A data-driven approach for estimating the time-frequency binary mask

Gibak Kim and *Philipos C. Loizou*

Department of Electrical Engineering, University of Texas at Dallas

imkgb27@gmail.com, loizou@utdallas.edu

10.21437/Interspeech.2009-257

Abstract

The ideal binary mask, often used in robust speech recognition applications, requires an estimate of the local SNR in each time-frequency (T-F) unit. A data-driven approach is proposed for estimating the instantaneous SNR of each T-F unit. By assuming that the *a priori* SNR and *a posteriori* SNR are uniformly distributed within a small region, the instantaneous SNR is estimated by minimizing the localized Bayes risk. The binary mask estimator derived by the proposed approach is evaluated in terms of hit and false alarm rates. Compared to the binary mask estimator that uses the decision-directed approach to compute the SNR, the proposed data-driven approach yielded substantial improvements (up to 40%) in classification performance, when assessed in terms of a sensitivity metric which is based on the difference between the hit and false alarm rates.

Index Terms: ideal binary mask, SNR estimation, Bayes risk

1. Introduction

The ideal binary mask is a technique explored in computational auditory scene analysis (CASA) and has also been applied to automatic speech recognition in noisy environments [1]. Recent studies have reported large gains in speech intelligibility using the ideal binary mask technique [2, 3]. The ideal binary mask is designed to retain the time-frequency (T-F) regions of the target signal that are stronger (i.e., SNR>0 dB) than the interfering noise (masker), and removes the regions that are weaker than the interfering noise (i.e., SNR$\leq$ 0 dB). Therefore, the binary mask assumes the value of 1 if the local SNR of the T-F unit is greater than a fixed threshold (e.g., 0 dB), and assumes the value of 0 otherwise. Fig. 1 shows as an example the synthesized signal obtained by applying the binary mask to a noise-masked sentence.

One way to estimate the binary mask is to compute the local SNR of the T-F unit and compare it against a threshold. A comparison of binary mask estimation techniques was provided in [4], where the local SNR of T-F unit was computed using the decision-directed approach [5] and various statistical estimators of the magnitude spectrum. Results in [4] indicated that the decision-directed approach performs poorly in terms of estimating the binary mask. In this paper, we aim to improve the performance of binary mask estimation by using data-driven optimization techniques to derive the instantaneous SNR. In [6, 7], data-driven methods were applied to estimate the magnitude spectrum of the clean signal by minimizing a certain spectral distortion measure. It was assumed that the gain function (applied to the noisy speech spectrum) was a function of two parameters, the *a priori* and *a posteriori* SNRs, and derived the gain function by minimizing various spectral distortion measures.

Rather than focusing to estimate the clean signal magnitude spectrum (as done in [6, 7]) from the noisy observations, we focus in the present paper on the estimation of the instantaneous SNR using a data-driven approach. Within a small region of the *a priori* SNR and *a posteriori* SNR values, we define a localized Bayes risk by choosing an appropriate cost function. This Bayes risk is minimized based on a data-driven approach that uses sentences corrupted by white noise as training data. The performance of the SNR estimator, in terms of classification accuracy, is compared against that obtained using the conventional decision-directed approach.

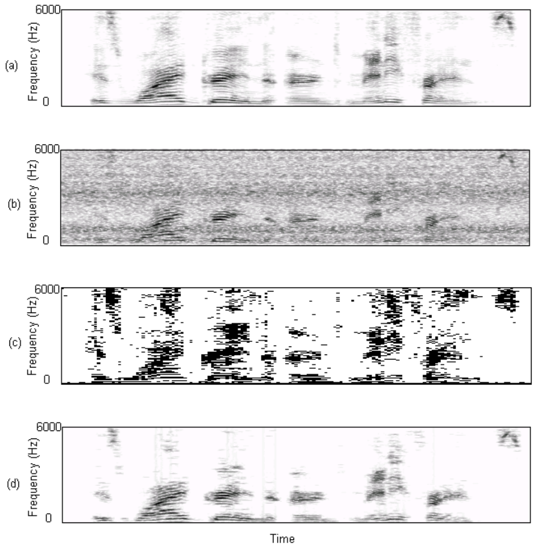

Figure 1: (a) Wide-band spectrogram of an IEEE sentence in quiet. (b) Spectrogram of corrupted sentence by speech-shaped noise at 0 dB SNR. (c) (Oracle) ideal binary mask, with black pixels indicating target-dominated T-F units and white pixels indicating masker-dominated T-F units. (d) Synthesized signal obtained by multiplying the binary mask shown in panel (c) with the corrupted signal shown in panel (b).

2. Binary mask estimation

2.1. Decision-directed method

The decision-directed method is a popular method used in the statistical-based speech enhancement algorithms for estimating the *a priori* SNR [5], which is defined as the ratio of speech spectral variance to noise spectral variance, i.e., ($\xi_k = \lambda_{x,k}/\lambda_{d,k}$). The decision-directed estimator of the *a priori*

SNR is given by:

$$\hat{\xi}_k(m) = \alpha \frac{\hat{A}_k^2(m-1)}{\lambda_{d,k}(m-1)} + (1-\alpha)\max\left[\gamma_k(m)-1, 0\right] \quad (1)$$

where $\hat{A}_k(m-1)$ is the estimate of the magnitude spectrum at time $(m-1)$ and frequency bin k, α is a smoothing constant (α= 0.98) and γ_k is the *a posteriori* SNR defined as the ratio of noisy speech spectral amplitude (R_k) to the estimated noise variance ($\gamma_k = R_k/\hat{\lambda}_{d,k}$). The estimate of the speech spectral amplitude $\hat{A}_k$ is obtained by applying a gain function $G(\cdot)$ to R_k as follows:

$$\hat{A}_k(m) = G\left(\hat{\xi}_k(m), \gamma_k(m)\right) \cdot R_k(m). \quad (2)$$

where the gain function $G(\cdot)$ is usually a non-linear function of *a priori* SNR $\hat{\xi}_k(m)$ and *a posteriori* SNR $\gamma_k(m)$.

2.2. Localized Bayes risk minimization

Previously, data-driven methods were used to derive the gain function in Eq. 2 needed for estimating the clean speech spectral amplitude [6, 7]. While Fingscheidt *et al.* [7] have used a large corpus of clean speech and noise data to train frequency-dependent gain functions for a specific noise environment, Erkelens *et al.* [6] have proposed frequency-independent and environment-independent gain functions trained with white noise corrupted data at a wide range of SNR levels. In both approaches, the gain functions were expressed as a function of the *a priori* and *a posteriori* SNRs and derived by minimizing a spectral distortion measure. The data-driven gain functions were stored in look-up tables indexed by the *a posteriori* and *a priori* SNRs, and retrieved to estimate the spectral amplitude of clean speech.

We propose to estimate the instantaneous SNR (ζ_k) based on information provided by the *a priori* SNR and *a posteriori* SNR values. We denote the estimator of the instantaneous SNR as $\hat{\zeta}_k$, and we let $\varepsilon = \zeta_k - \hat{\zeta}_k$ denote the error of the instantaneous SNR estimator. For a given cost function $\mathcal{C}(\varepsilon)$, the Bayes risk $\mathcal{R}$ is defined as [8]:

$$\begin{aligned} \mathcal{R} &= E[\mathcal{C}(\varepsilon)] & (3) \\ &= \int\int\int \mathcal{C}(\zeta_k - \hat{\zeta}_k) p(\hat{\xi}_k, \gamma_k, \zeta_k) d\hat{\xi}_k d\gamma_k d\zeta_k & (4) \\ &= \int\int\int \mathcal{C}(\zeta_k - \hat{\zeta}_k) p(\zeta_k|\hat{\xi}_k, \gamma_k) d\zeta_k p(\hat{\xi}_k, \gamma_k) d\hat{\xi}_k d\gamma_k & (5) \end{aligned}$$

where all the SNRs ($\zeta_k, \hat{\zeta}_k, \hat{\xi}_k, \gamma_k$) are represented in dB. We segment $\hat{\xi}_k$ and γ_k into a small region, and we assume that the joint density $p(\hat{\xi}_k, \gamma_k)$ is uniform within this small region. If $\hat{\xi}_k$ and γ_k fall into cell (i, j), we can write the localized Bayes risk as:

$$\mathcal{R}_{i,j} = \int \mathcal{C}(\zeta_k^{(i,j)} - \hat{\zeta}_k^{(i,j)}) p(\zeta_k^{(i,j)}|\hat{\xi}_k^{(i,j)}, \gamma_k^{(i,j)}) d\zeta_k^{(i,j)} \quad (6)$$

where $\zeta_k(i, j)$ denotes the SNR of region (i, j). The SNR estimator can be obtained by minimizing the localized Bayes risk ($\mathcal{R}_{i,j}$). The following cost function was chosen in our study:

$$\mathcal{C}(\zeta_k^{(i,j)} - \hat{\zeta}_k^{(i,j)}) = \left|\zeta_k^{(i,j)} - \hat{\zeta}_k^{(i,j)}\right|^{\frac{1}{\beta}} \quad (7)$$

where β is a compression coefficient of the error. If $\beta = 1/2$, the cost function is quadratic and the localized Bayes risk is the mean squared error. In this case, the mean of the posterior pdf is the MMSE estimator and is given by:

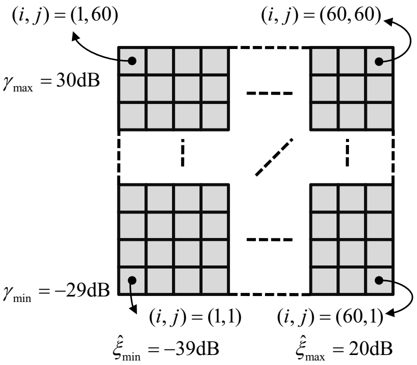

Figure 2: The segmentation of the *a priori* SNR and *a posteriori* SNR values into cells.

$$\hat{\zeta}_{k,MMSE}^{(i,j)} = \int \zeta_k^{(i,j)} p(\zeta_k^{(i,j)}|\hat{\xi}_k^{(i,j)}, \gamma_k^{(i,j)}) d\zeta_k^{(i,j)}. \quad (8)$$

In the case of $\beta = 1$, the cost function penalizes errors proportionally. As β increases, the cost function approaches a "hit-or-miss" cost function leading to a maximum *a posteriori* (MAP) estimator:

$$\hat{\zeta}_{k,MAP}^{(i,j)} = \arg\max_{\zeta_k^{(i,j)}} p(\zeta_k^{(i,j)}|\hat{\xi}_k^{(i,j)}, \gamma_k^{(i,j)}). \quad (9)$$

Without making any assumptions about the posterior pdf $p(\zeta_k^{(i,j)}|\hat{\xi}_k^{(i,j)}, \gamma_k^{(i,j)})$, we compute the localized Bayes risk using a data-driven approach as follows. First, we partition the $\hat{\xi}_k$ and γ_k values, ranging from -39 dB to 20 dB and -29 dB to 30 dB respectively, in steps of 1 dB. Thus, each ($\hat{\xi}_k$,γ_k) pair falls into one of 3600 cells (60 × 60) (see example in Fig. 2). As we are interested in deriving a frequency-independent SNR estimator, ($\hat{\xi}_k, \gamma_k$) pairs from different frequency bins can fall into the same cell. In the training stage, white noise is added to the clean signals at various SNR levels. The MMSE estimator [5] is used to obtain the magnitude spectrum, and from that, the SNR $\hat{\xi}_k$ is estimated using the decision-directed approach (Eq. 1). For each frame and each frequency bin, we have a ($\hat{\xi}_k^{(i,j)}, \gamma_k^{(i,j)}$) pair which falls into the cell (i, j). Concurrently, the true instantaneous SNRs ($\zeta_k^{(i,j)}$) are collected for the corresponding cell (i, j) from the whole training data set. Fig. 3 shows example posterior SNR distributions for different cells derived from the training data. We approximate the localized Bayes risk as follows:

$$\tilde{\mathcal{R}}_{i,j} = \frac{1}{M_{i,j}} \sum_{m=1}^{M_{i,j}} \left|\zeta^{(i,j)}(m) - \hat{\zeta}^{(i,j)}\right|^{\frac{1}{\beta}} \quad (10)$$

where $M_{i,j}$ is the number of data points falling into cell (i, j) during the training. In the data-driven approach, the expectation of the cost function with respect to the posterior pdf (Eq. 6) is approximated by the average of the cost function over the training data falling into cell (i, j). Note that the subscript 'k' was omitted in the above equation since a frequency-independent

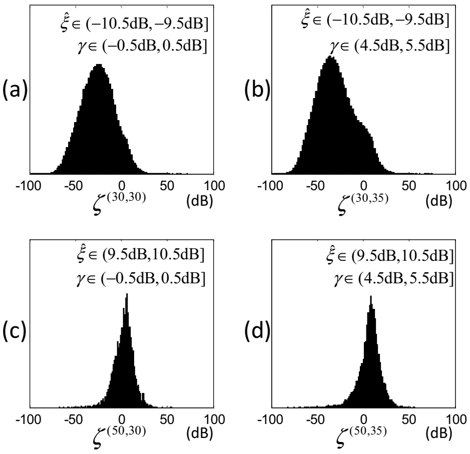

Figure 3: Example posterior SNR distributions $p(\zeta^{(i,j)}|\hat{\xi}^{(i,j)}, \gamma^{(i,j)})$ for different cells (i, j) derived from training data.

SNR estimator is pursued. Finally, a grid search is employed to find an SNR estimator $\hat{\zeta}^{(i,j)}$ that minimizes $\tilde{\mathcal{R}}_{i,j}$. The resulting SNR values $\hat{\zeta}^{(i,j)}$ are stored in a 60×60 matrix.

After the training stage, the following procedure is taken to estimate the binary mask. First, the *a priori* and *a posteriori* SNRs are estimated using the decision-directed approach (see Eq. 1) for each frequency bin. Then, the corresponding cell (i, j) indexed by the *a priori* and *a posteriori* SNR pairs is looked up to retrieve the corresponding SNR $\hat{\zeta}^{(i,j)}$. Finally, the time-frequency binary mask is estimated by comparing the estimated instantaneous SNR against a threshold. In our study, the SNR threshold was set to 0 dB. Note that during the training stage, some cells (i, j) might not be occupied frequently and $M_{i,j}$ may not be large enough to produce a reliable estimator. Hence, in those instances where the $M_{i,j}$ is too small (e.g., < 2500), the decision-directed approach is used.

3. Experimental results

3.1. Materials and procedure

Sentences were taken from the IEEE database [9]. The IEEE sentences are phonetically balanced with relatively low word-context predictability. The sentences were produced by one male speaker in a sound-proof booth using Tucker Davis Technologies (TDT) recording equipment. The recordings are available from [10]. The speech material was originally recorded at a sampling rate of 25 kHz and downsampled to 12 kHz.

In the training stage, white noise was added to 390 sentences ($\sim$17 min) at overall SNRs ranging from -15 dB to +25 dB, in steps of 5 dB. For the test, three types of noise were used: babble, factory noise, speech-shaped noise. The babble was taken from the Auditec CD (St. Louis, MO) which was recorded by 20 talkers with equal number of male and female speakers. The factory noise was taken from the NOISEX database [11] and the speech-shaped noise was stationary having the same long-term spectrum as the speech signal collected in the IEEE corpus. For testing, 60 sentences (not used in the training stage) were used. Noise was added to the IEEE sentences at 0 dB SNR.

The sentences were segmented (Hanning window) into overlapping frames of 384 samples (32 ms) with 50% overlap. The noise estimation algorithm proposed in [12] was used for estimating the noise variance in Eq. 1.

3.2. Evaluation

To quantify the accuracy of the proposed binary mask estimation algorithm, we computed the hit (HIT) and false alarm rates (FA). HIT and FA rates were computed by comparing the estimated binary mask against the (oracle) ideal binary mask. The performance is tabulated in Table 1 as a function of β, the compression coefficient of the error in the cost function (Eq. 7). For comparative purposes, Table 1 also includes the results obtained using the decision-directed approach, which is denoted as "DD". When the error is compressed with higher order of compression coefficient β, both the HIT and FA rates increase. Perceptually, the two types of errors that can be introduced, namely miss (=1-HIT) and false alarm, are not equivalent [3]. This is so, because the false alarm errors will possibly introduce more noise distortion, as T-F units that would otherwise be zeroed out (presumably belonging to the masker) would now be retained. The miss errors will likely introduce speech distortion, as theses errors are responsible for zeroing out T-F units that are dominated by the target signal and should therefore be retained. To account for the combined effect of both errors (miss and false alarm), we propose the use of the difference metric, HIT-FA, for predicting the intelligibility of speech synthesized using the estimated binary masks. A modestly high correlation ($r = 0.80$) between this metric and speech intelligibility was found in our previous work [13]. The difference metric (HIT-FA) is tabulated in Table 1 and also plotted in Fig. 4.

Other noise estimation methods (e.g., MCRA [14]) could potentially be used for estimating the noise variance in Eq. 1. A similar pattern in performance was observed as shown in Fig. 4 when the MCRA method [14] was used. Lower performance (in terms of HIT-FA) was observed, however, when the MCRA method [14] was used.

4. Conclusions

We proposed a binary mask estimation technique which compares the estimate of the instantaneous SNR of each T-F unit against a predefined threshold. The instantaneous SNR was derived by minimizing the localized Bayes risk using a data-driven approach. The proposed binary mask estimator was evaluated in terms of hit and false alarm rates, which were computed by comparing the estimated binary mask against the (oracle) ideal binary mask. The experimental results showed that the proposed method yields substantial improvements in the difference metric, HIT-FA, relative to the conventional decision-directed method used for estimating the SNR.

5. Acknowledgements

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

6. References

[1] D. L. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Hoboken, NJ: Wiley & IEEE Press, 2006.

[2] D. Brungart, P. Chang, B. Simpson, and D. Wang, "Isolating

Table 1: Hit (HIT) and false alarm rates (FA) obtained using the decision-directed (DD) and proposed data-driven methods.

Masker	Performance	DD	β=1/2	β=1	β=2	β=3	β=4	β=5	β=100
Babble	HIT(%)	29.85	27.15	31.14	33.77	34.69	34.93	35.32	36.00
	FA(%)	14.19	10.84	13.72	15.78	16.51	16.72	17.03	17.59
	HIT-FA(%)	15.66	16.31	17.42	17.99	18.18	18.21	18.29	18.41
Factory	HIT(%)	25.49	23.89	28.04	30.96	31.90	32.16	32.62	33.31
	FA(%)	8.53	5.61	7.92	9.71	10.39	10.57	10.88	11.42
	HIT-FA(%)	16.96	18.28	20.12	21.25	21.51	21.59	21.74	21.89
Speech-shaped	HIT	23.78	23.58	27.88	30.99	31.92	32.18	32.68	33.35
	FA	2.00	0.92	1.53	2.15	2.38	2.44	2.59	2.78
	HIT-FA	21.78	22.66	26.35	28.84	29.54	29.74	30.09	30.57

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] Y. Hu and P. C. Loizou, "Techniques for estimating the ideal binary mask," *The 11th International Workshop on Acoustic Echo and Noise Control*, September 2008.

[5] 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.

[6] J. Erkelens, J. Jensen, and R. Heusdens, "A data-driven approach to optimizing spectral speech enhancement methods for various error criteria," *Speech Commun.*, vol. 49, pp. 530–541, 2007.

[7] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-optimized speech enhancement," *IEEE Trans. Audio Speech Language Process.*, vol. 16, no. 4, pp. 825–834, 2008.

[8] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1993.

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

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

[11] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, 1993.

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

[13] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *Submitted to J. Acoust. Soc. Am.*, 2008.

[14] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Letters*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

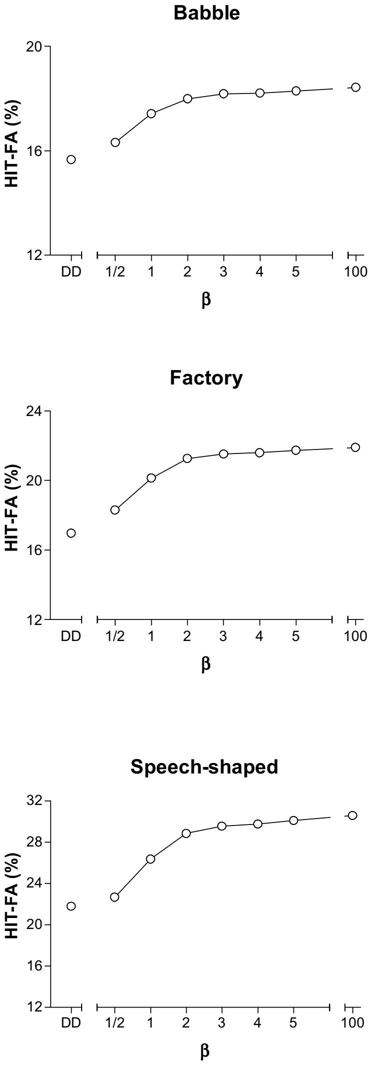

Figure 4: Performance, in terms of the difference between the hit and false alarm rates (HIT-FA), of the proposed method as a function of β. For comparison, the performance of the decision-directed method, indicated as "DD", is also shown.