Improving Speech Intelligibility in Noise Using a Binary Mask That Is Based on Magnitude Spectrum Constraints

Gibak Kim and Philipos C. Loizou, *Senior Member, IEEE*

***Abstract*—A new binary mask is introduced for improving speech intelligibility based on magnitude spectrum constraints. The proposed binary mask is designed to retain time-frequency (T-F) units of the mixture signal satisfying a magnitude constraint while discarding T-F units violating the constraint. Motivated by prior intelligibility studies of speech synthesized using the ideal binary mask, an algorithm is proposed that decomposes the input signal into T-F units and makes binary decisions, based on a Bayesian classifier, as to whether each T-F unit satisfies the magnitude constraint or not. Speech corrupted at low signal-to-noise (SNR) levels (−5 and 0 dB) using different types of maskers is synthesized by this algorithm and presented to normal-hearing listeners for identification. Results indicated substantial improvements in intelligibility over that attained by human listeners with unprocessed stimuli.**

***Index Terms*—Binary mask, speech enhancement, speech intelligibility.**

I. Introduction

RECENT studies with normal-hearing listeners have reported large gains in speech intelligibility using the SNR-based ideal binary mask technique [1], [2]. The binary mask was designed to retain 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) [3]. In our previous work [4], we demonstrated the potential of the binary mask technique to improve speech intelligibility when the SNR-based mask was estimated using a binary Bayesian classifier.

A different mask, that does not rely on the SNR criterion, can alternatively be constructed by imposing constraints on the two types of gain-induced distortion: amplification distortion occurring when the estimated (by a noise-suppression algorithm) magnitude is larger than the true magnitude, and attenuation distortion occurring when the estimated magnitude is smaller than the true magnitude [5]. The intelligibility listening studies in [5] showed that the processed speech was found to be substantially more intelligible than the noisy speech when appropriate constraints were imposed on the magnitude spectrum, particularly when the amplification distortions were either limited or eliminated. In that study, the ideal (oracle) magnitude-based mask was investigated. In this letter, we propose a noise-suppression algorithm, which *estimates* the binary mask, based on magnitude constraints, from the noisy observations. Listening tests are conducted with normal-hearing listeners to evaluate the proposed noise-suppression algorithm in terms of speech intelligibility benefits.

Manuscript received August 24, 2010; revised October 04, 2010; accepted October 06, 2010. Date of publication October 14, 2010; date of current version November 04, 2010. This work was supported by Grant R01 DC007527 from the National Institute of Deafness and other Communication Disorders, NIH. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jen-Tzung Chien.

G. Kim is with the School of Electronic Engineering, College of Information & Communication, Daegu University, Gyeongbuk 712-714, Korea (e-mail: imkgb27@gmail.com).

P. C. Loizou is with the Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: loizou@utdallas.edu).

Digital Object Identifier 10.1109/LSP.2010.2087412

II. Proposed Noise-Suppression Algorithm

Our previous study [5] demonstrated that large gains in intelligibility could be achieved when appropriate constraints are imposed on the gain-induced speech magnitude distortions. Based on the encouraging findings of our prior study, we propose a method to estimate this binary mask, which depends on magnitude-spectrum constraints. Fig. 1 shows the block diagram of the proposed algorithm, consisting of a training stage (bottom panel) and an intelligibility enhancement stage (top panel), both of which are described next.

A. A Binary Mask Based on Magnitude-Spectrum Constraints

In this section, we describe a binary mask that is based on magnitude spectrum constraints rather than the local SNR criterion [5]. Let $Y(k,t)$ denote the noisy spectrum at time frame t and frequency bin k. The estimate of the signal spectrum magnitude, $|\hat{X}(k,t)|$, is obtained by multiplying the magnitude of the noisy spectrum, $|Y(k,t)|$ with a gain function $G(k,t)$ as follows:

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

The (square-root) Wiener gain function [6], given by the following equation, was used:

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

where $\mathrm{SNR}_{\mathrm{prio}}$ is the *a priori* SNR estimated using the following equation:

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

where $\alpha = 0.98$ is a smoothing constant and $\hat{\lambda_D}(k,t)$ is the estimate of the background noise variance obtained by a noise

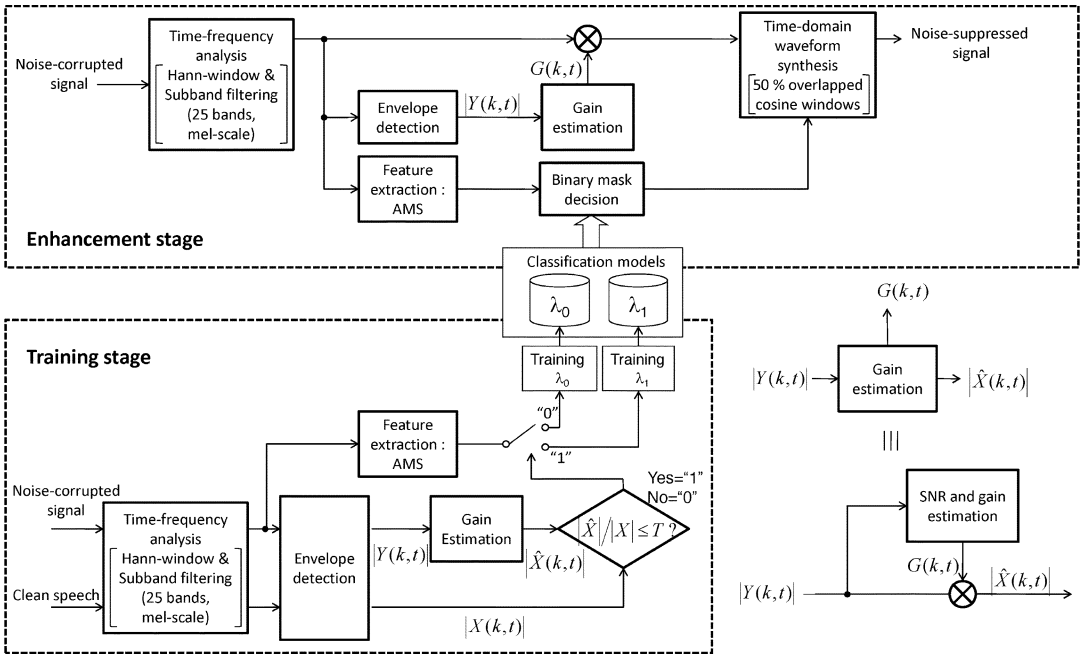

Fig. 1. Block diagram of the procedure used for constructing the proposed binary mask based on magnitude constraints.

estimation algorithm (the algorithm proposed in [7] was used in this letter). The estimate of speech spectrum magnitude was initialized as $|\hat{X}(k,0)| = 0$ and the estimate of the noise variance was initialized as $\hat{\lambda_D}(k,0) = \hat{\lambda_D}(k,1) = |Y(k,1)|^2$.

The binary mask is constructed by imposing constraints on the distortions introduced by the gain function. More precisely, the estimated magnitude spectrum $|\hat{X}(k,t)|$ is compared against the true speech magnitude $|X(k,t)|$ for each T-F unit (k,t), and T-F units satisfying the constraint are retained, while T-F units violating the constraints are removed. The new magnitude spectrum $|X_M(k,t)|$ is computed as follows:

$$|\hat{X}_M(k,t)| = \begin{cases} |\hat{X}(k,t)| & \text{if } \frac{|\hat{X}(k,t)|}{|X(k,t)|} \le T \\ 0 & \text{else} \end{cases} \tag{4}$$

where T denotes the threshold value. The above constraint was chosen since our previous listening study [5] indicated that the gain-induced amplification distortion (occurring when $|\hat{X}| > |X|$) had the most detrimental effect on speech intelligibility. The threshold, T, was set to 2 for the first 15 frequency bands (spanning 68–2186 Hz) and to 6 for the higher frequency bands. The higher threshold value for the higher frequency bands was found to better preserve low-energy consonants, since the use of higher thresholds tends to retain more T-F units. This was also done to be consistent with the SNR-based binary mask proposed in [4].

The above mask (4) was found to be quite effective in improving speech intelligibility [5]. It is, however, the ideal magnitude-constraints binary mask (IMBM), as it requires access to the clean magnitude spectrum, which we do not have. A method for estimating the above mask from noisy observations is presented next.

B. Estimating the Magnitude-Constraints Based Binary Mask

We used a Bayesian classifier similar to that used in [4] to identify T-F units as either satisfying or violating the constraint specified in (4). The noise-corrupted signal is first segmented into 20-ms frames, with 50% overlap between adjacent frames. Each speech frame is bandpass filtered into 25 channels according to mel-frequency spacing and followed by Hann windowing. In the training stage (Fig. 1), features are extracted, typically from a large speech corpus, and then used to train two Gaussian mixture models (GMMs) representing two feature classes: satisfying the constraint and violating the constraint. Amplitude modulation spectrograms (AMSs) are used as features, as they are neurophysiologically and psychoacoustically motivated [8]. A two-class Bayesian classifier was used to estimate the binary mask for each T-F unit. The distribution of the feature vectors of each class was represented with a GMM (256 mixtures) and the two classes were denoted as λ_0 for mask "0" and λ_1 for mask "1". The class λ_1 was trained with feature vectors composed of T-F units satisfying the given speech magnitude constraint (4). On the contrary, when a T-F unit violates the constraint, the corresponding feature vector was used for training λ_0. In the enhancement stage, the AMS features are first computed and the binary mask values of each T-F unit are estimated using a Bayesian classifier with models λ_0 and λ_1. The target speech envelopes of the 25 bands are estimated using the Wiener gain function [see (1)–(3)]. T-F units are retained or removed according to the estimated binary mask value for each T-F unit. Finally, the enhanced speech waveform is reconstructed by applying 50%-overlapped cosine windows to the Wiener-processed subband signals[1]. Detailed procedures of feature extraction, training, and enhancement can be found in [4].

III. Intelligibility Listening Tests

We conducted listening tests to assess the intelligibility of speech synthesized using the estimated binary mask

[1]Note that the proposed binary mask was applied to the Wiener-processed subband signals of the 25 channel (mel-spaced) filterbank (see Fig. 1). This is equivalent to applying the binary mask to the estimate of the speech magnitude spectrum in (4). In contrast, the SNR-based binary mask was applied to the noisy subband signals [4].

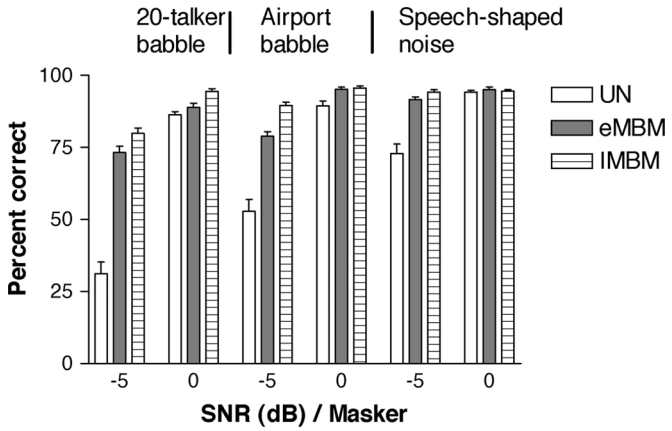

Fig. 2. Mean intelligibility scores for female-speaker sentences as a function of SNR level and masker type. The bars labeled as "UN" show the scores obtained with noise-corrupted (unprocessed) stimuli, while the bars labeled as "IMBM" show the scores obtained with sentences processed using the ideal magnitude binary mask (4). "eMBM" shows the scores for sentences synthesized with the magnitude binary mask estimated using the male-speaker trained GMMs. Error bars indicate standard errors of the mean.

(Section II-B). Sentences taken from the IEEE database [9] were used as test material[2]. The sentences in the IEEE database are phonetically balanced with relatively low word-context predictability. The sentences were produced by one male and one female speaker in a sound-proof booth using Tucker Davis Technologies (TDT) recording equipment. The sentences were originally recorded at a sampling rate of 25 kHz and down-sampled to 12 kHz. Three types of noise (20-talker babble, airport-babble, speech-shaped noise) were used as maskers. The (steady) speech-shaped noise was stationary having the same long-term spectrum as the sentences in the IEEE corpus. The airport-babble was recorded at an airport [11], and the 20-talker babble (20 talkers with equal number of female and male talkers) was taken from the Auditec CD (St. Louis, MO).

A total of 390 IEEE sentences (produced by a male talker) were used to train the GMM models. These sentences were corrupted by three types of noise at -5, 0 and 5 dB SNR. The maskers were randomly cut from the noise recordings and mixed with the target sentences at the prescribed SNRs. Each corrupted sentence had thus a different segment of the masker, and this was done to evaluate the robustness of the Bayesian classifier in terms of generalizing to different segments of the masker having possibly different temporal/spectral characteristics. The male-speaker data were used in the training and the female-speaker data were used in the listening tests. This was done to show the robustness of the binary classifier in terms of handling speakers not included in the training. The average fundamental frequencies (F0) for the male-speaker and female-speaker data were 122 and 241 Hz respectively. There was no overlap between the sentence lists used in the training and test data sets.

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 18 conditions ($=$ 2 SNR levels (-5, 0 dB) $\times$ 3 processing conditions $\times$ 3 types of maskers). The three processing conditions included the noise-corrupted (unprocessed) stimuli, speech processed using the ideal magnitude mask (4), denoted as IMBM, and the estimated binary mask (Section II-B), denoted as eMBM. The duration of each sentence was approximately 2.5 s. The experiments were performed 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 for every 30 minutes of listening.

[2]The sentence recognition test was chosen over a diagnostic rhyme test (DRT) [10], for the following reasons: sentence tests 1) better reflect real-world communicative situations, 2) are open-set tests, and as such scores may vary from a low of 0% correct to 100% correct. In contrast, the DRT test is a closed-set test, has a chance score of 50% and needs to be corrected for chance.

TABLE I
HIT AND FALSE ALARM RATES (FA) OBTAINED IN THE VARIOUS MASKER CONDITIONS

	20-talker babble		Airport babble		Speech-shaped noise	
	-5 dB	0 dB	-5 dB	0 dB	-5 dB	0 dB
HIT	75.65%	78.25%	73.46%	75.91%	72.34%	80.49%
FA	9.56%	12.96%	10.11%	11.69%	8.87%	11.23%
HIT-FA	66.09%	65.29%	63.45%	64.22%	63.47%	69.26%

IV. RESULTS AND DISCUSSION

Fig. 2 shows the results of the listening tests expressed in terms of the mean percentage of words identified correctly. A substantial improvement in intelligibility was obtained with the proposed algorithm (eMBM), compared to that attained with unprocessed (noise-corrupted) speech. The improvement (over 40% points in babble at -5 dB SNR) was more evident at -5 dB SNR levels for all three maskers tested. Performance at 0 dB SNR was limited in most cases by ceiling (plateau) effects. Performance with the eMBM approached in most cases the upper bound, i.e., performance with the oracle mask (IMBM).

To quantify the accuracy of the binary Bayesian classifier, we report the average hit (HIT) and false alarm (FA) rates for three test sets in Table I. HIT and FA rates were computed by comparing the estimated binary mask against the (oracle) IMBM. High hit rates (lowest with speech-shaped noise at -5 dB; 72.34%) and low false-alarm rates (highest with 20-talker babble at 0 dB; 12.96%) were obtained. Fig. 3(c) shows example spectrograms of signals synthesized using the proposed algorithm. The clean signal [Fig. 3(a)] was corrupted by airport babble at -5 dB SNR [Fig. 3(b)]. As can be seen from Fig. 3(c), the voiced/unvoiced boundaries are made clearer following the processing, and the residual noise is substantially reduced.

Both the SNR-based mask [4] and the proposed magnitude binary mask improved speech intelligibility. This raises the question as to whether one mask offers any advantages over the

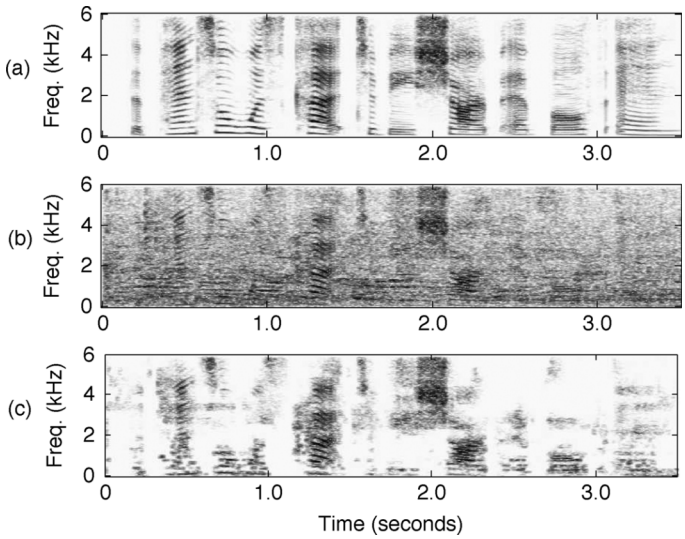

Fig. 3. (a) Narrowband spectrograms of the clean signal, (b) corrupted signal (airport babble, $\mathrm{SNR} = -5$ dB), and (c) synthesized signal obtained by the proposed algorithm.

TABLE II
SPEECH QUALITY EVALUATION, BASED ON PESQ SCORES, AND COMPARISON BETWEEN THE SNR-BASED MASK (SNR BM) AND MAGNITUDE-SPECTRUM BASED MASK (EMBM) IN THE VARIOUS MASKER CONDITIONS

	20-talker babble		Airport babble		Speech-shaped noise	
	-5 dB	0 dB	-5 dB	0 dB	-5 dB	0 dB
UN	0.97	1.35	0.95	1.22	0.78	1.17
SNR BM	1.03	1.36	0.89	1.27	0.92	1.32
eMBM	1.10	1.49	1.13	1.54	1.35	1.80

other, perhaps in terms of speech quality. Given that the proposed magnitude binary mask is applied to the Wiener processed spectra rather than the corrupted spectra, we hypothesized that the proposed magnitude binary mask yields better speech quality. To test this hypothesis, we used the Perceptual Evaluation of Speech Quality (PESQ) measure to assess the quality of the processed speech [12]. Table II compares the speech quality of speech synthesized using the SNR-based [4] and magnitude-based masks (eMBM). The PESQ scores for the proposed algorithm were found to be higher in all conditions, compared to the unprocessed noise-corrupted speech. Furthermore, the PESQ scores of speech synthesized using the estimated magnitude binary mask were found to be consistently higher than those computed using the SNR-based mask. The largest gain (0.48) in PESQ scores was obtained for speech-shaped noise at 0 dB SNR while the lowest gain (0.07) was obtained for the 20-talker babble at -5 dB SNR. We believe that the higher PESQ scores obtained with the proposed algorithm can be attributed to the fact that better noise suppression was achieved since the Wiener gain was applied to the noisy signal before binary masking.

V. SUMMARY

A new noise-suppression algorithm was designed to improve speech intelligibility using a binary mask that was based on magnitude spectrum constraints rather than on the local SNR criterion [1]. The binary mask for each T-F unit was estimated using a Bayesian classifier with the use of neurophysiologically motivated features (AMS). Each T-F unit was retained or removed according to the estimated binary mask. The proposed algorithm was evaluated using listening tests with normal-hearing listeners and results indicated large gains in intelligibility (Fig. 2). Objective evaluation (based on PESQ scores) of speech synthesized using the magnitude-based mask revealed better speech quality than that obtained with speech synthesized using the SNR-based mask.

REFERENCES

[1] 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. Amer.*, vol. 120, pp. 4007–4018, 2006.
[2] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *J. Acoust. Soc. Ameer.*, vol. 123, no. 3, pp. 1673–1682, 2008.
[3] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 181–197.
[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. Amer.*, vol. 126, no. 3, pp. 1486–1494, Sept. 2009.
[5] G. Kim and P. C. Loizou, "Why do speech enhancement algorithms not improve speech intelligibility?," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, 2010, pp. 4738–4741.
[6] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing*, 1996, pp. 629–632.
[7] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Commun.*, vol. 48, pp. 220–231, 2006.
[8] B. Kollmeier and R. Koch, "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," *J. Acoust. Soc. Amer.*, vol. 95, no. 3, pp. 1593–1602, 1994.
[9] "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. 17, no. 3, pp. 225–246, 1969.
[10] W. D. Voiers, "Evaluating processed speech using the diagnostic rhyme test," *Speech Technol.*, pp. 30–39, Jan./Feb. 1983.
[11] Online Sound Effects Database [Online]. Available: Online:http://labrosa.ee.columbia.edu/dpwe-bin/sfxlist.cgi
[12] "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU-T Rec. p. 862, 2000.