A Psychoacoustic Noise Reduction Method by Auditory Inhibition

Serajul Haque and Roberto Togneri
School of Electrical, Electronic and Computer Engg., University of Western Australia, Perth, Australia
Email: {serajul.haque, roberto.togneri}@uwa.edu.au

Sven Nordholm
Dept. Electrical and Computer Engg., Curtin University, Perth, Australia
Email: S.Nordholm@curtin.edu.au

Abstract—An auditory-motivated noise reduction methodology is proposed based on the inhibitory property of the human auditory system. The proposed psychoacoustic spectral subtraction method utilizes non-linearly and spatially distributed inhibitory weights which are determined based on the critical bands of the auditory system. Moreover, the weights are adjusted based on the segmental a posteriori SNR, such that higher inhibition is applied to the more noisy segments. Spurious noise perturbation in the vicinity of the dominant peaks are suppressed by the inhibitory spectral subtraction which improves speech degraded by additive and musical noise. The performance of the proposed psychoacoustic noise suppression method was evaluated by two perceptual test measures, PESQ and PEMO-Q using TIMIT utterances in clean and noise conditions, and shown to achieve better PESQ scores over the spectral subtraction method. It is further shown that by combining the proposed inhibitory weighting technique with the spectral subtraction method, higher performance benchmark in both PESQ and PEMO-Q may be obtained.

Index Terms—lateral inhibition, auditory system, psychoacoustics, speech processing, spectral subtraction

I. INTRODUCTION

In applications such as mobile phones, teleconferencing, automatic speech recognition, and hearing aids, speech may be corrupted by the environment noise during its transmission. In order to maintain the speech quality and perceptual intelligibility at the receiving end, it is required to reduce the corrupting noise in the speech without significantly affecting the original speech components. The most common method of single channel noise reduction is the spectral subtraction (SS) first proposed by Boll [1], in which the estimate of the noise magnitude spectrum is subtracted from the noisy speech magnitude spectrum. However, high variance of the noise magnitude spectrum estimates may result in under-subtraction, which produces sharp spurious random peaks in the magnitude spectrum of the processed speech, resulting in “musical noise”. On the other hand, strong over-subtraction may result in a reduction in the speech content, or may produce negative power spectral values. Several algorithms have been proposed to reduce these effects. For example, in the STSA-MMSE method [2] which is considered the state-of-the-art in speech enhancement, the a priori SNR is estimated from the noisy observations and the short term spectral amplitudes of the clean speech are estimated in a mean square error sense as Fourier series coefficients. Since it is a statistical method, the accuracy depends on the a priori noise estimation process.

An alternative method proposed for noise reduction is that based on the human auditory system. This is motivated by the fact that the human auditory system performs superbly even in severe noisy conditions. Several psychoacoustic methods for noise reduction have been proposed in the literature. These have been based mostly on the noise masking property of the auditory system and the computation of a masking threshold below which all noise components are masked. For example, in [3], a psychoacoustic spectral subtraction method is proposed in which the subtractive parameters are adjusted according to the noise masking threshold level to obtain optimum tradeoff between the amount of noise reduction, speech distortion and the level of musical noise. However, in this method the masking threshold calculation is difficult to perform on the noisy speech observations directly. This may lead to inaccuracies in estimating the threshold, particularly at higher frequencies and in presence of broadband noise. In this paper, we have proposed a new psychoacoustic method of spectral subtraction that has been based on the inhibitory property of the auditory system. The method does not require the computation of a masking threshold. The novelty of our approach is that the method utilizes a spatial weighting technique dependent on two parameters, (a) the critical bands (CB) of the auditory system, and (b) the segmental a posteriori SNR. Present studies suggest that the effects of lateral inhibition might be asymmetric at the cortical level and the inhibition is greater from low to high frequency compared to from high to low frequency [4]. We propose a method of spatial weighting based on the critical bands, which is compatible with the tonotopical arrangement of
the auditory nerves along the basilar membrane (BM) in the cochlea.

Moreover, the inhibitory weights can be parametrically adjusted based on the segmental a posteriori SNR of the noisy observations, depending on the severity and type of noise. This in effect applies higher inhibition where noise is more dominant than speech. The proposed psychoacoustic noise suppression method improves speech degraded by additive and residual noise by suppressing spurious peaks and noise perturbations in the neighborhood of dominant peaks. Since the method is based on a uniform inhibitory processing per CB, musical noise is reduced.

In the following sections, the paper is presented as follows. In Section II, we discuss the inhibitory characteristics of the auditory system and our proposed method of spectral subtraction using the inhibitory characteristics. In Section III, we present the algorithm of the method of spectral subtraction using the inhibitory characteristics of the auditory system and our proposed follow. In Section II, we discuss the inhibitory discussions and conclusions drawn from the experiments presented in Section IV followed by an overall posteriori.

II. LATERAL INHIBITORY EFFECT IN THE MAMMALIAN AUDITORY SYSTEM

In the human auditory system, an excited neuron may reduce the activity of its neighboring neurons due to the phenomenon of lateral inhibition [5]-[8]. This is the primary characteristics of the fusiform (pyramidal) cells in the dorsal cochlear nucleus. This effect may enhance the spatial profile and spectral contrasts in the same way the human vision increase contrasts to highlight edges.

We propose a noise reduction method based on this phenomenon as shown in Fig. 1. It consists of a critical band processing stage consisting of a set of band-pass filters tuned to the characteristic frequencies (CF) of the basilar membrane (BM), followed by an inhibitory model (LI) and a resynthesize module. For the inhibitory model, we have adopted the model in [5]. In this model, the lateral inhibitory effect may be implemented by a spatio-temporal filter having a low-pass effect with a time constant, τ = 1/(2πfτ) and a nonlinear transformation, G.

The LI block mimics a biological model to characterize the dynamical behavior of a neuron for a single channel of the CB when excited by an input stimulus. For an input E, the corresponding outputs, Y and X represent the membrane potential and the firing rate of the neuron, respectively, and G represents the nonlinear transformation between the membrane potential and the firing rate. The ensemble of auditory nerves may be considered in a spatial matrix (i,j) where the i-th neuron has a connectivity with the j-th neuron. E(i) for the i-th neuron is projected to the inhibitory layer Y(i) with a certain amount of divergence represented by the matrix v(i,j).

The input pattern G of the non-linear module is modified by the lateral inhibitory interactions by the weights, w(i,j) to obtain the output X(i). From Fig. 1, the dynamic behavior of the i-th neuron is given by

\[ \tau \frac{dY_i}{dt} + Y_i = \sum_j v(i, j) E_j \]  
\[ X_i = G(Y_i) - \sum_j w(i, j) G(Y_j) \]

where \( \tau \) represents the membrane time-constant, and \( w(i,j) \) is the coefficient of the inhibitory interactions between neuron \( i \) and neuron \( j \). For the following simplifying assumptions,

\[ G(Y_i) = Y_i \]
\[ v(i, j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \]
\[ \tau = 0 \]

And eliminating \( Y_i \) from (1) and (2),

\[ X_i = E_i - \sum_j w_{ij} E_j \]

Hence, the output may be represented as a subtraction of a weighted sum of neighboring processing cell responses from each input cell response.

For any value of \( \tau \) sufficiently small (e.g. 0.0001 seconds), the inhibitory LP filter has the characteristic such that at higher frequencies (>1/fc), the magnitudes are attenuated by the membrane capacitance, thus increasing the contrast of the dominant peaks. The spatio-temporal behavior of the neurons is implemented by the set of CBs considering that the CBs are arranged tonotopically along the BM of the cochlea, that is spatially along the frequency axis (y-axis), and temporal behavior is represented by spectral magnitudes along the time axis (x-axis).

Figure 1. The schematic of the noise reduction system with the inhibitory filter LI

Figure 2. The noise suppression and spectral enhancement due to spatially weighted spectral subtraction on the first two formants at 750 Hz and 1100 Hz of the synthesized phonemic vowel /æ/ in clean condition. The left figure shows the unenhanced spectral magnitudes and the right figure shows the enhancement by the proposed method using (6).
Fig. 2 demonstrates the inhibitory effects by spectral subtraction using (6) on the first two formants of the synthesized phonemic vowel /æ/ due to the inhibitory interaction as stated above. It is observed in the right figure that the spurious perturbations are inhibited, and at the same time the spectral magnitudes are enhanced.

III. SPATIO-TEMPORAL INHIBITORY WEIGHTS

Instead of using a single weight for all CBs, we further modify the weights spatially based on the critical bands [9] and the a posteriori SNR per frame. Additionally, an over-inhibition factor is introduced dependent on the segmental a posteriori SNR.

A. Spatial Weighting Based on the Critical Bands

It has been demonstrated in [4] that the effect of inhibition is greater at lower frequencies than at higher frequencies. This effect is similar and also consistent with the auditory simultaneous masking effect in which lower frequencies are more easily masked than higher frequencies relative to a masker, as observed from the masking threshold curve. Accordingly, the inhibition is adjusted such that higher inhibition is applied at higher critical band frequencies than at lower critical band frequencies, thus representing the weights \( w_{ij} \) as a function of the spatial-temporal dimensions of the cochlea in which the ensemble of neurons are arranged tonotopically along the BM.

To express \( w_{ij} \) as a function of the CB characteristic frequency, we represent it as \( w_{ij}(f) \). The frequency range from 10 Hz to 5300 Hz was divided into 23 critical bands, which are the important frequency segments in a band-limited application, e.g., in mobile telephony. It is observed from (6) that the higher the spatial weights, \( w_{ij}(f) \), the greater is the inhibitory effect on the input. Consequently, the weights are normalized to 1 at the lowest frequency and decreased linearly on the logarithmic frequency scale up to 5300 Hz such that \( 0\leq w_{ij}(f) \leq 1 \) within the above frequency range.

B. Inhibitory Weights Based on the Segmental SNR

For optimum performance under different noise conditions, a single set of weights, \( w_{ij}(f) \) is not applicable under all noise conditions. Therefore, a posteriori SNR is defined from the estimated noise and the noisy observation, the clean speech and the noise, \( s(n), d(n), v(n) \), is expressed as

\[ s(n) = d(n) + v(n) \tag{7} \]

where \( d(n) \) and \( v(n) \) each represent the time sequence of the sampled clean speech and the additive background noise. Taking the Fourier transform of (7),

\[ |S(\omega)|e^{j\alpha(\omega)} = |D(\omega)|e^{j\beta(\omega)} + |V(\omega)|e^{j\theta(\omega)} \tag{8} \]

where \( |S(\omega)|, |D(\omega)| \) and \( |V(\omega)| \) are the magnitude spectra of the noisy observation, the clean speech and the noise, and \( \alpha(\omega), \beta(\omega) \) and \( \theta(\omega) \) are the phase spectra of the noisy speech, clean speech and noise, respectively. For a discrete system, \( S_k = S(\omega)|_{\omega = 2\pi f_k} \), \( k \) being the spectral points at which the magnitudes are evaluated, and similarly for \( D_k \) and \( V_k \). Representing \( \lambda_d = E[|D_k|^2] \) and \( \lambda_v = E[|V_k|^2] \), where \( E(\cdot) \) is the expectation operator, as the variances of the \( k \)-th spectral components of the clean speech and noise, respectively, the a posteriori SNR, \( \gamma_k \) is defined as

\[ \gamma_k = \frac{|S_k|^2}{\lambda_v(k)} \tag{9} \]

which defines the SNR at each spectral point \( k \), and the segmental SNR per frame, \( \gamma_s \), was obtained as

\[ \gamma_s = \frac{1}{K} \sum_k \gamma_k \quad s = 1...M \tag{10} \]

where \( k = 1, \ldots, K \), \( K \) being the total number of spectral points, and \( M \) is the total number of frames. The noise variance, \( \lambda_v(k) \) was estimated by averaging the initial frames which are generally silent frames. For noise dominated frames, \( \gamma_s \) will have a lower value and for speech dominated frames, it will have a higher value. We utilize the a posteriori SNR, \( \gamma_k \) for determining the inhibitory weights, \( w_{ij}(f) \), normalized in each CB by using the nonlinear equation

\[ w(f_{CB}, \gamma_k) = \frac{\gamma_s - A(f_{CB})\xi}{\gamma_s} \tag{11} \]

where \( A \) and \( \xi \) are constants depending on the noise types (normally \( A = 1 \) and \( \xi = 1 \)). Representing \( f_{CB} \) by \( f \) for simplicity, and \( w_{ij} \) as \( w \) since a single weight is used per critical band, we modify (6) for our speech enhancement algorithm as

\[ X_{ik} = E_{ik} - \sum_k w_i(f, \gamma_s)E_{ik} \tag{12} \]

where \( i = 1, \ldots, C \), \( C \) is the number of critical bands, and \( k \) represents spectral points obtained from the DFT.

C. Over-inhibition Factor

It is observed that an over-inhibition term may improve the performance in noisy environments. Using (11), the inhibitory weights are consequently defined as

\[ w(f, \gamma_s) = \frac{\gamma_s - A(f)\xi}{\gamma_s} + \frac{g}{\gamma_s} \tag{13} \]

where \( g/\gamma_s \) is an over-inhibition term, and \( g \) is a scalable de-noising gain variable depending on the SNR. The over-inhibition term, \( g/\gamma_s \) increases the weight further if the segment is noisy corresponding to lower values of \( \gamma_s \) which will increase the inhibitory effect for the noisy segments. The value of \( g \) is set to a very low value, typically 0.006 for clean condition and is increased as the noise increases.
noise is reduced in the speech and non-speech segments, though a slight reduction in amplitude is observed in some weak segments, probably due to the scalable over-inhibition term in (13). The magnitude spectrum of noisy speech at 5 dB SNR white noise and the same speech enhanced by LI_SE is shown in Fig. 6.

IV. EVALUATION OF SPEECH QUALITY

Two perceptual objective tests were used to evaluate the performance of the proposed algorithm: Perceptual Evaluation of Speech Quality (ITU-T P.862.3 PESQ) [10], which is the standard method for the evaluation of speech quality in telephony systems, and the PEMO-Q [11], which is an objective assessment of audio quality using an auditory model. PEMO-Q defines a test-bench for the optimization of a noise reduction scheme, and is based on comparing the internal auditory representation between a
reference and a test speech by a perceptual similarity measure (psm). The proposed enhancement technique was compared with raw unenhanced speech, speech enhanced by spectral subtraction, (SS) as available in the Voicebox [12], speech enhanced by the proposed inhibitory weights (LI_SE), and enhancement by a combination of SS and LI_SE (SS_LI_SE). For noise estimation in the SS method, Martin’s method of minimum statistics was used with an over-subtraction factor of 3. Five TIMIT sentence

<table>
<thead>
<tr>
<th>Noise</th>
<th>SNR (dB)</th>
<th>Unenhanced</th>
<th>Specsub, SS</th>
<th>LI_SE</th>
<th>SS_LI_SE</th>
<th>Unenhanced</th>
<th>Specsub, SS</th>
<th>LI_SE</th>
<th>SS_LI_SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td></td>
<td>4.47</td>
<td>4.26</td>
<td>4.26</td>
<td>4.15</td>
<td>0.99</td>
<td>0.99</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>White</td>
<td>10</td>
<td>2.27</td>
<td>2.50</td>
<td>2.49</td>
<td>2.58</td>
<td>0.78</td>
<td>0.84</td>
<td>0.82</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1.88</td>
<td>2.14</td>
<td>2.14</td>
<td>2.20</td>
<td>0.74</td>
<td>0.80</td>
<td>0.78</td>
<td>0.80</td>
</tr>
<tr>
<td>Babble</td>
<td>10</td>
<td>1.55</td>
<td>1.79</td>
<td>1.81</td>
<td>1.81</td>
<td>0.69</td>
<td>0.76</td>
<td>0.72</td>
<td>0.76</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2.37</td>
<td>2.47</td>
<td>2.49</td>
<td>2.53</td>
<td>0.86</td>
<td>0.88</td>
<td>0.88</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1.99</td>
<td>2.09</td>
<td>2.11</td>
<td>2.15</td>
<td>0.81</td>
<td>0.84</td>
<td>0.83</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>1.44</td>
<td>1.68</td>
<td>1.68</td>
<td>1.70</td>
<td>0.75</td>
<td>0.78</td>
<td>0.76</td>
<td>0.78</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Two types of noise were added from the NOISEX 92: stationary white noise and real-world complex babble noise. For white noise, the inhibitory weights, \( w_i(f, y_s) \) and the \( y_s \) values are shown in figure 3. For babble noise the threshold for \( y_s \) used was 28 (compared to 160 for white noise), and 58 (compared to 260 for white noise).

Table I shows the PESQ and the PEMO-Q scores obtained from the experimental evaluations. It is observed that the LI_SE enhancement provides modest improvement in speech quality in noisy conditions over the baseline SS method as observed from the PESQ scores. However, PEMO-Q similarity scores for LI_SE degraded compared to SS. The higher PEMO-Q scores for the SS may be due to the increased tonal components attributed to the musical noise generated in SS. However, when the LI_SE is combined with the SS, that is, applying LI_SE on speech which has been preprocessed by the SS, the results are improved for both PESQ and PEMO-Q as observed in the SS_LI_SE scores.

V. DISCUSSION AND CONCLUSIONS

A psychoacoustic method of noise reduction by spectral subtraction is proposed using the inhibitory property of the auditory system. The method utilizes the critical bands to adjust the weights in the spatial dimension such that higher inhibition is applied in the higher frequency bands than the lower frequency bands. Moreover, the weights are adjusted depending on the \( a \) posteriori SNR such that higher inhibition is applied in the noise-dominated segments than the speech-dominated segments. The perceptually inspired speech enhancement method improves the quality of noisy speech by suppressing spurious peaks and noise in the neighbourhood of dominant peaks and performs low frequency spectral enhancement at the same time. The musical noise is also reduced because of the inhibitory effect, and a uniform processing strategy based on critical bands instead of a time-frequency cell processing as in the SS method. The proposed method has shown improved PESQ values over the spectral subtraction method. However, interactions between the CBs due to the inhibitory processing, and the effects of the previous frames are not considered in this work. The LI_SE computation time is slightly higher (~1.5 times) than the SS. It is shown that the method when combined with SS may give better performance over the SS method used alone. Informal listening trials verified that the perceived musical noise is also reduced compared to the SS. Inhibitory filter performance and parameter optimization are areas which may be emphasized as future research.

ACKNOWLEDGMENT

This research is partly funded by the Australian Research Council (ARC) grant no. DP1096348.

REFERENCES

Serajul Haque received the M.S. degree in electrical engineering from Louisiana Tech University and the Ph.D. degree in electrical engineering from the University of Western Australia. His doctoral research was in the area of perceptual features for speech recognition. He joined the University of Western Australia as a lecturer in July, 2008, and is currently a postdoctoral researcher at the same university. He is a member of the IEEE, and Australasian Speech Science and Technology Association (ASSTA). He is the author of a book “Auditory Features for Speech Recognition and Enhancement” published by VDM Verlag. While working for Beximco Computers Ltd, he developed a Bengali language processing BIOS for PC jointly with IBM Language Laboratory, Bangkok. His research interests include speech signal processing, auditory perception and cognitive process, machine learning, image processing and pattern recognition, biometric verifications, and visual speech processing.

Prof Roberto Togneri (M’89–SM’04) received the B.E. degree in 1985, and the Ph. D degree in 1989 both from the University of Western Australia. He joined the School of Electrical, Electronic and Computer Engineering at The University of Western Australia in 1988, where he is now currently a Professor. Prof Togneri is a member of the Signals and Systems Engineering Research Group and heads the Signal and Information Processing Lab. His research activities include signal processing and robust feature extraction of speech signals, statistical and neural network models for speech and speaker recognition, audio-visual recognition and biometrics, and related aspects of communications, information retrieval, and pattern recognition. He has published over 100 refereed journal and conference papers in the areas of signals and information systems, was co-author of the book "Fundamentals of Information Theory and Coding Design", Chapman & Hall/CRC, 2003 and is the chief investigator on two Australian Research Council Discovery Project research grants from 2010 to 2013.

Prof Sven Nordholm received his Ph.D. in Signal Processing in 1992, Lictenate of engineering 1989 and MscEE (Civilingenjor) 1983 all from Lund University, Sweden. He was one of the founding members of the Department of Signal Processing, Blekinge Institute of Technology, BTH, Sweden in 1990. At BTH he held positions as Lecturer, Senior Lecturer, Associate Professor and Professor. Since 1999 he has been at Curtin University of Technology in Perth, Western Australia. From 1999-2002 he was director of ATRI and Professor at Curtin University of Technology. From 2002-2009 he was director Signal Processing Laboratory, WATRI. Western Australian Telecommunication Research Institute, a joint institute between The University of Western Australia and Curtin University. Currently, he is Prof. Signal Processing and Head of Department, Dept. of Electrical and Computer Engineering, Curtin University. He is also Chief Scientist and co-founder of a start-up company Senseaer providing voice communication in extreme noise conditions. He is an associate editor Eurasip Journal on Advances in Signal Processing. He is a Senior Member of IEEE and a member of IEEE TC AASP.