AUDITORY MASKING THRESHOLD ESTIMATION FOR BROADBAND NOISE SOURCES WITH APPLICATION TO SPEECH ENHANCEMENT*  

Ruhi Sarikaya and John H. L. Hansen  
Robust Speech Processing Laboratory  
Center for Spoken Language Understanding  
University of Colorado at Boulder, Boulder, CO 80309  
http://cslu.colorado.edu  ruhi@cslu.colorado.edu  jhlh@cslu.colorado.edu

ABSTRACT  
This paper addresses issues encountered in the use of an Auditory Masking Threshold (AMT) for speech enhancement and proposes an algorithm to improve AMT estimation for broadband noise sources. We determined that while AMT estimation is fairly accurate, and hence an enhancement scheme based on AMT can suppress audible noise to a greater extent for low frequency colored noise sources, the algorithm fails to converge to the clean speech AMT for broadband communication channel noise. We propose a new AMT estimation scheme and incorporate the proposed algorithm into a previously developed enhancement framework [2]. We evaluate our algorithm on a set of sentences obtained from the standard TIMIT database for flat communications channel noise (FLN), and automobile highway noise (HWY) at 5 dB and 0 dB SNR levels, respectively. Evaluations were performed for 8 kHz and 16 kHz sampled speech and performance is measured with both objective and subjective assessment methods. The results show that the new AMT codebook based enhancement method is more effective than traditional AMT methods. Also, that traditional AMT methods may not be as effective for reduced bandwidth speech (4 kHz), or broadband interference, but that alternative AMT estimation methods can help improve convergence properties.

1. INTRODUCTION  
In voice communications systems, speech signal are often degraded by such adverse environments as automobile interior for hands-free cellular, a communications channel or aircraft cockpit noise. The problem of sufficient noise suppression while keeping perceptual processing artifacts to a minimal level remains an open problem.

A number of speech enhancement techniques have been proposed in the past. These approaches are generally based on some type of spectral subtraction [3, 4], Wiener filtering [6, 7], or a minimum mean square error estimate of the speech [5]. Several recent methods have also considered auditory based approaches [12, 9, 2] where properties of a hearing model have been exploited for improved speech perception. The enhancement framework proposed in [2] is based on suppression of audible noise. In this scheme, a noise masking mechanism has been exploited through the use of AMT which has been widely used in perceptual audio coding [1]. Although it was shown that perceptual speech quality is improved for a low frequency noise source for speech sampled at 16 kHz, the viability of the algorithm has not been demonstrated on broadband noise sources at the more common 8 kHz sample rate.

In this study we address the following issues:

1. How does the accuracy of the AMT estimate influence the resulting enhanced speech?  
2. What is the effect of speech enhancement with AMT when the noise type changes from low-frequency noise to broadband and/or sampling rate changes from 16 kHz to 8 kHz?  
3. What is the functionality of AMT in regards to processing artifacts vs. noise suppression?

Additionally, we formulate an algorithm for accurate estimation of AMT and demonstrate both subjective and objective improvement over the baseline enhancement system [2] for different noise types. The proposed method is based on a dual vector quantizer (VQ) Codebook of AMTs, one for clean speech and one for noisy speech that are obtained from a training set. These codebooks are generic in the sense that they are independent of both speaker and the SNR level. The basic idea is to find the closest noisy speech AMT for the incoming noisy AMT and use the corresponding clean AMT which had been tied to the noisy AMT during training. A similar codebook based scheme was previously proposed in [11] where the codebook entries are noisy and clean speech spectra. However, in that system, errors made during the selection of a noisy speech spectrum directly affects the enhanced speech since the corresponding incorrect clean spectrum is selected as the enhanced speech spectrum. This does not occur often but the effect is noticeable. Here we take advantage of the fact that the enhanced speech is not so sensitive to small errors during the selection process of the AMT since it enters into the enhancement scheme within an attenuation gain term. The remaining part of the paper is organized as follows. In Sec. 2, we present the derivation of the AMT and the baseline algorithm used for enhancement. Next, we address the  

*This work was supported in part by a grant from the U.S. Government.
issues listed above in Sec. 3, followed by a new algorithm formulation method for AMT estimation. In Sec. 5, evaluation results are presented. Finally, we draw conclusions in the last section together with possible future directions.

2. AMT ESTIMATION & THE BASELINE SPEECH ENHANCEMENT SYSTEM

2.1 AMT Estimation

AMT defines a spectral amplitude threshold below which all frequency components are masked in the presence of the masker signal. A detailed derivation of AMT can be found in [1]. Here, we summarize the main derivation steps in the calculation of AMT:

1. Obtain energies in speech critical bands (CB).
2. Convolves the spectrum of the critical band spectrum to obtain a spread masking threshold.
3. Compute offset term for spread masking thresholds.
4. Normalize and account for the absolute auditory thresholds.

Step 1:

\[ B_{P_{X_i}}(i) = \sum_{k=k_b}^{k_{b+1}} P_X(k,i), \quad 1 \leq b \leq B' \] (1)

The energy \( B_{P_{X_i}}(i) \), in CB \( b \) for the frame \( i \) is obtained from the power spectrum, \( P_X(k,i) \), of the speech, where \( k_b \) and \( k_{b+1} \) are the lower and upper limits of CB \( b \) and \( B' \) is the total number of CBs and is dependent on the sampling frequency of the signal.

Step 2:

CB energies are convolved with the basilar membrane spreading function \( S_{Pr}(\cdot) \) to obtain the spread critical band spectra, \( C_{P_{X_i}}(i) \).

\[ C_{P_{X_i}}(i) = \sum_{j=1}^{n'} S_{Pr}(b-j+B') B_{P_{X_j}}(i) \] (2)

Step 3:

In order to determine the tonalike/noiselike nature of the spectrum, a spectral flatness measure (SF M) is used. In the first branch of Eq. 3, SF M is defined as the ratio of the geometric mean of the power spectrum \( G_{P_{X}(k,i)} \) to the arithmetic mean \( A_{P_{X}(k,i)} \) of the power spectrum. In the second branch, \( \text{SF M}_{P_{X}}(k,i) \) is used to generate a measure of tonality with \( \text{SF M}_{\text{max}} \) which is equal to \(-60 \text{ dB} \) for a sine wave. For white noise only, the SF M is equal to \( 0 \text{ dB} \) and hence \( \text{ton}_{P_{X}}(k,i) = 0 \). An offset, \( O_{P_{X}}(i) \) is then estimated by which the threshold is adjusted to take signal tonality into account.

\[ \text{SF M}_{P_{X}}(k,i) = \frac{G_{P_{X}(k,i)}}{A_{P_{X}(k,i)}} \]

\[ \text{ton}_{P_{X}}(k,i) = \min \left\{ \frac{10 \log_{10} \left( \frac{\text{SF M}_{P_{X}}(k,i)}{\text{SF M}_{\text{max}}} \right)}{1} \right\} \]

\[ O_{P_{X}}(i) = \text{ton}_{P_{X}}(k,i)(14.5 + b) + (1 - \text{ton}_{P_{X}}(k,i)) 5.5, \quad 1 \leq b \leq B' \] (3)

Now the auditory masking threshold can be given as follows for a speech frame:

\[ T_{P_{X_i}}(i) = 10^{10 \log_{10}(C_{P_{X_i}}(i) - O_{P_{X_i}}(i)) / 10} \] (4)

Step 4:

\[ T_{P_{X_i}}(i) = \max \left\{ T_{\text{num}}(b), \frac{\sum_{j=1}^{n'} S_{Pr}(b-j+B')} {\sum_{j=1}^{n'} S_{Pr}(b-j+B')} \right\} \] (5)

The final AMT is given above after normalization and incorporating the absolute hearing threshold, \( T_{\text{num}}(b) \), for each critical band \( b \).

2.2 Audible Noise Suppression Based Speech Enhancement

This enhancement framework is proposed in [2] and combines AMT from [1] with a parametric nonlinear gain term from [10] in the minimization of audible noise. Audible noise is defined as the difference between an audible spectrum of noisy speech and the audible spectrum of clean speech assuming the clean spectrum is known [2].

\[ A_d(k,i) = A_{P_{X}}(k,i) - A_{P_{X}}(k,i), \quad 0 \leq k \leq K - 1 \] (6)

Therefore the criterion for enhancement is,

\[ A_d(k,i) \leq 0, \quad 0 \leq k \leq K - 1 \] (7)

We refer the reader for a detailed derivation of the method to [2]. However, we briefly present the solution which satisfies this criterion.

\[ T_b(i) = \text{AMT}(P_{X}^{-1}(k,i)) \]

\[ a_b(i) = D_b + T_b(i) \left( \frac{D_b}{T_b(i)} \right)^{1/v_b(i)} \]

\[ P_{X}^{-1}(k,i) = \frac{\left( P_{X}^{-1}(k,i) \right)^{v_b(i)} + (P_{X}^{-1}(k,i))^{v_b(i)}}{v_b(i) + (P_{X}^{-1}(k,i))^{v_b(i)}} \] (8)

where \( D_b \) is the mean power spectrum of the noise in critical band \( b \), which is updated at each iteration of the algorithm. \( P_{X}^{-1}(k,i) \) is the estimate of the clean power speech spectrum at iteration \( j \). Here \( a_b(i) \) defines a threshold below which frequency components of the noisy speech are highly attenuated, whereas \( v_b(i) \) controls the rate of suppression.

3. ISSUES RELATED TO ENHANCEMENT & IMPACT OF AMT ESTIMATION ON PERFORMANCE

We believe that the use of AMT allows one to better balance the introduction of artifacts when noise levels are low. The AMT cannot directly suppress noise, but indirectly influences an integrated enhancement filter such as a spectral subtraction or Wiener-like noise suppression filter.

We experimentally observed that the performance of the speech enhancement algorithm greatly suffers when AMT is not properly estimated. The function of AMT is rather to shape artifacts as a result of the noise removal process in a perceptually inaudible manner. Although noise can be removed from speech in its original form, it reappears in an equally disturbing manner in the form of tonal artifacts when the AMT is not properly estimated.
The impact of AMT accuracy can be adequately appreciated when speech is bandlimited to 4 kHz rather than 8 kHz and corrupted with broadband noise sources instead of low-frequency noise sources. For 4 kHz bandwidth speech, the baseline enhancement algorithm does not perform well in estimating the clean speech spectrum and consequently the AMT estimate does not converge to the target (clean) speech AMT. However, for the case of a 16 kHz sampling rate and low-frequency noise sources, the vast majority of the CBs are reasonably clean which is very important for high frequency consonants, hence the estimated AMT is typically very close to the target AMT.

Our contribution in this study is to formulate an algorithm (next section) for estimation of AMT without relying on the poor estimate of the clean speech when all or most of the CBs are highly degraded by noise, and when the speech is band limited to 4 kHz which is typical for virtually all voice communications systems.

4. ALGORITHM DESCRIPTION

The accuracy in the estimation of AMT solely depends on the estimation of the clean speech spectrum. The degree to which speech is degraded is not only a function of the SNR level but also noise type. For equivalent SNR levels, a comparison of low frequency HWY and broadband communications channel noise (FLN) noises, show that FLN is more perceptually disturbing.

Our method is based on a dual VQ-Codebook of clean and noisy speech masking thresholds. The clean codebook entries are obtained from data collected using 3 speakers from the TIMIT database where each speaker contributed 4 sentences. We note that the data was not rich enough to cover all phones and possible acoustical variations for the AMT space. Also, the codebook was not optimized both in terms of phone coverage and size, since each phone has a different set of AMTs depending on the temporal context within the phrase where an AMT is estimated. However, even this limited data set was enough to show some improvement over the baseline method. Although the size of the clean AMT codebook is 2048 in our system, it can be increased to larger sizes at the expense of increased complexity.

Noisy codebook entries were obtained by degrading clean speech at 5dB SNR with FLN noise. There is a high degree of algorithm flexibility, since the noisy codebook of AMTs can be arranged in a sub-codebook tree structure depending on phone and noise type. In the evaluations, we investigate the system’s sensitivity to noise when a noisy codebook is used to enhance speech corrupted by another noise type at a different SNR level.

A block diagram describing the codebook based algorithm is given in Fig. 1. For frame $i$ of a noisy incoming speech AMT, $\hat{T}(i)$ is estimated. The distance between $\hat{T}(i)$ and the $m_{1b}$ entry of codebook $\mathbf{cb}_j$, is given as,

$$d(\hat{T}(i), T^{cb}_{mb}) = \| \hat{T}(i) - T^{cb}_{mb} \|$$  \hspace{1cm} (9)

The index of the entry which gives the minimum distance is selected and used to extract the corresponding clean speech AMT. There is a one-to-one mapping between all sub-codebooks and the clean speech codebook.

$$j = \text{arg min}_{m,n} (d(\hat{T}(i), T^{cb}_{mb})) \quad 1 \leq m \leq M, \quad 1 \leq n \leq N.$$  \hspace{1cm} (10)

The selected clean AMT entry, $j$, is used in the Eq. 8 enhancement iterations. Some of the advantages of this scheme are as follows. First, when it is difficult to estimate the correct clean speech spectrum, the baseline system estimates the AMT incorrectly resulting in enhanced speech that contained artifacts. Second, the proposed scheme can produce arbitrarily close clean speech when the codebook size is reasonably large. Computationally, this scheme is very efficient compared to the baseline system. The initial spectral subtraction step in the baseline enhancement method can be skipped in the new scheme since it serves as the initial estimate for clean speech spectrum and hence clean AMT estimate. Furthermore, the new scheme requires only 1 to 3 iterations to suppress noise and minimize artifacts whereas the baseline system requires 3 to 4 iteration to obtain the best enhanced speech.

5. EVALUATIONS

The two algorithms are evaluated using objective speech measures [13] according to the protocol described in [8] as well as informal listening tests. Additive HWY and FLN noises are added to 8 kHz sampled, randomly selected, 8 TIMIT sentences at 0 and 5 dB SNRs. For each sentence, the segmental SNR (seg-SNR) measure was calculated for both degraded and enhanced signals for baseline and the proposed algorithm. The global utterance quality is computed by taking the average of the frame based seg-SNR measures across speech only sections of each utterance.

For our evaluations, we selected FLN noise at 5 dB and HWY noise at 0 dB SNR since these settings have roughly the same level of perceptual distortions. During the enhancement process, the noise power spectrum is estimated from the first 10 frames of the speech signal. Note that a one time estimate is used since the noise is not assumed time varying.

The results of the simulations across broad speech phone classes are summarized in Tables 1 and 2. Here we show the seg-SNR measure for degraded speech, speech enhanced using both baseline and the proposed codebook based enhancement schemes. The number of iterations in the two enhancement procedures was determined experimentally based on best perceptual quality. Therefore objective quality scores are supported by informal listening tests.

Although there is only small seg-SNR improvement for FLN, there is audible perceptual improvement over the baseline algorithm. For HWY noise, there is significant improvement (e.g., average seg-SNR is improved from 0.74 to 1.09). It is interesting to point out that while the proposed algorithm showed improvement for consonants (e.g., stops, fricatives, etc.) for HWY noise, there was more improvement for vowels (e.g., semi-vowels, vowels, diphthongs) in FLN noise. As both algorithms are perceptually based,
the improvement can in fact be assessed more accurately through listening tests. Currently, we have been performing a formal listening test for both algorithms. The results of the formal listening test will be presented in the conference.

<table>
<thead>
<tr>
<th>Segmental SNR Measure</th>
<th>Sound Type</th>
<th>Degraded</th>
<th>Baseline</th>
<th>VQ-CH</th>
<th>% of frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>stops</td>
<td>-7.519</td>
<td>-1.600</td>
<td>-0.318</td>
<td>12.2</td>
<td></td>
</tr>
<tr>
<td>fricatives</td>
<td>-8.933</td>
<td>-1.654</td>
<td>-1.168</td>
<td>19.5</td>
<td></td>
</tr>
<tr>
<td>affricates</td>
<td>-6.567</td>
<td>-2.557</td>
<td>-1.904</td>
<td>2.8</td>
<td></td>
</tr>
<tr>
<td>nasals</td>
<td>-7.110</td>
<td>1.082</td>
<td>0.846</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td>semi-vowels</td>
<td>-0.430</td>
<td>3.244</td>
<td>3.028</td>
<td>9.5</td>
<td></td>
</tr>
<tr>
<td>vowels</td>
<td>-1.220</td>
<td>2.281</td>
<td>2.244</td>
<td>27.0</td>
<td></td>
</tr>
<tr>
<td>diphthongs</td>
<td>2.735</td>
<td>3.763</td>
<td>3.553</td>
<td>9.6</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>-3.777</td>
<td>0.746</td>
<td>1.000</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Objective Speech Quality vs Broad Phoneme Classification. TIMIT sentences were Degraded with Additive HWY noise (0 dB SNR).

<table>
<thead>
<tr>
<th>Segmental SNR Measure</th>
<th>Sound Type</th>
<th>Degraded</th>
<th>Baseline</th>
<th>VQ-CH</th>
<th>% of frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>stops</td>
<td>-6.342</td>
<td>-0.583</td>
<td>0.154</td>
<td>12.2</td>
<td></td>
</tr>
<tr>
<td>fricatives</td>
<td>-7.085</td>
<td>0.557</td>
<td>0.168</td>
<td>19.5</td>
<td></td>
</tr>
<tr>
<td>affricates</td>
<td>-4.794</td>
<td>0.080</td>
<td>-0.550</td>
<td>2.8</td>
<td></td>
</tr>
<tr>
<td>nasals</td>
<td>-4.648</td>
<td>2.494</td>
<td>2.483</td>
<td>4.7</td>
<td></td>
</tr>
<tr>
<td>semi-vowels</td>
<td>3.232</td>
<td>3.791</td>
<td>3.905</td>
<td>9.5</td>
<td></td>
</tr>
<tr>
<td>vowels</td>
<td>1.958</td>
<td>3.362</td>
<td>3.362</td>
<td>27.0</td>
<td></td>
</tr>
<tr>
<td>diphthongs</td>
<td>6.166</td>
<td>4.281</td>
<td>4.394</td>
<td>9.6</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>-3.063</td>
<td>2.007</td>
<td>2.146</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Objective Speech Quality vs Broad Phoneme Classification. TIMIT sentences were Degraded with Additive FLN noise (5 dB SNR).

6. CONCLUSIONS AND FUTURE WORK

The issues of Auditory Masking Threshold estimation with application to speech enhancement has been addressed. Furthermore a codebook based AMT estimation algorithm was formulated which showed a better approximation to the clean speech AMT than the baseline AMT estimation when speech is bandlimited to 4 kHz, and corrupted with broadband as well as low frequency noise sources. Both objective and subjective improvement was demonstrated for both noise types at 0 dB and 5 dB SNRs. Currently, we are building a phonetically balanced AMT codebook that represents speech better than the current codebook which was obtained from a limited TIMIT sentence set. We believe that this will further improve the performance of the proposed algorithm.

ACKNOWLEDGEMENT

The authors thank Dr. John Mourjopoulou for providing speech files to verify the baseline system.

1. REFERENCES