Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 59 (2014) 22–30

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

The analysis of the simplification from the ideal ratio to binary mask in signal-to-noise ratio sense

Shan Liang, WenJu Liu *, Wei Jiang, Wei Xue

National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China

Received 8 March 2013; received in revised form 11 November 2013; accepted 2 December 2013
Available online 31 December 2013

Abstract

For speech separation systems, the ideal binary mask (IBM) can be viewed as a simplified goal of the ideal ratio mask (IRM) which is derived from Wiener filter. The available research usually verify the rationality of this simplification from the aspect of speech intelligibility. However, the difference between the two masks has not been addressed rigorously in the signal-to-noise ratio (SNR) sense. In this paper, we analytically investigate the difference between the two ideal masks under the assumption of the approximate W-Disjoint Orthogonality (AWDO) which almost holds under many kinds of interference due to the sparse nature of speech. From the analysis, one theoretical upper bound of the difference is obtained under the AWDO assumption. Some other interesting discoveries include a new ratio mask which achieves higher SNR gains than the IRM and the essential relation between the AWDO degree and the SNR gain of the IRM.

Keywords: Ideal binary mask; Ideal ratio mask; W-Disjoint Orthogonality

1. Introduction

The problem of speech separation which aims to remove or attenuate interference has been widely studied for decades. Computational auditory scene analysis (CASA) which is inspired by research on human auditory perception (Bregman, 1990) is one promising approach to this problem (Weintraub, 1985; Cooke, 1993; Brown and Cooke, 1994; Wang and Brown, 1999; Hu and Wang, 2004; Wang and Brown, 2006). Due to the non-stationary nature of speech, the time-domain signals are firstly decomposed into time–frequency (T–F) domain by using discrete short-time Fourier transform (DSTFT) (Mallat, 1998) or auditory filtering (Patterson et al., 1988). Each element of T–F representation is called as a T–F unit corresponding to a certain time and frequency index. Then, CASA techniques generally approach speech separation by two main stages: segmentation and grouping. The ideal binary and ratio masks are two conventional computational goals in CASA (Barker et al., 2000; Hu and Wang, 2001; Srinivasan et al., 2006). Several works show that the two ideal masks have different advantages (Brungart et al., 2006; Li and Loizou, 2008; Peharz and Pernkopf, 2012; Liang et al., 2012). However, the difference between the two ideal masks in terms of signal-to-noise (SNR) has not been rigorously addressed. In this paper, this difference is studied analytically and experimentally under approximate W-Disjoint Orthogonality (WDO) assumption (Yilmaz and Rickard, 2004).

The IBM proposed in Hu and Wang (2001, 2004) is a 0–1 matrix along time and frequency indexes with which we classify all the time–frequency (T–F) units into reliable and unreliable classes. The reliable units are dominated by the target speech, while the unreliable units are dominated by the interference. Several CASA techniques, such as (Brown, 1993; Brown and Cooke, 1994; Ellis, 1996;

* Corresponding author. Tel.: +86 1082614505.
E-mail addresses: sliang@nlpr.ia.ac.cn (S. Liang), lwj@nlpr.ia.ac.cn (W. Liu), wjiang@nlpr.ia.ac.cn (W. Jiang), wxue@nlpr.ia.ac.cn (W. Xue).

0167-6393/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2013.12.002

Wang and Brown, 1999; Hu and Wang, 2001; Kim et al., 2009), and some blind speech separation techniques (Yilmaz and Rickard, 2004; Melia, 2007; Sawada et al., 2011) use the IBM as the computational goal. The IRM defined in Srinivasan et al. (2006) is a soft masking strategy. It is closely related to the Wiener filter (Wiener, 1949) whose frequency response is $P_x/(P_x+P_n)$, where P_x and P_n are the energy density of the target and interference signals respectively. The IBM can also be obtained by quantizing the Wiener filter at each T–F unit to the closest binary value. Intuitively, the IRM achieves higher SNR gain over the IBM because the Wiener filter minimizes the mean-square error (MSE) for stationary signals.

Although the IBM is a simplified form of the IRM, however, many separation systems prefer the IBM as the computational goal due to its three main desirable properties. First, previous works have demonstrated that the IBM could improve speech intelligibility significantly (Roman et al., 2003; Brungart et al., 2006; Li and Loizou, 2008). Moreover, psychoacoustic experiments in Li and Loizou (2008) demonstrated that binary masks that deviate from the IBM degrade the intelligibility performance gradually. In the work (Loizou and Kim, 2011), they explained why existing speech enhancement algorithms can not improve speech intelligibility and provided an analytical proof that the IBM can maximize the average of the spectral SNRs. They further proved that maximizing the geometric average of SNRs is equivalent to maximizing a simplified form of the articulation index which is an objective measure used for predicting speech intelligibility (Kryter, 1962). Second, noise tracking is the fundamental task for the IRM estimation. But the common noise tracking algorithm, such as (Martin, 2001; Rangachari and Loizou, 2006), can not track highly non-stationary real world noise well. By contrast, many auditory features which are robust to the effects of noise have been proposed for the IBM estimation, such as pitch-based features (Brown and Cooke, 1994; Ellis and Rosenthal, 1995; Seltzer et al., 2004; Hu and Wang, 2004; Hu and Wang, 2010; Han and Wang, 2012; Liang et al., 2013) and amplitude modulation spectrum (AMS) (Kim et al., 2009). Noise tracking is not necessary for the IBM estimation. Therefore, it can be well generalized to non-stationary noise. Third, the complex noise spectrum estimation task can be simplified into a binary classification task with the IBM estimation. While the IRM estimation requires the relative energy ratio of the two signals, the IBM estimation is considerably simpler than the IRM estimation (Li and Wang, 2009). Bayesian classifier based IBM estimation can be traced back to Seltzer et al. (2004). Recently, many different variations of the Bayesian classifier and other statistical classification methods have been used in this task (Kim et al., 2009; Hu and Wang, 2010; Han and Wang, 2012; Liang et al., 2013).

In the IBM based resynthesis, the energy lying in unreliable units is totally removed. It may cause too many nonlinear distortions (musical noise) in the extracted signal (Ma et al., 2010). In practice, some inevitable errors in the IBM estimation may further increase the distortion. On one hand, conventional automatic speech recognition (ASR) systems are extremely sensitive to the distortions. Using ratio mask in the range [0.0,1.0] is one approach to minimize the effect of distortions on recognition (Barker et al., 2000). We should note that the ratio mask defined in Barker et al. (2000) indicates the degree of confidence on whether or not the T–F unit is reliable. Therefore, it is a different concept with the IRM. Other approaches include missing data imputation techniques (Cooke et al., 2001; Raj et al., 2004). On the other hand, the separation results in Peharz and Pernkopf (2012) show that ratio mask usually results in better perceptual quality, while the binary mask achieves higher interference suppression. In Liang et al.'s work (2012), they propose a method for smoothing the binary mask based speech cochleagram estimation. The separation results show that the ratio mask achieves better performance on suppressing artifacts.

Since the SNR measure produces a single ratio making it easy to evaluate the performance of a separation system, it remains a widely used performance metric. Theoretically, the IRM gets higher SNR gain relative to the IBM. Experiments in Li and Wang (2009) showed that the IBM gets slightly lower SNR results than the IRM even with non-sparse interference, such as white noise. But they have not explained why the difference is so small. Furthermore, there is not yet a rigorous conclusion about the upper bound of the difference. Strictly speaking, the IBM is equivalent to the IRM only when the target and interference signals subject to W-Disjoint Orthogonality (WDO) property (Yilmaz and Rickard, 2004). The WDO property means that the T–F representations corresponding to the target and interference signals rarely overlap. If both of the target and interference are sufficiently sparse, such as speech signal, the energy overlap is very small with high probability. In this case, the WDO property is approximately satisfied. Other typical blind speech separation algorithms using the IBM estimation as the computational goal include (Melia, 2007; Sawada et al., 2011). But the difference between the two ideal mask frameworks under approximate WDO property has not been rigorously addressed.

Also in this paper, we do not concerned with how to estimate the IBM and the IRM. We analytically investigate the SNR gain of the IBM and the IRM with DSTFT (Mallat, 1998) based T–F representation. With SNR performance as the optimal goal, three key points are found during the analysis. First, the IBM is the optimal binary mask while the T–F decomposition is orthogonal. This result is consistent with the theorem given in Li and Wang (2009). Second, although the IRM is not the optimal linear mask model in theory, it approximates to the optimal model under approximate WDO assumption. Third, the difference of the two ideal masks is no more than $10\log_{10}2$dB. Experiments with ten kinds of real world noise further show the difference is always smaller than 1 dB. Finally, we propose an explanation why the difference is so small.

The paper is organized as follows. In the next section, some notations and definitions are introduced. SNR gain is discussed analytically in Section 3. Speech separation experiments are further used to verify the discussion in Section 4. The last section gives some conclusions.

2. Notation and Definition

Let $x(t)$ and $n(t)$ denote the $T-$ length speech and interference signals in time domain, respectively. In the classical additive noise model, the noisy speech is given by:

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

Speech separation systems attempt to estimate the original speech $x(t)$ from mixture $y(t)$ as accurately as possible.

Since speech and many real world noises are non-stationary, the time domain signal has to be decomposed into time–frequency domain by DSTFT (Mallat, 1998) or auditory filtering (Patterson et al., 1988). Suppose that $S_\cdot(\tau,f)$ denotes the DSTFT coefficient of a signal in τ 'th time frame and fth frequency index. With a real and symmetric window function, $g(t) = g(-t)$, the DSTFT of $y(t)$ is given by:

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

The power spectrum density is $P_y(\tau,f) = |S_y(\tau,f)|^2$.

Since DSTFT is complete and stable, $x(t)$ can be reconstructed from $S_x(\tau,f)$ by inverse discrete short-time Fourier transform (IDSTFT) (Mallat, 1998). In other words, both of speech separation and enhancement tasks can be transformed into the problem of $S_x(\tau,f)$ estimation. Suppose that $\widehat{S}_x(\tau,f)$ is an estimation of $S_x(\tau,f)$, the estimation in time domain $\hat{x}(t)$ is:

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

2.1. Ideal binary and ratio mask

Let $P_x(\tau,f)$ and $P_n(\tau,f)$ denote the power spectrum densities of the target speech and the interference, respectively. The IBM is defined as follows:

$$M_B(\tau,f) = \begin{cases} 1, & \text{if} \quad P_x(\tau,f) - P_n(\tau,f) > \theta \\ 0, & \text{else} \end{cases}, \tag{4}$$

where θ is a threshold. If $M_B(\tau,f) = 1$, the T–F unit is called reliable; otherwise it's called unreliable. With the binary mask matrix, the estimation of $S_x(\tau,f)$ can be further written as $\hat{S}_x(\tau,f) = M_B(\tau,f)S_y(\tau,f)$ (Yilmaz and Rickard, 2004). Ideal ratio mask which is closely related to Wiener filter (Wiener, 1949) is defined as follows (Srinivasan et al., 2006):

$$M_R(\tau,f) = \frac{P_x(\tau,f)}{P_x(\tau,f) + P_n(\tau,f)}. \tag{5}$$

Similarly, $\tilde{S}_x(\tau,f) = M_R(\tau,f)S_y(\tau,f)$. Fig. 1 shows an example of the IBM and IRM for a speech signal mixed with white noise at 0 dB input SNR. The spectrograms of a male utterance and white noise are shown in Fig. 1 (a) and (b), while the IBM with $\theta = 0$ and the IRM matrices are shown in (c) and (d). Many speech enhancement algorithms, such as the Wiener-type speech-enhancement algorithm (Hu and Loizou, 2004; Rangachari and Loizou, 2006), are based on ratio mask strategy. From the two definitions, we can find that quantizing the IRM to the closest binary mask will result in the IBM. Ellis (2006) proposes an argument to the optimality of the IBM. Since the IBM with 0 dB threshold is the closest binary value to the Wiener filter which achieves the minimum mean square error (MMSE) for stationary signals (Wiener, 1949), it may be the optimal binary mask in terms of the mean square error (MSE) and the SNR gain.

2.2. W-Disjoint Orthogonality (WDO)

The WDO property is derived from the sparse nature of speech signal in the T–F domain, where sparse means that a small percentage of the T–F units contain a large percentage of the signal energy. As shown in Fig. 1 (a), a large percentage of the speech energy is contained in the harmonic structure. If both of the target and interference signals are sparse, the T–F units containing significant energy of the two signals rarely overlap. A rigorous definition of WDO is given by Yilmaz and Rickard (2004):

$$S_x(\tau,f)S_n(\tau,f) = 0, \quad \forall\tau,\ f. \tag{6}$$

Obviously, the WDO property is a mathematical idealization. Previous work (Yilmaz and Rickard, 2004) showed that the energy overlap, $|S_x(\tau,f)S_n(\tau,f)|$, is a very small value with high probability while the interference is a different speech signal. A more rigorous statement is that $|S_x(\tau,f)S_n(\tau,f)|$ is much smaller than $P_x(\tau,f)$ if this unit is reliable, and $P_n(\tau,f)$ otherwise. This property is called as approximate WDO (AWDO) in this paper. A general metric has been proposed (Melia, 2007) to measure the WDO degree:

$$WDOM = \frac{\sum_{\tau,f}|S_x(\tau,f)S_n(\tau,f)|}{\sqrt{\sum_{\tau,f}P_x(\tau,f)\sum_{\tau,f}P_n(\tau,f)}}. \tag{7}$$

Lower WDOM value indicates higher AWDO degree.

3. SNR gain of the ideal binary and ratio masks

The SNR gain which is closely related to the MSE is defined as:

$$L(\hat{x},x) = \sum_{t=0}^{T-1}[\hat{x}(t) - x(t)]^2 = \sum_{t=0}^{T-1} r(t)^2. \tag{8}$$

According to Parsevals equality (Mallat, 1998), the MSE is given by:

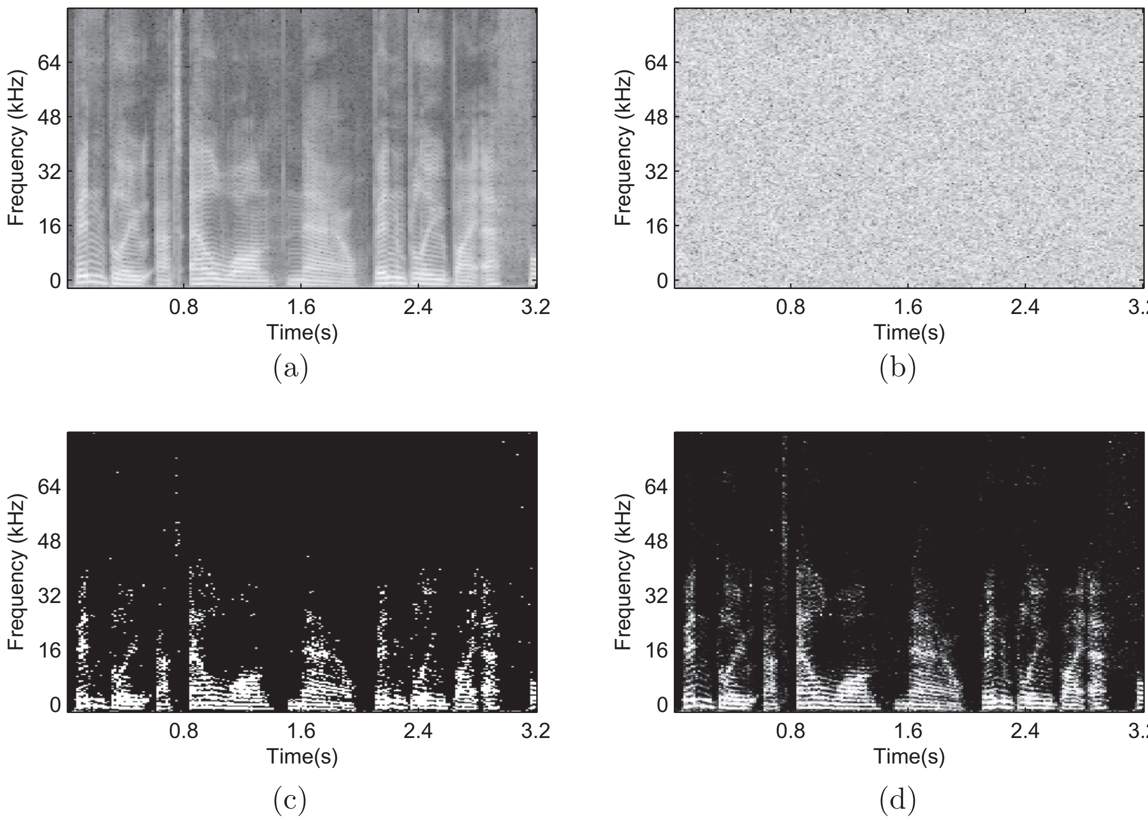


Fig. 1. An example of the IBM and IRM. (a) The spectrogram of the target speech, (b) The spectrogram of the white noise, (c) The IBM, (d) The IRM.

$$\sum_{t=0}^{T-1} r(t)^2 = \frac{1}{T}\sum_{\tau=0}^{T-1}\sum_{f=0}^{T-1} |\hat{S}_x(\tau,f) - S_x(\tau,f)|^2. \tag{9}$$

The definition of SNR is given by:

$$\begin{aligned} SNR &= 10\log_{10}\left(\frac{\sum_t x(t)^2}{\sum_t r(t)^2}\right) \\ &= 10\log_{10}\left(\sum_t x(t)^2\right) - 10\log_{10}\left(\sum_t r(t)^2\right). \end{aligned} \tag{10}$$

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

3.1. SNR gain of the ideal binary mask

To simplify the representation, one-dimension coordinate, $k \doteq (\tau, f)$, is used in the following discussions. The MSE corresponding to the IBM can be written as:

$$\begin{aligned} L(\hat{x},x) &= \sum_t \hat{r}(t)^2 = \frac{1}{T}\sum_k |\hat{S}_x(k) - S_x(k)|^2 \\ &= \frac{1}{T}\sum_k |M_B(k)[S_x(k) + S_n(k)] - S_x(k)|^2 \\ &= \frac{1}{T}\sum_{k\in U} P_x(k) + \frac{1}{T}\sum_{k\in R} P_n(k), \end{aligned} \tag{11}$$

where U and R denote the unreliable and reliable T–F sets, respectively. We can find that the IBM with $\theta = 0$ is the optimal binary mask in terms of SNR gain under DSTFT based T–F representation. For a random reliable unit where $P_x(k_0) > P_n(k_0)$, i.e., if we flip the IBM to $M_B(k_0) = 0$, the new MSE will be updated to:

$$\hat{L}(\hat{x},x) = L(\hat{x},x) + \frac{1}{T}(P_x(k_0) - P_n(k_0)). \tag{12}$$

Obviously, the MSE corresponding to the IBM, $L(\hat{x},x)$, is always smaller than $\hat{L}(\hat{x},x)$. Moreover, similar result can be obtained for any unreliable units where $P_x(k_0) < P_n(k_0)$. Therefore, the IBM minimizes the MSE over all binary masks. This result is consistent with the conclusion given in Li and Wang (2009).

We should note that the optimality may not hold if the T–F transform is non-orthogonal. In previous work, Li and Wang (2009) showed several counterexamples with gammatone filtering (Patterson et al., 1988) based T–F representation.

3.2. SNR gain of the ideal ratio mask

The MSE corresponding to the IRM, $L(\tilde{x},x)$, is given by (Appendix A):

$$L(\tilde{x},x) = \sum_t \tilde{r}(t)^2 = \frac{1}{T}\sum_k \left|\frac{P_x(k)S_n(k) - P_n(k)S_x(k)}{P_x(k) + P_n(k)}\right|^2. \tag{13}$$

As shown in Appendix A, $L(\tilde{x}, x)$ can be further approximated by the following equation under WDO or AWDO condition:

$$\sum_t \tilde{r}(t)^2 \approx \frac{1}{T}\sum_k \left(\frac{P_x(k)P_n(k)}{P_x(k)+P_n(k)}\right) = \frac{1}{T}\sum_k \varphi(k). \tag{14}$$

Particularly, "≈" can be re-written to "=" under WDO condition. From Eq. (14), we can find that the MSE of the IRM, $\sum_t \tilde{r}(t)^2$, will increase with the increasing of energy overlap. Since the energy overlap is directly correlated with AWDO metric which is defined in Eq. (7), the SNR gain of the IRM also hinges upon the AWDO degree.

In this paper, we regard the framework, $\overline{S}_x(k) = \gamma(k)S_y(k)$, as linear mask model (LMM) where $\gamma(k)$ is a real value. We can find that the IRM is the optimal linear mask model in the SNR sense if the cross term is totally ignored. To verify this conclusion, we firstly give the MSE of a general LMM as follows:

$$\begin{aligned}\sum_t r(t)^2 &= \frac{1}{T}\sum_k |\gamma(k)S_y(k) - S_x(k)|^2 \\ &= \frac{1}{T}\sum_k |(\gamma(k)-1)S_x(k) + \gamma(k)S_n(k)|^2.\end{aligned} \tag{15}$$

As in Appendix A, the MSE can be re-written as:

$$\begin{aligned}&\sum_t r(t)^2 = \frac{1}{T}\sum_k \Big[(\gamma(k)-1)^2 P_x(k) + \gamma(k)^2 P_n(k) \\ &+2\gamma(k)(\gamma(k)-1)\Re\big(S_x(k)S_n^*(k)\big)\Big].\end{aligned} \tag{16}$$

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

$$\begin{aligned}\frac{\partial\left(\sum_t r(t)^2\right)}{\partial\gamma(k)} &= \frac{2}{T}\big(P_x(k)[\gamma(k)-1] + P_n(k)\gamma(k) \\ &\quad + [2\gamma(k)-1]\Re\big(S_x(k)S_n^*(k)\big)\big) = 0.\end{aligned} \tag{17}$$

The optimal LMM, $\gamma_{opt}(k)$, is given by:

$$\gamma_{opt}(k) = \frac{P_x(k) + \Re\big(S_x(k)S_n^*(k)\big)}{P_x(k)+P_n(k)+2\Re\big(S_x(k)S_n^*(k)\big)}. \tag{18}$$

If $\Re(S_x(k)S_n^*(k))$ is equal to 0 for $\forall k$, the IRM is equivalent to γ_{opt}. In other words, the optimality of the IRM is also closely related to the degree of the AWDO property.

The following theorem summarizes above analytical results:

Theorem 1. *The IRM minimizes the MSE and consequently maximizes the SNR over all linear mask models under WDO condition. Under AWDO assumption, the IRM approximates to the optimal model. The increasing of* $|S_x(k)S_n(k)|$ *will degrade the SNR gain gradually.*

3.3. The difference between the IBM and the IRM in SNR gain

$\frac{P_n(k)}{P_x(k)+P_n(k)}$ is always no more than 1 for any unreliable unit, while $\frac{P_x(k)}{P_x(k)+P_n(k)}$ is always no more than 1 for any reliable unit. Therefore, $\sum_t \tilde{r}(t)^2 \leqslant \sum_t \hat{r}(t)^2$. This means that the IRM gets higher SNR gain than the IBM. Let ΔSNR denotes the difference:

$$\Delta SNR = SNR_{IRM} - SNR_{IBM} = 10\log_{10}\left(\frac{\sum_t \hat{r}(t)^2}{\sum_t \tilde{r}(t)^2}\right). \tag{19}$$

Under strict WDO condition given in Eq. (6), $P_n(k)$ is equal to zero for any reliable unit, while $P_x(k)$ is equal to zero for any unreliable unit. Therefore, the IBM and IRM get the same MSE and SNR results which means $\Delta SNR = 0$ dB.

Under AWDO condition, combine Eqs. (11) and (14):

$$\begin{aligned}&\sum_t \hat{r}(t)^2 - \sum_t \tilde{r}(t)^2 \approx \frac{1}{T}\left(\sum_{k\in U}\phi_1(k) + \sum_{k\in R}\phi_2(k)\right), \\ &\phi_1(k) = \frac{P_x^2(k)}{P_x(k)+P_n(k)}, \quad \phi_2(k) = \frac{P_n^2(k)}{P_x(k)+P_n(k)}.\end{aligned} \tag{20}$$

According to the definition of the IBM with $\theta = 0$, the following inequality can be derived:

$$\begin{aligned}&0 \leqslant \phi_1(k)/\varphi(k) = P_x(k)/P_n(k) < 1, \quad \forall k \in U, \\ &0 \leqslant \phi_2(k)/\varphi(k) = P_n(k)/P_x(k) < 1, \quad \forall k \in R.\end{aligned} \tag{21}$$

Therefore,

$$\begin{aligned}&\sum_{k\in U}\phi_1(k) + \sum_{k\in R}\phi_2(k) < \sum_k \varphi(k), \\ &\Rightarrow \sum_t \hat{r}(t)^2 - \sum_t \tilde{r}(t)^2 \leqslant \sum_t \tilde{r}(t)^2 \\ &\Rightarrow \sum_t \hat{r}(t)^2 \leqslant 2\sum_t \tilde{r}(t)^2.\end{aligned} \tag{22}$$

According to Eq. (19), ΔSNR is no more than $10\log_{10}2$ dB. The following theorem summarizes above analytical result:

Theorem 2. *The IBM is equivalent to the IRM in strict WDO condition. In AWDO condition, the IRM obtains higher SNR gain over the IBM. But the difference,* ΔSNR, *is no more than* $10\log_{10}2$ dB.

Experiments in Li and Wang (2009) have showed that ΔSNR is around 0.7 dB ($10\log_{10}2 \approx 3.01$) even with white interference. As is well known, white noise is non-sparse because the energy density is a constant. That is, the energy equally lies on all T–F units. However, it is very difficult to further quantify the upper bound of ΔSNR in theory. Here, we give a brief discussion about 0 dB input SNR condition which means that $\sum_k P_x(k) = \sum_k P_n(k)$. A relative MSE measure is defined as follows:

$$\Delta MSE = \left(\sum_{k\in U}\phi_1(k) + \sum_{k\in R}\phi_2(k)\right)\Big/\sum_k \varphi(k). \tag{23}$$

We can find that $P_x(k)$ is much smaller than $P_n(k)$ for most unreliable units due to the sparse nature of speech. Similarly, $P_n(k)$ is much smaller than $P_x(k)$ for most reliable units. This conclusion can be convinced by experiments in Section 4.1. Therefore, ΔMSE may be much smaller than 1 and consequently $\sum_t \hat{r}(t)^2 \ll 2\sum_t \tilde{r}(t)^2$. This result also implies that $\sum_t \tilde{r}(t)^2$ increases much faster than $\sum_t \hat{r}(t)^2 - \sum_t \tilde{r}(t)^2$ with the increasing of energy overlap. We suspect that this is the main reason why the ΔSNR is so small.

4. Experimental results

The above discussions are verified experimentally with different kinds of background interference. 40s-length speech signals are randomly taken from the training set provided by the Grid corpus (Cooke et al., 2006). The signals are down-sampled to 16 kHz before mixing interferences. The interference signals are collected by Cooke (1993), which includes 10 different types of real world noise. The speech and noise signals are mixed with 0 dB input SNR. To compute the DSTFT, the noisy time domain signals are divided in frames of 512 samples with an overlap of 50%.

4.1. Distributions of $\phi_1(k)/\varphi(k)$ and $\phi_2(k)/\varphi(k)$

To simplify the representation, we define a random variable as follows:

$$\epsilon(k) = \begin{cases} \phi_1(k)/\varphi(k) = P_x(k)/P_n(k), & \text{if} \quad k \in U \\ \phi_2(k)/\varphi(k) = P_n(k)/P_x(k), & \text{if} \quad k \in R \end{cases}. \tag{24}$$

As given in Eqs. (19) and (23), $\epsilon(k)$ is closely related to ΔMSE and ΔSNR. In this paper, we use histogram model to describe the probability density function (PDF), $p(\epsilon)$. We still take white noise with 0 dB input SNR as an example. 160625 T–F units which are randomly selected are firstly classified into reliable and unreliable sets. Then, two histograms with 30 bins are estimated for the two sets, respectively.

The two histograms are shown in Fig. 2. We can find that $\epsilon(k) \ll 1$ for most reliable and unreliable units. We further use exponential distribution with parameter $\lambda = 1/E(\epsilon(k))$ to fit the histogram where $E(\cdot)$ denotes the expectation. As shown in Fig. 2, the two histograms can be approximated by exponential distributions. In addition, the expectation corresponding to unreliable units is much smaller than reliable units ($1/25.41 \ll 1/3.43$). We suspect that this is due to the fact that target speech is quite sparse. Therefore, only a small percentage of speech energy is contained in unreliable units, while relatively more noise energy is contained in reliable units.

4.2. SNR results under AWDO condition

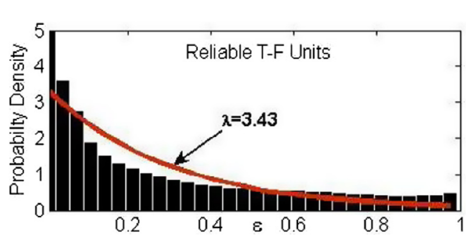
The average values of SNR_{IRM}, ΔSNR, $WDOM$ and ΔMSE are shown in Table 1. Overall, the difference, ΔSNR,is always smaller than 1 dB over the ten kinds of interference. Besides, we get three interesting findings from Table 1. First, the $WDOM$ corresponding to N0, N2, N5 and N6 are much smaller than others. The four types of noise are typical impulse noise which is sparser than speech signal. Second, SNR_{IRM} decreases with the increasing of $WDOM$. This result implies that MSE, $\sum_t \tilde{r}(t)^2$, increases under non-sparse interference. It is consistent with Theorem 1. The last point is that all the ΔMSE lies on a narrow interval 0.2∼0.4 although $WDOM$ lies on a wide range. Since ΔMSE is directly related to ΔSNR, ΔSNR always lies on a narrow interval no matter that the interference is sparse or not. This result is agreement with the discussion in last paragraph in Section 3.3. Besides, ΔSNR is nearly proportional to ΔMSE , as is shown in Fig. 3.

4.3. SNR results of the optimal linear mask model

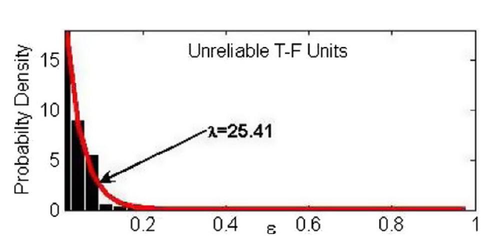
The average SNR results corresponding to the IRM and the optimal linear mask model (OLMM), $\gamma_{opt}(k) = \frac{P_x(k)+\Re\left(S_x(k)S_n^*(k)\right)}{P_x(k)+P_n(k)+2\Re\left(S_x(k)S_n^*(k)\right)}$, are shown in Table 2. Compared to

Table 1
Average SNR gain with respect to different types of noise. Noise types: N0, 1-kHz pure tone; N1, white noise; N2, noise bursts; N3, cocktail party noise; N4, rock music; N5, siren; N6, trill telephone; N7, female speech; N8, male speech and N9, female speech.

Noise Type	SNR_{IRM} (dB)	ΔSNR (dB)	$WDOM$ (%)	ΔMSE (%)
N0	24.61	0.52	3.33	24.70
N1	13.66	0.65	20.41	29.36
N2	17.82	0.61	8.54	28.96
N3	9.24	0.71	43.64	38.03
N4	14.04	0.65	19.14	33.23
N5	21.08	0.82	5.04	32.30
N6	23.79	0.65	3.76	26.66
N7	16.09	0.68	12.80	33.95
N8	15.32	0.62	14.04	32.25
N9	12.93	0.88	23.28	35.51



(a)Reliable T-F units



(b)Unreliable T-F units

Fig. 2. Histograms of ϵ for reliable and unreliable units. Thick line represents the probability density of exponential distribution.

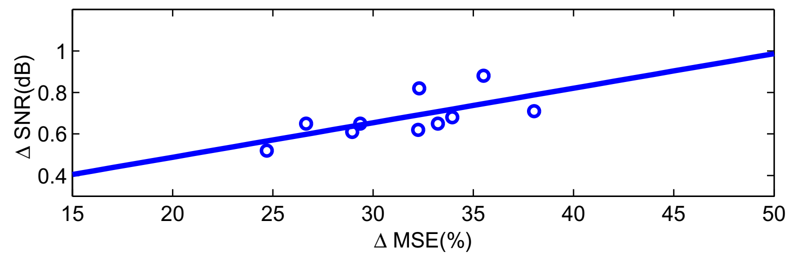


Fig. 3. Representation of the correlation between Δ*MSE* and Δ*SNR*. The line is the least-square error (LSE) rule based estimation.

Table 2
Average SNR results of the IRM and the optimal linear mask model (OLMM).

Mixture	N0	N1	N2	N3	N4	N5	N6	N7	N8	N9
IRM	24.61	13.66	17.82	9.24	14.04	21.08	23.79	16.09	15.32	12.93
OLMM	27.10	16.18	20.38	11.44	16.27	23.17	26.32	18.23	17.62	15.28

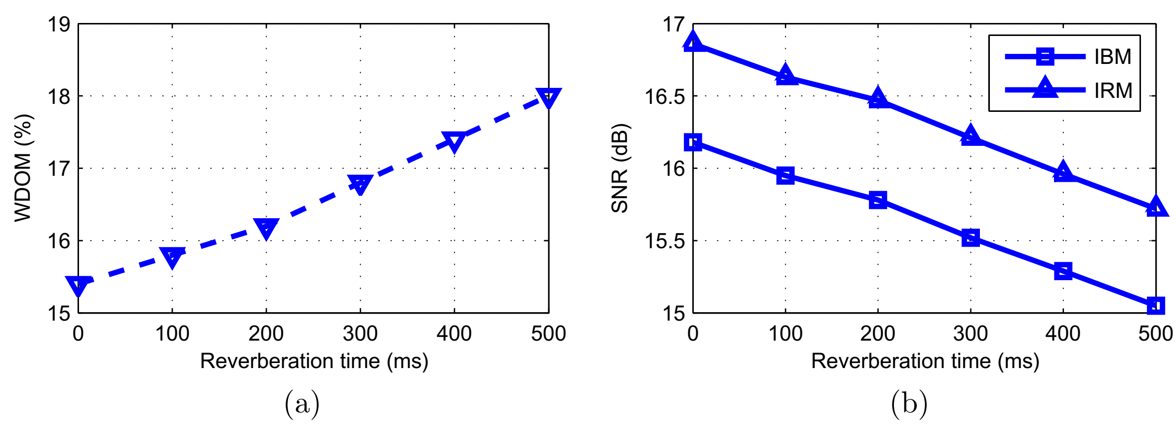


Fig. 4. Average WDOM and separation results with various room reverberation times.

the IRM, the OLMM obtains relatively higher SNR results for all types of noise. According to the AWDO assumption, $\Re(S_x(k)S_n^*(k))$ is relatively minor compared to $P_x(k)$ for most reliable units. However, this simplification is no longer valid for unreliable units. We suspect that this is the main reason for the difference.

4.4. Separation results in reverberation environments

Obviously, the reverberation in real world environment may weaken the AWDO degree to some extent. Since the AWDO degree is directly correlated with the SNR gains of the IBM and the IRM, we provide qualitative comparisons between the two masks under reverberation conditions. The reverberation impulse response is generated by the image-source method (ISM) described in Lehmann and Johansson (2008). The sound propagation is simulated for a small room with 3 m×4 m×4 m size. The reverberation time (T_{60}) ranges from 0 to 500 ms with 100 ms step. The mixture is generated with 0 dB input SNR. The average values of the WDOM and SNR gains are shown in Fig. 4 (a) and (b), respectively. We can find that the AWDO degree is weakened gradually with the increase of T_{60}. Furthermore, both of the SNR gains corresponding to the IRM and the IBM decrease gradually. This result is consist with Theorem 1 that the decreasing of AWDO degree will degrade the SNR gain of the IRM. Specifically, the average value of WDOMs is about 15.4% with T_{60} =0 ms. While T_{60} increases to 500 ms, it only increases to 18.0%. This result indicates the AWDO assumption still hold even in room reverberation environments. Consequently, the difference, Δ*SNR*, has not increased or decreased too much.

Besides, three types of real world noises which are recorded in cafeteria, square and subway environments[1] are used to test the AWDO assumption. For each type of

[1] Signal Separation Evaluation Campaign, 2011. Online: http://sisec.wiki.irisa.fr/tiki-index.php?page=Two-channel+mixtures+of+speech+and+real-world+background+noise.

Table 3
Separation results in real world environments. Ca: cafeteria, Sq: square, Su: subway.

Noise	*WDOM* (%)	SNR_{IRM} (dB)	SNR_{IBM} (dB)	ΔSNR (dB)
Ca	33.0	10.98	10.43	0.55
Sq	23.5	13.54	12.89	0.65
Su	14.0	16.64	16.15	0.49

noises, a 20 s-length signal is selected. The average *WDOM* value and SNR result are shown in Table 3. Overall, the *WDOM* values of the three types of noise are lower than the value corresponding to cocktail party noise. Therefore, the AWDO assumption still holds in the three classical environments. Furthermore, all the values of ΔSNR are lower than 1 dB. This result implies that the IBM can be considered as an effective simplification of the IRM even in many real world environments.

5. Conclusions

In this paper, we investigate the SNR gain of the IBM and the IRM, respectively. Due to the sparse nature of speech, the WDO or AWDO property is valid under many real world noises. Under WDO and AWDO conditions, we prove two theorems to describe the SNR gains of the IBM and the IRM. Then, we find that one upper bound of the difference between the two ideal masks is $10\log_{10}2$ dB. Experimental results on a speech separation database further show that the difference is smaller than 1 dB. We believe that the sparse nature of speech is the fundamental reason that the difference is so small. Therefore, the IBM is a reasonable approximation of the IRM even in the SNR sense.

It is worth reminding that both of the two ideal masks are derived from precisely accurate estimations of power spectrum which are almost impossible to be achieved in practice. The SNR gains of binary and ratio masks will certainly degraded by the error in spectrum estimation. It is no denying that the spectrum estimation is not necessary in many present binary mask estimators. But, only from the perspective of spectrum estimation, the IBM seems to be more robust to the spectrum errors. Take a reliable units as example, all the spectrum estimations which subject to $\hat{P}_x(k) > \hat{P}_n(k)$ result in correct binary mask. In other words, precisely accurate spectrum is not necessary in the IBM estimation. By contrast, the ratio mask is more sensitive to the spectrum estimation. Substantial effort is needed in our future work to further quantify the degradation.

Acknowledgements

This research was supported in part by the China National Nature Science Foundation (No. 91120303, No. 61273267 and No. 90820011).

Appendix A. The MSE corresponding to the IRM can be written as:

$$L(\tilde{x},x) = \sum_t \tilde{r}(t)^2 = \frac{1}{T}\sum_k |\widetilde{S}_x(k) - S_x(k)|^2 = \frac{1}{T}\sum_k |M_R(k)[S_x(k)+S_n(k)] - S_x(k)|^2. \tag{A.1}$$

According to the definition of the IRM, $L(\tilde{x},x)$ is given by:

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

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

Under WDO assumption, $\left|S_x(k)S_n^*(k)\right|$ is equal to 0 for any units. Under AWDO assumption, it is much smaller than $P_x(k)$ and $P_n(k)$ for most of the reliable and unreliable units, respectively. Moreover, $\Re\left(S_x(k)S_n^*(k)\right) \leqslant \left|S_x(k)S_n^*(k)\right|$. This means that the cross term, $\Re\left(S_x(k)S_n^*(k)\right)$, is relatively smaller compared to $P_x(k)+P_n(k)$. Therefore, $L(\tilde{x},x)$ can be approximated by the following equation:

$$L(\tilde{x},x) \approx \frac{1}{T}\sum_k \frac{P_x(k)P_n(k)}{P_x(k)+P_n(k)}. \tag{A.3}$$

References

Barker, J., Josifoveski, L., Cooke, M., Green, P., 2000. Soft decisions in missing data techniques for robust automatic speech recognition. In: Proc. ICSLP, 2000, Beijing, China, pp. 373–376.

Bregman, S., 1990. Auditory Scene Analysis. MIT Press, MA.

Brown, G.J., 1993. Computational auditory scene analysis: a representational approach. J. Acoust. Soc. Amer. 94, 2454.

Brown, G.J., Cooke, M., 1994. Computational auditory scene analysis. Comput. Speech Lang. 8, 297–336.

Brungart, D., Chang, P.S., Simpson, B.D., Wang, D.L., 2006. Isolating the energetic component of speech-on-speech masking with an ideal binary time-frequency mask. J. Acoust. Soc. Amer. 120, 4007–4018.

Cooke, M.P., 1993. Modeling Auditory Processing and Organization. Cambridge University, U.K.

Cooke, M.P., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Commun. 34, 267–285.

Cooke, M.P., Barker, J., Cunningham, S., Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Amer. 120 (5), 2421–2424 (<http://spandh.dcs.shef.ac.uk/gridcorpus/>).

Ellis, D.P.W., Rosenthal, D.F., 1995. Mid-level representations for computational auditory scene analysis. In working notes of the workshop on Comp. Aud. Scene Analysis at the Intl. Joint Conf. on Artif. Intel., Montreal, Canada, pp. 111–117.

Ellis, D.P.W., 1996. Prediction-driven Computational Auditory Scene Analysis. MIT Press, Cambridge, MA (Ph.D. Thesis).

Ellis, D.P.W., 2006. Model-based scene analysis. In: Wang, D.L., Brown, G.J. (Eds.), Computational Auditory Scene Analysis: Principles, Algorithms and Application. Wiley/IEEE Press, Hoboken, NJ, pp. 15–146.

Han, K., Wang, D.L., 2012. A classification based approach to speech segregation. J. Acoust. Soc. Amer. 132, 3475–3483.

Hu, G.N., Wang, D.L., 2001. Speech segregation based on pitch tracking and amplitude modulation. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, October 21–24, NY, USA, pp. 79–82.

Hu, G.N., Wang, D.L., 2004. Monaural speech segregation based on pitch tracking and amplitude modulation. IEEE Trans. Neural Networks 15 (5), 1135–1150.

Hu, G.N., Wang, D.L., 2010. A tandem algorithm for pitch estimation and voiced speech segregation. IEEE Trans. Audio Speech Lang. Process. 18, 2067–2079.

Hu, Y., Loizou, P.C., 2004. Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE. Trans. Speech Audio Process. 12 (1), 59–67.

Kim, G., Lu, Y., Hu, Y., Loizou, P.C., 2009. An algorithm that improves speech intelligibility in noise for normal-hearing listers. J. Acoust. Soc. Amer. 126, 1486–1494.

Kryter, K., 1962. Methods for calculation and use of the articulation index. J. Acoust. Soc. Amer. 34 (11), 1689–1697.

Lehmann, E., Johansson, A., 2008. Prediction of energy decay in room impulse responses simulated with an image-source model. J. Acoust. Soc. Amer. 124 (1), 269–277.

Li, N., Loizou, P.C., 2008. Factors influencing intelligibility of ideal binary-masked speech: implications for noise reduction. J. Acoust. Soc. Amer. 123, 1673–1682.

Li, Y.P., Wang, D.L., 2009. On the optimality of ideal binary time-frequency masks. Speech Commun. 51, 1486–1501.

Liang, S., Liu, W.J., Jiang, W., 2012. Integrating binary mask estimation with MRF priors of cochleagram for speech separation. IEEE Signal Process. Lett. 19 (10), 627–630.

Liang, S., Liu, W.J., Jiang, W., 2013. A new Bayesian method incorporating with local correlation for IBM estimation. IEEE Trans. Audio Speech Lang. Process. 21 (3), 476–487.

Loizou, P.C., Kim, G., 2011. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. IEEE Trans. Audio Speech Lang. Process. 19 (1), 47–56.

Ma, W.Y., Yu, M., Xin, J., Osher, S., 2010. Reducing musical noise in blind source separation by time-domain sparse filters and split bregman method. In: Proc. Interspeech, September 26–30, 2010, Chiba, Japan, pp. 402–405.

Mallat, S., 1998. A Wavelet Tour of Signal Processing. Academic Press (Ch. 4).

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Acoust. Speech Signal Process. 9 (5), 504–512.

Melia, T., 2007. Underdetermined blind source separation in echoic environments using linear arrays and sparse representations, Ph.D Dissertation, University College Dublin, National University of Ireland.

Patterson, R.D., Holdsworth J., Nimmo-Smith, I., Rice, P., 1988. An efficient auditory filterbank based on the gammatone function. Rep. 2341, MRC Applied Psychology Unit.

Peharz, R., Pernkopf, F., 2012. On linear and mixmax interaction models for single channel source separation. In: Proc. IEEE Internat. Conf. on Acoust. Speech Signal Process, March 25–30, Kyoto, Japan, pp. 249–252.

Raj, B., Seltzer, M.L., Stern, R.M., 2004. Reconstruction of missing features for robust speech recognition. Speech Commun. 43, 275–296.

Roman, N., Wang, D.L., Brown, G.J., 2003. Speech segregation based on sound localization. J. Acoust. Soc. Amer. 114 (4), 2236–2252.

Sawada, H., Araki, S., Makino, S., 2011. Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. IEEE Trans. Audio Speech Lang. Process. 19 (3), 516–527.

Seltzer, M.L., Raj, B., Stern, R.M., 2004. A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition. Speech Commun. 43, 379–393.

Srinivasan, S., Roman, N., Wang, D.L., 2006. Binary and ratio time-frequency masks for robust speech recognition. Speech Commun. 48, 1486–1501.

Rangachari, S., Loizou, P.C., 2006. A noise-estimation algorithm for highly non-stationary environments. Speech Commun. 48, 220–231.

Wang, D.L., Brown, G.J., 1999. Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans. Neural Networks 10, 684–697.

Wang, D.L., Brown, G.J., 2006. Fundamentals of computational auditory scene analysis. In: Wang, D.L., Brown, G.J. (Eds.), Computational Auditory Scene Analysis: Principles, Algorithms, and Application. Wiley/IEEE Press, Hoboken, NJ, pp. 1–44.

Weintraub, M., 1985. A theory and computational model of auditory monaural sound separation. Ph.D. Dissertation, Stanford University Department of Electrical Engineering.

Wiener, N., 1949. Extrapolation, Interpolation, and Smoothing of Stationary Time Series. MIT Press, Cambridge, MA.

Yilmaz, O., Rickard, S., 2004. Blind separation of speech mixtures via time-frequency masking. IEEE Trans. Signal Process. 52 (7), 1830–1846.