A new binary mask based on noise constraints for improved speech intelligibility

Gibak Kim and Philipos C. Loizou

Department of Electrical Engineering, University of Texas at Dallas, USA

imkgb27@gmail.com, loizou@utdallas.edu

Abstract

It has been shown that large gains in speech intelligibility can be obtained by using the binary mask approach which retains the time-frequency (T-F) units of the mixture signal that are stronger than the interfering noise (masker) (i.e., SNR>0 dB), and removes the T-F units where the interfering noise dominates. In this paper, we introduce a new binary mask for improving speech intelligibility based on noise distortion constraints. A binary mask is designed to retain noise overestimated T-F units while discarding noise underestimated T-F units. Listening tests were conducted to evaluate the new binary mask in terms of intelligibility. Results from the listening tests indicated that large gains in intelligibility can be achieved by the application of the proposed binary mask to noise-corrupted speech even at extremely low SNR levels (-10 dB).

Index Terms: speech intelligibility, noise estimation, speech enhancement

1. Introduction

Though large advances have been made in the development of enhancement algorithms that can suppress background noise and improve speech quality, considerably smaller progress has been made in designing algorithms that can improve speech intelligibility [1]. Recent studies with normal-hearing listeners have reported large gains in speech intelligibility using the ideal binary mask technique [2, 3]. The 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). In our previous work [4], we demonstrated the potential of the binary mask technique to improve speech intelligibility when the mask was estimated using a binary Bayesian classifier.

The ideal binary mask is based on the SNR criterion for retaining or discarding T-F units. A different mask can alternatively be constructed by imposing constraints on the two types of speech distortion that can be introduced by the gain (suppression) function [5]. When the gain function is applied to the noisy spectrum, the resulting (magnitude) spectral amplitudes could be smaller than the true spectral amplitudes, hence, attenuation distortion is introduced, or could be larger, hence amplification distortion is introduced. The listening studies in [5] showed that of the two distortions, the amplification distortion (in excess of 6 dB) was most detrimental to speech intelligibility. Enhanced speech containing only attenuation distortion was found to be substantially more intelligible than the noisy speech. To construct speech containing only attenuation (or amplification) distortion a binary mask had to be applied to the enhanced spectrum.

In this paper, we propose a new binary mask for improving speech intelligibility by extending the idea examined in [5] to noise spectrum estimation. By examining the effects of noise-spectrum over- or under-estimation, one can construct a new binary mask. The experiments reported in this paper examine the individual contributions of the distortions introduced by noise spectrum overestimation or underestimation to speech intelligibility. This is important since many existing noise-estimation algorithms under-estimate the noise power-spectrum density (psd). The minimum-statistics algorithm [6], for instance, is based on tracking the minimum of short-term psd estimate in individual sub-bands, and as such the minimum noise psd estimate is a biased estimate of the mean psd. To evaluate the proposed noise-based binary mask, intelligibility listening tests are conducted with normal-hearing listeners. The results of the listening tests indicated that 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.

2. Binary mask criterion based on noise constraints

In this section, we describe the new binary mask based on noise constraints. The new time-frequency mask is constructed by imposing constraints on the noise spectrum estimate, and is applied to the enhanced spectrum.

2.1. Estimation of speech and noise magnitude spectra

Fig. 1 shows the block diagram of the steps involved in the construction of the proposed binary mask. Noisy sentences were first segmented into 20-ms frames, with 50% overlap between adjacent frames. Each speech frame was Hann-windowed and a 500-point (corresponding to 20-ms for the sampling rate of 25 kHz) discrete Fourier transform (DFT) is computed. The estimate of the speech magnitude spectrum is obtained by multiplying the magnitude of the observed noisy spectrum, denoted as $Y(k,t)$, with a gain function as follows:

$$\hat{X}(k,t) = G(k,t) \cdot Y(k,t) \tag{1}$$

where $G(k,t)$ denotes the gain function, and $\hat{X}(k,t)$ denotes the estimate of the clean speech (magnitude) spectrum at time frame t and frequency bin k. In this paper, we use a conventional Wiener algorithm as a gain function [7]. The Wiener algorithm was chosen as it is easy to implement, requires little computation and has been shown by [8, 9] to be equally effective, in terms of speech quality and intelligibility, as other more sophisticated noise-reduction algorithms. The (square-root) Wiener gain function is calculated based on the following

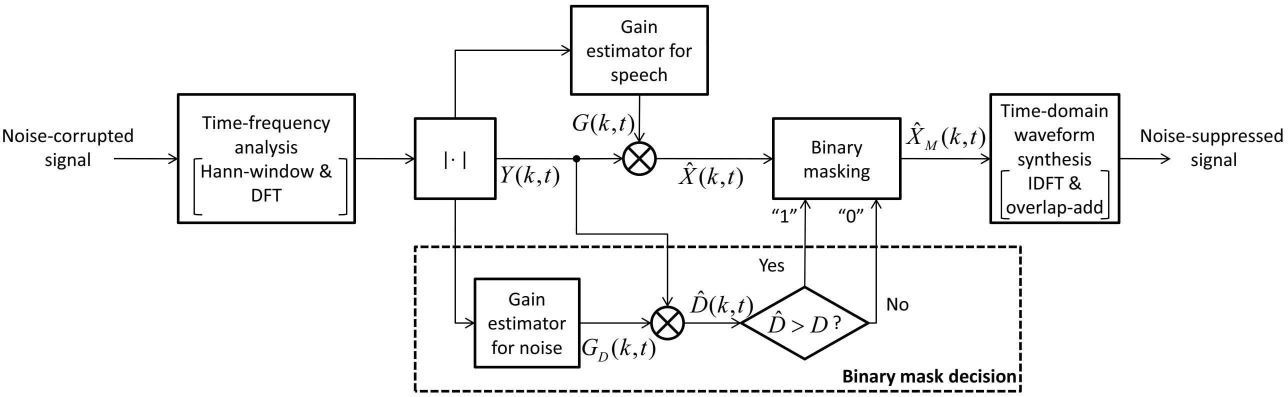

Figure 1: *Block diagram of the procedure used for constructing the proposed binary mask based on noise constraints.*

equation:

$$G(k,t) = \sqrt{\frac{SNR_{prio}(k,t)}{1 + SNR_{prio}(k,t)}} \tag{2}$$

where SNR_{prio} is the *a priori* SNR estimated using the following recursive equation [10]:

$$SNR_{prio}(k,t) = \alpha \cdot \frac{\hat{X}^2(k,t-1)}{\hat{\lambda_D}(k,t-1)} + (1-\alpha) \cdot \max\left[\frac{Y^2(k,t)}{\hat{\lambda_D}(k,t)} - 1, 0\right] \tag{3}$$

where $\alpha = 0.98$ is a smoothing constant and $\hat{\lambda_D}$ is the estimate of the background noise variance. The noise estimation algorithm proposed in [11] was used for estimating the noise variance $\hat{\lambda}_D$ in Eq. 3. Similar to Eq. 1, the estimate of the noise spectral magnitude $\hat{D}(k,t)$ is obtained as follows:

$$\hat{D}(k,t) = G_D(k,t) \cdot Y(k,t) \tag{4}$$

where G_D is the noise-equivalent Wiener gain function computed as [12]:

$$G_D(k,t) = \sqrt{\frac{1}{1 + SNR_{prio}(k,t)}}. \tag{5}$$

2.2. Construction of binary mask

Following the computation of the estimated noise magnitude spectrum, $\hat{D}(k,t)$, the binary mask is constructed by limiting (and controlling) the distortions introduced by the errors in estimating the noise magnitude spectrum. In particular, if $\hat{D}(k,t) > D(k,t)$ it is denoted as noise overestimation distortion, and if $\hat{D}(k,t) < D(k,t)$ it is denoted as noise underestimation distortion. In general, processed speech will contain both. In order to assess the effect of noise overestimation distortion alone or underestimation distortion alone on speech intelligibility, we imposed noise overestimation/underestimation constraints on the estimated speech spectral magnitude.

More precisely, the estimate of the noise magnitude spectrum $\hat{D}(k,t)$ was first compared against the true noise magnitude spectrum $D(k,t)$ for each time-frequency (T-F) unit (k,t), and T-F units satisfying the constraint were retained, while T-F units violating the constraints were zeroed out. For the implementation of the noise overestimation constraint, for instance, the modified magnitude spectrum $\hat{X}_M(k,t)$, was computed as follows:

$$\hat{X}_M(k,t) = \begin{cases} \hat{X}(k,t) & \text{if } \hat{D}(k,t) > D(k,t) \\ 0 & \text{else} \end{cases} . \tag{6}$$

Following the above selection of T-F units, an inverse DFT was applied to the modified spectrum $\hat{X}_M(k,t)$ using the phase of the noisy speech spectrum. The overlap-and-add technique was finally used to synthesize the noise-suppressed signal containing noise-overestimation distortion only. A similar procedure was taken to synthesize noise-suppressed signals containing noise-underestimation distortion only.

Fig. 2 shows example spectrograms of signals synthesized using the proposed binary mask based on noise overestimation constraints. The clean signal (panel a) was corrupted by babble at -5 dB SNR (panel b). The corrupted signal was filtered by the Wiener algorithm and is shown in panel c. The synthesized signal based on the overestimation constraint is shown in panel d. Although we see some attenuation distortion in the synthesized signal (panel d), it is clear that the voiced/unvoiced boundaries are more evident and the formants are for the most part preserved. Intelligibility listening tests were conducted next to validate the proposed binary mask.

3. Intelligibility listening tests

3.1. Methods and procedure

Listening tests were conducted to assess the intelligibility of speech processed using the proposed binary mask based on the two noise constraints. The sentences were taken from the IEEE database [13]. The IEEE sentences are phonetically balanced with relatively low word-context predictability and organized into lists of 10 sentences each. All sentence lists were designed to be equally intelligible, thereby allowing us to assess speech intelligibility in different conditions without being concerned that a particular list is more intelligible than another. The sentences were recorded at a sampling rate of 25 kHz by one male speaker in a sound-proof booth using Tucker Davis Technologies (TDT) recording equipment. The recordings are available

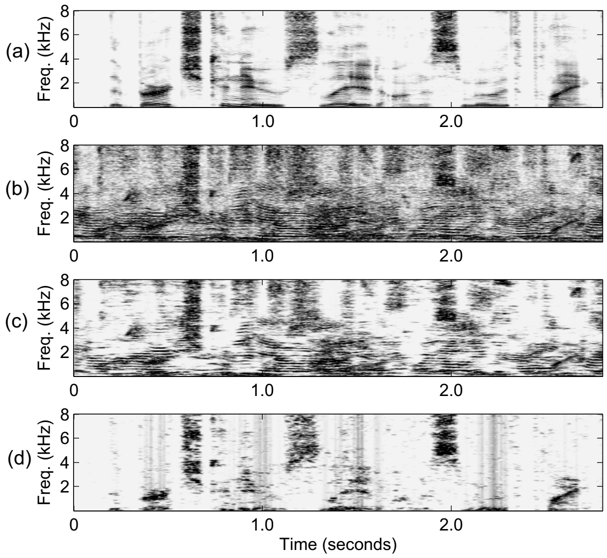

Figure 2: *Wideband spectrograms of the clean signal (panel a), corrupted signal (babble, SNR=-5 dB) (panel b), Wiener-processed signal (panel c), and binary masked signal with noise overestimation constraints applied to the enhanced signal (panel d).*

from [1]. Noisy speech was generated by adding babble noise at -10, -5, and 0 dB SNRs. The babble noise was produced by 20 talkers with equal number of female and male talkers.

In order to assess the full potential on speech intelligibility when the proposed binary mask is applied, we assumed knowledge of the true noise spectral magnitude $D(k,t)$. Thus, the binary mask was determined by comparing the true noise spectral magnitude $D(k,t)$ against the estimate of the noise magnitude $\hat{D}(k,t)$. In practice, the binary mask can be estimated using model-based classification or non-parametric decision rules (e.g., [4]).

Ten normal-hearing listeners were recruited for the listening experiments. They were all native speakers of American English, and were paid for their participation. The listeners participated in a total of 12 conditions (=3 SNR levels (-10, -5, 0 dB) × 4 processing conditions). The four processing conditions included speech processed using the Wiener algorithm with the noise overestimation mask ($\hat{D} > D$) and noise underestimation mask ($\hat{D} < D$), Wiener-processed speech without constraints, and the noise-corrupted (unprocessed) stimuli. The listening tests were conducted in a sound-proof room and stimuli were played to the listeners monaurally through Sennheiser HD 485 circumaural headphones at a comfortable listening level. The listening level was controlled by each individual but was fixed throughout the test for each subject. Prior to the sentence test, each subject listened to a set of noise-corrupted sentences to get familiar with the testing procedure. Two lists (20 sentences) were used per condition and none of the sentences were repeated across conditions. The order of the conditions was randomized across subjects. Listeners were asked to write down the words they heard, and intelligibility performance was assessed by counting the number of words identified correctly. The whole listening test lasted for about 2 hrs. Five-minute breaks were given to the subjects every 30 minute intervals.

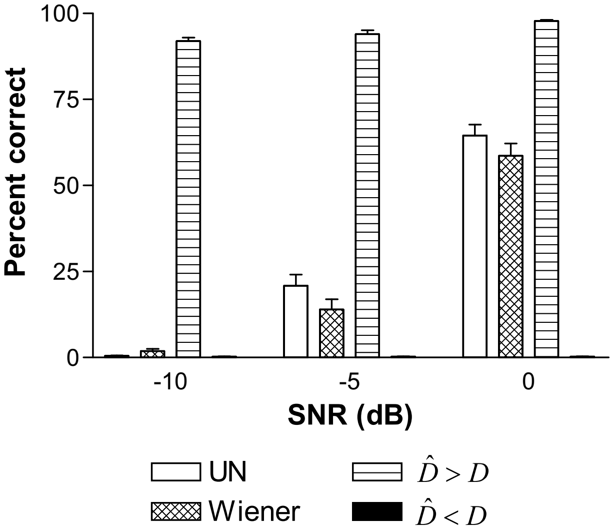

Figure 3: *Mean intelligibility scores as a function of SNR level and type of noise-estimation distortion. The bars labeled as "UN" show the scores obtained with noise-corrupted (unprocessed) stimuli, while the bars labeled as "Wiener" show the baseline scores obtained with the Wiener algorithm (no mask applied). The intelligibility scores obtained with the proposed noise-overestimation mask ($\hat{D} > D$) and noise underestimation mask ($\hat{D} < D$) applied to the enhanced spectra are also shown. Error bars indicate standard errors of the mean.*

3.2. Results and Discussion

Fig. 3 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 Fig. 3, performance improved dramatically when the proposed binary mask ($\hat{D} > D$) was applied. Performance at -10, -5, and 0 dB SNRs improved from near 0%, 21% and 65% with un-processed stimuli to 92%, 94% and 98% correct respectively when the proposed mask was applied. In contrast, performance degraded to near zero when the mask with noise underestimation constraints was applied. It is clear from Fig. 3 that the proposed noise-based binary mask performed as well as the known binary mask that uses the SNR selection criterion [2, 3].

To better understand the benefit of the proposed binary mask with noise overestimation constraints, we plotted the SNR histograms of all frequency bins falling in the noise underestimated and overestimated regions (see top panel of Fig. 4). It is clear from these histograms that the SNR of the noise-underestimated frequency bins are for the most part negative, thus explaining the poor performance (near 0% correct) obtained with the noise under-estimated mask. In contrast, the SNRs of the noise-overestimated frequency bins are more favorable and are distributed across both positive and negative SNR regions.

We further examined the estimated gain functions of the noise-overestimated and noise-underestimated frequency bins (Fig. 4). All data in Fig. 4 were obtained using twenty sentences corrupted by babble at -5 dB SNR. The middle and lower panels in Fig. 4 show the averaged gain functions plotted against

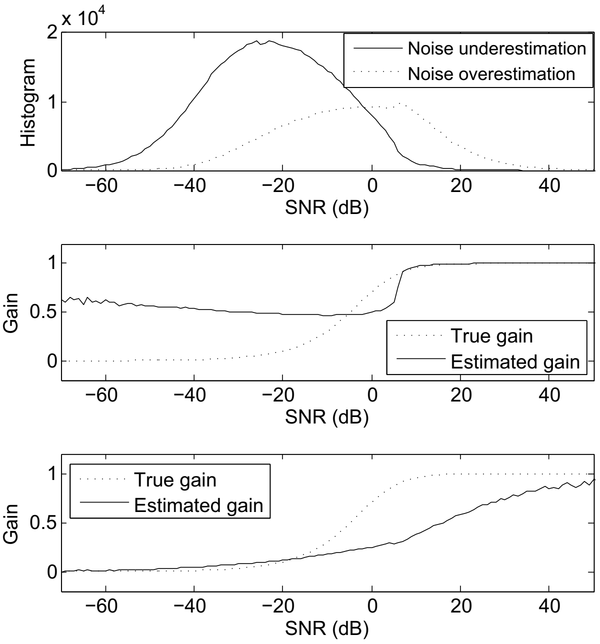

Figure 4: *Top panel shows histogram of SNRs for T-F units falling in noise-underestimated region and noise-overestimated region (top panel). Middle and bottom panels show true and averaged estimated Wiener gain for the noise-underestimated region (middle) and noise-overestimated region (bottom).*

the true SNR values. The true Wiener gain function is indicated in dashed lines and is plotted for comparative purposes. As shown in the middle panel, when the noise spectrum is underestimated, the estimated gain function overestimates the true Wiener gain function at negative SNR levels. This suggests that frequency bins that should otherwise be discarded (since their SNR<0 dB) are retained. In contrast, when the noise spectrum is over-estimated (bottom panel) the estimated gain function is in good agreement with the true Wiener gain function, at least for negative SNR levels. At positive SNR levels, the gain function is underestimated thereby introducing some attenuation distortion to the speech signal. This limited distortion, however, did not seem to impair speech intelligibility (see Fig. 3). In brief, from Fig. 4 we can conclude that the overestimation of the gain function at low SNR levels (< 0 dB) is harmful to speech intelligibility as it introduces noise-masked frequency bins.

4. Conclusions

A new binary mask was proposed for improving speech intelligibility. The proposed mask retains noise overestimated T-F units and removes noise underestimated T-F units. The proposed binary mask was evaluated using listening tests with normal-hearing listeners and the results demonstrated significant improvements in intelligibility even at extremely low SNR levels (-10 dB). The present study demonstrated that the commonly used binary mask based on the SNR criterion [3, 14] is not the only mask that can improve speech intelligibility. Binary masks based on either signal spectrum constraints [5] or noise constraints (present work) can also yield substantial gains in intelligibility.

5. Acknowledgements

This research was supported by Grant No. R01 DC007527 from National Institute of Deafness and other Communication Disorders, NIH.

6. References

[1] P. C. Loizou, *Speech enhancement: Theory and Practice*. Boca Raton: FL: CRC Press, 2007.

[2] D. Brungart, P. Chang, B. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *J. Acoust. Soc. Am.*, vol. 120, pp. 4007–4018, 2006.

[3] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Am.*, vol. 123, no. 3, pp. 1673–1682, 2008.

[4] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1486–1494, September 2009.

[5] G. Kim and P. Loizou, "Why do speech enhancement algorithms not improve speech intelligibility?" in *Proc. of IEEE Intern. Conf. on Acoust., Speech, Signal Processing*, 2010, pp. 4738–4741.

[6] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," vol. 9, pp. 504–512, Jul. 2001.

[7] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. of IEEE Intern. Conf. on Acoust., Speech, Signal Processing*, 1996, pp. 629–632.

[8] Y. Hu and P. C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Am.*, vol. 122, pp. 1777–1786, 2007.

[9] ——, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, pp. 588–601, 2007.

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust. Speech Signal Process.*, vol. ASSP-32, pp. 1109–1121, 1984.

[11] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Commun.*, vol. 48, pp. 220–231, 2006.

[12] Y. Lu and P. Loizou, "Speech enhancement by combining statistical estimators of speech and noise," in *Proc. of IEEE Intern. Conf. on Acoust., Speech, Signal Processing*, 2010, pp. 4754–4757.

[13] IEEE, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio and Electroacoustics*, pp. 225–246, 1969.

[14] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, pp. 267–285, 2001.