IEEK Transactions on Smart Processing and Computing

Binary Mask Criteria Based on Distortion Constraints Induced by a Gain Function for Speech Enhancement

Gibak Kim

School of Electrical Engineering, Soongsil University / Seoul, South Korea imkgb27@ssu.ac.kr

***** Corresponding Author: Gibak Kim

Received May 13, 2013; Revised May 28, 2013; Accepted June 12, 2013; Published August 31, 2013

** Regular Paper*

Abstract: Large gains in speech intelligibility can be obtained using the SNR-based binary mask approach. This approach retains the time-frequency (T-F) units of the mixture signal, where the target signal is stronger than the interference noise (masker) (e.g., SNR > 0 dB), and removes the T-F units, where the interfering noise is dominant. This paper introduces two alternative binary masks based on the distortion constraints to improve the speech intelligibility. The distortion constraints are induced by a gain function for estimating the short-time spectral amplitude. One binary mask is designed to retain the speech underestimated T-F units while removing the speech overestimated T-F units. The other binary mask is designed to retain the noise overestimated T-F units while removing noise underestimated T-F units. Listening tests with oracle binary masks were conducted to assess the potential of the two binary masks in improving the intelligibility. The results suggested that the two binary masks based on distortion constraints can provide large gains in intelligibility when applied to noise-corrupted speech.

Keywords: Binary mask, Speech intelligibility, Speech enhancement

1. Introduction

Despite the substantial advances in the development of methods for suppressing of background noise and improving the speech quality with a single microphone, there has been less progress in designing algorithms that can improve the speech intelligibility [1, 2]. Recent studies with normal-hearing listeners reported large gains in speech intelligibility using the SNR-based ideal binary mask technique [3, 4]. The binary mask was designed to leave the time-frequency (T-F) units untouched, where the target speech dominates the masker (noise), and discard the T-F units, where the masker is dominant over the target speech. Whether the target speech or masker is dominant is determined by examining the local SNR in each T-F unit. That is, if a local SNR at a T-F unit is greater than a pre-determined threshold, the T-F unit is considered to be a target-dominant region and the binary mask at this T-F unit is set to “1”. In a previous study [5], the SNR-based binary mask was estimated using a Bayesian classifier and the potential of the binary mask-based noise reduction for improving the speech intelligibility in noise was reported.

A different mask can alternatively be constructed by imposing constraints on speech or noise distortion after estimating the speech or noise magnitude [6, 7]. These masks do not rely on the SNR criterion and are different from the SNR-based binary mask in that the retained T-F units are multiplied by a gain function used in noise reduction techniques. Therefore, when a gain function is applied to a noisy spectrum, the resulting spectral amplitudes can be smaller than the true spectral amplitudes, i.e. attenuation distortion is introduced, or can be larger, resulting in amplification distortion. In the same manner, the noise spectral magnitudes are estimated and the effects of the noise spectrum over- or under-estimation can be examined.

To evaluate the binary masks based on the speech/noise spectrum overestimation/underestimation in view of intelligibility, listening tests were conducted with normal-hearing listeners. The results of the listening tests suggested that the binary mask based on a speech underestimation or noise overestimation constraint can

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education, Science and Technology (No. 2013027840).

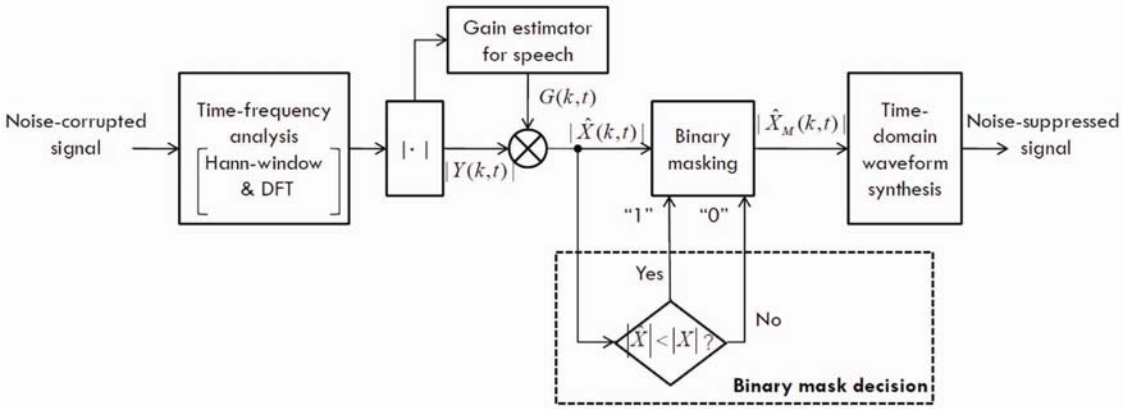

Fig. 1. Block diagram of the procedure for constructing the binary mask based on speech constraints.

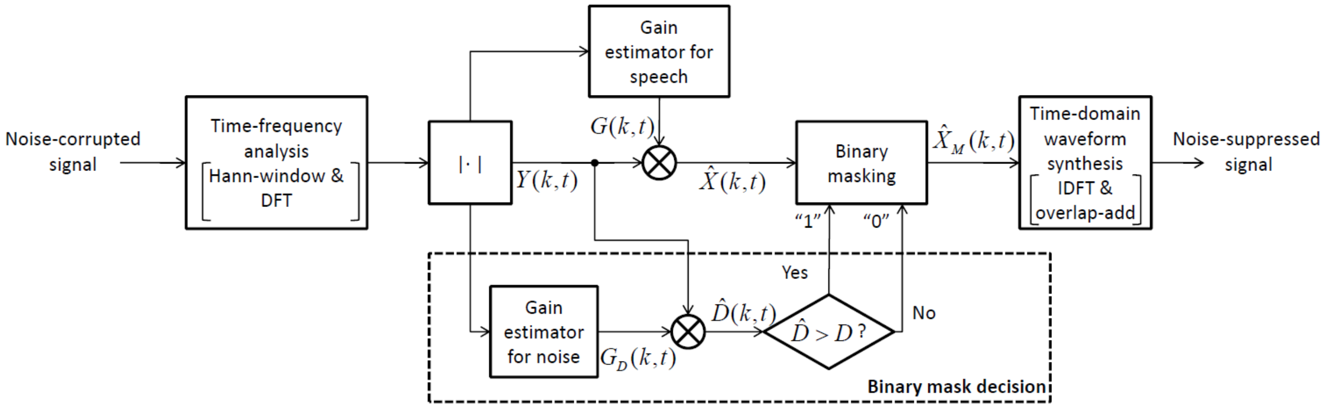

Fig. 2. Block diagram of the procedure for constructing the binary mask based on noise constraints.

improve substantially the speech intelligibility for speech waveform corrupted by background noise at low SNR levels.

2. SNR-based binary mask

SNR-based binary mask can be implemented by applying time-frequency analysis and selecting only the T-F units satisfying the following:

$$\hat{X}(k,t)=\begin{cases} Y(k,t) & \text{if } SNR_I(k,t)>\gamma \\ 0 & \text{else} \end{cases} \qquad (1)$$

where $Y(k,t)$, $\hat{X}(k,t)$ and $SNR_I(k,t)$ denote the noisy signal, binary masked signal, and instantaneous SNR at time frame t and frequency bin k, respectively. Typical values for the threshold γ range from 0 dB to -20 dB [4]. The number of T-F units retained increases with decreasing threshold, which produces a more conservative mask with less distortion. In principle, the original unprocessed mixture corresponds to a binary masked signal with a SNR threshold of $-\infty$. On the other hand, increasing the threshold reduces the total number of T-F units retained.

The SNR criterion has been used extensively in computational auditory scene analysis (CASA) studies [8], and the SNR-based binary mask was proposed as a computational goal for CASA [9].

Large gains in intelligibility have been reported by multiplying the ideal binary mask to the speech signal corrupted by noise, even at an extremely low (-5, -10 dB) SNR [3, 4]. In [5], the SNR-based binary mask was estimated using a Bayesian classifier. During the training stage, the true local SNR values were calculated and divided into two sets: one with $SNR>\gamma$ and one with $SNR<\gamma$. Two sets were modeled using Gaussian mixture models (GMMs) and the two classes were determined for test T-F units in the enhancement stage. The listening tests revealed the potential of the binary mask-based noise reduction for improving the speech intelligibility in noise.

3. Binary mask criteria based on speech/noise constraints

This section describes the binary mask based on speech or noise constraints. The time-frequency mask is constructed by imposing constraints on the speech or noise spectrum estimate, and is applied to the enhanced spectrum. These masks are different from the binary masks described in the previous section because the retained T-F units are multiplied by a gain function, which is determined to estimate the short-time spectral magnitude.

3.1 Estimation of speech and noise magnitude spectra

Figs. 1 and 2 show block diagrams of the procedure involved in the construction of the binary mask based on speech/noise constraints. The noisy speech waveform is first segmented into frames with a 20 ms duration and a 50 % overlap between adjacent frames. Each frame is applied by the Hann window and a 500-point (corresponding to 20 ms for the sampling rate of 25 kHz) discrete Fourier transform (DFT) is calculated. An estimate of the speech magnitude spectrum is obtained by multiplying the magnitude of the observed noisy spectrum, which is denoted as $|Y(k,t)|$, with a gain function for speech enhancement given by the following:

$$|\hat{X}(k,t)| = G(k,t)\cdot|Y(k,t)| \tag{2}$$

where $G(k,t)$ denotes the gain function, and $|\hat{X}(k,t)|$ is an estimate of the clean speech (magnitude) spectrum at time frame *t* and frequency bin *k*.

In this paper, a conventional square-root Wiener algorithm is used as a gain function [10]. The square-root Wiener gain is used because it is easy to implement, requires little computation, and has been reported to be equally effective, in terms of the speech quality and intelligibility, as other more sophisticated noise reduction algorithms [2, 11]. The square-root Wiener gain function is calculated based on the following equation:

$$G(k,t) = \sqrt{\frac{SNR_{prio}(k,t)}{1+SNR_{prio}(k,t)}} \tag{3}$$

where SNR_{prio} is the *a priori* SNR estimated using the following recursive equation [12]:

$$\begin{aligned} SNR_{prio}(k,t) &= \alpha\cdot\frac{|\hat{X}(k,t-1)|}{\hat{\lambda}_D(k,t-1)} \\ &\quad +(1-\alpha)\cdot\max\left[\frac{|Y(k,t)|^2}{\hat{\lambda}_D(k,t-1)}-1,0\right] \end{aligned} \tag{4}$$

where α is a smoothing constant chosen as 0.98 and $\hat{\lambda}_D$ is an estimate of the background noise variance. The noise estimation algorithm proposed in [13] is used to estimate the noise variance. For the binary mask based on the noise constraint in Fig. 2, the estimate of the noise spectral magnitude $|\hat{D}(k,t)|$ is obtained in a similar manner to (2) as follows:

$$|\hat{D}(k,t)| = G_D(k,t)\cdot|Y(k,t)| \tag{5}$$

where G_D is the noise-equivalent Wiener gain function expressed as [14]

$$G_D(k,t) = \sqrt{\frac{1}{1+SNR_{prio}(k,t)}} \tag{6}$$

3.2 Construction of binary masks based on speech constraints

After calculating the estimated speech magnitude spectrum, $|\hat{X}(k,t)|$, the binary mask is constructed by limiting (or controlling) the distortions introduced by errors in estimating the speech magnitude spectrum. In particular, $|\hat{X}(k,t)|>|X(k,t)|$ and $|\hat{X}(k,t)|<|X(k,t)|$ denote the speech overestimation distortion and speech underestimated distortion, respectively. To assess the effect of a speech overestimation distortion alone or an underestimation distortion alone on the speech intelligibility, the speech overestimation/underestimation constraints are imposed on the estimated speech spectral magnitude.

More precisely, the estimated speech magnitude spectrum is first compared with the true speech magnitude spectrum for each T-F unit. The T-F units satisfying the constraint are retained, whereas the T-F units violating the constraints are removed. For example, to estimate the speech underestimation constraint, the modified magnitude spectrum $|\hat{X}_M(k,t)|$, is calculated as follows:

$$|\hat{X}_M(k,t)| = \begin{cases} |\hat{X}(k,t)| & \text{if } |\hat{X}(k,t)|<|X(k,t)| \\ 0 & \text{else} \end{cases}. \tag{7}$$

3.3 Construction of binary masks based on noise constraints

A binary mask based on noise constraints can also be defined in a similar manner. First, the estimated noise magnitude spectrum, $|\hat{D}(k,t)|$, is computed and the binary mask is constructed by imposing the constraints on the distortions introduced by the errors in estimating the noise magnitude spectrum. If the estimated noise magnitude is greater than the true noise magnitude, (i.e. $|\hat{D}(k,t)|\geq|D(k,t)|$), it is denoted as noise overestimation distortion. Noise underestimation distortion occurs in the opposite case ($|\hat{D}(k,t)|<|D(k,t)|$). The new modified magnitude spectrum $|\hat{X}_M(k,t)|$ by imposing the noise overestimation constraint is expressed as

$$|\hat{X}_M(k,t)| = \begin{cases} |\hat{X}(k,t)| & \text{if } |\hat{D}(k,t)|\geq|D(k,t)| \\ 0 & \text{else} \end{cases}. \tag{8}$$

According to the above selection of T-F units, an inverse DFT is applied to the modified spectrum $|\hat{X}_M(k,t)|$ at each time frame and frequency bin using the phase of the noisy speech spectrum. The noise suppressed signal is synthesized by applying the overlap-and-add technique.

4. Intelligibility listening tests

4.1 Methods and procedure

Listening tests were conducted to assess the intelligibility of the speech processed using the binary mask based on the two noise constraints. The sentences were taken from the IEEE database [15]. The IEEE sentences were balanced phonetically 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 the speech intelligibility to be assessed under 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 from [1]. Noisy speech was generated by adding babble noise at -5 and 0 dB SNRs, which was produced by 20 talkers (10 men and 10 women).

Knowledge of the true speech and noise spectral magnitudes $\left(|X(k,t)|, |D(k,t)|\right)$ was assumed to assess the full potential on the speech intelligibility when the distortion constraint based binary mask was applied. Therefore, the binary mask was determined by comparing the true speech (or noise) spectral magnitude with an estimate of the speech (or noise) magnitude. In practice, the binary mask can be estimated using a model-based classification or non-parametric decision rules (e.g., [5]).

Seven normal-hearing listeners were recruited for the listening experiments for the noise constraint-based binary mask. Another ten listeners were recruited to test the speech constraint based binary mask. All were native speakers of American English, and were paid for their participation. Each listener participated under a total of 8 conditions (=2 SNR levels (-5, 0 dB) × 4 processing conditions). The four processing conditions included speech processed using the Wiener algorithm with the speech (or noise) overestimation mask, $|\hat{X}|>|X|$ (or $|\hat{D}|\geq|D|$) and speech (or noise) underestimation mask, $|\hat{X}|\leq|X|$ (or $|\hat{D}|<|D|$), Wiener-processed speech without constraints, and the noise-corrupted (unprocessed) stimuli.

The listening tests were conducted in a sound-proof room and the 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. Before the sentence test, each subject listened to a set of noise-corrupted sentences to become familiar with the testing procedure. Two lists (20 sentences) were used per condition and none of the sentences were repeated across the conditions. The order of the conditions was randomized across subjects. The listeners were asked to write down the words that they heard. The intelligibility performance was assessed by counting the number of words identified correctly. The entire listening test lasted for approximately 1.5 hrs. Five-minute breaks were given to the subjects at every 30 minute intervals.

4.2 Results

Figs. 3 and 4 present the results of the listening test expressed in terms of the mean percentage of words identified correctly by normal-hearing listeners. The bars indicated “UN” show the scores obtained with noise-corrupted (un-processed) stimuli, whereas the bars labeled “Wiener” show the scores obtained using the square-root

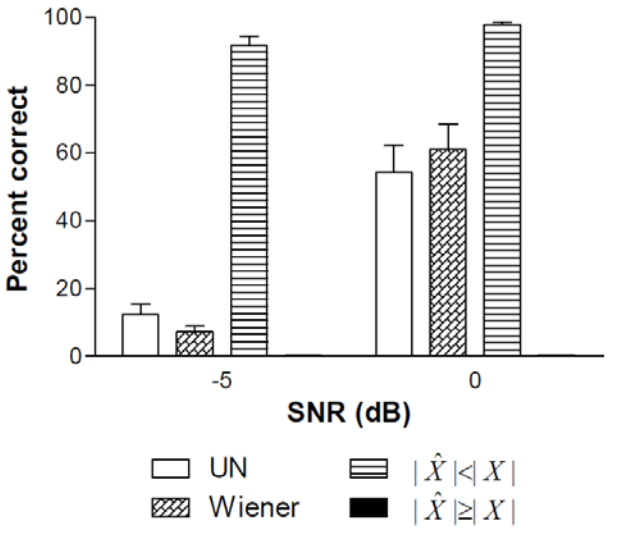

Fig. 3. Mean intelligibility scores as a function of the SNR level and type of speech estimation distortion.

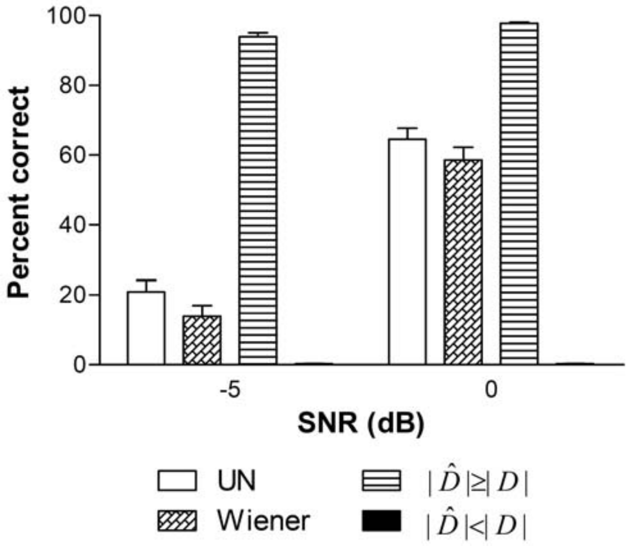

Fig. 4. Mean intelligibility scores as a function of the SNR level and type of noise estimation distortion.

Wiener algorithm (no mask applied). The error bars indicate the standard error of the mean. As shown in Figs. 3 and 4, the performance improved dramatically when binary masks $\left(|\hat{X}|<|X|, |\hat{D}|\geq|D|\right)$ were applied. The performances at -5, and 0 dB SNRs were improved from approximately 12 % and 54 % with unprocessed stimuli to 92 % and 98 % correct, respectively, when the mask based on a speech underestimation constraint $(|\hat{X}|<|X|)$ was applied. In the case that the binary mask based on a noise overestimation constraint $(|\hat{D}|\geq|D|)$ was applied, the performances were improved from 21 % and 65 % with un-processed stimuli to 94 % and 98 % correct, respectively at -5, and 0 dB SNRs. In contrast, the performance degraded to near zero when the mask with speech overestimation or noise underestimation constraints was applied. This is consistent with the results reported elsewhere [6]. Note that the Wiener processed speech showed lower performance in intelligibility than unprocessed speech due to the introduced distortion, which match the results in [2]. Figs. 3 and 4 show that the distortion constraint-based binary masks performed as well as the known binary mask that uses the SNR selection criterion [3, 4].

5. Conclusion

Two binary masks were evaluated for their ability to improve the speech intelligibility. The masks were induced by applying a speech distortion constraint or noise distortion constraint. The speech constraint-based mask retains the speech underestimated T-F units and the noise constraint-based mask retains the noise overestimated T-F units.

Listening tests with normal-hearing listeners were conducted to evaluate the binary masks. The results revealed significant improvements in intelligibility at low SNR levels (-5 and 0 dB). The present study showed that the commonly used binary mask based on the SNR criterion [3, 4] is not the only mask that can improve speech intelligibility. Binary masks based on either signal spectrum constraints or noise constraints can also produce substantial gains in intelligibility.

References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, CRC Press, 2013. Article (CrossRef Link)

[2] Y. Hu and P. C. Loizou, “A comparative intelligibility study of single-microphone noise reduction algorithms,” *J. Acoust. Soc. Am.*, Vol. 22, No. 3, pp. 1777-1786, 2007. Article (CrossRef Link)

[3] 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. Article (CrossRef Link)

[4] 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. Article (CrossRef Link)

[5] 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. Article (CrossRef Link)

[6] P. C. Loizou and G. Kim, “Why current speech-enhancement algorithms do not improve speech intelligibility: analysis and suggested solutions,” *IEEE trans. Audio, Speech and Language Processing*, Vol. 19, No. 1, pp. 47-56, January 2011. Article (CrossRef Link)

[7] G. Kim and P. C. Loizou, “A new binary mask based on noise constraints for improved speech intelligibility,” *Proc. Interspeech*, pp. 1632-1635, Makuhari, Chiba, Japan, September 2010. Article (CrossRef Link)

[8] D. Wang and G. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, New York: Wiley/IEEE Press, 2006. Article (CrossRef Link)

[9] D. Wang, On ideal binary mask as the computational goal of auditory scene analysis, in P. Divenyi (ed.), *Speech Separation by Humans and Machines*, Dordrecht, the Netherlands: Kluwer Academic, pp. 181-187. Article (CrossRef Link)

[10] P. Scalart and J. Filho, “Speech enhancement based on a priori signal to noise estimation,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 629-632, 1996. Article (CrossRef Link)

[11] Y. Hu and P. C. Loizou, “Subjective comparison and evaluation of speech enhancement,” *Speech Commun.*, Vol. 49, pp. 588-601, 2007. Article (CrossRef Link)

[12] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. Acoust. Speech Signal Processing*, Vol. ASSP-32, pp. 1109-1121, 1984. Article (CrossRef Link)

[13] S. Rangachari and P. C. Loizou, “A noise-estimation algorithm for highly non-stationary environments,” *Speech Commun.*, Vol. 48, pp. 220-231, 2006. Article (CrossRef Link)

[14] Y. Lu and P. C. Loizou, “Speech enhancement by combining statistical estimators of speech and noise,” *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp.4754-4757, 2010. Article (CrossRef Link)

[15] IEEE Subcommittee, “IEEE Recommended Practice for Speech Quality Measurements,” *IEEE Trans. Audio and Electroacoustics*, Vol. AU-17, No. 3, pp. 225-246, 1969. Article (CrossRef Link)

Gibak Kim received his B.S. and M.S. degrees in electronics engineering and Ph.D. degree in electrical engineering from Seoul National University, Seoul, Korea, in 1994, 1996 and 2007, respectively. From 1996 to 2000, he was with the Machine Intelligence Group, Department of the Information Technology, LG Electronics, Inc., Seoul, Korea. From 2000 to 2003, he also worked as a Senior Research Engineer involved at Voiceware, Ltd. in the development of the automatic speech recognizer. From 2007 to 2010, he was a Research Associate at the University of Texas at Dallas, Richardson. He was an Assistant Professor at Daegu University, Daegu, Korea from 2010 to 2011. He is currently Assistant Professor of School of Electrical Engineering at Soongsil University, Seoul, Korea. His general research interests include speech/image processing.