNR condition and for the LCR noise. In the hearing-impaired group, the only significant main effect was SNR. Enhancement did not significantly affect intelligibility scores in either group.

4. DISCUSSION AND CONCLUSIONS

In this study, we have considered the problem of speech enhancement in diverse environmental conditions using a speech enhancement scheme that employs an auditory masking threshold (AMT) to balance the degree of noise suppression versus perceived processing artifacts. The goals of this study have been to (i) modify the suppression structure to incorporate the modified generalized minimum mean square error (GMMSE) estimators, and (ii) establish a working framework for speech enhancement which directly incorporates the hearing response of individual hearing-impaired listeners. This approach was motivated by the earlier study that resulted in the TMK algorithm [1], which showed a substantial level of intelligibility improvement as measured by the DRT (diagnostic rhyme test) for individuals with normal-hearing. Motivated by this first demonstration of intelligibility improvement in the speech enhancement literature, we previously developed an approach which improved on the estimation of the AMT [9] and also evaluated the improved procedure using quality measures and formal DRT testing [9]. We saw that an approach that improves on the estimation of the AMT and integrates this into a generalized MMSE noise suppression algorithm [10, 11] does improve quality, but the level of intelligibility improvement was only modest for normal-hearing individuals [9]. Even so, we feel that these prior studies served as an important foundation to develop improved noise suppression schemes for hearing-impaired persons, and, in theory, should offer the potential to develop more effective automatic speech processing algorithms for digital hearing aids, which could both improve quality and intelligibility.

The present study has considered a revised formulation that is more suitable for hearing aid applications and incorporated the following processing phases: (i) a modified generalized minimum mean square error estimator (GMMSE) was employed, (ii) the frequency resolution of the cochlea was represented using the auditory filter equivalent rectangular bandwidths (ERBs) rather than the critical band scale, (iii) estimation of the auditory masking threshold and spreading functions for masking were adjusted to address the elevated thresholds and broader auditory filters that result from sensorineural hearing loss, and (iv) the current algorithm did not include the tonality offset developed for use in MPEG-4 audio coding applications, since it is based more on the harmonic structure of sounds associated with music. After developing the GMMSE-AMT[ERB] noise suppression scheme, we specialized the approach to those with normal hearing and hearing impaired listeners (i.e., NH and HI algorithm versions). The output level of the speech waveform was set to different levels for normal and hearing-impaired individuals. The algorithm was evaluated using large crowd room noise and flat communications channel noise at two separate SNRs. Using objective speech quality measures, the output SegSNR performance improved from 2.44 to 3.32 dB over the

Table 6: Summary of the main effects (processing, noise, SNR) from the analysis of variance carried out for the five attributes of quality using HINT sentences for each listener group: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$. F-values are also reported for significant interactions.

	Clarity	Pleasant	Background noise	Loudness	Overall
Normal-hearing group	F (1,5)	F (1,5)	F (1,5)	F (1,5)	F (1,5)
Processing	20.3**	4.5	16.6**	2.0	25.7**
Noise	41.4***	30.9**	7.7*	10.5*	48.3***
SNR	15.2**	6.6*	0.30	24.0**	15.6*
(Processing × SNR)	—	—	9*	—	—
Noise X SNR	—	—	—	—	13*
Hearing-impaired group	F (1,5)	F (1,5)	F (1,5)	F (1,5)	F (1,5)
Processing	4.5*	0.886	25***	5.3*	6.8**
Noise	0.74	1.4	3.1	2.6	0.745
SNR	8.5*	0.488	13.1*	5.12*	6.3*
(Processing × SNR)	—	—	10.9***	—	—
(Processing × noise)	—	—	—	4.2*	—

Table 7: Summary of main effects for ANOVA for NST scores for each listener group (NH, HI) with factors of processing, noise, and signal-to-noise ratio (SNR). Significant interaction effects are also listed. $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

Source	NH group		HI group	
	df	F (NH)	df	F (HI)
Processing	1,3	1.7	2,18	2.9
Noise	1,3	31.4*	1,9	2.9
SNR	1,3	11.8*	1,9	23.2**
Processing × noise	1,3	28.3*	—	—
Processing × SNR	—	—	2,18	4.7*

original degraded corpus. Using the Itakura-Saito objective quality measure, the level of distortion was measurably reduced from an initial degraded level of 2.38–4.23 down to 1.63–2.45, improvements ranging from 0.75 to 1.78. This improvement came within the acoustic phoneme space primarily in nasals, vowels, diphthongs, and semi-vowels, with the same performance for stops and fricatives.

Next, formal listener evaluations using 6 normal and 10 hearing-impaired individuals were performed for quality using HINT sentences and intelligibility using the CUNY nonsense syllable test. For subjective quality tests, a measurable level of speech quality improvement and background noise reduction were obtained with GMMSE-AMT[ERB-NH] for NH and HI listeners. The GMMSE-AMT[ERB-HI] version of the enhancement algorithm also showed quality improvement over the original degraded materials. However, results with GMMSE-AMT[ERB-HI] and GMMSE-AMT[ERB-NH] were similar. Customization of the AMT did not show significant advantages over the uncustomized (default NH version) method in listener ratings of quality.

Formal intelligibility evaluations using NST materials showed either a slight improvement, the same, or a slight reduction across the four noise conditions for GMMSE-AMT[ERB-HI] and GMMSE-AMT[ERB-NH] algorithm configurations. This is in stark contrast to the level of intelligibility improvement reported in [1] for normal-hearing individuals. As addressed in [9], possible reasons for discrepancies reported between [1] and our work include (i) differences in sampling rate/bandwidth, (ii) use of a voice activity detector with noise spectral update in [1] versus a single initial noise estimate for our studies, (iii) differences in linguistic backgrounds (Greek versus English) of listeners, and (iv) procedures used for listener evaluations. Finally, while the present study established a framework for customization, the customized implementation was not significantly better for hearing-impaired listeners. In the present formulation, two steps are crucial for speech enhancement: these include the particular method for estimating the AMT, and second the particular method used to perform the noise suppression given the AMT. Given the results from the present study, it is natural to ask if

(i) the noise suppression was not capable of taking full advantage of the customization for individual hearing responses; and/or

(ii) whether there remains an error in how the AMT estimation is performed for HI listeners; and finally,

(iii) whether there is additional knowledge or information, either separate or in addition to the AMT, needed to perform effective customized noise suppression for HI listeners.

In future studies, it would be useful to consider these three issues. Also, we maintained a single noise spectral estimate across the speech sentence, and engaging the voice activity detector to update noise estimates as well as α in the GMMSE enhancement scheme could improve performance. We believe that it would be possible to incorporate a codebook-based AMT scheme such as that in [31] for individuals with cochlear hearing loss. Such an approach would require extensive modeling of the particular types of hearing loss for each listener, and to incorporate this bias into the AMT codebook entry selection process.

ACKNOWLEDGMENT

This work was sponsored by a Grant from the Whitaker Foundation, and in part from the US Navy through SPAWAR Systems under Grant no. N66001-03-1-8905.

REFERENCES

[1] D. E. Tsoukalas, J. N. Mourjoupoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech Audio Processing*, vol. 5, no. 6, pp. 497–514, 1997.

[2] J. H. L. Hansen and S. Nandkumar, "Robust estimation of speech in noisy backgrounds based on aspects of the auditory process," *Journal of the Acoustical Society of America*, vol. 97, no. 6, pp. 3833–3849, 1995.

[3] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 2, pp. 126–137, 1999.

[4] M. Klein and P. Kabal, "Signal subspace speech enhancement with perceptual post-filtering," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '02)*, vol. 1, pp. 537–540, Orlando, Fla, USA, May 2002.

[5] Y. Hu and P. C. Loizou, "A perceptually motivated approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 5, pp. 457–465, 2003.

[6] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 11, no. 6, pp. 700–708, 2003.

[7] J. D. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE J. Select. Areas Commun.*, vol. 6, no. 2, pp. 314–323, 1988.

[8] R. D. Patterson and B. C. J. Moore, "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selectivity in Hearing*, pp. 123–177, Academic Press, London, UK, 1986.

[9] K. H. Arehart, J. H. L. Hansen, S. Gallant, and L. Kalstein, "Evaluation of an auditory masked threshold noise suppression algorithm in normal-hearing and hearing impaired listeners," *Speech Communications*, vol. 40, no. 4, pp. 575–592, 2003.

[10] V. Radhakrishnan, "Speech enhancement based on generalized minimum mean square error estimation & masking property of the human auditory system," Master's thesis, University of Colorado, Boulder, Colo, USA, 2002.

[11] J. H. L. Hansen, V. Radhakrishnan, and K. H. Arehart, "Speech enhancement based on generalized minimum mean square error estimators and masking properties of the auditory system," to appear in *IEEE Trans. Speech & Audio Proc.*

[12] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[13] B. C. J. Moore, *Perceptual Consequences of Cochlear Damage*, Oxford Psychology Series, Oxford University Press, Oxford, UK, 1995.

[14] B. C. J. Moore and B. R. Glasberg, "A model of loudness perception applied to cochlear hearing loss," *Auditory Neuroscience*, vol. 3, pp. 289–311, 1997.

[15] D. Byrne and H. Dillon, "The national acoustic laboratories (NAL) new procedure for selecting the gain and frequency response of a hearing aid," *Ear and Hearing*, vol. 7, no. 7, pp. 257–265, 1986.

[16] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 345–349, 1994.

[17] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square short time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[18] H. Buckholz, *The Confluent Hypergeometric Function*, Springer, New York, NY, USA, 1969.

[19] L. Burget and P. Moticek, "Noise estimation for efficient speech enhancement and robust speech recognition," in *Proc. 7th International Conference on Spoken Language Processing (ICSLP '02)*, vol. 2, pp. 1033–1036, Denver, Colo, USA, September 2002.

[20] M. R. Schroeder, B. S. Atal, and J. L. Hall, "Optimizing digital speech coders by exploiting masking properties of the human ear," *Journal of the Acoustical Society of America*, vol. 66, no. 6, pp. 1647–1652, 1979.

[21] S. R. Quakenbush, T. P. Barnwell, and M. A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1988.

[22] J. H. L. Hansen, "Speech enhancement," in *Encyclopedia of Electrical and Electronics Engineering*, vol. 20, pp. 159–175, John Wiley & Sons, New York, NY, USA, 1999.

[23] J. R. Deller Jr., J. H. L. Hansen, and J. G. Proakis, *Discrete Time Processing of Speech Signals*, IEEE Press, New York, NY, USA, 2nd edition, 2000.

[24] J. H. L. Hansen and L. M. Arslan, "Robust feature-estimation and objective quality assessment for noisy speech recognition using the credit card corpus," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 3, pp. 169–184, 1995.

[25] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," *Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1085–1099, 1994.

[26] S. B. Resnick, J. R. Dubno, S. Hoffnung, and H. Levitt, "Phoneme errors on a nonsense syllable test," *Journal of the Acoustical Society of America*, vol. 58, S114, Suppl. 1, 1975.

[27] A. C. Neuman, M. H. Bakke, C. Mackerise, S. Hellman, and H. Levitt, "The effect of compression ratio and release time on categorical rating of sound quality," *Journal of the Acoustical Society of America*, vol. 103, no. 5, pp. 2273–2281, 1998.

[28] A. Gabrielson, B. Hagerman, T. Bech-Kristensen, and G. Lundberg, "Perceived sound quality of reproductions with different frequency and sound levels," *Journal of Acoustical Society of America*, vol. 88, no. 3, pp. 1359–1366, 1990.

[29] J. R. Dubno and D. D. Dirks, "Evaluation of hearing-impaired listeners using a Nonsense-Syllable Test. I. Test reliability," *Journal of Speech and Hearing Research*, vol. 25, no. 1, pp. 135–141, 1982.

[30] G. A. Studebaker, "A "rationalized" arcsin transformation," *Journal of Speech and Hearing Research*, vol. 28, no. 3, pp. 455–462, 1985.

[31] R. Sarikaya and J. H. L. Hansen, "Auditory masking threshold estimation for broadband noise sources with application to speech enhancement," in *Proc. European Conference on Speech Communication and Technology (EUROSPEECH '99)*, vol. 6, pp. 2571–2574, Budapest, Hungary, September 1999.

Ajay Natarajan was born in New Delhi, India. In 2000, he received his B.S. degree in electrical engineering from Karnataka Regional Engineering College, India. He was a systems engineer at Wipro Technologies, Bangalore, before joining the University of Colorado, Boulder, in the fall of 2001. He received his M.S. degree in electrical and computer engineering from the University of Colorado. His research interests include digital speech processing, psychoacoustics, and speech recognition. He was a Research Assistant in the Speech, Language, and Hearing Sciences Department (SLHS) with Professors Hansen and Arehart. He was also a software programmer at SLHS, where he was responsible for implementing DSP-related software for auditory tests. He developed a Matlab package for auditory processing. He was the Webmaster for *the International Speech Conference ICSLP 2002*. He received the graduate-level interdisciplinary certification in human language technology from the Center for Speech Language Research (CSLR) and Java certification from SUN Microsystems. He is currently working as an Associate IVR Developer *VoiceLog*, Broomfield, Colorado.

John H. L. Hansen received the Ph.D. and M.S. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, Georgia, in 1988 and 1983, and the B.S.E.E. degree from Rutgers University, New Brunswick, New Jersey, in 1982. He is the Department Chairman and Professor in the Electrical Engineering Department, where he holds the Distinguished Chair in telecommunications engineering, Erik Jonsson School of Engineering and Computer Science, and Professor in speech and hearing, School of Brain and Behavioral Sciences, University of Texas at Dallas, Richardson, Texas. At UTD, he established the Center for Robust Speech Systems (CRSS) which is part of the Human Language Technology Institute. Previously, he served as Department Chairman and Professor in the Department of Speech, Language and Hearing Sciences (SLHS), and Professor in the Department of Electrical & Computer Engineering, University of Colorado at Boulder, where he cofounded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities in CRSS at UTD. He is serving as the IEEE Signal Processing Society Distinguished Lecturer for 2005, and Member of the IEEE Speech Technical Committee, and has served as Technical Advisor to US Delegate for NATO (IST/TG-01), Associate Editor for the IEEE Transactions on Speech & Audio Processing (1992–1999), and an Associate Editor for the IEEE Signal Processing Letters (1998–2000). His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human-computer interaction. In 2005, he was the recipient of the University of Colorado Teacher Recognition Award, and was the General Chairman for the International Conference on Spoken Language Processing, ICSLP-2002.

Kathryn Hoberg Arehart received her B.S. degree in biological sciences from Stanford University in 1984. She received an M.S. degree in 1987 and a Ph.D. degree in 1992, both in speech and hearing sciences from the University of Washington, Seattle. She also has a clinical certification in audiology from the American Speech-Language-Hearing Association. In 1992, she joined the Faculty of the Speech, Language, and Hearing Sciences, the University of Colorado, Boulder, where she now is an Associate Professor. Her research interests include auditory perception by listeners with cochlear hearing loss and design and evaluation of signal processing algorithms for hearing aids.

Jessica Rossi-Katz is a certified audiologist and a doctoral candidate in the Speech, Language and Hearing Sciences Department, the University of Colorado, Boulder. Her research investigates the process by which listeners with and without hearing loss selectively attend to competing speech information.