Enhancement of Speech Intelligibility using Binary Mask Based on Noise Constraints

Ramesh Nuthakki, Sreenivasa Murthy A

Abstract: The primary aim of this paper is to examine the application of binary mask to improve intelligibility in most unfavorable conditions where hearing impaired/normal listeners find it difficult to understand what is being told. Most of the existing noise reduction algorithms are known to improve the speech quality but they hardly improve speech intelligibility. The paper proposed by Gibak Kim and Philipos C. Loizou uses the Weiner gain function for improving speech intelligibility. Here, in this paper we have proposed to apply the same approach in magnitude spectrum using the parametric wiener filter in order to study its effects on overall speech intelligibility. Subjective and objective tests were conducted to evaluate the performance of the enhanced speech for various types of noises. The results clearly indicate that there is an improvement in average segmental signal-to-noise ratio for the speech corrupted at -5dB, 0dB, 5dB and 10dB SNR values for random noise, babble noise, car noise and helicopter noise. This technique can be used in real time applications, such as mobile, hearing aids and speech–activated machines.

Keywords: speech intelligibility, noise estimation, Speech enhancement, objective and subjective performance measures, spectrograms.

I. INTRODUCTION

Speech communication is the important medium of oral communication. Speech often gets degraded in daily life because of the poor listening conditions such as background noise, electronic transmissions etc. Because of the presence of the background noise, the expected quality and intelligibility of the speech signal gets affected [1-5].

Recent studies have proved that large gains can be attained using the ideal binary mask technique [7] in speech intelligibility. The binary mask technique is proposed to retain the time frequency (T-F) units where the target speech subdues the masker signal and removes the T-F units where the masker signal subdues. The capability of the binary mask technique in improving speech intelligibility is indicated by using a Bayesian classifier. The removal or retaining of T-F bins using in binary mask is dependent on the noise spectrum overestimation or underestimation criterion. By taking the gain function into consideration a separate mask can be created by applying constraints on the two types of distortions. The size of the spectral amplitudes varies from the original spectral amplitudes with the application of the gain function. As a result, the attenuation and amplification distortions take place. It has been proved that the amplification distortion is more damaging compared to attenuation distortion. The obtained enhanced speech which contains attenuation distortion is proved to be more intelligible to that of noisy speech. For constructing a speech that contains only attenuation (or amplification) distortion, binary mask has to be applied to the enhanced speech spectrum. The contribution of each distortion initiated by the noise spectrum over estimation or under estimation is taken into consideration. The results made it clear that the new binary mask technique is capable of improving speech intelligibility for distinct background noises even at very low SNR levels [2]. This paper is organized as follows. Section II describes the binary mask scenarios based on noise constraints. Section III shows overall intelligibility and quality measures and section IV gives a conclusion.

II. BINARY MASK SCENARIOS BASED ON NOISE CONSTRAINTS

Numerous studies have proved that high gains in speech intelligibility can be attained using the binary mask technique. The binary mask takes the values of zero and one. It is one of the selection criteria on the basis of noise overestimation or noise underestimation constraints. When a binary mask is applied to the time-frequency representation of a mixture signal, it eliminates portions of a signal that are assigned to value ‘zero’ and preserves the values that are assigned to value ‘one’. Studies have shown that the speech synthesized from the binary mask is highly intelligible even in reverberant conditions [1].

A. Evaluation of noise and speech spectra

A binary mask was discussed in this module depending on the noise constraints. By regulating the distortions introduced by noise spectrum estimate, a new time frequency mask was constructed and is applied to the enhanced spectrum. The clean speech signal is considered as c(n) and zero mean noise process as d(n). Then the degraded speech signal y(n) can be written as follows:

\[ y(n) = c(n) + d(n) \]  

The steps involved in the construction of the binary mask are clearly shown in the Fig. 1. The noisy speech sentences were divided into 20-ms frames with 50% of overlap connecting the adjoining frames. Each single speech frame is Hann-windowed and a 320-point FFT was measured. Let \( Y(k, n_i) \) denote the noisy spectrum at time frame index \( n_i \) and frequency bin \( k \). Then, the estimate of the speech magnitude spectrum \( \tilde{C}(k, n_i) \) can be obtained by multiplying \( Y(k, n_i) \) with the parametric
Enhancement of speech intelligibility using Binary Mask based on noise constraints

The Parametric Weiner gain function $G(k, n_i)$ is demonstrated in terms of priori SNR, $\hat{C}(k, n_i)$ indicates the estimate of clean speech magnitude spectrum at frame index $n_i$ and frequency bin $k$. In this paper we use wiener algorithm as the gain function, since it is capable of reducing the estimation errors. Unlike the other noise reduction algorithms, wiener algorithm is found to be accurate in terms of quality and intelligibility [1].

$$G(k, n_i) = \sqrt{\frac{SNR_{pri1}(k, n_i)}{Y + SNR_{pri0}(k, n_i)}}$$

(3)

The Parametric Weiner gain function $G(k, n_i)$ is demonstrated in terms of priori SNR, $\hat{C}(k, n_i)$ indicates the estimate of clean speech magnitude spectrum at frame index $n_i$ and frequency bin $k$. In this paper we use wiener algorithm as the gain function, since it is capable of reducing the estimation errors. Unlike the other noise reduction algorithms, wiener algorithm is found to be accurate in terms of quality and intelligibility [1].

SNR$^{pri}_0$ can be calculated using the recursive equation as

$$SNR_{pri0}(k, n_i) = \alpha \cdot \frac{\hat{C}^2(k, n_i)}{\hat{F}^2(k, n_i)} + (1-\alpha) \cdot \max \left\{ \frac{\hat{C}^2(k, n_i)}{\hat{F}^2(k, n_i)} - 0, 0 \right\}$$

(4)

Where $\alpha$ is the smoothing constant whose value is 0.98, $\hat{F}^2(k, n_i)$ represents the estimates of the background noise variance. In order to estimate the noise variance, we use the noise estimation algorithm suggested in [6, 14]. The noise spectrum magnitude $\hat{F}(k, n_i)$ is estimated as follows

$$\hat{F}(k, n_i) = G_0(k, n_i) Y(k, n_i)$$

(5)

$$G_0(k, n_i) = \sqrt{\frac{1}{Y + SNR_{pri0}(k, n_i)}}$$

(6)

Where $G_0(k, n_i)$ indicates the noise equivalent Parametric wiener gain function [1-2,17].

B. Formation of the binary mask

The estimated noise spectrum $\hat{F}(k, n_i)$ is first calculated, then a binary mask was constructed by limiting the distortions. When $\hat{F}(k, n_i) > E(k, n_i)$, the noise over estimation distortion occurs and when $\hat{F}(k, n_i) < E(k, n_i)$ the noise under estimation distortion occurs. The processed speech comprises of both the distortions. To find out the effect of noise over estimation or under estimation distortion alone on speech intelligibility, constraints were applied on the estimated speech spectral magnitude. For each single T-F unit, the estimated noise spectrum $\hat{E}(k, n_i)$ is determined against the original noise magnitude spectrum $E(k, n_i)$. Only the T-F units that are fulfilling the constraints were retained and the rest were removed. Then the modified magnitude spectrum $\hat{C}_M(k, n_i)$ is calculated as follows

$$\hat{C}_M(k, n_i) = \begin{cases} \hat{C}(k, n_i) & \text{if } \hat{E}(k, n_i) > E(k, n_i) \\ 0 & \text{else} \end{cases}$$

(7)

After this, to the above selection of T-F units, an inverse IFFT was applied to the modified spectrum $\hat{C}_M(k, n_i)$ using the phase of the noisy speech spectrum. Finally the enhanced signal was computed in equation

$$\begin{align*}
\text{SNR} &= \frac{10}{M} \sum_{i=0}^{M-1} 10 \log_{10} \left( \frac{\sum_{n=0}^{N-1} c(n)^2}{\sum_{n=0}^{N-1} (c(n) - \overline{c}(n))^2} \right) \\
\text{INTELLIGIBILITY} &= \text{SNR} - \text{DISCERNIBILITY}
\end{align*}$$

(8)

Where $C$ and $\hat{C}$ denotes the clean and enhanced speech signal, $M$ denotes the number of frames and $i$ denotes the frame length. The average segmental SNR is evaluated by positioning the clean and the enhanced speech signals for various types of noises and for different input SNR values. This measure yields good results using parametric wiener filter for stationary as well as non-stationary types of noises.

B. Subjective Measures

For this test we have taken a set of 10 listeners, 5 male and 5 female and they were asked to listen to the enhanced and noisy speech signals randomly. These subjective tests were mostly based on parameters like background quality (BAK),

III. OVERALL INTELLIGIBILITY AND QUALITY MEASURES

A. Objective Measures

The objective measure used in this paper is average segmental SNR(segSNR). Although there are many objective measures we have chosen this because of the accuracy it provides when compared with other parameters. It is one of the extensively used objective measure. Higher the value of segSNR, the enhanced speech signal carries more signal power as against noise power. The objective measures for speech quality can be performed by dividing the speech signal into 20 ms frames. After that the distortion measure was calculated between the original and the processed speech signal. The speech distortion was estimated by equating the distortion measures of each frame in the time domain. In this paper we have considered the SSNR in the time domain as an objective measure. Clearly table-1 shows the improvement in SSNR values for random noise, babble noise, car noise and helicopter noise. During the computation of SSNR, the signal energy at the intervals of silence was relatively low leading to large negative SSNR values. The average segmental SNR[1,15-16] is written as:

$$SSNR = \frac{10}{M} \sum_{i=0}^{M-1} 10 \log_{10} \left( \frac{\sum_{n=0}^{N-1} c(n)^2}{\sum_{n=0}^{N-1} (c(n) - \overline{c}(n))^2} \right)$$

Where $C$ and $\hat{C}$ denotes the clean and enhanced speech signal, $M$ denotes the number of frames and $i$ denotes the frame length. The average segmental SNR is evaluated by positioning the clean and the enhanced speech signals for various types of noises and for different input SNR values. This measure yields good results using parametric wiener filter for stationary as well as non-stationary types of noises.
signal quality (SIG) and overall signal quality. They were asked to give scores from 1 to 5 for the above parameters. It has been found that, at -5dB, 0dB, 5dB, 10dB SNR levels the speech signals were corrupted by various kinds of noises such as babble noise, random noise, car noise and helicopter noise etc. The total scores given by the listeners are collected and shown in the form of tables 2 and 3. The listening tests showed a clear improvement in the speech quality [8-10, 15] for random noise, babble noise, helicopter noise and car noise.

C. Spectrograms

A spectrogram is a visual representation of sound. It displays the amplitude of the frequency components of the signal over time. It gives an account of speech signal’s relative energy concentration in frequency as a result of time and it displays the time-varying properties of the speech wave form. The red regions are associated to the energy signal. The voiced regions are indicated by the striped display and the form. The red regions are associated to the energy signal. The spectrograms it is c

Table- 3: Subjective measures using parametric wiener filter

<table>
<thead>
<tr>
<th>Noise</th>
<th>I/P SNR(dB)</th>
<th>BA K</th>
<th>SIG</th>
<th>OVL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Noise</td>
<td>10 (γ=5, β=0.2)</td>
<td>3.9</td>
<td>4.5</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>5 (γ=4, β=0.2)</td>
<td>3.9</td>
<td>4.3</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>0 (γ=1.5, β=0.5)</td>
<td>3.8</td>
<td>4.0</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>-5 (γ=4.2, β=0.4)</td>
<td>2.2</td>
<td>2.3</td>
<td>2.9</td>
</tr>
<tr>
<td>Babble Noise</td>
<td>10 (γ=0.7, β=0.3)</td>
<td>4.3</td>
<td>4.5</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>5 (γ=0.7, β=0.3)</td>
<td>3.9</td>
<td>4.3</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>0 (γ=3, β=0.2)</td>
<td>3.9</td>
<td>4.2</td>
<td>4.3</td>
</tr>
<tr>
<td></td>
<td>-5 (γ=0.4, β=0.9)</td>
<td>2.6</td>
<td>2.7</td>
<td>3.0</td>
</tr>
<tr>
<td>Car Noise</td>
<td>10 (γ=2.5, β=0.3)</td>
<td>4.3</td>
<td>4.6</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>5 (γ=2.5, β=0.3)</td>
<td>3.9</td>
<td>4.3</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>0 (γ=3, β=0.5)</td>
<td>3.8</td>
<td>4.2</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td>-5 (γ=5.5, β=0.4)</td>
<td>3.7</td>
<td>4.1</td>
<td>4.4</td>
</tr>
<tr>
<td>Helicopter Noise</td>
<td>10 (γ=2.5, β=0.3)</td>
<td>4.1</td>
<td>4.3</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>5 (γ=2.5, β=0.3)</td>
<td>3.9</td>
<td>4.1</td>
<td>4.6</td>
</tr>
<tr>
<td></td>
<td>0 (γ=3, β=0.4)</td>
<td>3.2</td>
<td>3.7</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td>-5 (γ=3.5, β=0.4)</td>
<td>2.3</td>
<td>2.9</td>
<td>4.0</td>
</tr>
</tbody>
</table>

D. Intelligibility hearing Tests

In order to assess the intelligibility [11-13] of the processed speech, listening tests were conducted. The sentences were taken from Indian English data base and IEEE. Noisy speech was generated by adding helicopter noise, car noise, random noise and babble noise at -5dB, 0dB, 5dB, 10dB SNRs. Before conducting the test, each listener was made to listen to a set of noise-corrupted sentences. After that they were asked to identify the words from the estimated speech signal. The performance was evaluated by counting the number of words identified correctly and is shown in Fig.5 and Fig.6.


E. Results and Discussion

From the subjective results shown in the table 2 and 3, it is evident that there is an improvement in intelligibility for the random noise, babble noise, helicopter noise and car noise at 0dB, -5dB, +5dB and 10 dB SNR values respectively. The results are obtained by considering the mean percentage of words identified correctly by the normal hearing listeners. When $\hat{E}(k, n_t) > E(k, n_t)$, the intelligibility has improved with parametric wiener filter. The unprocessed speech scores are indicated by UN. From the figures 5 and 6, it is clear that when the proposed noise constraints were applied, the performance at -5dB, 0dB,+5dB and 10 dB SNR levels improved from 21%,50%,65% and 80% with unprocessed speech to 88%,92%,94% and 97% respectively using wiener filter and reduced to nearly zero with noise underestimated constraints. Similarly in parametric wiener filter too, the performance at SNR levels -5dB, 0dB, +5dB and 10dB improved from 21%,50%,65% and 80% with the unprocessed speech to 96%,98%,99% and 100% respectively and reduced to closely zero with noise underestimated constraints. When we compare the two figures 5 and 6, with SNR levels -5dB, 0dB, +5dB, +10dB there is an improvement in word count from 88% to 96%, 92% to 98%, 94% to 99% and 97% to 100%. These results demonstrate that the proposed binary mask yields good in parametric wiener filter against wiener filter for the random noise, babble noise car noise and helicopter noise.

Table-1: Objective measures using parametric wiener and wiener Filter

<table>
<thead>
<tr>
<th>Noise</th>
<th>I/P SNR (dB)</th>
<th>$\gamma$</th>
<th>$\beta$</th>
<th>Seg.SNR(dB) using Parametric Wiener Filter</th>
<th>Seg.SNR(dB) using Wiener Filter ($\gamma=1$, $\beta=1$, Kim’s approach)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Noise</td>
<td>10</td>
<td>5</td>
<td>0.2</td>
<td>17.8190</td>
<td>14.5425</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>4</td>
<td>0.2</td>
<td>13.0516</td>
<td>9.8073</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1.5</td>
<td>0.5</td>
<td>12.9847</td>
<td>12.8214</td>
</tr>
<tr>
<td></td>
<td>-5</td>
<td>4.2</td>
<td>0.4</td>
<td>8.7793</td>
<td>8.6038</td>
</tr>
<tr>
<td>Babble Noise</td>
<td>10</td>
<td>0.7</td>
<td>0.3</td>
<td>14.8044</td>
<td>11.0596</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.7</td>
<td>0.3</td>
<td>11.0111</td>
<td>5.0853</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>3</td>
<td>0.2</td>
<td>5.2701</td>
<td>3.0390</td>
</tr>
<tr>
<td></td>
<td>-5</td>
<td>0.4</td>
<td>0.9</td>
<td>1.0528</td>
<td>0.4857</td>
</tr>
<tr>
<td>Helicopter Noise</td>
<td>10</td>
<td>2.5</td>
<td>0.3</td>
<td>22.7748</td>
<td>21.4692</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2.5</td>
<td>0.3</td>
<td>18.4329</td>
<td>15.6423</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>3</td>
<td>0.5</td>
<td>14.5684</td>
<td>12.6549</td>
</tr>
<tr>
<td></td>
<td>-5</td>
<td>5.5</td>
<td>0.4</td>
<td>10.4940</td>
<td>8.3977</td>
</tr>
<tr>
<td>Car Noise</td>
<td>10</td>
<td>2.5</td>
<td>0.3</td>
<td>16.9248</td>
<td>13.5615</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2.5</td>
<td>0.3</td>
<td>14.6185</td>
<td>12.5752</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>3.5</td>
<td>0.4</td>
<td>13.0179</td>
<td>10.1321</td>
</tr>
<tr>
<td></td>
<td>-5</td>
<td>3.5</td>
<td>0.4</td>
<td>11.1581</td>
<td>7.2035</td>
</tr>
</tbody>
</table>
Fig. 1. Steps used for constructing the Proposed binary mask depending on noise constraints

Fig. 2. Spectrograms showing the helicopter noise (SNR= 0 and -5 dB)
Enhancement of speech intelligibility using Binary Mask based on noise constraints

Fig. 3. Spectrograms showing the babble noise (SNR= 0 and 5 dB)

Fig. 4. Spectrograms showing the car noise (SNR= 0 and -5 dB)
Fig. 5. Mean intelligibility scores using helicopter noise

Fig. 6. Mean intelligibility scores using helicopter noise
Enhancement of speech intelligibility using Binary Mask based on noise constraints

IV. CONCLUSION

We have used the binary mask approach for parametric wiener gain filter using MATLAB. Subjective and objective tests were conducted for different values of γ and β for various background noises at 10dB, 5dB, 0dB and -5dB SNR values. The objective tests clearly indicate improvement in values of SSNR for random noise, babble noise, car noise and helicopter noise at 10dB, 5dB, 0dB and -5dB SNR values. The subjective results also shows an overall improvement in speech quality as well as intelligibility for random noise, babble noise, car noise and helicopter noise at 10dB, 5dB, 0dB and -5dB SNR values. The results shows a significant improvement in single channel speech intelligibility even at low SNR values (-5dB).

REFERENCES


AUTHORS PROFILE

Ramesh Nuthakki received the B.Tech degree in ECE and master’s degree in Digital systems and communication engineering from R.E.C (NIT) Calicut University, Calicut in 1999. From 1999-2005 he worked as a senior engineer in VSNL (TCL), Chennai and later joined the MNC Wipro Technologies and worked as a senior software engineer from 2005-2008 and he had also been to Canada and deployed as a project lead for Nortel Networks. After that he worked in the Esteemed Organization IBM as a Associated Project Manager from 2008-2011. In the year 2012 he joined as an Asst. Professor in Atira Institute of Technology, Bangalore and is continuing till date. His area of interest includes speech signal processing, and Networking. He is member of IEEE and IETE.

Dr. Sreenivasa Murthy is a research guide/supervisor in department of Electronics and Communication Engineering, University Visvesvaraya College of Engineering, Bangalore. Initially he worked in industry ( BEL, Bangalore ) for a period of 7 years. He is currently working as a Professor of ECE and mentoring M.Tech students and Ph D scholars. He has a teaching experience of more than three decades. He has completed Ph D in IISc, Bangalore. His area of research includes Digital Signal processing, Speech processing, Image processing, probability theory & stochastic processes and information theory & coding.