TECHNIQUES FOR ESTIMATING THE IDEAL BINARY MASK

Yi Hu and Philipos C. Loizou

Center for Robust Speech Systems
Department of Electrical Engineering
University of Texas at Dallas
Richardson, TX, USA.
{loizou,yihuyxy}@utdallas.edu

ABSTRACT

This paper provides a comparison of binary mask estimation techniques, based on different ways of estimating the instantaneous SNR. The effect of six different gain functions and three noise estimation algorithms on estimating the SNR, and subsequently the binary mask was assessed. New criteria are proposed for classifying time-frequency bins as belonging to the target or masker signals. Sentences from the NOIZEUS corpus embedded at 0-10 dB SNR levels in four types of noise were used for evaluation. Performance of the binary mask estimation algorithms was evaluated in terms of hit rate and false alarm. Results indicated that the use of different SNR estimation techniques affects primarily the false alarm rate.

Index Terms— Speech intelligibility, SNR estimation, ideal binary mask

1. INTRODUCTION

The ideal binary mask has been set as a computational goal in computational auditory scene analysis (CASA) algorithms and has also been adopted in "missing feature" speech recognition techniques [1]. Recently, a number of studies demonstrated high gains in speech intelligibility using the ideal binary mask technique [2–4]. The ideal binary mask takes values of zero and one, and is constructed by comparing the local SNR in each time-frequency (T-F) bin against a preset threshold. In the intelligibility studies, it is usually applied to the time-frequency representation of a mixture signal and eliminates portions of a signal (those assigned to a "zero" value) while preserving others (those assigned to a "one" value).

Accurate estimates of the ideal binary mask are thus important not only for robust speech recognition but also for algorithms aimed at improving speech intelligibility. Various methods have been proposed to estimate the binary mask and include methods based on a Bayesian classification of speech-specific features [5], pitch continuity information [2], sound-localization cues [6] and estimates of the *posterior* SNR [7].

The present study analyzes various other techniques that do not require training of a large labeled corpus, and do not rely on having access to auditory-grouping cues or access to binaural inputs. All proposed techniques are centered around estimation of the instantaneous SNR from noisy observations (no training is required).

This paper is organized as following. Section 2 describes the ideal binary mask estimation techniques, Section 3 presents the results and Section 4 presents the conclusions.

2. BINARY MASK ESTIMATION TECHNIQUES

The ideal binary mask is typically derived by comparing the true instantaneous SNR to a preset threshold (e.g. 0 dB). In practice, we do not have access to the true instantaneous SNR and have to estimate it from the noisy observations. A common method used in the speech-enhancement literature for estimating the SNR is the decision-directed approach [8]. More precisely, the so called *a priori* SNR (denoted as ξ) at frame m for frequency bin k is often estimated using the decision directed approach as follows [8]:

$$
\begin{aligned}
\hat{\xi}_k(m) &= \alpha \frac{(G(k, m-1)Y_k(m-1))^2}{\hat{D}_k^2(m-1)} + \\
&\quad (1-\alpha)\max(\gamma_k(m) - 1, 0)
\end{aligned} \tag{1}
$$

where $\alpha = 0.98$, $G(k, m-1)$ is the gain function of frame $m-1$ at frequency bin k, $\hat{D}_k$ is the estimate of the noise magnitude spectrum, Y_k is the noisy-speech magnitude spectrum and $\gamma_k = Y_k^2/\hat{D}_k^2$ is the *posterior* SNR. It is clear from the above equation that the accuracy of the SNR estimate depends on the gain function ($G(k, m)$) and the estimate of the noise spectrum ($\hat{D}_k$). In [8], the MMSE gain function was used, but other gain functions (e.g., Wiener gain function) could potentially be used. Similarly, there exist many methods for estimating the noise spectrum and include voice activity detection (VAD) algorithms which update the noise spectrum during silent periods and noise-estimation algorithms which update the noise spectrum continuously even during speech activity [7, Ch. 9].

Since the choice of the gain function as well as the type of noise spectrum estimation method used can potentially affect the accuracy of the binary mask, we investigate the performance of six different gain functions and three different methods for estimating the noise spectrum. The following six gain functions were considered: the Wiener algorithm (Wiener) [9], the Minimum Mean Square Error (MMSE) algorithm [8], the MMSE algorithm with speech presence uncertainty (MMSE-SPU) [8], the log Minimum Mean Square Error (logMMSE) algorithm [10], the Perceptually Motivated Bayesian Estimators of the Magnitude Spectrum (pMMSE) [11], and the spectral subtraction algorithm (SpecSub) [12]. Detailed description and Matlab implementations of all algorithms can be found in [13].

The performance of two different noise estimation methods [14, 15] and one voice activity detection (VAD) algorithm [16] was also examined. The threshold value for voice activity was set to 0.25 in the VAD algorithm. These algorithms were used to update the noise spectrum $\hat{D}_k$ in Eq. 1.

Based on Eq. 1, we can declare a T-F unit as being target dominated if $\xi > 1$ (i.e., local SNR > 0 dB) and masker-dominated if $\xi \leq 1$. Aside from using the SNR value $\hat{\xi}$ from Eq. 1 as a criterion, we also considered four other criteria. These criteria are summarized in Table 1 and include: the *posterior* SNR (γ) as per [7], the SNR criterion as per [17], combined γ and ξ, and the conditional probability of speech-presence, $p(H_1|Y(\omega_k))$, where $Y(\omega_k)$ denotes the complex noisy spectrum. The value of $p(H_1|Y(\omega_k))$ was determined as per [8], with $q = 0.3$. In our study, a T-F unit was declared target-dominated if $p(H_1|Y(\omega_k)) > 0.9$. Criterion C2 (Table 1) is a new criterion that is proposed in this study.

3. EVALUATION OF BINARY MASK ESTIMATION TECHNIQUES

The NOIZEUS database, comprising of 30 sentences produced by six speakers, was used in the evaluation of the ideal binary-mask techniques [18]. Four types of noise were included: multi-talker babble, car, street, and suburban train noise. The noise signals were added to the speech signals at 0-10 dB SNR. The sentences were processed using the FFT applied to 20-ms Hanning-windowed frames, with 50% overlap between frames.

Performance was assessed using two probability values: probability of correct detection (P_D) (i.e., hit rate) and probability of false alarm (P_F). P_D measures the accuracy in classifying correctly target dominated T-F units, while the false-alarm measure (P_F) provides the probability that a masker-dominated T-F unit was wrongly classified as target-dominated T-F unit. Clearly, we would like P_F to be low (close to 0) and P_D to be high (close to 1). The plot of P_D vs. P_F provides the receiver operating characteristics (ROC) of the various binary-mask estimation techniques.

Two experiments were run to assess the influence of the

gain function ($G(k, m)$) in Eq. 1, and the influence of the algorithm used to update the noise spectrum $\hat{D}_k$ in Eq. 1. In these experiments, the binary mask was estimated by comparing the estimated SNR $\hat{\xi}$ (Eq. 1) to 1 (e.g., 0 dB). An additional experiment was run to assess the influence of alternative criteria (see list in Table 1) other than $\hat{\xi}$.

3.1. Effect of gain function

Figures 1-2 show the performance (hit rate vs. false alarm) of Eq. 1 in estimating the binary mask using six different gain functions. The noise-estimation algorithm proposed in [15] was used in this Experiment for updating the noise spectrum. Overall, performance with the statistical-model based algorithms (MMSE, MMSE-SPU, logMMSE, pMMSE) was better than performance with the Wiener and spectral subtractive algorithms. The difference in performance was more evident in terms of the false-alarm (P_F) measure. The intelligibility study in [4] showed that performance was affected the most by the false alarm rate rather than the value of the hit rate. That is, a lower false alarm was found to be perceptually more desirable than having a higher hit rate [4].

3.2. Effect of noise spectrum estimation

Fig. 3 shows the performance of Eq. 1 in estimating the binary mask using three different methods for updating the noise spectrum $\hat{D}_k$: a VAD algorithm [16], Martin's noise estimation algorithm [14] and the noise-estimation algorithm proposed in [15]. The MMSE gain function was used in this experiment. Overall, Martin's noise estimation algorithm performed the worst, while the other two algorithms performed better (and comparably well). This outcome may be attributed to the fact that Martin's algorithm is conservative and does not respond quickly to rapid changes in the SNR level.

3.3. Effect of different criteria

Fig. 4 shows the performance of five different criteria (Table 1) used for classifying target-dominated T-F units. Note that criterion C1 was used in the previous two experiments. The MMSE gain function was used in this experiment along with the noise-estimation algorithm proposed in [15]. Criterion C4 [17] performed consistently better than the other criteria. The proposed criterion C2 performed slightly better (lower false alarm rate) than criterion C1, particularly in babble. Criterion C3 performed the worst, but could potential yield better performance had a different threshold value was chosen [7]. The new criterion C5 performed as well or better than criterion C1.

4. CONCLUSIONS

This paper provided a comparison of several binary mask estimation techniques. All techniques were centered on different

C1	C2	C3	C4	C5
$\xi > 1$	$\xi > 1$ & $\gamma > 2$	$\gamma > 1$	$\hat{S}_k > 0.707 Y_k$, where $\hat{S}_k = \max(0, Y_k - \hat{D}_k)$	$p(H_1|Y) > 0.9$

**Table 1**. Five different criteria for classifying target-dominated T-F units.

ways to estimate the instantaneous SNR. The effect of various gain functions and noise estimation algorithms on estimating the SNR was assessed. The potential of using different criteria for classifying target-dominated T-F units was also investigated. Performance of binary mask estimation algorithms was evaluated in terms of hit rate and false alarm. Results indicated that the use of different SNR estimation techniques affects primarily the false alarm rate. Performance generally improves (lower false alarm rate) as the SNR level increases (from 0 to 10 dB).

5. REFERENCES

[1] D. Wang, and G. Brown, *"Computational Auditory Analysis"* John Wiley & Sons, 2006.

[2] N. Roman and D. Wang, "Pitch-based monaural segregation of reverberant speech," *The Journal of the Acoustical Society of America*, vol. 120, pp. 458–469, 2006.

[3] D. Brungart, P. Chang, B. Simpson, and D. Wang, "Isolating the energetic component of speech-on-speech masking with ideal time-frequency segregation," *The Journal of the Acoustical Society of America*, vol. 120, pp. 4007–4018, 2006.

[4] N. Li and P. C. Loizou, "Factors influencing intelligibility of ideal binary-masked speech: Implications for noise reduction," *The Journal of the Acoustical Society of America*, vol. 123, pp. 1673–1682, 2008.

[5] M. Seltzer, B. Raj and R. Stern, "A Bayesian classifier for spectrographhic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, pp. 379–393, 2004.

[6] N. Roman, D. Wang and G. Brown, "Speech segregation based on sound localization" *J. Acoust. Soc. Am.*, vol. 114, pp. 2236–2252, 2003.

[7] P. Renevey and A. Drygajlo, "Detection of reliable features for speech recognition in noisy conditions using a statistical criterion," in *Proc. Consistent and Reliable Acoustic Cues for Sound Analysis Workshop*, 2001, pp. 71–74.

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

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

[10] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-33, pp. 443–445, 1985.

[11] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Proc.*, pp. 857–869, Sept. 2005.

[12] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. ASSP-27, pp. 113–120, 1979.

[13] P. C. Loizou, *Speech enhancement: Theory and Practice*, CRC Press, 2007.

[14] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Proc.*, vol. 9, pp. 504–512, July 2001.

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

[16] J. Sohn, N. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, pp. 1–3, Jan. 1999.

[17] A. Vizinho, P. Green, M. Cooke, and L. Josifovski, "Missing data theory, spectral subtraction and signal-tonoise estimation for robust asr: an integrated study," in *Eurospeech*, 1999, pp. 2407–2410.

[18] Y. Hu and P. C. Loizou, "Subjective evaluations and comparisons of speech enhancement methods," *Speech Communication, Special issue on Speech Enhancement*, vol. 49, pp. 588–601, 2007.
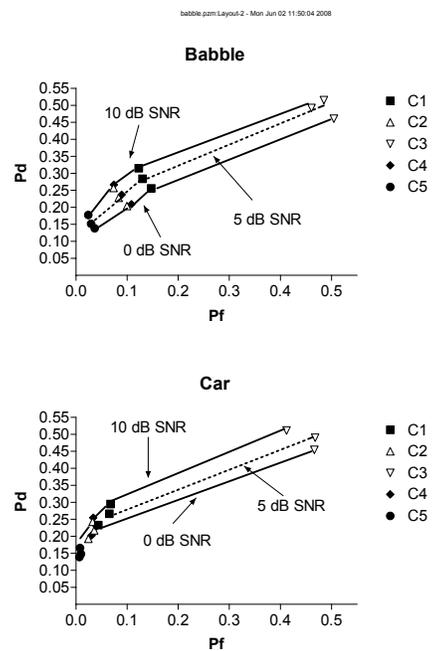
Babble

Babble

Train

Car
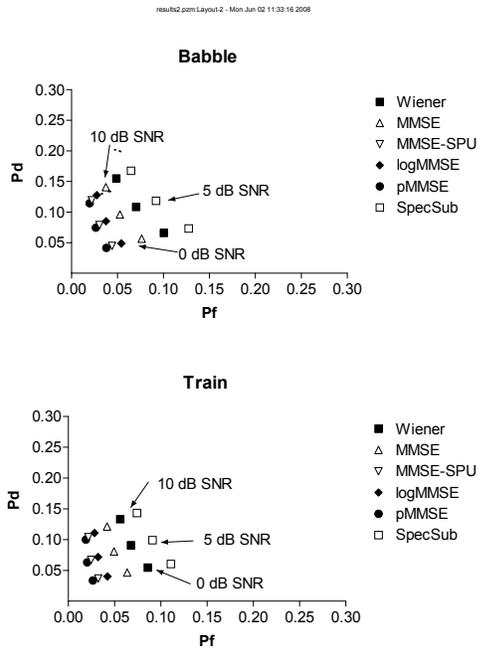
Fig. 1. Performance of binary mask estimation techniques for six different gain functions. Speech was corrupted by babble and train noise at 0 dB, 5 dB and 10 dB.
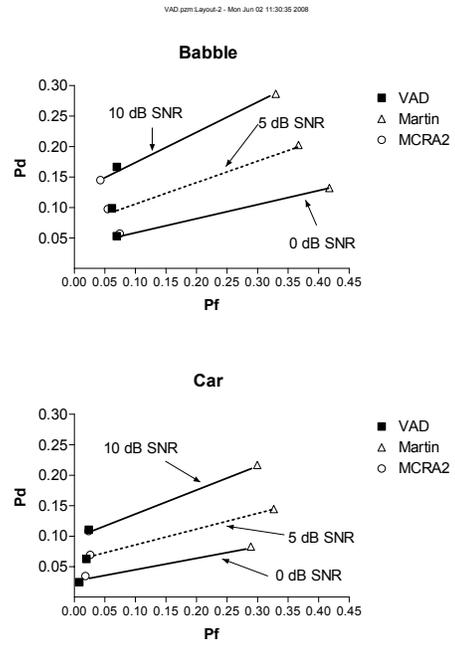
Fig. 3. Performance of binary mask estimation techniques for three different noise estimation algorithms. Speech was corrupted by babble and car noise at 0 dB, 5 dB and 10 dB.

Car

Babble

Street

Car
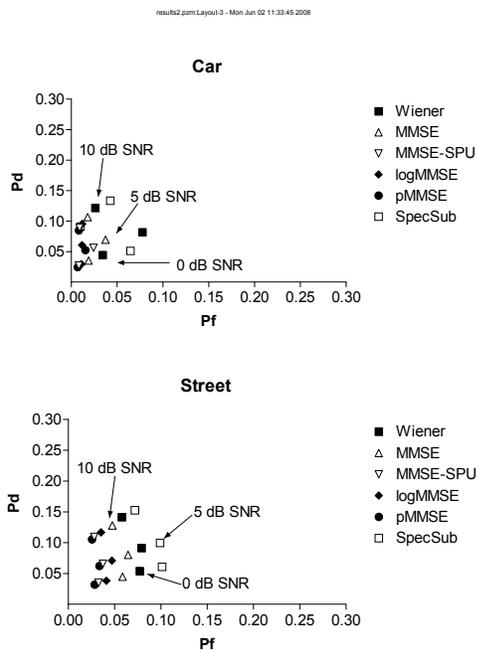
Fig. 2. Performance of binary mask estimation techniques for six different gain functions. Speech was corrupted by car and street noise at 0 dB, 5 dB and 10 dB.

Fig. 4. Performance of binary mask estimation techniques for five different criteria (Table 1). Speech was corrupted by babble and car noise at 0 dB, 5 dB and 10 dB.