INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume 5, Issue 3)
Available online at: www.ijariit.com

Speech enhancement using binary mask technique

Chetan S. Gowda
chetansg69@gmail.com
Atria Institute of Technology, Bangalore, Karnataka

Anubhav Singh
anubhav1620@gmail.com
Atria Institute of Technology, Bangalore, Karnataka

Goutham Kumar V.
gouthamlegend1996@gmail.com
Atria Institute of Technology, Bangalore, Karnataka

Jayanth Kumar M. P.
jayanthjk23@gmail.com
Atria Institute of Technology, Bangalore, Karnataka

Ramesh Nuthakki
nuthakki.ramesh@atria.edu
Atria Institute of Technology, Bangalore, Karnataka

ABSTRACT

Many noise reduction algorithm improves overall speech quality but little progress has been made to improve speech intelligibility. The paper proposed by Gibak Kim and Philipos C. Loizou uses new binary mask approach to improve speech intelligibility, for our paper we have used the same approach but used parametric Wiener filter algorithm for the gain function computation. Subjective and objective measures tests were conducted to evaluate the overall speech enhancement quality and speech intelligibility. For the objective measure, the segmental signal-to-noise ratio parameter was calculated. The results indicate improvement in segmental signal-to-noise ratio for speech corrupted by noise at 0dB and -5 dB SNR levels for Helicopter noise and Car noise.

Keywords— *Speech intelligibility, Noise estimation*

1. INTRODUCTION

Speech Enhancement has been a subject of intensive research since 1970 to enhance noisy speech that is corrupted by additive noise, multiplicative noise or convolution noise, these noises can be stationary or non- stationary. The Enhancement improves the perceptual quality of the degraded speech signal using audio processing techniques. The enhancement may or may not include intelligibility. The intelligibility is a measure of how comprehensible speech is in given conditions. It is affected by the level and quality of speech signal, the type and level of background noise, reverberation. Although many advances are made in developing enhancement algorithm that suppresses background noise and improves overall speech quality, considerably less progress is made in developing an algorithm that improves speech intelligibility. Algorithms that improve intelligibility in a noisy environment is extremely useful in cell phone application and hearing aids devices.

One way of improving intelligibility is by constructing a binary mask in speech enhancement algorithm. The ideal binary mask has shown to improve speech intelligibility. The binary mask is designed as to retain the time-frequency (T-F) units where the speech is present (local SNR > 0dB) and discard the units where masker (noise) is present (local SNR < 0dB). This ideal binary mask is constructed using binary Bayesian classifiers. Alternatively, a new binary mask can be proposed by applying conditions on the two types of speech distortions that can be introduced by the gain function. These two distortions are attenuation distortion and amplification distortion. If the estimated spectral amplitude is less than the true spectral amplitude then it is called as attenuation distortion and, if the estimated spectral amplitude is more than the true spectral amplitude then it is called as amplification distortion. Studies have shown that amplification distortion (in excess of 6 dB) is more harmful to target speech intelligibility than attenuation distortion, to synthesize a speech encompassing only attenuation distortion, the new binary mask conditions to be imposed on the enhanced speech spectrum.

2. BINARY MASK CONDITIONS BASED ON NOISE CONSTRAINTS

In this section, the description of the proposed new binary mask is given, the new binary mask is based on noise constraint, and conditions are applied on the noise spectrum to derive the new time-frequency mask that is applied to the enhanced speech spectrum.

2.1 Noise and speech estimate

The construction of the new binary mask follows the steps shown in the figure 1 block diagram. Speech corrupted by noise was divided into 20ms frames with 50% overlap between the adjacent frames. Each frame goes through Hann- windowing which smoothens the transitions, then 320-point Fast Fourier Transform (FFT) is computed. The estimate of the speech can be derived by multiplying the observed noisy spectrum N (k, n) with a gain function as follows:

$$E^{'} (k, n) = \varphi (k, n). N (k, n) \qquad (1)$$

Φ (k, n) represents the Parametric Wiener gain Function,
E’ (k, n) represents the estimation of clean speech
k represents the frequency bin and n represents the frame index.

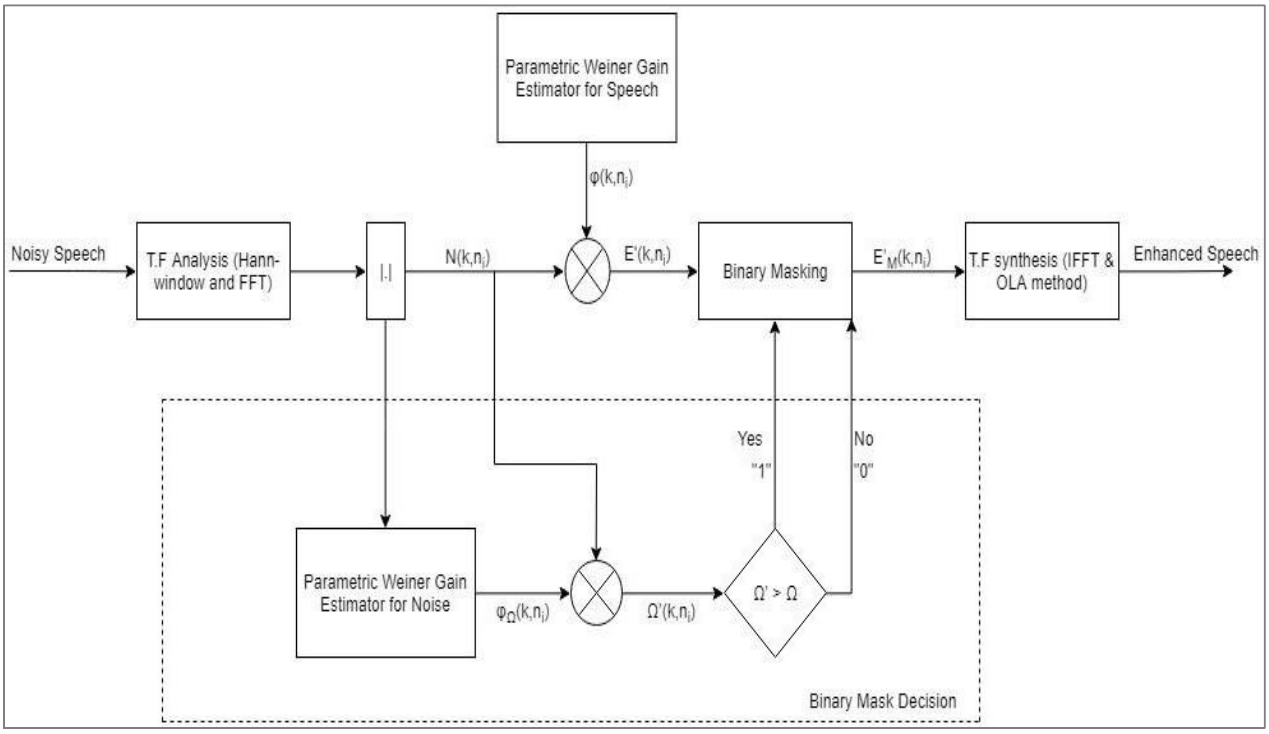
Parametric Weiner Gain Estimator for Speech
φ(k,n_i)
Noisy Speech
T.F Analysis (Hann-window and FFT)
|.|
N(k,n_i)
E'(k,n_i)
Binary Masking
E'_M(k,n_i)
T.F synthesis (IFFT & OLA method)
Enhanced Speech
Yes
"1"
No
"0"
Parametric Weiner Gain Estimator for Noise
φ_Ω(k,n_i)
Ω'(k,n_i)
Ω' > Ω
Binary Mask Decision

Fig. 1: Block diagram of the procedure used for constructing the proposed binary mask based on noise constraints.

Table 1: Subjective measure analysis

Noise	I/P SNR (dB)	BAK	SIG	OYL
Babble	0 dB	3.5	3.5	3.6
	-5 dB	2.5	2.8	2.8
Car	0 dB	3.9	4	4.3
	-5 dB	3.8	3.9	3.8
Helicopter	0 dB	4	4.1	4.3
	-5 dB	3.1	3.8	4
Random	0 dB	3.3	3.2	3.4
	-5 dB	2.7	2.8	2.9

Table 2: Objective measures analysis (Reference)

Noise	SNR(dB)	α	β	SSNR(dB)
Random Noise	0	1	1	12.8214
	-5	1	1	8.6038
Babble Noise	0	1	1	3.0390
	-5	1	1	0.4857
Helicopter Noise	0	1	1	12.6549
	-5	1	1	8.3977
Car Noise	0	1	1	6.2559
	-5	1	1	6.1319

Table 3: Objective measures analysis (Obtained)

Noise	SNR(dB)	α	β	SSNR(dB)
Random Noise	0	0.7	0.2	13.1928
	-5	1	0.6	9.7287
Babble Noise	0	0.3	0.9	3.2092
	-5	0.3	0.9	0.8927
Helicopter Noise	0	0.7	0.4	15.0257
	-5	1	0.3	12.2669
Car Noise	0	1	0.3	12.4440
	-5	1	0.4	10.0480

The parametric wiener filter was chosen because it is easy to implement, it is more consistent in-terms of speech intelligibility and speech enhancement compared to other sophisticated algorithms and requires less computation. The parametric wiener gain function is calculated based on the following equation:

$$\varphi(k,n) = \left(\frac{SNRprio(k,n)}{\alpha + SNRprio(k,n)}\right)^{\beta} \qquad (2)$$

Where $\mathrm{SNR}_{\mathrm{prio}}$ is the priori SNR estimated using the following recursive equation.

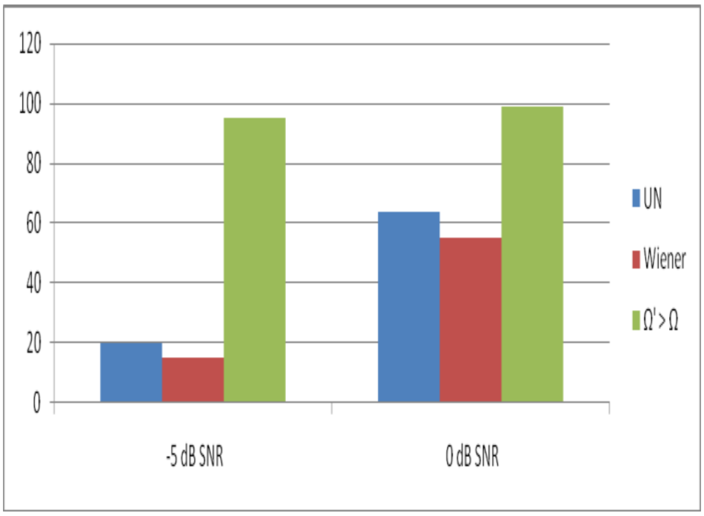

Fig. 2: Intelligibility scores in magnitude domain

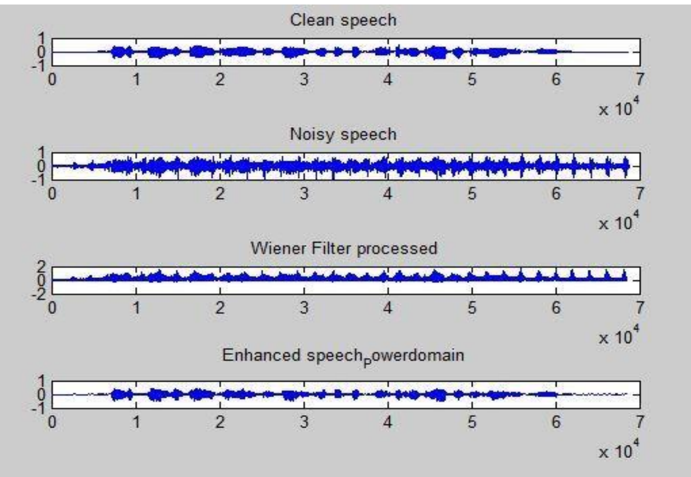

Fig. 3: Results of enhancement of speech degraded by helicopter noise

$$SNRprio(k,n) = \alpha\frac{E'^2(k,n-1)}{\gamma'_\Omega(k,n-1)} + (1-\alpha).max[\frac{N^2(k,n)}{\gamma'_\Omega(k,n-1)} - 1, 0] \quad (3)$$

Where α is the smoothing factor and its value equals to 0.98,

γ'Ω is the estimate of the background noise variance. The estimate of the noise spectral magnitude Ω' (k, n) can be given by

$$\Omega'(k, n) = \varphi\Omega(k, n) . N(k, n) \quad (4)$$

Where φΩ is noise-equivalent wiener gain function given as

$$\varphi_\Omega(k,n) = \frac{1}{1 + SNRprio(k,n)} \quad (5)$$

2.2 Binary mask construction

For the construction of the binary mask, out of the two distortion case for the estimate of noise magnitude spectra, the noise overestimation (Ω' (k, n) > Ω (k, n)) was chosen because it is less detrimental to speech intelligibility and speech enhancement compared to the noise underestimation (Ω' (k, n) < Ω (k, n)). The processed estimated speech will contain both distortion cases.

The estimate of the noise magnitude spectrum Ω' (k, n) was first compared against the true noise magnitude spectrum Ω (k, n) for each time-frequency (T-F) unit (k, n), and T-F units satisfying the constraint were retained, while T-F units violating the constraints were zeroed out. The modified estimated magnitude spectrum E'M (k, n) is computed as

$$E'_M(k,n) = \begin{cases} E'(k,n) & if \quad \Omega'(k,n) > \Omega(k,n) \\ 0 & else \end{cases} \quad (6)$$

Inverse FFT was applied to the modified estimated magnitude spectrum E'M (k, n). The overlap-and-add technique was finally used to synthesize the noise-suppressed signal containing noise-overestimation distortion only.

3. INTELLIGIBILITY AND OVERALL QUALITY MEASURES

3.1 Subjective Measures

Listening tests were conducted with a group of 5 listeners, 4 male and 1 female, who were asked to listen to the quality of enhanced speech signal with respect to the noisy speech signal. They were asked to give the ratings between 1 to 5. These subjective measures are mainly based on the parameters like Background Quality (BAK), Signal Quality (SIG) and Overall Signal Quality (OVL).

The speech signals were corrupted by noises like babble noise, random noise, car noise and helicopter noise at 0 dB and -5 dB SNR levels. The overall scores given by the listeners are gathered and are represented in the form of a table as shown. From table 1, it is clear that there is an improvement in the overall speech quality for car noise and helicopter noise.

3.2 Objective measures

An objective measure of overall speech quality was implemented by first segmenting the speech signal into 20ms frames, and then computing a distortion measure between the original and processed signal. The speech distortion was computed by averaging the distortion measures of every speech frame. The distortion measure computation was done in the time domain, for the project we have chosen Segmental Signal-To-Noise Ratio (SSNR) in the time domain as the objective measure. The objective measure table 2 and table 3 suggests improvement in Segmental signal-to-noise ratio values.

3.3.1 Segmental Signal-To-Noise Ratio Measures: SSNR in the time domain is the simplest objective measure used to evaluate speech enhancement algorithm. The original and processed signals were aligned in time and all the phase errors were corrected. Since SNRseg measure is based on the geometric mean of the signal-to-noise ratios across all frames of the speech signal. One problem we faced with the estimation of SNRseg was that the signal energy during intervals of silence in the speech signal (which was abundant in conversational speech) was very small resulting in large negative SNRseg values, which biased the overall measure. Therefore we used an alternative equation which eliminates the large SNRseg due to silent frames, which is given as

$$SNRseg = \frac{10}{M}\sum_{m=0}^{M-1} log_{10}(1 + \frac{\sum_{n=Nm}^{Nm+N-1} e^2(n)}{\sum_{n=Nm}^{Nm+N-1}(e(n) - e'(n))^2}) \quad (7)$$

e (n) is the clean signal e'(n) is the enhanced signal
N is the frame length (20ms)
M is the number of frames in the signal

3.3 Spectral analysis

The spectrogram is a graphical display of the power spectrum of speech as a function of time. The spectrogram describes the speech signal's relative energy concentration in frequency as a function of time and, as such, it reflects the time-varying properties of the speech waveform. The red regions are correlated to the energy signal. The voiced regions are specified by striped display because of the periodicity of the time wave form whereas the unvoiced are fully covered in. Colours indicate the magnitude of the spectrogram i.e red colour indicates the high energy and blue colour indicates the low energy. The spectrogram of car noise and helicopter noise is shown in fig.3 and fig.4 and it is clear and evident from the spectrogram that there is an improvement in speech intelligibility and overall speech quality for random noise and helicopter noise.

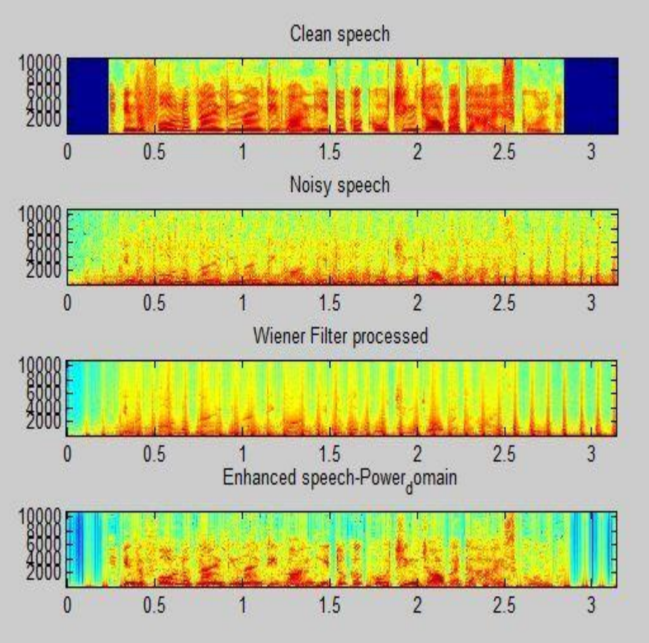

Fig. 3: Spectrogram showing magnitude spectral domain for Helicoptor Noise (SNR= -5dB)

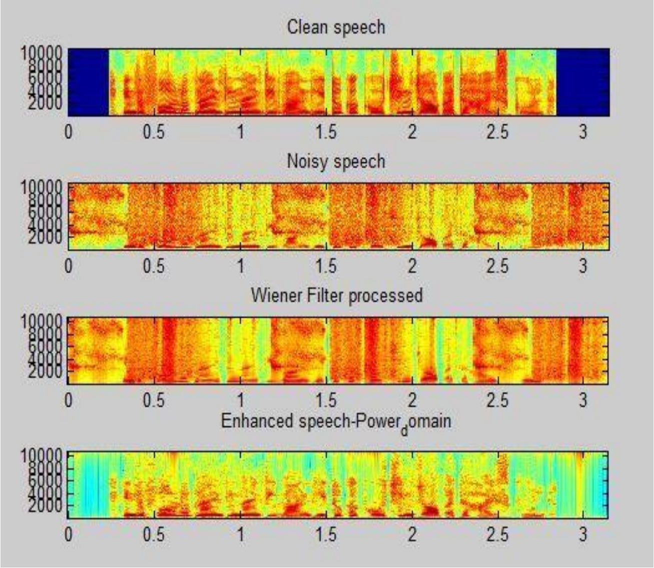

Fig. 4: Spectrogram showing magnitude spectral domain for Car Noise (SNR= -5dB)

4. RESULTS AND DISCUSSION

It is clear from the subjective measures table that there is an improvement in enhanced speech quality at 0dB and -5dB for car noise and helicopter noise. The improvement in SSNR value for helicopter noise and car noise in shown in the SSNR table. The figure shows the results, expressed in terms of the mean percentage of words identified correctly, by normal-hearing listeners. The bars indicated as "UN" show the scores obtained with noise-corrupted (un-processed) stimuli. As shown in figure 4, performance improved dramatically when the proposed binary mask (Ω'> Ω) was applied. Performance at -5, and 0 dB SNRs improved from near, 20% and 64% with un-processed stimuli to 95% and 99% correct respectively when the proposed mask was applied. In brief, we can conclude that the proposed binary mask constraints improve the intelligibility at 0dB and - 5dB SNR levels.

5. CONCLUSION

The new binary mask approach proposed by Gibak Kim and Philipos C. Loizou [2] was implemented but for parametric wiener gain filter using MATLAB. Subjective and objective tests were conducted, for objective tests, the parameter calculated was segmental signal-to-noise ratio in the time domain. The tests were run for different combinations of α and β of parametric wiener gain filter for different background noises at 0 dB and -5dB SNR levels. The objective results clearly indicate improvement in values of segmental signal-to-noise ratio for speech corrupted by Helicopter noise and Car noise at 0 dB and -5dB SNR levels. The subjective results also show improvement in overall speech enhancement quality and speech intelligibility for speech corrupted by Helicopter noise and Car noise at 0 dB and -5dB SNR levels.

6. REFERENCES

[1] P. C. Loizou, Speech Enhancement: Theory and Practice 2. Boca Raton: FL: CRC Press, 2014.

[2] A new binary mask based on noise constraints for improved speech intelligibility Gibak Kim and Philipos C. Loizou, 2010.

[3] A noise-estimation algorithm for highly non- stationary environments Sundarrajan Rangachari, Philipos C. Loizou, 2006.

[4] Single channel speech enhancement using a new binary mask in the power spectral domain, Ramesh Nuthakki, A Sreenivasa Murthy, Naik D C, 2017.

[5] G. Kimet al., "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Am., vol. 126, no. 3, pp.1486–1494, Sep. 2009

[6] An algorithm that improves speech intelligibility in noise for normal-hearing listeners Gibak Kim, Yang Lu, Yi Hu, and Philipos C. Loizoua Department of Electrical Engineering, the University of Texas at Dallas, Richardson, Texas 7508

[7] T. Lee and F. Theunissen, "A single microphone noise reduction algorithm based on the detection and reconstruction of spectro temporal features," Proc. R. Soc. A vol. 471, no. 2184, Dec. 2015.

[8] Liu, P. Smaragdis, and M. Kim, "Experiments on Deep Learning for Speech De noising," in INTERSPEECH, 2014.

[9] Kumar and D. Florencio, "Speech Enhancement In Multiple-Noise Conditions using Deep Neural Networks,"arXiv: 1605.02427 [cs], May 2016, arXiv: 1605.02427.

[10] Cohen, "Noise spectrum estimation in adverse environments: im-proved minima controlled recursive averaging," IEEE/ACM Trans. Audio, Speech, Language Process., vol. 11, no. 5, pp. 466– 475, Sep. 2003.