The optimal ratio time-frequency mask for speech separation in terms of the signal-to-noise ratio

Shan Liang, Wenju Liu,[a] Wei Jiang, and Wei Xue
National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, People's Republic of China
sliang@nlpr.ia.ac.cn, lwj@nlpr.ia.ac.cn, wjiang@nlpr.ia.ac.cn, wxue@nlpr.ia.ac.cn

Abstract: In this paper, a computational goal for a monaural speech separation system is proposed. Since this goal is derived by maximizing the signal-to-noise ratio (SNR), it is called the optimal ratio mask (ORM). Under the approximate W-Disjoint Orthogonality assumption which almost always holds due to the sparse nature of speech, theoretical analysis shows that the ORM can improve the SNR about $10\log_{10}2$ dB over the ideal ratio mask. With three kinds of real-world interference, the speech separation results of SNR gain and objective quality evaluation demonstrate the correctness of the theoretical analysis, and imply that the ORM achieves a better separation performance.

PACS numbers: 43.72.Dv, 43.60.Ac, 43.60.Cg [DOS]
Date Received: July 23, 2013 **Date Accepted:** September 25, 2013

1. Introduction

Speech separation or enhancement is one of the key problems in speech processing. Available approaches to this problem include computational auditory scene analysis (CASA),[1,2] noise tracking,[3,4] blind speech separation (BSS),[5,6] and so on. Since speech signals are non-stationary, all these algorithms firstly decompose the time-domain signal into time-frequency (T-F) domain by discrete short-time Fourier transform (DSTFT) or auditory filtering.[7] Then, a binary mask or ratio mask is estimated and further used to estimate the spectrum or cochleagram of the original speech. Finally, the time-domain speech is synthesized by inverse discrete short-time Fourier transform (IDSTFT) or the method described in Ref. 8. All these algorithms can be integrated into a framework which is regarded as a linear mask model in this paper.

Many CASA based separation systems[1,2] and BSS algorithms[5,6] set the ideal binary mask (IBM) as the computational goal, while many noise tracking based enhancement methods[3,4] use the Wiener-type ratio mask. The Wiener-type ratio mask, which is called the ideal ratio mask (IRM),[9] is motivated by the frequency response of the Wiener filter,[10] which achieves the optimal signal-to-noise ratio (SNR) gain for stationary signals. However, the speech signals and many real-world noises are non-stationary. The optimality of the IRM in terms of the SNR has not been rigorously addressed for non-stationary signals.

In this paper, a new ratio mask which achieves the optimal SNR gain over all the ratio masks is derived analytically. We further find that the IRM is a simplification of the derived optimal ratio mask (ORM). Furthermore, we theoretically analyze the SNR improvement from the IRM to the ORM. The expectation of the improvement is deduced under the approximate W-Disjoint Orthogonality (AWDO) assumption. Finally, we verify the AWDO assumption with experiments on three kinds of real-world noise. The separation results show that the average improvement is very close to the derived expectation.

[a]Author to whom correspondence should be addressed.

The rest of the paper is organized as follows. In Sec. 2, we present some necessary notations and definitions. The SNR gains of the ORM and IRM are discussed in Sec. 3. Speech separation experiments are further used to verify the discussions in Sec. 4. Section 5 gives some conclusions.

2. Notation and definition

Suppose that the T-length speech and interference signals in time domain are denoted by $x(t)$ and $n(t)$, respectively. The additive mixture $y(t)$ is given by

$$y(t) = x(t) + n(t). \tag{1}$$

The speech separation problem aims to estimate speech signal $x(t)$ from the mixture $y(t)$ as accurately as possible. Let $S_\cdot(\tau, f)$ denote the DSTFT of a signal in τth time frame and fth frequency index. The power spectrum density (PSD) is $P_\cdot(\tau, f) = |S_\cdot(\tau, f)|^2$. With a real and symmetric window function $g(t) = g(-t)$, we take the DSTFT of $y(t)$ for example:

$$S_y(\tau, f) = \sum_{t=0}^{T-1} y(t) g(t-\tau) \exp\left(\frac{-i2\pi ft}{T}\right). \tag{2}$$

Since DSTFT is complete and stable, $x(t)$ can be reconstructed from $S_x(\tau, f)$ by IDSTFT.[11] In other words, the speech separation problem can be transformed into the problem of $S_x(\tau, f)$ estimation. Letting $\hat{S}_x(\tau, f)$ denote an estimation of $S_x(\tau, f)$, the time domain signal $\hat{x}(t)$ is

$$\hat{x}(t) = \frac{1}{T} \sum_{\tau=0}^{T-1} g(t-\tau) \sum_{f=0}^{T-1} \hat{S}_x(\tau, f) \exp\left(\frac{i2\pi ft}{T}\right). \tag{3}$$

The mean square error (MSE) is defined as

$$L(\hat{x}, x) = \sum_{t=0}^{T-1} [\hat{x}(t) - x(t)]^2 = \sum_{t=0}^{T-1} r(t)^2, \tag{4}$$

where $r(t) = \hat{x}(t) - x(t)$. According to Parseval's equality,[11] the MSE is given by

$$\sum_{t=0}^{T-1} r(t)^2 = \frac{1}{T} \sum_{\tau=0}^{T-1} \sum_{f=0}^{T-1} |\hat{S}_x(\tau, f) - S_x(\tau, f)|^2. \tag{5}$$

The SNR is defined as follows:

$$\mathrm{SNR} = 10 \log_{10} \left(\frac{\sum_t x(t)^2}{\sum_t r(t)^2} \right). \tag{6}$$

Since the total energy of target speech, $\sum_t x(t)^2$, is a constant, the SNR gain is inversely proportional to the MSE. This means that maximizing the SNR gain is equivalent to minimizing the MSE.

3. Linear mask model

To simplify the representation, one-dimension coordinate notation, $k \equiv (\tau, f)$, is used in the following discussions. Let $P_x(k)$ and $P_n(k)$ denote the PSDs of the target speech and interference at the kth T-F unit, respectively. In this paper, we are not concerned with how to estimate $P_x(k)$ and $P_n(k)$, which is the task of noise tracking.[4] The main

purpose is discussing the optimal computational goal and proving the importance of taking into account the phase information when filtering signals with $P_x(k)$ and $P_n(k)$ as the prior information.

In CASA techniques[1,2] and BSS algorithm,[5,6] the computational goal is the IBM estimation which is defined as follows:

$$B(k) = \begin{cases} 1, & \text{if } P_x(k) > P_n(k) \\ 0, & \text{else.} \end{cases} \tag{7}$$

The units labeled by 1 are called reliable or speech dominated units, while the others are called unreliable or noise dominated units. The work of Li and Wang[12] proves that the IBM achieves the minimum MSE over all binary masks if the T-F decomposition is orthonormal. The estimation of $S_x(k)$ is given by $\hat{S}_x(k) = B(k)S_y(k)$.

The IBM is a simplification form of the IRM which is defined in[9]

$$R(k) = \frac{P_x(k)}{P_x(k) + P_n(k)}. \tag{8}$$

The IRM is the instantaneous approximation of the Wiener filter[10] which is the optimal filtering model for stationary signals in terms of the MSE. Similarly, the estimation of $S_x(k)$ is given by $\hat{S}_x(k) = R(k)S_y(k)$. Many speech enhancement systems, such as the Wiener-type enhancement algorithms,[3,4] are based on ratio mask strategy.

These algorithms can be integrated to one framework, $\hat{S}_x(k) = \gamma(k)S_y(k)$, where $\gamma(k)$ is a real value. In this paper, this framework is regarded as a linear mask model.

3.1 The ORM

The MSE corresponding to the linear mask model is given as follows:

$$\begin{aligned} L(\hat{x}, x) &= \frac{1}{T}\sum_k |\gamma(k)S_y(k) - S_x(k)|^2 = \frac{1}{T}\sum_k |(\gamma(k)-1)S_x(k) + \gamma(k)S_n(k)|^2 \\ &= \frac{1}{T}\left(\sum_k [(\gamma(k)-1)^2 P_x(k) + \gamma(k)^2 P_n(k)] + \sum_k 2\gamma(k)(\gamma(k)-1)\Re(S_x(k)S_n^*(k))\right), \end{aligned} \tag{9}$$

where superscript "*" denotes the conjugate operator and $\Re(\cdot)$ returns the real component of a complex number.

Minimize the MSE by the partial derivative of $\gamma(k)$ as follows:

$$\frac{\partial L(\hat{x}, x)}{\partial \gamma(k)} = \frac{2}{T}[[\gamma(k)-1]P_x(k) + \gamma(k)P_n(k) + [2\gamma(k)-1]\Re(S_x(k)S_n^*(k))]. \tag{10}$$

The ORM, $\gamma_{\text{opt}}(k)$, is obtained when $\partial L(\hat{x}, x)/\partial\gamma(k) = 0$:

$$\gamma_{\text{opt}}(k) = \frac{P_x(k) + \Re(S_x(k)S_n^*(k))}{P_x(k) + P_n(k) + 2\Re(S_x(k)S_n^*(k))}. \tag{11}$$

By comparing Eqs. (8) and (11), we can find that the IRM is a simplification of the ORM. Furthermore, the IRM is equivalent to the ORM if $\Re(S_x(k)S_n^*(k))$ is equal to 0 for all units. $\Psi(k) \equiv S_x(k)S_n^*(k)$ is further used to simplify the representation.

Furthermore, we further find that the ORM can be estimated without phase information. Since the following equation always holds:

$$P_y(k) = P_x(k) + P_n(k) + 2\Re(\Psi(k)) \Rightarrow 2\Re(\Psi(k)) = P_y(k) - [P_x(k) + P_n(k)], \tag{12}$$

the ORM given in Eq. (11), $\gamma_{\text{opt}}(k)$, can be re-written as

$$\gamma_{\text{opt}}(k) = \frac{P_y(k) + P_x(k) - P_n(k)}{2P_y(k)}. \tag{13}$$

The two versions of the ORM defined by Eqs. (11) and (13) achieve identical separation results.

The MSE corresponding to $\gamma_{\text{opt}}(k)$ is given by

$$L(\hat{x}, x) = \frac{1}{T}\sum_k |\gamma_{\text{opt}}(k)(S_x(k) + S_n(k)) - S_x(k)|^2 = \frac{1}{T}\sum_k \frac{P_x(k)P_n(k) - [\Re(\Psi(k))]^2}{P_x(k) + P_n(k) + 2\Re(\Psi(k))}. \tag{14}$$

For comparison, we also give the MSE corresponding to the IRM as follows:

$$\begin{aligned} L(\tilde{x}, x) &= \frac{1}{T}\sum_k |R(k)(S_x(k) + S_n(k)) - S_x(k)|^2 \\ &= \frac{1}{T}\sum_k \frac{P_x(k)P_n(k)[P_x(k) + P_n(k) - 2\Re(\Psi(k))]}{[P_x(k) + P_n(k)]^2}. \end{aligned} \tag{15}$$

3.2 Expectation of the SNR improvement

The SNR improvement from the IRM to the ORM is discussed under the AWDO assumption. The W-Disjoint Orthogonality (WDO) property is derived from the sparsity of speech signal in the T-F domain, where sparsity means that a small percentage of the T-F units contain a large percentage of the signal energy. The rigorous definition of WDO[5] is given by

$$S_x(k)S_n(k) = S_x(k)S_n^*(k) = 0, \quad \forall k. \tag{16}$$

Obviously, the WDO property is a mathematical idealization. In a general case, $|\Psi(k)|$ is very small with high probability.[5] A more rigorous statement is that $|\Psi(k)|$ is much smaller than $(P_x(k) + P_n(k))/2$ for most units. An experiment will be conducted to verify this property in Sec. 4. We regard this property as the AWDO property.

On one hand, $\Re(\Psi(k)) \le |\Psi(k)|$. On the other hand, according to the AWDO assumption $2|\Psi(k)|$ is much smaller than $P_x(k) + P_n(k)$ for most units. Therefore, the MSE of the IRM can be further approximated by

$$L(\tilde{x}, x) \approx \frac{1}{T}\sum_k \frac{P_x(k)P_n(k)}{P_x(k) + P_n(k)}. \tag{17}$$

Similarly, the MSE of the ORM can be further approximated by the following equation:

$$L(\hat{x}, x) \approx \frac{1}{T}\sum_k \frac{P_x(k)P_n(k) - [\Re(\Psi(k))]^2}{P_x(k) + P_n(k)} = \frac{1}{T}\sum_k \frac{P_x(k)P_n(k)\sin^2(\theta(k))}{P_x(k) + P_n(k)}, \tag{18}$$

where $\Psi(k) = \sqrt{P_x(k)P_n(k)}\exp(i\theta(k))$ and $\theta(k) = \angle(S_x(k)) + \angle(S_n^*(k))$. Operator $\angle(\cdot)$ returns the angle of a complex number. Since we have no prior information on $\theta(k)$, it is assumed to be uniformly distributed in the interval $[0, 2\pi]$. Since $\int_0^{2\pi} \sin^2(\theta(k))d\theta(k) = \pi$, the expectation of $L(\hat{x}, x)$ is given by

$$\mathrm{E}[L(\hat{x},\, x)]_{\theta(k)} \approx \frac{1}{T}\sum_{k}\frac{P_x(k)P_n(k)}{P_x(k)+P_n(k)}\int_0^{2\pi}\sin^2(\theta(k))\frac{1}{2\pi}d\theta(k) = \frac{1}{2T}\sum_{k}\frac{P_x(k)P_n(k)}{P_x(k)+P_n(k)}. \tag{19}$$

In other words, the expectation of $L(\hat{x},\, x)$ approximates a half of $L(\tilde{x},\, x)$.

From the definition of the SNR given in Eq. (6), the SNR improvement from the IRM to the ORM is given by

$$\Delta\mathrm{SNR} = \mathrm{SNR_O} - \mathrm{SNR_I} = 10\log_{10}\left(\frac{L(\tilde{x},\, x)}{L(\hat{x},\, x)}\right), \tag{20}$$

where $\mathrm{SNR_O}$ and $\mathrm{SNR_I}$ correspond to the SNR gains of the ORM and the IRM, respectively. Therefore, the expectation of ΔSNR is about $10\log_{10}2$ dB.

4. Experimental results

To verify the above analysis experimentally, 40 s length speech signals are randomly taken from the training set provided by the Grid speech corpus.[13] Three different types of real world noise which are recorded in cafeteria, square, and subway environments[14] are selected as the interferences. For each type of noise, a 20 s length signal is selected. All the signals are down-sampled to 16 kHz. The target and interference signals are mixed with the input SNR ranging from −3 to 9 dB with 3 dB step. To compute the DSTFT, the noisy time domain signals are divided into frames of 512 samples with an overlap of 50%.

4.1 AWDO test

If both the target and interference signals are speech, the AWDO property has been verified.[5] In this section, we further validate this property while target speech signals mix with the real world interferences. The WDO degree is measured with the metric proposed in:[15]

$$\mathrm{WDOM} = \frac{\sum_k |S_x(k)S_n(k)|}{\sqrt{\sum_k P_x(k)\sum_k P_n(k)}}. \tag{21}$$

A lower WDOM value indicates a higher AWDO degree.

A 20 s length speech signal is assumed to be the target, while the other 20 s length signal is assumed to be the interference. For comparison, we also compute the WDOM corresponding to the real world interferences. The results given in Table 1 show that the WDOMs of the cafeteria noise and the speech signal are very close to each other. Therefore, the AWDO property still holds even when the interference is non-sparse. Particularly, for subway noise, the WDOM is relatively low compared to the other types of interference. If both the target and interference are non-sparse, the cross term, $|S_x(k)S_n(k)|$, may be very high. Consequently, the AWDO property may not hold in this case.

4.2 Separation results of the IRM and ORM

The average SNR results of the IRM and the ORM are shown in Table 2. As shown in Table 2, the ORM always achieves a higher SNR gain over the IRM with respect to different types of noise and input SNR conditions. The average of the ΔSNR, about 3.61 dB, is very close to the expectation ($10\log_{10}2 \approx 3.01$ dB) which is derived analytically.

Table 1. Average WDOM (%) with respect to different kinds of interference. Sp: Speech signal; Ca: Cafeteria; Sq: Square; Su: Subway.

Noise type	Sp	Ca	Sq	Su
WDOM (%)	34.1	34.9	24.9	14.3

Table 2. Average SNR gain (dB) under different kinds of interference. Ca: Cafeteria; Sq: Square; Su: Subway.

Input SNR	Noise	Ca	Sq	Su	Avg.
−3 dB	SNR_I	9.30	11.83	15.18	12.10
—	SNR_O	11.70	14.62	19.46	15.26
0 dB	SNR_I	10.97	13.51	16.65	13.72
—	SNR_O	13.54	16.53	21.43	17.16
3 dB	SNR_I	12.76	15.30	18.13	15.39
—	SNR_O	15.48	18.45	23.31	19.09
6 dB	SNR_I	14.61	17.10	19.61	17.10
—	SNR_O	16.82	19.59	23.97	20.12
9 dB	SNR_I	16.54	18.91	21.10	18.85
—	SNR_O	19.68	22.39	26.99	23.00

The algorithm proposed by Hu and Loizou[16] is further used to evaluate the objective quality of the separated speech signals. This algorithm converts several composite objective measures to a mean opinion score-like listening quality score, which ranges from 1 to 5. The higher the score, the better the perceptual quality. Let OQS_I and OQS_O denote the quality scores corresponding to the IRM and ORM, respectively. The average results are presented in Table 3. Under all the noise types and input SNR conditions, the ORM achieves consistently higher objective quality scores than the IRM. Especially, we can see that the improvement is very significant under low input SNR conditions. Six files of the mixture and the separated speech signals by the IRM and ORM are given in Mm.1–Mm.6, respectively.

Mm. 1. A speech signal mixing with cafeteria noise at −3 dB input SNR. This is a file of type "wav" (313 Kb).

Mm. 2. The speech signal separated from Mm. 1 by the IRM. This is a file of type "wav" (313 Kb).

Mm. 3. The speech signal separated from Mm. 1 by the ORM. This is a file of type "wav" (313 Kb).

Mm. 4. A speech signal mixing with square noise at −3 dB input SNR. This is a file of type "wav" (313 Kb).

Mm. 5. The speech signal separated from Mm. 4 by the IRM. This is a file of type "wav" (313 Kb).

Table 3. Average scores of objective quality under different kinds of interference. Ca: Cafeteria; Sq: Square; Su: Subway.

Input SNR	Noise	Ca	Sq	Su	Avg.
−3 dB	OQS_I	3.75	3.93	4.28	3.99
—	OQS_O	4.33	4.44	4.59	4.45
0 dB	OQS_I	3.91	4.07	4.39	4.12
—	OQS_O	4.40	4.51	4.64	4.52
3 dB	OQS_I	4.06	4.20	4.50	4.25
—	OQS_O	4.47	4.56	4.69	4.57
6 dB	OQS_I	4.20	4.33	4.59	4.37
—	OQS_O	4.54	4.62	4.75	4.64
9 dB	OQS_I	4.33	4.44	4.67	4.48
—	OQS_O	4.62	4.67	4.81	4.70

Mm. 6. The speech signal separated from Mm. 4 by the ORM. This is a file of type "wav" (313 Kb).

In this experiment, both the IRM and ORM are computed using P_x and P_n as prior information. While the precisely accurate PSDs of the speech and noise are given, the phase difference can be easily derived by Eq. (12). We should note that it is impossible to achieve precisely accurate PSDs by the actual estimator. The error in PSD estimation will weaken the performance of the two ratio masks to some extent. Since the IRM and ORM are directly based on the PSD estimation, a substantial effort is needed in the future to analyze the robustness of the error in PSD estimation.

5. Conclusion

In this paper, we propose the ORM in terms of the SNR. Under the AWDO assumption, we further derive an approximate expectation of the SNR improvement from the IRM to the ORM. Separation experiments show that the average SNR improvement is very close to the theoretical expectation. The results also show that the ORM can improve the speech quality, especially for low input SNR conditions. We think that the phase information, which has not received much attention in many present speech separation and enhancement systems, is the main reason for the improvement. This work implies that the phase estimation may be a valuable research topic in the future.

Acknowledgments

This research was supported in part by the China National Nature Science Foundation (Grant Nos. 91120303, 61273267, and 90820011).

References and links

[1]G. N. Hu and D. L. Wang, "Speech segregation based on pitch tracking and amplitude modulation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, NY (2001).
[2]G. N. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," IEEE Trans. Neural Networks **15**(5), 1135–1150 (2004).
[3]Y. Hu and P. C. Loizou, "Speech enhancement based on wavelet thresholding the multitaper spectrum," IEEE. Trans. Speech Audio Process. **12**(1), 59–67 (2004).
[4]S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," Speech Commun. **48**, 220–231 (2006).
[5]O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," IEEE Trans. Signal Process. **52**(7), 1830–1846 (2004).
[6]H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," IEEE Trans. Audio, Speech, Lang. Process. **19**(3), 516–527 (2011).
[7]R. D. Patterson, J. Holdsworth, I. Nimmo-Smith, and P. Rice, "An efficient auditory filterbank based on the gammatone function," Report No. 2341, MRC Applied Psychology Unit (1988).
[8]M. Weintraub, "A theory and computational model of auditory monaural sound separation," Ph.D. dissertation, Department of Electrical Engineering, Stanford University, Stanford, CA, 1985.
[9]S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," Speech Commun. **48**, 1486–1501 (2006).
[10]N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series* (MIT Press, Cambridge, MA, 1949).
[11]S. Mallat, *A Wavelet Tour of Signal Processing* (Academic, New York, 1998), Chap. 4.
[12]Y. P. Li and D. L. Wang, "On the optimality of ideal binary time-frequency masks," Speech Commun. **51**, 230–239 (2009).
[13]M. P. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," J. Acoust. Soc. Am. **120**(5), 2421–2424 (2006).
[14]Signal Separation Evaluation Campaign (2011). Online: http://sizec.wiki.irisa.fr/tiki-index.php?page=Two-channel+mixtures+of+speech+and+real-world+background+noise.
[15]T. Melia, "Underdetermined blind source separation in echoic environments using linear arrays and sparse representations," Ph.D dissertation, University College Dublin, National University of Ireland, 2007.
[16]Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," IEEE Trans. Audio, Speech, Lang. Process. **16**(1), 229–238 (2008).