A new noise-tracking algorithm for generalizing binary time-frequency (T-F) masking to ratio masking

Shan Liang, Wei Jiang, Wenju Liu
National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, 100190, China
{siang, wjiang, lwj}@nlpr.ia.ac.cn

Abstract

In this paper, we attempt to generalize the ideal binary mask (IBM) estimation to the ideal ratio mask (IRM) estimation. Under binary masking, the error in IBM estimation may greatly distort the original speech spectrum. The main purpose of this paper is using ratio mask to smooth this negative impact. Since the key issue is the noise tracking, we firstly use exponential distributions to model the distribution of noise power with binary mask and mixture power as condition. Then, we use a Gaussian distribution to model the correlation of noise estimation between adjacent T-F units. As the IBM of majority units can be estimated correctly, the correlation model could reduce the impact introduced by the error in IBM estimation. Systematic experiments show that our algorithm outperforms a common binary masking based method in terms of SNR gain and PESQ scores.

Index Terms: Ideal Binary Mask, Ideal Ratio Mask, Markov Chain Monte Carlo, Bayesian rule

1. Introduction

Recently computational auditory scene analysis (CASA) techniques which are inspired by the human auditory perception have received attention due to their advantage in dealing non-stationary noise [1-5]. In such techniques, binary time-frequency (T-F) masking had applied to extract target signal. The ideal binary mask (IBM) proposed by Hu and Wang in [1,4] is a binary matrix along time and frequency. Specifically, for each time-frequency (T-F) unit, if the power of speech is greater than the noise, the unit is labeled as 1 and otherwise it is labeled as 0.

The optimality of ideal binary masks in terms of signal-to-noise ratio (SNR) gain has been rigorously discussed by Li and Wang in [2]. They find that the IBM is optimal only while the T-F decomposition is orthogonal. That is, there is no overlap between the T-F representations of the target and noise. Obviously, this is an over-idealization assumption. According to Wiener filtering, ideal ratio mask (IRM) [2] which is a more general form of IBM may achieve higher SNR gains. Experimental results in [2] have proved this point. However, the improvement is quite small in ideal condition.

With the additive mixture spectrum, the key issue in IRM estimation is the noise power tracking [2]. As well known, it’s very difficult to track the non-stationary noise via common noise-tracking algorithms [6-7]. In contrast, the IBM estimation requires only binary decisions. Recently, statistical learning methods have been used in such a binary-classification task, including Batesian classifier [3] and Multilayer Perceptron (MLP)[5]. One main advantage of these methods is that they could well handle highly non-stationary noise. Therefore, relatively simple computational cost and the excellent capacity of generalization to non-stationary noise are another reasons for the existing CASA techniques [3-5] to prefer the IBM over IRM as the computational goal.

However, it’s almost impossible to estimate the IBM with one-hundred-percent accuracy. Both of the Miss and False Alarm (FA) error rates defined in [3] may greatly distort the original speech spectrum with binary mask based resynthesized strategy. In this paper, we firstly track the noise power with the IBM estimation and then use ratio mask to smooth the negative impact introduced by the error in IBM estimation. We should point out that the proposed algorithm just generalize the IBM estimation to some extent. Therefore, it inherits the good capacity to handle non-stationary noise. Systematic experiments show that the ratio mask estimated by the proposed algorithm could improve the original IBM estimation obtained by method [3] effectively in terms of SNR gain and PESQ scores. Indeed, evaluation on noise-tracking shows that our algorithm could estimate the noise power with relatively high accuracy over a well-known speech-presence probability based method [6].

The paper is organized as follows. In the next section, we present an overview of the proposed method. The details are discussed in section 3. We evaluate the proposed algorithm in section 4 and compare with the methods proposed in [3,6]. The last section gives some conclusions.

2. Framework Overview

Let \( X(t, c) \), \( S(t, c) \) and \( N(t, c) \) denote the power of noisy speech, speech and noise at frequency index \( c \) and time frame \( t \) respectively. Since the speech and noise signals are statistically independent, the additive-noise spectrum can be modeled as:

\[
X(t, c) \approx S(t, c) + N(t, c). \tag{1}
\]

As in [2], the IRM is defined as:

\[
M_B(t, c) = \frac{S(t, c)}{S(t, c) + N(t, c)} \approx 1 - \frac{N(t, c)}{X(t, c)}. \tag{2}
\]

Let \( M_B(t, c) \) denotes the estimation of IBM. The key issue in generalizing the binary to ratio masks is the noise-tracking with \( M_B \) and \( X \) as observation. The overall Bayesian framework can be represented by the following equation:

\[
p(N|X, M_B) \propto p(N) \prod_{t,c} p(N(t,c)|X(t,c), M_B(t,c)). \tag{3}
\]

We use a conditional distribution labeled by \( A \) to depict the correlation between the noise power and binary mask for each unit.
independently. The correlation between the noise power within adjacent units is further taken into account in the prior model $p(N)$. In the Minimum Mean-Square Error (MMSE) sense, the noise power is estimated by the conditional expectation of Eq. (3). To avoid the extraordinarily complex computation introduced by direct integration, we use a Markov Chain Monte Carlo (MCMC) method [8] to approach the expectation. With the noise estimation, the IRM is then obtained by Eq.(2).

3. System Description

3.1. Conditional Distribution

From the definition of IRM, the noise power can be rewritten as:

$$N(t, c) = (1 - M_R(t, c))X(t, c) = \tilde{M}_R(t, c)X(t, c).$$ (4)

The probability density function can be rewritten as:

$$p(N(t, c)|X(t, c), M_B(t, c)) = p(\tilde{M}_R(t, c)|M_B(t, c)).$$ (5)

20 sentences degraded by 3 types of noise at -5.0 and 5 dB SNR are used to quantize the two distributions. The sentences are selected from TIMIT database [11] and the noise includes factory, babble [12] and cocktail party noise [4]. The histograms with 20 bins are shown in Fig.1.

![Figure 1: Histograms of the two condition distributions](image1)

As shown in Fig.1, the two histograms can be well approached by exponential distribution. That is,

$$p(\tilde{M}_R|M_B(t, c) = 1) = \begin{cases} \lambda_1 e^{-\lambda_1 \tilde{M}_R} & \text{if } \tilde{M}_R \in [0, 1] \\ 0 & \text{otherwise} \end{cases},$$ (6)

$$p(\tilde{M}_R|M_B(t, c) = 0) = \begin{cases} \lambda_2 e^{-\lambda_2 \tilde{M}_R} & \text{if } \tilde{M}_R \in [0, 1] \\ 0 & \text{otherwise} \end{cases},$$ (7)

where $\lambda_1$ and $\lambda_2$ generally refer to the rate parameters. According to Maximum Likelihood (ML) rule, the two parameters can be learned from corpus by $\lambda_1 = \frac{1}{\tilde{M}_R^L}$ and $\lambda_2 = \frac{\tilde{M}_R^L}{\tilde{M}_R^L}$. $\tilde{M}_R^L$ and $\tilde{M}_R^L$ denote the statistical average of $\tilde{M}_R$ and $M_R$ corresponding to the samples in training corpus. In this paper, the estimation of the two parameters are $\lambda_1 = 5.27$ and $\lambda_2 = 13.48$.

3.2. Noise Prior Model

In the Eq.(4), the noise is estimated for each T-F unit independently. It is almost impossible to obtain the IBM in separation stage. The error in IBM estimation may have a great impact on the noise estimation. One way to reduce this negative impact is to take the local correlation between the noise estimation of adjacent T-F units into account.

We compute the local noise level $\mu_N(t, c)$ by a smoothing stage as follows:

$$\mu_N(t, c) = \frac{1}{2L} \sum_{l=1}^{L} [N(t + l, c) + N(t - l, c)],$$ (8)

where $L$ defines the length of smoothing window. It is reasonable to assume that the true noise power is distributed at a narrow interval around this average level. So, we further assume the noise obey to a Gaussian prior distribution:

$$p(N(t, c)) = G(N(t, c); \mu_N(t, c), \beta \mu_N(t, c))$$ (9)

where $\beta$ is a positive constant.

In order to validate this assumption, we define a random variable,

$$\epsilon(t, c) = (N(t, c) - \mu_N(t, c))/\mu_N(t, c).$$ (10)

According to the assumption, $\epsilon(t, c)$ should obey to a Gaussian distribution with mean 0 and standard deviation $\beta$. With the test corpus mentioned in the previous subsection, the mean and $\beta$ are estimated as -0.02 and 0.39 by ML rule respectively. We plot the Gaussian probability density function (PDF) and the histogram with 20 bins in Fig. 2. We can find that the Gaussian PDF with estimated parameters approach to the histogram for majority part.

![Figure 2: Histogram of $\epsilon(t, c)$](image2)

A prior model is further proposed as follows:

$$p(N) \propto \prod_{t, c} p(N(t, c))$$ (11)

We should point out that the local correlation is taken into account to compute $\mu_N$. Therefore, $p(N(t, c))$ is statistic dependent with $p(N(t \pm l, c)), 1 \leq l \leq L$. That is, $p(N)$ is very high-dimensional function.

3.3. Expectation of Noise Power

It is computationally very difficult to obtain the expectation by direct integration of high-dimensional posterior distribution. Therefore, we use Markov Chain Monte Carlo (MCMC) method [8] which is one popular approach to simplify the computation complexity. The key idea of MCMC is to simulate direct draws
from a complex distribution by generating a sequence of candidates.

Firstly, the initial value \( \hat{N}^{(0)} \) is generated by a well-known noise-tracking method [6]. Let \( \hat{N}^{(t)} \) denote the candidates generated in \( t \)th iteration.

Then, we select a unit \( \mu(i,c) \) randomly. A random variable \( \delta \in [0, X(i,c)] \) subjecting to uniform distribution is generated as a new noise estimation in candidate \( \hat{N} \). Indeed, the estimation in the other units remains the same as that in previous iteration \( N^{(t)} \).

From the generation of new candidate, we can find the proposal distribution in MCMC is symmetric. Therefor, the Metropolis-acceptance ratio [8] can be computed as follows:

\[
r = \min \left\{ 1, \frac{p(\hat{N}|X,M_B)}{p(N^{(t)}|X,M_B)} \right\}
\]

Since the estimation of the units except \( \mu(i,c) \) remains the same, the acceptance ratio is simplified as follows:

\[
p(\hat{N}|X,M_B) \sim \frac{p(\hat{N}(t,c)|X(t,c),M_B(t,c))p(\hat{N}(t,c))}{p(N^{(t)}(t,c)|X(t,c),M_B(t,c))p(N^{(t)}(t,c))}
\]

Then, a random variable \( \theta \in [0,1] \) subjecting to uniform distribution is generated. If \( \theta < r \), \( \hat{N} \) is retained as the new candidate \( N^{(t+1)} \). Otherwise, the \( N^{(t+1)} \) is equal to \( N^{(t)} \).

The sequence of candidates will converge to true posterior distribution given in Eq.(3) after several iterations of above process. Finally, the expectation of noise power \( N_{est} \) is obtained by the statistical average of the candidates.

### 4. Experimental Results

The proposed algorithm is tested with 30 sentences taken from TIMIT database [11]. All the sentences are down-sampled to 16 kHz before noises are added. Three types of noise are used, namely babble, factory [12] and cocktail party noise [4]. The noisy speech signals are generated with SNR in the range of -5 to 5 dB with 5-dB steps. We firstly decompose the input signal into frequency domain using 64-channel gammatone filters from 50 Hz to 8000 Hz [9]. Then, the response is divided into 20-ms time frames with 10-ms overlapping for each frequency channel. It produces a new T-F representation, cochleagram, which is similar to the spectrogram obtained by short-time Fourier transform (STFT).

We use the Bayesian classifier proposed in [3] to estimate the IBM. A set of auditory features are extracted to exploit the degree of corruption in each T-F unit, including amplitude modulation spectrum (AMS) [3] and pitch related features [5]. The pitch counters are estimated from noisy speech by RAPT [10]. The distribution of the features corresponding to each class is represented with a Gaussian Mixture Model (GMM). Another 20 sentences are used to train a 32-mixture Gaussian for each class and each channel. The IBM is estimated by maximum a posteriori (MAP) rule. The average accuracy in the test corpus is 73.9\%. The average Miss and FA error rates are 39.3\% and 19.4\% respectively.

All the parameters discussed in Sec.3 are estimated under IBM condition. Appreciate adjustments to the parameters to improve the robustness to the error in IBM estimation. The parameters are set as: \( \lambda_1 = 4, \lambda_2 = 10, \beta = 1, L = 3 \). The maximum number of MCMC iterations is set as 100,000.

Since the key issue in IRM estimation is the noise tracking, we compare the our method with a well known noise tracking method [6] firstly. Fig.3 shows the noise-tracking results generated by our algorithm and the method [6] for a speech degraded by factory noise at 0 dB SNR. Although the true noise power varies very fast along time, the proposed method could approximate to the true value for majority part. In contrast, the method [6] could only track the slow-varying components of original noise.

To systematically evaluate the performance, we use a symmetric segmental logarithmic estimation error as an objective measure [7]. It is defined as:

\[
\text{LogErr} = \frac{1}{TC} \sum_{t,c} \left| 10 \log \frac{N(t,c)}{N_{est}(t,c)} \right|
\]

where \( N(t,c) \) and \( N_{est}(t,c) \) denote the true and estimated noise power respectively. \( T \) and \( C \) represent the number of time frames and frequency channels.

As shown in Fig.4, our algorithm achieves relatively low LogErr for all conditions. The average LogErr of the proposed
Table 1: Average SNR gain (dB) with respect to different SNR levels of input mixtures. EBM: the binary mask estimated by method [3], NT: the ratio mask estimated by the noise-tracking method [6], pro: the proposed algorithm.

<table>
<thead>
<tr>
<th>Environments</th>
<th>SNR (dB)</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise</td>
<td>EBM</td>
<td>NT</td>
</tr>
<tr>
<td>Babble</td>
<td>-5</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>5.03</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>8.72</td>
</tr>
<tr>
<td>Factory</td>
<td>-5</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>5.04</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>8.43</td>
</tr>
<tr>
<td>Cocktail</td>
<td>-5</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>4.96</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>8.62</td>
</tr>
<tr>
<td>Avg.</td>
<td></td>
<td>4.65</td>
</tr>
</tbody>
</table>

Table 2: Average PESQ scores with respect to different SNR levels of input mixtures. EBM: the binary mask estimated by method [3], NT: the ratio mask estimated by the noise-tracking method [6], pro: the proposed algorithm.

<table>
<thead>
<tr>
<th>Environments</th>
<th>SNR (dB)</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise</td>
<td>EBM</td>
<td>NT</td>
</tr>
<tr>
<td>Babble</td>
<td>-5</td>
<td>1.28</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1.47</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1.77</td>
</tr>
<tr>
<td>Factory</td>
<td>-5</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1.30</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1.63</td>
</tr>
<tr>
<td>Cocktail</td>
<td>-5</td>
<td>1.10</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1.40</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1.73</td>
</tr>
<tr>
<td>Avg.</td>
<td></td>
<td>1.41</td>
</tr>
</tbody>
</table>

For more complete comparison, the proposed algorithm is also evaluated by the widely used ITU-PESQ scores [13] which is highly correlated with the subjective scores. It ranges from -0.5–4.5. The higher the score, the better the perceptual quality. As shown in Table 2, the proposed algorithm achieves relatively high score for all conditions.

5. Conclusions

In this paper, we propose a new algorithm that generalizes the IBM estimation to IRM estimation. The improvements on SNR gain and PESQ scores show the effectiveness of the proposed algorithm for reducing the distortion introduced by the Miss and FA rates in IBM estimation. Indeed, systematic evaluations also show that the proposed algorithm achieves higher accuracy in noise-tracking over a well known method [6].

6. Acknowledgements

This research was supported in part by the China National Nature Science Foundation (No.91120303, No.90820011 and No.90820303), and the National Grand Fundamental Research 973 Program of China (No. 2004CB318105).

7. References