Perceptual Based Speech Enhancement For Normal-Hearing & Hearing-Impaired Individuals*

Ajay Natarajan[1,2], *John HL Hansen*[1,2], *Kathryn Arehart*[2], *Jessica A Rossi-Katz*[2]

Center for Spoken Language Research[1], Department of Speech, Language and Hearing Sciences[2]
University of Colorado, Boulder, USA

[nataraja,hansen,arehart,rossija]@colorado.edu

Abstract

This paper describes a new noise suppression scheme with the goal of improving speech-in-noise perception for hearing-impaired listeners. Following the work of Tsoukalas et al. (1997) [4], Arehart et al (2003) [3] implemented and evaluated a noise suppression algorithm based on an approach that used the auditory masked threshold in conjunction with a version of spectral subtraction to adjust the enhancement parameters based on the masked threshold of the noise across the frequency spectrum. That original formulation was based on masking properties of the normal auditory system, with its theoretical underpinnings based on MPEG-4 audio coding [6]. We describe here a revised formulation, which is more suitable for hearing aid applications and which addresses changes in masking that occur with cochlear hearing loss. In contrast to previous formulations, the algorithm described here is implemented with generalized minimum mean square error estimators, which provide improvements over spectral subtraction estimators [1]. Second, the frequency resolution of the cochlea is described with auditory filter equivalent rectangular bandwidths (ERBs) [2] rather than the critical band scale. Third, estimation of the auditory masked thresholds and masking spreading functions are adjusted to address elevated thresholds and broader auditory filters characteristic of cochlear hearing loss. Fourth, the current algorithm does not include the tonality offset developed for use in MPEG-4 audio coding applications. The scheme also shows an overall improvement of 11% in the Itakura-Saito distortion measure.

1. Introduction

Individuals with cochlear hearing loss have much more difficulty understanding speech, than normal-hearing people. This effect is compounded in noisy environments. This increased difficulty 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 cochlear hearing loss [10].

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. Digital signal processing hearing aids may also include algorithms for feedback cancellation and for active noise reduction. Spectral subtraction is one possible noise-reduction algorithm for hearing aid applications. Generally, noise reduction circuits employing spectral subtraction use mathematical criteria based on the estimated speech-to-noise ratio. A primary issue is achieving a balance between pure noise suppression within a signal-plus-noise model using a mathematical-based criteria such as signal-to-noise ratio and the musical noise artifacts caused by the processing techniques.

Tsoukalas et al. [4] used a spectral subtraction technique based on aspects of the auditory process. Their method considers an enhancement approach that uses the auditory masked threshold (AMT) [6] in conjunction with a version of spectral subtraction to adjust the parameters used in the subtraction process based on the masked threshold of the noise across the frequency spectrum. The AMT is calculated in a four step process: 1) obtain energies in speech critical band (CB) analysis 2) convolve the spreading function [12] with the CB spectrum to obtain a spread masking threshold 3) compute an offset term for spread masking thresholds to take into account signal tonality and 4) normalize/compare and account for absolute auditory thresholds.

Based on the work in [4], Arehart et al. [3] implemented an AMT-noise suppression algorithm and evaluated its effectiveness in improving speech-perception in noise for both normal-hearing and hearing-impaired listeners. The AMT-NS algorithm yielded significantly better quality ratings and significantly better intelligibility scores in both normal hearing and hearing impaired listeners in some but not all of the test conditions.

The algorithm implemented in [4] and [3] is based on masking properties of the normal auditory system, with its theoretical underpinnings based on MPEG-4 audio coding [6]. Alternate processing strategies that specifically consider hearing aid applications and the effects of cochlear hearing loss may optimize the AMT-NS approach to speech enhancement for hearing-impaired listeners.

This study describes a new noise suppression scheme, with the goal of improving speech-in-noise perception by hearing-impaired listeners. This paper first describes the algorithm development, then presents objective measures of algorithm performance, and finally discusses details about the algorithm formulation for hearing-impaired listeners.

2. Algorithm Development

The flowchart of the algorithm is given in Fig 1. The algorithm can be broken down into 1) Enrollment, 2) AMT estimation, and 3) Noise suppression. The steps for the normal-hearing and

*This work has been supported by a grant from the Whitaker Foundation and National Science Foundation under the cooperative agreement IIS-9817485

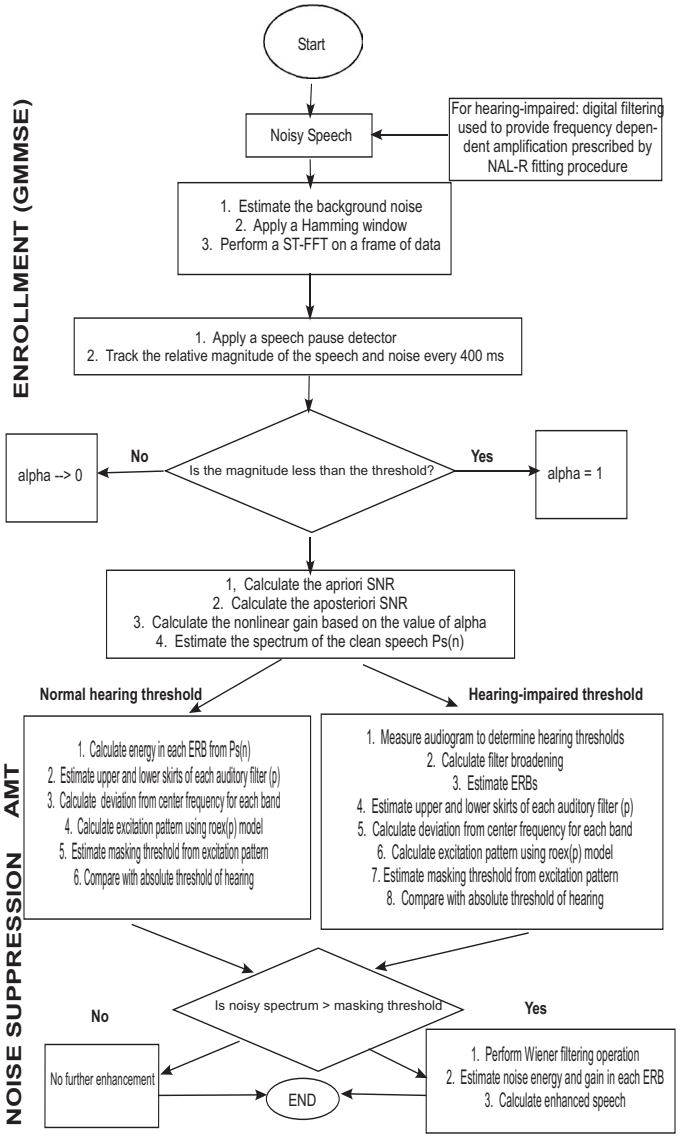

Figure 1: Flow chart of the enhancement algorithm

hearing-impaired differ in a) the estimation of the AMT i.e. the broadening of the excitation pattern due to broader filters and elevated thresholds, and b) frequency-dependent amplification approximating the linear gain prescribed by the NAL-R hearing aid fitting procedure [13].

2.1. GMMSE-Enrollment

The purpose of this step is to obtain an estimate of the clean speech power spectrum, that is needed to calculate the AMT. The speech is assumed to be degraded with additive noise and the speech and noise segments are uncorrelated as in Eqn(1).

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

The short term power spectrum is calculated by applying a Hamming window to a frame of speech. Under this assumed model, one can obtain a family of MMSE speech spectral estimators as,

$$\hat{X_p} \quad = \quad \left(E\left\{X_p^{\alpha}|Y_p\right\}\right)^{1/\alpha}. \tag{2}$$

A small value of $\alpha \rightarrow 0$ is suitable for increased noise suppression and improving the segmental SNR [1], whereas a higher value of $\alpha, \alpha \rightarrow 1$ reduces the amount of musical processing artifacts and speech distortions. This suggests a method to dynamically change the value of α, rather than using 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 then we change the value of $\alpha \rightarrow 0$, and in regions of speech, the value α is set to 1. The speech/pause algorithm [5] used to dynamically change α is described below. Let P_{nk} be the power spectrum of the noise for the k^{th} subband, and P_{xk} be the power spectrum of the noisy speech signal for the k^{th} subband. The values of P_{nk} and P_{xk} are calculated as below

$$P_{nk}[n] = \gamma P_{nk}[n-1] + \frac{1-\gamma}{1-\beta}(P_{xk}[n] - \beta P_{xk}[n-1]) \tag{3}$$

$$P_{xk}[n] = \alpha P_{xk}[n-1] + (1-\alpha)\left(X_k|[n]|^2\right) \tag{4}$$

where $\alpha = 0.7\ \beta = 0.998\ \gamma = 0.45$. The speech pause detector algorithm is applied as shown below.

$$NX_{relk}[n] = \frac{NX_k[n] - NX_{mink}[n]}{NX_{maxk}[n] - NX_{mink}[n]} \tag{5}$$

where $NX_k[n] = \frac{P_{nk}[n]}{P_{xk}[n]}$. The values of NX_{mink} and NX_{maxk} are calculated looking back onto the previous 400 ms of speech signal. The value of P_{nk} is modified if NX_{relk} is less than a predetermined threshold. We then apply a non linear gain term, based on the value of α, the a priori SNR and the a posteriori SNR, to the noisy power spectrum to get the estimate of the clean power spectrum.

2.2. AMT Estimations and the ERB

The steps for calculating the AMT are :

1. Determine the auditory filter broadening in normal and impaired ears
2. Calculate the total energy in each auditory filter (ERB)
3. Compute the excitation pattern based on the auditory filter characteristics
4. Compare the excitation pattern with the absolute threshold of hearing

Masking occurs when the audibility of one sound is increased due to the presence of another sound. The auditory masked threshold is the dB level at which one sound is just audible in the presence of another sound. Masking in the peripheral auditory system has been explained by a power spectrum model [7] that assumes that (a) the auditory system consists of a series of overlapping bandpass filters, (b) that detection of a signal occurs through a filter centered at or near the signal frequency, and (c) that only the frequency components of the masker that pass through the filter centered at or near the signal frequency will contribute to the masking of that signal [10].

The series of bandpass filters has often been described in terms of the critical band or Bark scale [9]. More recently, the peripheral auditory filters have been described in terms of equivalent rectangular bandwidth (ERB). Whereas previous AMT-NS formulations have used the critical band/Bark scale to approximate the auditory filters, we use the ROEX model - ERB scale for two reasons: 1) recent experimental literature has shown that the ERB scale better approximates cochlear filtering (e.g. changes in shape as a function of level) and 2)

changes in auditory filtering caused by cochlear hearing loss can be described by the ROEX/ERB model [11]. The ERB and CB are quite similar in the mid frequency region (1-4 kHz) but the ERBs are more closely space in lower frequencies, which is indicative of higher resolution in this region. The ERB values, for a normal-hearing individual, over the whole frequency range are described by the following equation:

$$ERB_{bandwidth} = 24.7(4.37F + 1), \tag{6}$$

where $ERB_{bandwidth}$ is in Hz and F is the center frequency in kHz. For the hearing impaired individual the $ERB_{bandwidth}$ is equal to 24.7(4.37F+1)*B, where B ($B \geqslant 1$) is the broadening term and is described below. Total threshold for HI listeners is a combination of threshold loss due to outer and inner hair cell damage. The assumption made here is that the outer hair loss corresponds to 80% of the total hearing loss. The signal power for calculating the excitation pattern is recalculated to account for the headphone correction factors. The broadening of the auditory filters due to hearing loss can be described by [11],

$$B = 10^{0.01757(HL_{ohc}-22)*\left([1-(f_c-1)^2]/3.09\right)} \tag{7}$$

up to 1 kHz frequencies, and

$$B = 10^{0.01757(HL_{ohc}-22)} \tag{8}$$

for higher frequencies, where f_c is the center frequency in kHz. The constant 0.01757 is chosen so that the B has a value of 3.8 for a HL_{ohc} = 55dB. This corresponds to the maximum value of broadening. For normal individuals, the value of B is set to be equal to 1. Once the filter shapes are defined, the signal power in each auditory filter is calculated as X_{ERB}. Next, 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

$$W(g) = (1 + pg)exp(-pg) \tag{9}$$

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

$$g = \left(\frac{|f - f_c|}{f_c}\right) \tag{10}$$

The parameter p describes both the bandwidth and the slope of the skirts of the auditory filter. The lower frequency skirt, (p_l) of the auditory becomes less sharp with increasing level. p_l varies with Broadening and level, as

$$p_{l(x)} = p_{l(51)} - \left(0.35 - \frac{(B-1)}{3}\right)\left(\frac{p_{l(51)}}{p_{l(51,1k)}}\right)(X_{ERB} - 51) \tag{11}$$

where $p_{l(51)}$ is the center frequency for an equivalent noise level of 51 dB/ERB , $p_{l(51,1k)}$ is the value of p_l at 1 kHz for noise level of 51dB/ERB and X_{ERB} is 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

$$p_u = \frac{4 * fc}{24.7(4.37F + 1).} \tag{12}$$

It can be seen in Fig. 2, that the excitation pattern and filter shapes vary with the signal level. At lower levels the filter shapes for hearing-impaired individuals are broader. The excitation pattern is compared with the absolute threshold of hearing and the AMT is set as the greater of the two.

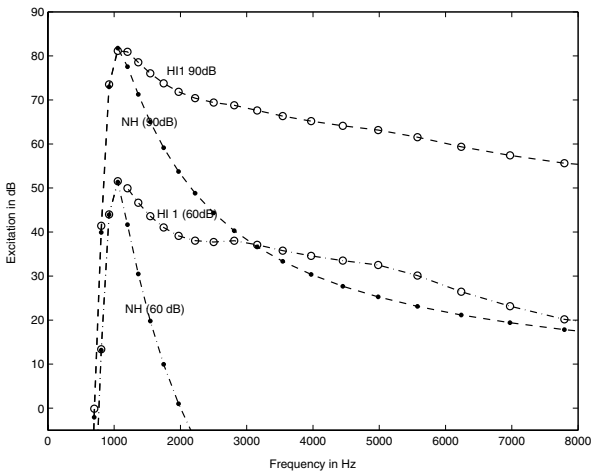

Figure 2: Excitation pattern for normal-hearing (NH) subject and hearing- impaired (HI) subject for a 1000 Hz tone at 60 & 90 dB SPL. HI has a high frequency hearing loss.

2.3. Audible Noise Suppression

The noisy power spectrum is compared with the AMT that was calculated in the previous section. The clean speech spectrum is calculated using a nonlinear gain function that is derived using a Wiener filtering operation [4]. The enhanced signal is renormalized and converted to the time domain.

3. Algorithm Evaluations

The 192 single sentence core test set in the TIMIT database with both male and female speakers was degraded with both colored and white additive noise sources. The noise level was set at 5dB SNR. A plot of clean, degraded and enhanced speech for one sentence, "In wage negotiations the industry bargains as a unit with the single union" is shown in Fig. 3. Fig. 4 describes the shapes of the audible masked threshold, noise power spectrum and clean power spectrum for voiced speech.

3.1. Objective Qualitative Measure

The following is the set of results comparing the segmental SNR and Itakura-Saito distortion measure between the GMMSE scheme with $\alpha \rightarrow 0$, $\alpha = 1$ and the modified GMMSE using the speech pause detector algorithm, where the value of alpha dynamically changed (GMMSE-Enrollment with masking disabled stage shown in Fig 1). These results are illustrated in Table 1. The speech was degraded with White Gaussian Noise with 5dB SNR. It can be seen that the modified GMMSE algorithm has an improved segmental SNR over the scheme with $\alpha = 1$ and SNR is nearly same for when $\alpha \rightarrow 0$. There is, however a considerable improvement over these schemes in the IS distortion measure.

Table 1: Objective quality measures for GMMSE schemes

	Degraded	$\alpha \rightarrow 0$	$\alpha = 1$	Mod GMMSE
SNR	-2.087	1.498	-0.276	1.337
IS	3.504	3.841	2.564	2.557

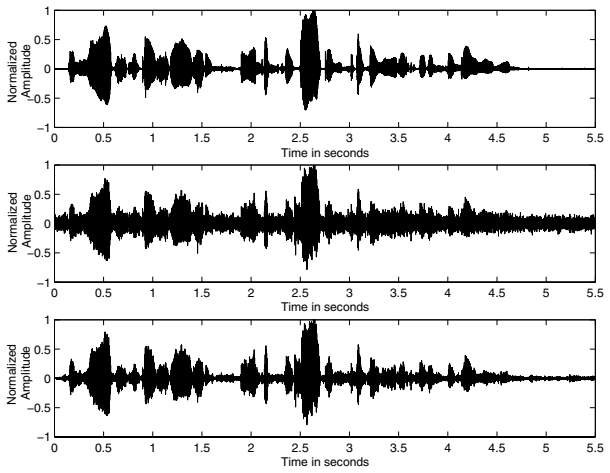

Figure 3: [Time waveforms for a single speech file] (a) Clean Spectrum (b) Degraded large crowd noise (LCR) at 5dB SNR (c) Enhanced speech waveform using GMMSE with ERB scheme.

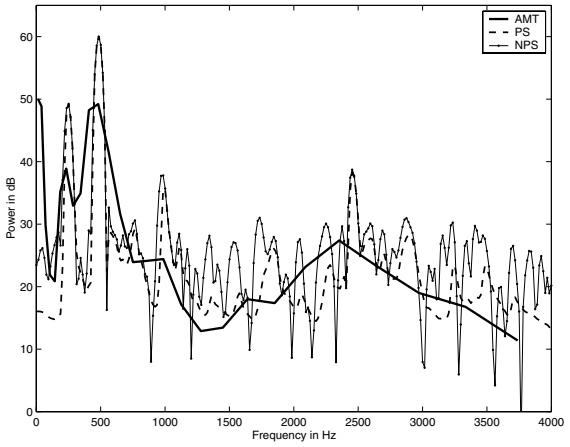

Figure 4: AMT, Noisy Spectrum (NPS) and Clean Spectrum (PS)of a voiced speech signal

A second set of experiments was performed with the modified GMMSE used in the AMT enrollment phase, and the audible noise suppression scheme (i.e., Wiener filtering with AMT masking engaged and tonality offset disabled) employed in the enhancement phase. The Itakura-Saito (IS) distortion measures and the Segmental SNR for the enhanced and degraded speech is shown in Table 2. The enhancement was done for normal-hearing individuals. It can be seen that there is a clear improvement in the segmental SNR, and a reduction in the the IS distortion measure in the enhanced speech, over speech degraded with four different noise sources. (AWG: White Gaussian Noise, FLN: Flat Channel Communication Noise, LCR: Large Crowded Room Noise, and HWY Highway Noise)

4. Conclusions & Future Work

The ERB scale provides us a framework to estimate the filter characteristics of hearing-impaired individuals which is not addressed in the a traditional critical band framework. The performance of the enhancement scheme is clearly illustrated from the plots and quality measures shown in the previous section.

Table 2: Objective quality measures for the new scheme

	SegSNR		IS	
NOISE (SNR)	DEG	ENH	DEG	ENH
AWG (5dB)	-2.087	0.703	3.50	2.00
FLN (5dB)	-2.09	0.87	3.35	1.90
LCR (5dB)	-1.85	0.59	2.38	1.63
HWY (5dB)	-1.547	0.48	0.81	0.69

Subjective listening tests are the most appropriate way to assess quality and intelligibility for HI listeners. We are currently assessing intelligibility and quality of the degraded and enhanced speech in a group of HI listeners with the algorithm customized for each individual listener. The methods employed for these listener evaluations are similar to those used in [3] and include the Nonsense Syllable Test (NST) and 10 point rating scales for quality. The formulation here has clearly demonstrated the ability of reformulating a previous proposed GMMSE-AMT enhancement scheme to one that is more appropriate for individuals with hearing loss.

5. References

[1] Radhakrishnan, V. and Hansen, J.H.L. "Speech enhancement based on generalized minimum mean square error estimation and masking property of the human auditory system." *IEEE Trans. Speech& Audio Proc.,* Submit: Dec 2002.

[2] Moore, B.C.J. and Glasberg B.R. "Derivation of auditory filter shapes from notched-noise data." *Hear Research.*, Aug 4 47(1-2):103–138, 1990.

[3] Arehart, K.H., Hansen, J.H.L., Gallant, S. and Kalstein, L. "Evaluation of an auditory masked threshold noise suppression algorithm in normal-hearing and hearing-impaired listeners." *Speech Commun.,* Vol 40 (4): 575–592 June 2003.

[4] Tsoukalas, D.E. Mourjoupoulos, J. and Kokkinakis, G. "Speech enhancement based on audible noise suppression." *IEEE Trans. Speech& Audio Proc.,* 5(6):497–514, 1997.

[5] Burget, L. and Moticek, P. "Noise estimation for efficient speech enhancement and robust speech recognition." *Proc., ICSLP 2002, Denver* vol 3:1033–1036, Sept-2002.

[6] Johnston, J.D. "Transform coding of audio signals using perceptual noise criteria." *IEEE J. Select Areas Commun.,* 6:314–323, 1988.

[7] Fletcher, H. "Auditory Patterns." *Review of Modern Physics,* 5(6) 47-65, 1940.

[8] Patterson, R.D., and Moore, B.C.J. "Auditory filters and excitation patterns are representations of frequency resolutions." In *Frequency Selectivity in Hearing,* pp 123-177. Academic Press, London.

[9] Scharf, B. ch 5 in "Foundations of Modern Auditory Theory" New York Academic, 1970.

[10] Moore, B.C.J. *Perceptual Consequences of Cochlear Damage.* Oxford Psychology Series 1995.

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

[12] Schroeder, M.R., Hall, J.H. and Atal, B.S. Optimizing digital speech coders by exploiting the masking properties of the human ear. *JASA*, 66(6):1647–1652, 1979.

[13] Bryne, D. and Dillon, H. The National Acoustic Laboratories (NAL) new procedure for selecting the gain and frequency response of a hearing aid. *Ear and Hearing 7*, (7): 257–265, 1986.

[14] Ephraim, Y. and Malah, D. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-32(6):1109–1121, 1984.