International Journal of Research in Engineering, Science and Management
Volume-3, Issue-1, January-2020
www.ijresm.com | ISSN (Online): 2581-5792

Speech Enhancement Using Ideal Binary Mask Based on Channel Selection Criteria

Akshaya Hari[1], Ayesha Afnan[2], Junaid Abbas[3], Faisal Ahmed Shariff[4], Ramesh Nuthakki[5]
[1,2,3,4]*Student, Department of Electronics and Communication Engineering, Atria Institute of Technology, Bangalore, India*
[5]*Assistant Professor, Department of Electronics and Communication Engineering, Atria Institute of Technology, Bangalore, India*

***Abstract*: Over the last 40 years' researchers/engineers have proposed quite a many speech enhancement algorithms to reduce noise, but little efforts have been made to improve speech intelligibility. The primary aim of this project is to examine the application of binary mask to improve speech quality and intelligibility in most unfavorable conditions where hearing impaired/normal listeners find it difficult to understand what is being told. While noise tracking or voice activity detection algorithms have been found to perform well in steady background noise, they do not perform well in non-stationary types of noise like multi-talker babble. Most algorithms make use of soft gain function for suppressing noise. Since the gain function is soft, it will have limitations in as far as its ability to improve intelligibility. Only algorithms that can improve the overall Signal-to-Noise ratio computed across all bands can improve speech intelligibility. One strategy for improving the overall SNR is to discard bands with extremely low SNRs while retaining bands with favorable SNRs.**

***Keywords*: Binary mask, Noise estimation, Speech intelligibility**

1. Introduction

Speech, being a primary form of communication, plays an important role in human life. Human speech communication typically degrades due to various surrounding environmental conditions. The most common factor that causes the degradation of speech quality and intelligibility is the background noise, which can be stationary or non-stationary and is assumed uncorrelated and additive to the speech signal. Speech enhancement aims to improve speech quality by using various algorithms. The objective of enhancement is improvement in intelligibility and/or overall perceptual quality of degraded speech signal using audio signal processing techniques. The intelligibility is a measure of how 'comprehensible' or 'understandable' speech is in given conditions. 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.

Recent studies with normal-hearing listeners have reported large gains in speech intelligibility using the ideal binary mask technique. Ideal binary mask identifies speech dominated and noise dominated units and is computed and applied to the noisy input spectrum to get the noise-suppressed spectrum. This binary mask was designed to retain the time-frequency (T-F) regions where the target speech dominates the masker (noise) (e.g., local SNR >0 dB) and remove T-F units where the masker dominates (e.g., local SNR < 0 dB). 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 a binary mask is dependent on the noise spectrum overestimation or underestimation criterion. This is important since many existing noise-estimation algorithms under-estimate the noise power-spectrum density (psd). A different mask can be constructed instead by applying constraints on the two types of speech distortion that can be initiated by the gain function. A variation in spectral amplitudes is obtained with the application of the gain function which helps create a separate mask by applying constraints on the two types of distortions. As a result, the attenuation and amplification distortions take place. Studies 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. Hence, a binary mask is applied to the enhanced speech spectrum to construct a speech signal comprising of only the attenuation distortion.

The proposed binary mask technique can improve substantially speech intelligibility even for sentences corrupted by background noise at SNR levels as low as -10 dB SNR. This method of speech enhancement cancels the interference of noise and other speech from targeted speech using binary Ideal Mask and improves SNR of speech. It helps to increase the intelligibility of the speech and to make the corrupted speech more pleasant to the listener. Hence, improving the efficiency of speech communication by reducing listener fatigue and increasing listening comfort.

2. Literature Survey

A different mask is constructed by applying constraints on the two types of speech distortion that can be initiated by the

gain function. In order to construct speech containing only amplification distortion, a new binary mask was proposed for improving speech intelligibility based on noise distortion constraints in power spectral domain. Results showed that the proposed binary mask produced enhanced speech quality in power domain compared to the magnitude domain [1]. The method proposed by Boll, noise is estimated from the initial few frames of silence region and is subtracted from noisy speech spectral magnitude. Concentration was more on the methods spectral subtraction and by considering the over subtraction factor (α) and spectral floor factor (β). Results for negative SNR's and for helicopter noise gives good results by using α and β [2]. An algorithm that estimates the binary mask for noise-corrupted speech based on the instantaneous SNR estimation was proposed. White noise was used for the training of the SNR estimator and the binary mask estimation was performed for babble, factory, speech-shaped noise. The experimental results showed that the proposed algorithm estimates the binary mask better than the binary mask estimation using the a priori SNR estimator. Assessment of the

Table 1
Literature Survey

Author (Year)	Title	Technique	Advantage/Scope
[1] Ramesh Nuthakki A, Sreenivasa Murthy and Naik D C(2018)	Single channel speech enhancement using a new binary mask in power spectral domain (2018)	New binary mask was proposed for improving speech intelligibility based on noise distortion constraints in power spectral domain.	The proposed binary mask produced enhanced speech quality in power spectral domain compared to the magnitude domain.
[2] Ramesh Nuthakki A, Sreenivasa Murthy and Naik D C (2017)	Modified Magnitude Spectral Subtraction Methods for Speech Enhancement	Noise estimated from the initial few frames of silence region is subtracted by considering the over subtraction factor (α) and spectral floor factor (β) using spectral subtraction method.	In perception test, for negative SNR's and helicopter noise, it gives good results by using α and β.
[3] Gibak Kim(2015)	Binary mask estimation for noise reduction based on instantaneous SNR estimation using Bayes risk minimization	SNR estimation is done by minimizing the Bayes risk. White noise was used for training SNR estimator and binary mask estimation was done for babble, factory noise.	Substantial improvements compared to the method using DD SNR estimation.
[4] Fei Chen, Philipos C. Loizou (2011)	Impact of SNR and gain-function over- and under-estimation on speech intelligibility	Upper bound on the gain function was derived to constrain the values of gain function to ensure SNR over-estimation errors were minimized.	Enhancement algorithms that can limit the values of gain function to fall within this upper bound will improve speech intelligibility.
[5] Gibak Kim and Philipos C. Loizou(2011)	Gain-induced speech distortions and the absence of intelligibility benefit with existing noise-reduction algorithms	Two interrelated reasons for absence of intelligibility improvement. The correct determination of the background noise spectrum acts as a parameter to improving the intelligibility	The gain induced distortions are confined to be of attenuation type in order for the noise suppression algorithms to work on the improvement of intelligibility
[6] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen(2011)	An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech	STOI parameters is introduced which shows high correlation with intelligibility of noise and T-F weighted noisy speech.	STOI is based on shorter time segments (386ms) resulting in high correlation with intelligibility.
[7] Gibak Kim and Philipos C. Loizou(2010)	Why do speech-enhancement algorithms not improve speech intelligibility?	Two interrelated reasons for absence of intelligibility improvement.	If amplification distortions are regularized, then the speech intelligibility can be improvised.
[8] Gibak Kim and Philipos C. Loizou(2010)	Improving Speech Intelligibility in Noise Using a Binary Mask That Is Based on Magnitude Spectrum Constraints	Binary mask was introduced for improving speech intelligibility based on magnitude spectrum constraints.	PESQ revealed better speech quality than that obtained with speech synthesized using the SNR based mask.
[9] Gibak Kim and Philipos C. Loizou(2010)	A new binary mask based on noise constraints for improved speech Intelligibility	New binary mask based on noise distortion constraints was used which retains noise overestimated T-F units and removes noise underestimated T-F units.	Large gains were obtained even for the speech which is corrupted by the noise at extremely low SNR levels (-10 dB).
[10] Jianfen Ma, Yi Hu and Philipos C. Loizou.(2009)	Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions	To calculate the performance of new speech based STI measures and AI based measures operating on short (30ms) intervals in realistic noisy conditions.	The measure was found to predict modestly (γ=0.68-0.83) with the intelligibility of speech embedded in fluctuating maskers when the proposed BIF's were used.
[11] Ning Li and Philipos C. Loizou.(2008)	Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction	The effects of the local SNR threshold input SNR level, masker type and errors introduced in estimating the ideal binary mask.	Performance was affected the most when the masker dominates T-F units, were wrongly labeled as target dominated T-F units.
[12] Sundarrajan Rangachari, Philipos C. Loizou(2006)	A noise-estimation algorithm for highly non-stationary environments	Noise estimate is updated by averaging the noisy speech power spectrum using time and frequency dependent smoothing factors.	The segmental SNR values and the LLR values were found to be more consistent.
[13] Yi Hu, Philipos C. Loizou ,Elsevier B.V. (2006)	Subjective comparison and evaluation of speech enhancement algorithms	The speech enhancement algorithms were chosen to encompass four different classes of noise reduction methods: spectral subtractive, subspace, statistical-model based and Wiener-type algorithms	The Wiener performed well in some conditions. The subspace algorithms performed poorly.

effect of SNR estimation errors on gain function estimation via sensitivity analysis was done. Intelligibility listening studies were conducted to validate the sensitivity analysis [3], [4].

Focus was on 2 types of distortions being, Amplification distortion and Attenuation distortion. There was an incremented benefit in the intelligibility in the steady noise conditions in the attenuation distortions. Short-time objective intelligibility measure (STOI) shows high correlation with the intelligibility of noisy and time–frequency weighted noisy speech. STOI is based on shorter time segments (386 ms) [5], [6]. Analysis of potential factors that influence the intelligibility of processed speech was done. They divide the speech distortions into: Region 1: Only attenuation distortion occurs. Region 2: Amplification distortions up to 6.02 dB. Region 3: Amplification distortions of 6.02 dB or greater. Mixture envelopes are classified into 2 classes, those where SNRESI > 0 (regions 1 & 2), which are retained, and those with SNRESI < 0 (region 3) which is eliminated. We can conclude by saying that if amplification distortions are regularized, then the speech intelligibility can be improvised [5], [7].

A binary mask was introduced which was designed to retain time-frequency (T-F) units of the mixture signal satisfying a magnitude constraint, while discarding T-F units violating the constraint. This mask that does not rely on the SNR criterion was constructed by imposing the constraints on the two types of gain-induced distortion: amplification distortion and attenuation distortion [1,], [5], [8]. A new binary mask based on noise distortion constraints was propounded. The mask retains noise overestimated Time-Frequency units and removes noise underestimated T-F units. It retains the time-frequency (T-F) regions where the target speech dominates the masker (noise) (local SNR >0 dB) and remove T-F units where the masker dominates (local SNR < 0 dB) [1], [8], [9]. To evaluate the performance of new speech-based STI measures, modified coherence-based measures, and AI-based measures operating on short-term (30 ms) intervals in realistic noisy conditions. The measure was found to predict modestly (r=0.68–0.83) well the intelligibility of speech embedded in fluctuating maskers when the proposed BIFs were used [10].

The factors influencing intelligibility of ideal binary-masked speech are examined. Specifically, the effects of the local SNR threshold, input SNR level, masker type, and errors introduced in estimating the ideal mask. Consistent with previous studies, intelligibility of binary-masked stimuli is quite high even at −10 dB SNR for all maskers tested [1], [8], [9], [11]. A noise-estimation algorithm was proposed for highly non-stationary noise environments. The noise estimate is updated by averaging the noisy speech power spectrum using time and frequency dependent smoothing factors. For the triplet noise case, the proposed method had higher preference scores compared to all the other methods [12].

Report on the subjective comparison and evaluation of 13 speech enhancement algorithms using the ITU-T P.835 methodology is done. The speech enhancement algorithms were chosen to encompass four different classes of noise reduction methods: spectral subtractive, subspace, statistical-model based and Wiener-type algorithms. These algorithms were evaluated using a newly developed noisy speech corpus (NOIZEUS) suitable for evaluation of speech enhancement algorithms. In terms of overall quality and speech distortion, the following algorithms performed the best: MMSESPU, logMMSE, logMMSE-ne, pMMSE and MB [13].

3. Block diagram

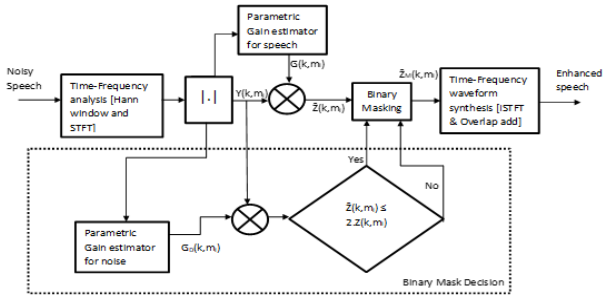

Fig. 1. Procedure for the binary mask depending on the signal residual selection criterion

We consider clean speech signal z(n) disturbed by a zero mean noise process d(n), (uncorrelated with z(n)), then the resulting degraded speech signal y(n):

$$y(n) = z(n) + d(n) \quad (1)$$

We consider clean speech signal z(n) disturbed by a zero mean noise process d(n), (uncorrelated with z(n)), then the resulting degraded speech signal y(n):
y(n) = z(n) + d(n) ----(1)

The above Fig. 1 clearly shows the steps used in the construction of the proposed binary mask in magnitude spectral domain. Noisy speech sentences are divided into 20-ms frames, along 50% overlap connecting the adjoining frames. Each individual speech frame is Hann-Windowed and Short Time Fourier transform is computed. The estimate of the speech spectra is obtained by multiplying the spectrum of the observed noisy spectrum, denoted by $Y(k, m_i)$ with $G(k, m_i)$ which represents the gain function expressed in terms of a priori SNR, and $\hat{Z}(k, m_i)$ denotes the estimate of the clean speech spectrum at frame index mi, and frequency bin k. After computing the estimated noise spectrum, the binary mask is formulated by limiting the distortions initiated by the errors in estimating the noise spectrum. Especially if $\hat{Z}(k, m_i) \leq 2Z(k, m_i)$, binary mask lets the spectrum pass through and masks the spectrum if vice versa. Usually, processed speech will contain both noise underestimation and overestimation. The estimate of the noise spectrum is first differentiated against the real noise magnitude spectrum for each time-frequency (T-F) unit, the T-F units satisfying the constraints are retained and those not satisfying are removed. The enhanced speech in time domain is obtained

by applying ISTFT.

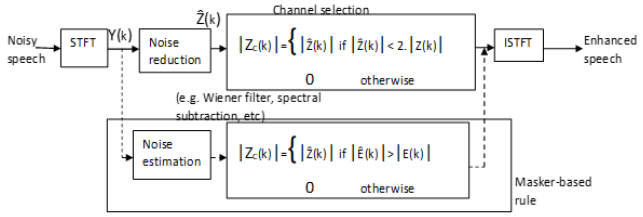

Fig. 2. Block diagram of two different channel selection algorithms

Above, Figure 2 shows the block diagram of the processing involved with the aforementioned SNRESI based algorithm. Unlike the SNR rule, the SNRESI rule selects channels from the enhanced (noise-suppressed) spectrum rather than from the noise corrupted spectrum. The noise-reduction block shown may include any conventional noise reduction algorithm. The choice of algorithm will not influence performance, at least in terms of intelligibility.

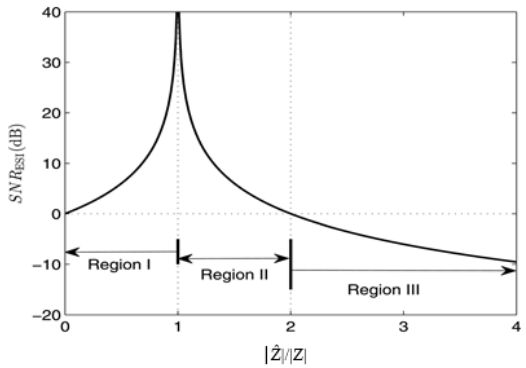

Fig. 3. Plot showing the relationship between SNR_{ESI} and the ratio of enhanced ($|\hat{Z}|$) to clean ($|Z|$) spectra

The above Fig. 3, focused on assessing the impact of these gain induced distortions on speech intelligibility in competing talker and steady noise conditions. Using a conventional noise reduction algorithm (square root weiner), they were confined into three regions.

Region I: Containing only attenuation distortion.

Region II: Containing only amplification distortion smaller than 6dB.

Region III: Containing amplification distortion greater than 6dB.

In order to get clarity about the impact of these two distortions, a Parametric Weiner filter is to be proposed in which the gain function varies for certain parameters δ and Ω. By changing these parameters, we get different types of Weiner filters with distinct attenuation characteristics. When δ=1 and Ω=1, we get the square root Weiner filter. When δ=1 and Ω=2, we get the Weiner filter.

$$G(k, m_i) = \sqrt{\left(\frac{SNRprio\ (k,mi)}{\delta + SNRprio\ (k,mi)}\right)}^{\Omega}$$

4. Conclusion

A new binary mask is to be implemented in magnitude and power spectral domain using MATLAB. Speech quality and intelligibility should be improved and to be measured as per ITU-T standards. The objective parameters should be evaluated in terms of STOI (Short Time Objective Intelligibility) and SSNR (segmental signal to noise ratio) for speech corrupted by stationary / non-stationary noise such as Helicopter noise, Car noise, random noise, Multitalker babble at various input SNR levels. The subjective results should show improvement in overall MOS (Mean Opinion Score).

5. Future Scope

In this survey the following considerations are made: Improvement in speech intelligibility and overall speech enhancement quality for signal degraded by noise as low as -10dB SNR levels. Improvement in values for other objective measure parameters as well as increased comfort in hearing aids.

References

[1] "Single channel speech enhancement using a new binary mask in power spectral domain" by Ramesh Nuthakki A, Sreenivasa Murthy and Naik D C. Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018).

[2] "Modified Magnitude Spectral Subtraction Methods for Speech Enhancement" by Ramesh Nuthakki A, Sreenivasa Murthy and Naik D C. 2017 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 2017.

[3] "Binary mask estimation for noise reduction based on instantaneous SNR estimation using Bayes risk minimisation" by Gibak Kim, Electronics Letters 19th March 2015 Vol. 51 No. 6 pp. 526–528.

[4] "Impact of SNR and gain-function over- and under-estimation on speech intelligibility" by Fei Chen, Philipos C. Loizou, Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75083-0688, USA, accepted 8 September 2011.

[5] "Gain-induced speech distortions and the absence of intelligibility benefit with existing noise-reduction algorithms" by Gibak Kim and Philipos C. Loizou Department of Electrical Engineering, University of Texas at Dallas, Richardson, Texas. 75080 VC 2011 Acoustical Society of America, pp. 1581–1596.

[6] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," IEEE Transactions on audio, speech, and language processing, vol. 19, no. 7, September 2011.

[7] Gibak Kim and Philipos C. Loizou, "Why do speech-enhancement algorithms not improve speech intelligibility?" ICASSP 2010.

[8] "Improving Speech Intelligibility in Noise Using a Binary Mask That Is Based on Magnitude Spectrum Constraints" by Gibak Kim and Philipos C. Loizou IEEE Signal processing letters, vol. 17, no. 12, December 2010.

[9] Gibak Kim and Philipos C. Loizou, "A new binary mask based on noise constraints for improved speech Intelligibility" INTERSPEECH 2010.

[10] Jianfen Ma, Yi Hu and Philipos C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," Acoustical Society of America, 2009.

[11] Ning Li and Philipos C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," Acoustical Society of America, 2008.

[12] Sundarrajan Rangachari, Philipos C. Loizou, "A noise-estimation algorithm for highly non-stationary environments" by Speech Communication, vol. 48, 2005.

[13] Yi Hu, and Philipos C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," 2006.