Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 51 (2009) 230–239
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

On the optimality of ideal binary time–frequency masks

Yipeng Li [a,*], DeLiang Wang [b]

[a] Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210-1277, USA
[b] Department of Computer Science and Engineering and Center of Cognitive Science, The Ohio State University, Columbus, OH 43210-1277, USA

Received 19 February 2008; received in revised form 26 August 2008; accepted 1 September 2008

Abstract

The concept of ideal binary time–frequency masks has received attention recently in monaural and binaural sound separation. Although often assumed, the optimality of ideal binary masks in terms of signal-to-noise ratio has not been rigorously addressed. In this paper we give a formal treatment on this issue and clarify the conditions for ideal binary masks to be optimal. We also experimentally compare the performance of ideal binary masks to that of ideal ratio masks on a speech mixture database and a music database. The results show that ideal binary masks are close in performance to ideal ratio masks which are closely related to the Wiener filter, the theoretically optimal linear filter.

Keywords: Ideal binary mask; Ideal ratio mask; Optimality; Sound separation; Wiener filter

1. Introduction

Recently monaural and binaural sound separation have received attention. A promising approach to the problem, called *computational auditory scene analysis* (CASA) (Wang and Brown, 2006), is inspired by the perceptual theory of *auditory scene analysis* (ASA) (Bregman, 1990), which attempts to explain the remarkable capability of the human auditory system in segregating an acoustic signal into streams that correspond to different sound sources. The majority of CASA systems developed so far (Brown and Cooke, 1994; Wang and Brown, 1999; Roman et al., 2003; Li et al., 2006; Deshmukh et al., 2007) have applied binary time–frequency (*T–F*) masking to extracting a target sound. Typically, in such systems a signal is first transformed to a *T–F* representation such as a spectrogram. Then an element of such a representation, called a *T–F* unit corresponding to a certain time and frequency, is assigned 1 if its energy is considered as from the target or 0 otherwise. Hu and Wang (2001, 2004) proposed a binary mask, called the *ideal binary mask* (IBM), where a *T–F* unit is assigned 1 if in that unit the target energy exceeds the interference energy and 0 otherwise. Specifically, consider a mixture $z[n] = x[n] + y[n]$, where n denotes discrete time, $x[n]$ the target signal and $y[n]$ the interference signal. Denote $\mathbf{Z}$, $\mathbf{X}$, and $\mathbf{Y}$ as the *T–F* representations of $z[n]$, $x[n]$, and $y[n]$ obtained from some *T–F* transformation, respectively. The IBM $\mathbf{M}$ for the target signal $x[n]$ is defined as following:

$$\mathbf{M}_{cm} = \begin{cases} 1, & \text{if } |\mathbf{X}_{cm}|^2 > |\mathbf{Y}_{cm}|^2, \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where $\mathbf{X}_{cm}$ and $\mathbf{Y}_{cm}$ are the spectral values of $\mathbf{X}$ and $\mathbf{Y}$ at a *T–F* unit u_{cm} indexed by frequency c and time m, respectively. Note that the construction of the IBM requires the premixed target and interference signals. Given $\mathbf{Z}$ and $\mathbf{M}$, an estimate of the target signal $x[n]$ can be reconstructed from the element-wise product of $\mathbf{Z}$ and $\mathbf{M}$.

Fig. 1 shows an example of the IBM for a speech signal mixed with a babble noise. The magnitude spectrogram of a female utterance is shown in Fig. 1a. Fig. 1b shows the magnitude spectrogram of a 20-talker babble noise.

* Corresponding author. Tel.: +1 614 292 7402; fax: +1 614 292 2911.
E-mail addresses: li.434@osu.edu, liyip@cse.ohio-state.edu (Y. Li).

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.09.001

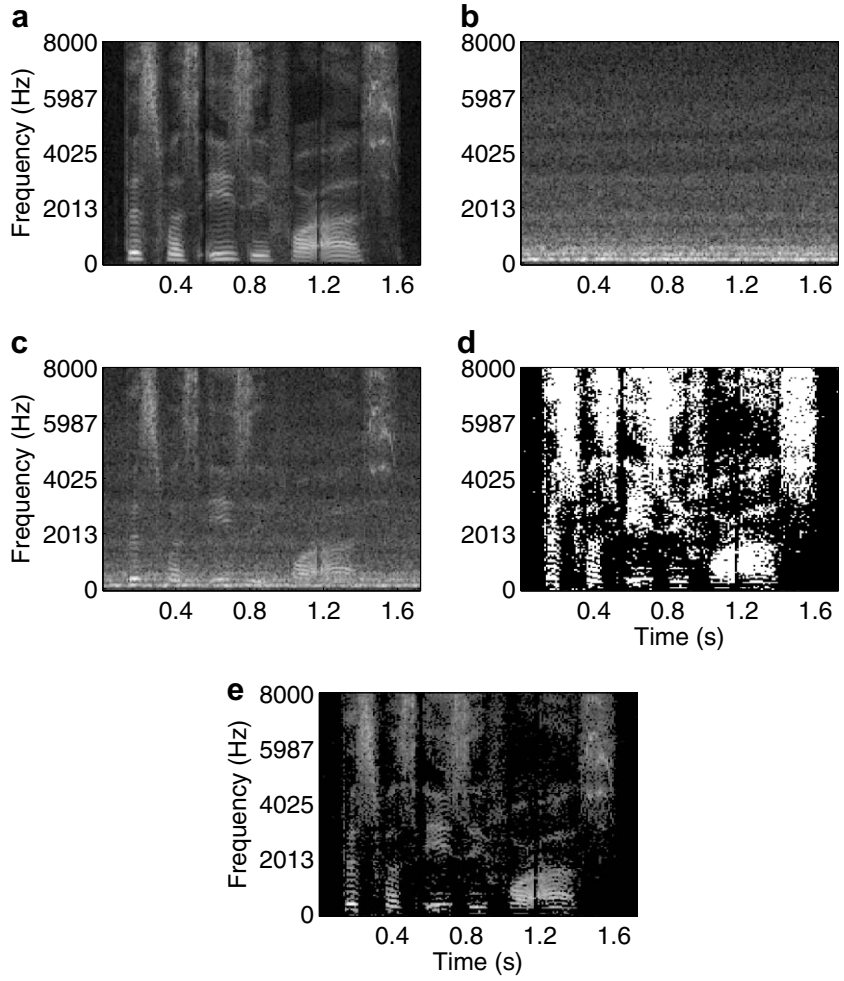

Fig. 1. An example of the IBM. (a) The magnitude spectrogram of a female utterance "this was easy for us". (b) The magnitude spectrogram of a babble noise. (c) The magnitude spectrogram of the mixture. (d) The IBM. (e) The magnitude spectrogram of the mixture masked by the IBM.

Fig. 1c shows the magnitude spectrogram of the mixture of speech and noise mixed with equal overall energy. The IBM for this mixture is shown in Fig. 1d, where white indicates 1 and black 0. Fig. 1e shows the magnitude spectrogram of the mixture masked by the IBM.

The IBM has several desirable properties as a computational goal of CASA (Wang, 2005), including direct correspondence to the auditory masking phenomenon, flexibility in constructing different IBMs out of the same mixture depending on what the target is, and well-definedness regardless of the number and types of signals in the mixture. The IBM has also been shown to be important for human speech intelligibility and automatic speech recognition. A number of recent psychoacoustic experiments have demonstrated that target speech reconstructed from the IBM can dramatically improve the intelligibility of speech masked by different types of noise, even in very noisy conditions (Roman et al., 2003; Brungart et al., 2006; Li and Loizou, 2008). Li and Loizou (2008) further demonstrate that binary masks that deviate from the IBM show gradual degradation of intelligibility performance. The IBM has also been shown to improve the performance of automatic speech recognition by a large margin (Srinivasan et al., 2006).

A widely used metric for performance measure in sound separation is signal-to-noise ratio (SNR). For sound separation it is defined as

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_n x^2[n]}{\sum_n (\hat{x}[n] - x[n])^2}, \tag{2}$$

where $\hat{x}[n]$ is the estimated target signal. A purported property of the IBM is that it is globally optimal, i.e., the IBM produces an output with the highest SNR gain among *all binary masks* (Hu and Wang, 2004; Ellis, 2006). Partly due to this claim, many recent computational systems have used the IBM as a measure of ceiling performance for sound source separation (Li et al., 2006; Kim et al., 2006; Harding et al., 2006; Radfar et al., 2007; Deshmukh et al., 2007; Reddy and Raj, 2007). However, the global optimality of the IBM has not been rigorously addressed. In this paper we theoretically examine the optimality of the IBM. Note that this paper is not intended to advocate the use of SNR as the performance measure of sound sep-

aration systems. Instead we analyze an assumed property of the IBM that is given in terms of SNR. Also in this paper, we are not concerned with how to estimate the IBM, which is the task of sound separation.

It has been noted that the IBM is locally optimal in the SNR sense, i.e., flipping a T–F unit's assignment in the IBM always yields a lower SNR for that T–F unit. There exist two arguments for the global optimality of the IBM. Hu and Wang (2004) argue for the global optimality based on its local optimality. At each T–F unit, the IBM either maximally retains target energy or removes interference energy. As a result, the sum of missing target energy that is discarded by the mask and interference energy that gets through the mask, i.e., the denominator in (2), is minimized. Therefore the IBM would achieve the highest SNR. This argument is flawed in that SNR calculation is nonlinear: the denominator in (2) is not equal to the linear combination of energy retained or removed in each individual T–F unit. Ellis (2006) makes an argument from the viewpoint of Wiener filtering. According to Wiener filtering, optimal SNR can be achieved by the Wiener filter whose frequency response is $P_T/(P_T+P_I)$, where P_T and P_I are the power spectrum densities of target and interference signals, respectively. Quantizing the Wiener filter at each T–F unit to the closest binary value results in the IBM which would produce the optimal binary mask. However, this argument suffers the same drawback as the one by Hu and Wang since it is still based on the local optimality of the IBM: the optimal quantization is performed on each T–F unit.

To closely examine the optimality of the IBM, we consider the optimality of the IBM at three levels: the T–F unit level, the time frame level, and the global level, and find that local optimality does not translate to global optimality. In Section 2 we show that, at each level, the IBM can be optimal under certain conditions imposed on T–F decomposition. We also give counterexamples showing that the IBM is not optimal when these conditions are violated. In Section 3 we compare SNR gain of the IBM to that of ideal ratio masks which are closely related to the Wiener filter. Conclusion and discussion are presented in Section 4.

2. The optimality of the ideal binary mask at different levels

2.1. T–F unit level

Two types of T–F transformations are commonly used in sound separation systems. The first one, such as the short time Fourier transform (STFT), first divides a signal into successive frames and then transforms each frame to the frequency domain. This is called block transform (Princen and Bradley, 1986). The second one, such as a gammatone filterbank (see Section 2.4 for details), first filters a signal by a filterbank and then divides the output of each filter into successive frames. This belongs to the paradigm of filterbank-based transform (Princen and Bradley, 1986). At the T–F unit level, when the block transform is used, only a single spectral value is observed at each T–F unit. As a result, the conventional definition of SNR in (2) is not applicable. Since the SNR defined in (2) essentially is the ratio of the energy of the target to the energy of the estimation error, we can extend it to the T–F unit level

$$\mathrm{SNR}_{cm} = 10\log_{10}\frac{|\mathbf{X}_{cm}|^2}{|\hat{\mathbf{X}}_{cm}-\mathbf{X}_{cm}|^2}, \tag{3}$$

where $\hat{\mathbf{X}}_{cm}$ is the estimated spectral value of the target at T–F unit u_{cm}. When the filterbank-based transform is used, in each T–F unit a time-domain signal is observed (when decimation is not used). In that case, we can still apply the SNR definition as (2).

We first show that at the T–F unit level, the IBM is optimal with respect to SNR defined in (3) for the block transform. At each T–F unit, the IBM takes value 1 if the energy of the target is stronger than that of the interference within the unit, and 0 otherwise. Consequently, the spectral estimate of $\mathbf{X}_{cm}$ is (assuming that the T–F transformation is linear)

$$\hat{\mathbf{X}}_{cm} = \begin{cases} \mathbf{Z}_{cm} = \mathbf{X}_{cm} + \mathbf{Y}_{cm}, & \text{if } |\mathbf{X}_{cm}|^2 > |\mathbf{Y}_{cm}|^2, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Consider the case where $|\mathbf{X}_{cm}|^2 > |\mathbf{Y}_{cm}|^2$, i.e., the target is stronger in energy than the interference at u_{cm}. If u_{cm} is assigned 1, as in the IBM (see (1)), then the denominator in (3) is

$$|\hat{\mathbf{X}}_{cm}-\mathbf{X}_{cm}|^2 = |\mathbf{X}_{cm}+\mathbf{Y}_{cm}-\mathbf{X}_{cm}|^2 = |\mathbf{Y}_{cm}|^2. \tag{5}$$

On the other hand, if u_{cm} is assigned 0, different from the IBM, then the denominator is

$$|\hat{\mathbf{X}}_{cm}-\mathbf{X}_{cm}|^2 = |0-\mathbf{X}_{cm}|^2 = |\mathbf{X}_{cm}|^2. \tag{6}$$

Since $|\mathbf{Y}_{cm}|^2 < |\mathbf{X}_{cm}|^2$, the denominator is smaller than when u_{cm} is assigned according to the IBM.

Similarly, if $|\mathbf{X}_{cm}|^2 \leqslant |\mathbf{Y}_{cm}|^2$ (i.e., the target is not stronger in energy than the interference) and u_{cm} is assigned 0 as in the IBM, then the denominator becomes

$$|\hat{\mathbf{X}}_{cm}-\mathbf{X}_{cm}|^2 = |0-\mathbf{X}_{cm}|^2 = |\mathbf{X}_{cm}|^2. \tag{7}$$

If u_{cm} is assigned 1, then

$$|\hat{\mathbf{X}}_{cm}-\mathbf{X}_{cm}|^2 = |\mathbf{X}_{cm}+\mathbf{Y}_{cm}-\mathbf{X}_{cm}|^2 = |\mathbf{Y}_{cm}|^2. \tag{8}$$

Since $|\mathbf{X}_{cm}|^2 \leqslant |\mathbf{Y}_{cm}|^2$, the IBM always minimizes the denominator and consequently maximizes the SNR. Therefore we conclude that the IBM is optimal at the T–F unit level for the SNR defined in (3) for the block transform.

For the filterbank-based transform, according to the IBM, the time-domain estimation of $x_{cm}[n], \hat{x}_{cm}[n]$, in each u_{cm} is

$$\hat{x}_{cm}[n] = \begin{cases} z_{cm}[n] = x_{cm}[n] + y_{cm}[n], & \text{if } \sum_n x_{cm}^2[n] > \sum_n y_{cm}^2[n], \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

where $x_{cm}[n], y_{cm}[n]$, and $z_{cm}[n]$ are the filtered and framed target signal, interference signal, and the mixed signal at u_{cm}, respectively. Following the same procedure as (5)–(8), it can be easily shown that $\sum_n (\hat{x}_{cm}[n] - x_{cm}[n])^2$ is minimized when the mask is determined according to the IBM. Therefore we can conclude that the IBM is optimal at the *T–F* unit level for the SNR defined in (2) for the filterbank-based transform.

2.2. Time frame level

We now consider $x_m[n]$, the time-domain target signal at frame m. Without loss of generality, we assume that the index of n is from 0 to $N-1$. We first consider the discrete Fourier transform (DFT) of $x_m[n]$

$$\mathbf{X}_{cm} = \sum_{n=0}^{N-1} x_m[n] \mathrm{e}^{-\frac{2\pi cn}{N} j}, \quad c = 0, \ldots, N-1.$$

The SNR of $\hat{x}_m[n]$, the estimate of $x_m[n]$, with respect to $x_m[n]$ can be calculated using (2) with summation of n from 0 to $N-1$. It is clear from (2) that maximizing the SNR is the same as minimizing the denominator, the energy of the error signal $\hat{x}_m[n] - x_m[n]$. According to the Parseval's theorem (Oppenheim et al., 1999), the energy of the error signal can be equivalently calculated in the frequency domain when the transform is DFT, i.e.,

$$\sum_{n=0}^{N-1} (\hat{x}_m[n] - x_m[n])^2 = \frac{1}{N} \sum_{c=0}^{N-1} |\hat{\mathbf{X}}_{cm} - \mathbf{X}_{cm}|^2. \tag{10}$$

In Section 2.1, we have shown that the IBM minimizes $|\hat{\mathbf{X}}_{cm} - \mathbf{X}_{cm}|^2$ for each c. Therefore the IBM also minimizes the summation $\sum_{c=0}^{N-1} |\hat{\mathbf{X}}_{cm} - \mathbf{X}_{cm}|^2$. As a result, the IBM yields the highest SNR among all binary masks.

The key step in the above proof is applying Parseval's theorem to equate the energy summation in the time domain to that in the spectral domain. This is possible because the bases used in DFT are orthonormal, i.e., orthogonal and the length of each basis is 1. In general, Parseval's theorem holds for any orthonormal frequency decomposition. This can be seen clearly when a set of orthonormal bases are used for frequency decomposition: Let $\{\mathbf{e}_i\}$ be a complete set of bases with $\langle \mathbf{e}_i, \mathbf{e}_j \rangle = \delta_{ij}$, where $\langle \cdot, \cdot \rangle$ denotes the inner product and δ_{ij} the Dirac delta function. If $\langle \mathbf{x}, \mathbf{e}_i \rangle = a_i$, the projection of vector $\mathbf{x}$ on basis $\mathbf{e}_i$, then we have

$$\langle \mathbf{x}, \mathbf{x} \rangle = \left\langle \sum_i a_i \mathbf{e}_i, \sum_j a_j \mathbf{e}_j \right\rangle = \sum_i \sum_j a_i a_j \langle \mathbf{e}_i, \mathbf{e}_j \rangle = \sum_i a_i^2. \tag{11}$$

Therefore we can conclude that a sufficient condition for the IBM to be optimal at the time frame level is orthonormal frequency decomposition.

2.3. Global level

For the entire target signal $x[n]$, we first consider STFT. $x[n]$ can be written as

$$x[n] = \sum_{m=0}^{M-1} x_m[n]/A[n], \tag{12}$$

where m is the frame index and M the number of frames. $A[n]$ is the normalization factor and $A[n] = \sum_{m=0}^{M-1} w[n - m\tau]$, where w is a window function with length N and τ is the frame shift. $x_m[n]$ is a windowed signal of $x[n]$ at frame m. Therefore $x_m[n] = 0$ for $n < m\tau$ and $n \geqslant m\tau + N$. Similarly we can write the entire estimated signal as

$$\hat{x}[n] = \sum_{m=0}^{M-1} \hat{x}_m[n]/A[n]. \tag{13}$$

Again $\hat{x}_m[n] = 0$ for $n < m\tau$ and $n \geqslant m\tau + N$.

The energy of the entire error signal is

$$\begin{aligned} \sum_n (\hat{x}[n] - x[n])^2 &= \sum_n \left(\sum_m \hat{x}_m[n]/A[n] - \sum_m x_m[n]/A[n] \right)^2 \\ &= \sum_n \frac{1}{A^2[n]} \left(\sum_m (\hat{x}_m[n] - x_m[n]) \right)^2 \\ &= \sum_n \frac{1}{A^2[n]} \left(\sum_m (\hat{x}_m[n] - x_m[n])^2 \right. \\ &\quad \left. + 2 \sum_{m_1} \sum_{m_2 > m_1} (\hat{x}_{m_1}[n] - x_{m_1}[n])(\hat{x}_{m_2}[n] - x_{m_2}[n]) \right), \end{aligned} \tag{14}$$

where m_1 and m_2 are frame indices.

If consecutive frames do not overlap, for a particular n, either $\hat{x}_{m_1}[n] - x_{m_1}[n]$ or $\hat{x}_{m_2}[n] - x_{m_2}[n]$ is zero. This is because a frame is zero outside of its corresponding window and $m_1 \neq m_2$. In this case, the cross terms in (14) do not contribute to the overall error energy and (14) becomes

$$\sum_n (\hat{x}[n] - x[n])^2 = \sum_n \frac{1}{A^2[n]} \sum_m (\hat{x}_m[n] - x_m[n])^2. \tag{15}$$

Assume $A[n]$ is constant for all n, we have

$$\sum_n (\hat{x}[n] - x[n])^2 = \frac{1}{A^2} \sum_m \sum_n (\hat{x}_m[n] - x_m[n])^2. \tag{16}$$

Note that in the above equation, the order of summation is also switched.

Since the IBM minimizes $\sum_n (\hat{x}_m[n] - x_m[n])^2$ for each frame m when DFT is used for frequency decomposition as discussed in Section 2.2, it also minimizes the energy of the entire error signal. Consequently, the IBM is optimal. Given non-overlapping consecutive frames, the window function must be rectangular in order for $A[n]$ to be constant.

If consecutive frames overlap, the cross terms also contribute to the overall energy of the error signal. In this case, a *T–F* unit couples with *T–F* units in the overlapping

frames. For example, if the overlap is 50%, it can be shown that a T–F unit will couple with every other T–F unit in the successive frame. It is in general difficult to quantify the contribution of the cross terms and compare it with the square terms. However, because of the nonlinearity in the SNR calculation, we suspect that IBM may not be optimal in the overlapping case. In the next subsection we will show that other binary masks can indeed give higher SNR in this case.

We have shown that for the IBM to be optimal when STFT is used for frequency decomposition, the consecutive frames have to be non-overlapping and the window is rectangular. However the requirements of non-overlapping frames and rectangular windowing are not critical for the IBM to be optimal. The key to optimality is that the T–F decomposition bases are orthonormal. For STFT, non-overlapping frames and rectangular windowing simply ensure that the T–F bases are orthonormal. STFT can be considered as applying DFT to each frame consecutively. This can be represented as matrix multiplication

$$\mathbf{X}_r = \mathbf{W}\mathbf{x} = \begin{bmatrix} \mathbf{W_1} & & & \mathbf{0} \\ & \mathbf{W}_2 & & \\ & & \ddots & \\ \mathbf{0} & & & \mathbf{W}_m \end{bmatrix} \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N_g] \end{bmatrix}, \tag{17}$$

where $\mathbf{W}_i$ is the DFT matrix applied to frame i, N_g is the length of the entire signal, and $\mathbf{X}_r$ is the reshaped version of $\mathbf{X}$: $\mathbf{X}_r$ is a column vector where the DFT of each frame is stacked together. Here the windowing is included in the DFT matrix. Rectangular windowing guarantees that the rows of $\mathbf{W}_i$ remain orthonormal. When consecutive frames do not overlap, $\mathbf{W}_i$ does not share columns with $\mathbf{W}_{i+1}$. As a result, the rows (also the columns) of $\mathbf{W}$ are orthonormal. Therefore, the IBM can be optimal using overlapping frames and non-rectangular windowing given that the decomposition matrix $\mathbf{W}$ is orthonormal. Modified discrete cosine transform (MDCT) is such an example (Vincent et al., 2007). It has an orthonormal bases while allowing overlap in consecutive frames. It has been shown that the IBM is optimal for MDCT (Vincent et al., 2007).

The following summarizes the main analytical result in the form of a theorem.

Theorem 1. *A sufficient condition for the ideal binary mask to be globally optimal is that the time–frequency decomposition is orthonormal.*

2.4. Counterexamples

In this section we show several counterexamples in which the IBM is not optimal. Note that it is not difficult to come up with such counterexamples. This suggests that the IBM is probably not optimal when the condition stated before, i.e., orthonormal T–F decomposition, is not satisfied. In all examples, signals are sampled at 20 kHz and are mixed to 0 dB SNR to create mixtures for analysis.

We first present a counterexample showing that the IBM is not optimal when a non-orthogonal gammatone filterbank is used for frequency decomposition. The gammatone filterbank has been widely used in CASA systems for frequency decomposition (Wang and Brown, 2006). The impulse response of a gammatone filter is

$$g[n] = \begin{cases} (nT)^{l-1}\exp(-2\pi bnT)\cos(2\pi fnT), & n \geqslant 0, \\ 0 & \text{otherwise}, \end{cases} \tag{18}$$

where T is the sampling interval, $l = 4$ is the order of the filter, b is the equivalent rectangular bandwidth (ERB), and f is the center frequency of the filter. Typically, a gammatone filterbank consists of 32 to 128 filters with f quasi-logarithmically spaced, based on the ERB-rate. The gammatone filterbank does not provide orthogonal frequency decomposition of a signal because the polyphase matrix of the gammatone filterbank is not paraunitary (Strang and Nguyen, 1996). Fig. 2 shows an example that the IBM is not optimal with the gammatone filterbank for a single frame. In this example, the gammatone filterbank has 128 channels and the center frequencies are linearly spaced from 50 to 8000 Hz on the ERB-rate scale. The top two panels show two musical signals with 2048 data points. The lower left is the IBM and the lower right is a binary mask obtained with a local SNR threshold (LC) (Brungart et al., 2006) of 1 dB, i.e., u_{cm} is labeled 1 if and only if $10\log_{10}\frac{\sum_n x_{cm}^2[n]}{\sum_n (\hat{x}_{cm}[n]-x_{cm}[n])^2} > 1$, where $x_{cm}[n]$ is the time-domain signal underlying u_{cm} and $\hat{x}_{cm}[n]$ is the estimate. In all the illustrations in this subsection, white indicates that a T–F unit is labeled 1 and black 0. The estimated signals are reconstructed from the two binary masks using a technique introduced by Weintraub (1985) (also see Wang and Brown, 2006). Since the resynthesis procedure is an integrated part of the gammatone filterbank-based analysis, we do not attempt to isolate its contribution to the SNR gain. In this case, the IBM gives a 7.0 dB SNR gain while the other binary mask gives a 7.3 dB SNR gain.

The second counterexample, illustrated in Fig. 3, shows that the IBM is not optimal when a non-rectangular window is used with STFT. In this example, consecutive frames do not overlap. The top two panels plot two musical signals. When a hamming window with a length of 512 samples is applied, the SNR gain of the IBM (lower left) is 3.97 dB while the SNR gain of a mask (lower right) with a LC of 0.4 dB is 3.99 dB. One of the noticeable differences between the two masks is indicated by a circle.

If consecutive frames overlap, the IBM may not be optimal even with a rectangular window when STFT is used. Fig. 4 shows such an example with the same musical signals as in Fig. 3. The frame length is 512 and the overlap is 50%. The SNR gain of the IBM (lower left) is 16.7 dB while the SNR gain for a mask obtained with a LC of 0.4 dB is

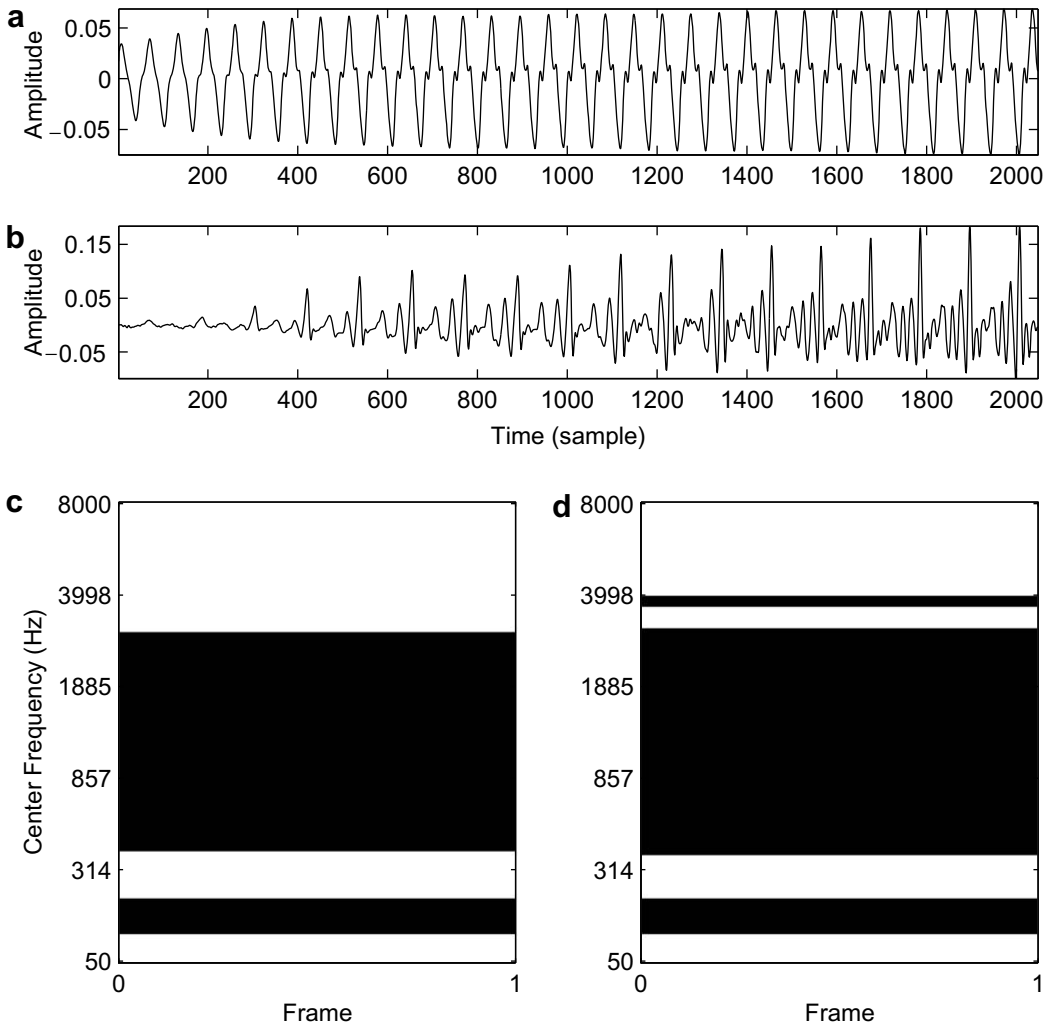

Fig. 2. An example showing that the IBM is not optimal when a gammatone filterbank is used for frequency decomposition for one frame. (a) The waveform of a target music signal. (b) The waveform of an interference music signal. (c) The IBM. (d) A mask generated with a local SNR threshold of 1 dB (see text).

16.9 dB (lower right). The circle marks one noticeable difference between the two masks. This example shows the effect of coupling between overlapping frames.

3. The ideal binary mask and the ideal ratio mask

Most sound separation systems decompose a signal into overlapping frames to reduce boundary effects caused by windowing. In this case, based on the discussion in Sections 2.3 and 2.4, the IBM may not be optimal. On the other hand, its SNR gain is close to that of ideal ratio masks (IRM). The IRM is defined as (Srinivasan et al., 2006)

$$R_{cm} = \frac{|\mathbf{X}_{cm}|^2}{|\mathbf{X}_{cm}|^2 + |\mathbf{Y}_{cm}|^2} \tag{19}$$

for each c and m. The IRM is closely related to the Wiener filter, the optimal linear filter in the minimum mean-square error sense (Wiener, 1949). Moreover, if a target signal, an interference signal, and their mixture are jointly Gaussian, the Wiener filter is the optimal filter among all possible filters, linear or nonlinear (van Trees, 1968). Additionally, given that the causality of a filter is not required and the target signal and the interference signal are uncorrelated, the Wiener filter amounts to the same ratio as (19) with spectral values replaced by power spectral densities (van Trees, 1968). The conditions for the Wiener filter to be a ratio mask can be satisfied in many cases: the non-causality of the filter can be allowed since most sound separation systems operate offline; the uncorrelatedness can also be assumed since sound sources are generally independent.

One can show that the IRM always leads to a local SNR gain no smaller than the IBM in the filterbank-based transform. For u_{cm}, consider three underlying signals: the target $x_{cm}[n]$, the interference $y_{cm}[n]$, and the mixture $z_{cm}[n]$. With linear frequency decomposition, $z_{cm}[n] = x_{cm}[n] + y_{cm}[n]$. The ratio mask can be defined using the energy of time-domain signals as

$$r = \frac{\sum_n x_{cm}^2[n]}{\sum_n x_{cm}^2[n] + \sum_n y_{cm}^2[n]}. \tag{20}$$

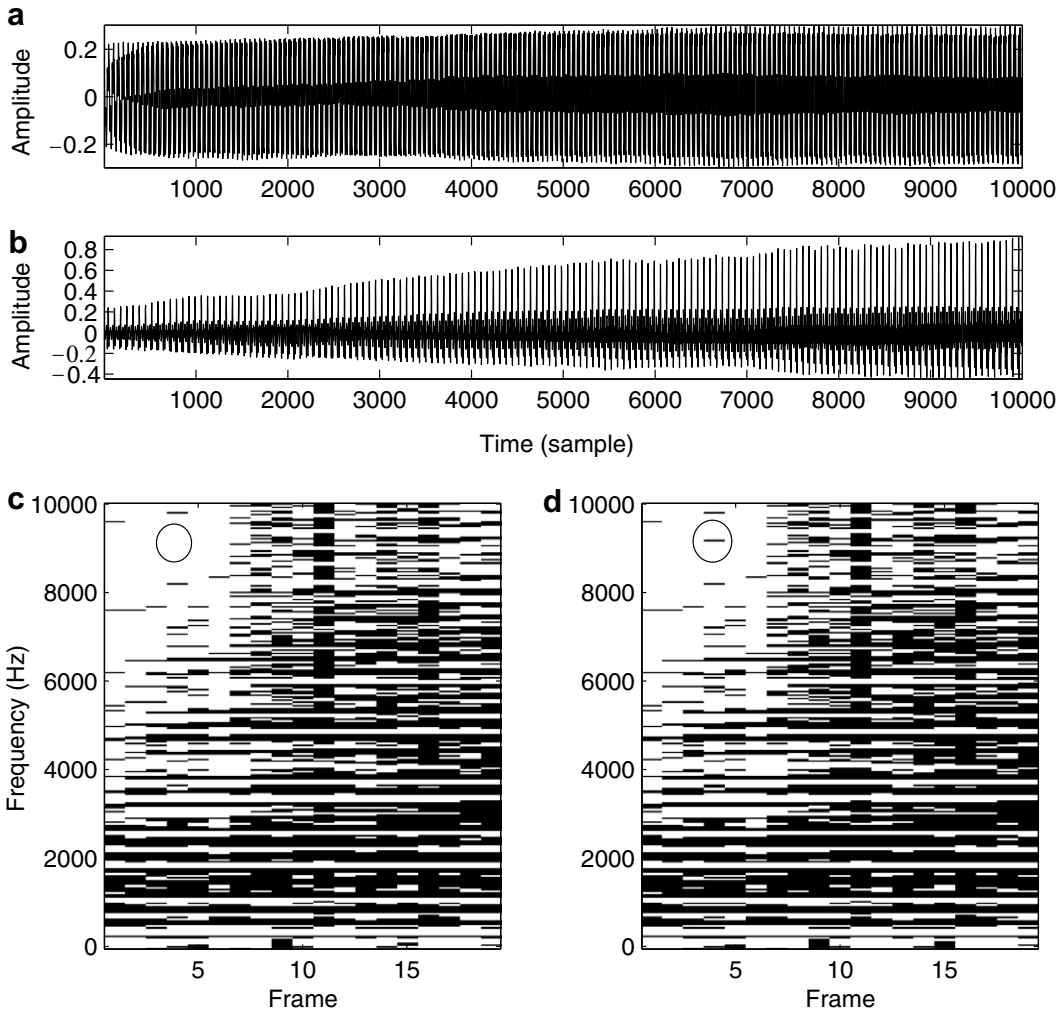

Fig. 3. An example showing that the IBM is not optimal when a hamming window is used for orthogonal *T–F* decomposition. (a) The waveform of a target music signal. (b) The waveform of an interference music signal. (c) The IBM. (d) A mask generated with a LC of 0.4 dB.

Denote $E = \sum_n x_{cm}^2[n] + \sum_n y_{cm}^2[n]$ and we have $\sum_n x_{cm}^2[n] = rE$ and $\sum_n y_{cm}^2[n] = (1-r)E$. Consider the case where $\sum_n x_{cm}^2[n] > \sum_n y_{cm}^2[n]$, the target stronger than the interference. When applying the IBM, the *T–F* unit is assigned 1 and $z_{cm}[n]$ is retained. In this case, the local SNR is

$$\mathrm{SNR} = 10\log_{10}\frac{\sum_n x_{cm}^2[n]}{\sum_n (z_{cm}[n] - x_{cm}[n]])^2}. \tag{21}$$

When the IRM is applied, the new SNR is

$$\mathrm{SNR}' = 10\log_{10}\frac{\sum_n x_{cm}^2[n]}{\sum_n (rz_{cm}[n] - x_{cm}[n])^2}. \tag{22}$$

Now compare the denominators in (21) and (22)

$$\begin{aligned}
&\sum_n (z_{cm}[n] - x_{cm}[n])^2 - \sum_n (rz_{cm}[n] - x_{cm}[n])^2 \\
&\quad = \sum_n y_{cm}^2[n] - \sum_n (ry_{cm}[n] + (r-1)x_{cm}[n])^2 \\
&\quad = (1-r^2)\sum_n y_{cm}^2[n] - 2r(r-1)\sum_n x_{cm}[n]y_{cm}[n] \\
&\qquad - (r-1)^2\sum_n x_{cm}^2[n] \\
&\quad = (1-r^2)(1-r)E - (r-1)^2 rE = (r-1)^2 E.
\end{aligned} \tag{23}$$

Note that in the above derivation we assume $\sum_n x_{cm}[n]y_{cm}[n] = 0$, which is roughly equivalent to uncorrelatedness between the two signals. Since $(r-1)^2E \geqslant 0, \sum_n (z_{cm}[n] - x_{cm}[n])^2 \geqslant \sum_n (rz_{cm}[n] - x_{cm}[n])^2$. This shows that compared to IBM, IRM gives an equal or smaller denominator and therefore the same or better SNR. The equal sign holds when $r = 1$, i.e., when the interference is absent at u_{cm}. Similarly we can show that when $\sum_n x_{cm}^2[n] \leqslant \sum_n y_{cm}^2[n]$, the IRM also achieves an SNR that is at least as good as the IBM. In this case, the equal sign holds when $r = 0$, i.e., when the target is absent at u_{cm}.

In the above discussion, we show that the IRM is locally no worse in terms of SNR compared to the IBM. However it is difficult to theoretically quantify the global difference between the two. We investigate this issue experimentally using mixtures of interest. In particular, we use a speech mixture database and a music database. The speech mixture database is collected by Cooke (1993), which includes different types of interference that are commonly encountered in real environments. It also has premixed target and interference, which makes the construction of the

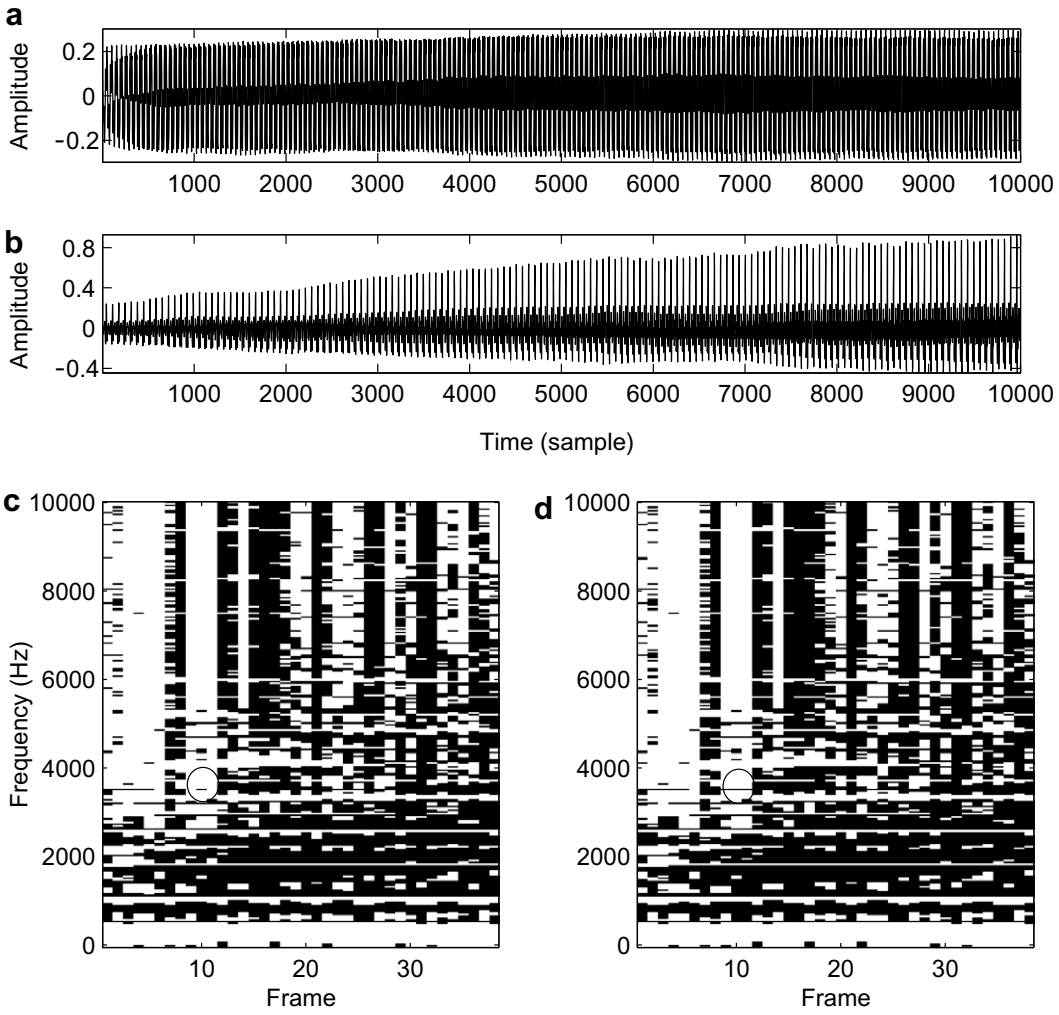

Fig. 4. An example showing that the IBM is not optimal when frames overlap even with a rectangular window. (a) The waveform of a target music signal. (b) The waveform of an interference music signal. (c) The IBM. (d) A mask generated with a LC of 0.4 dB.

IBM and the IRM possible. For music, we use a database constructed for musical sound separation. The database is synthesized from the tenor and the alto line of string quartets by J.S. Bach. Each line is constructed based on MIDI data using instrument samples from the RWC database (Goto et al., 2003). Details of synthesis can be found in (Li and Wang, 2007). For each database, we consider the SNR gain of the IBM and the IRM over two different kinds of frequency decomposition—DFT and the gammatone filterbank (GF) as described in Section 2.4. In each case, the frame length is 512 points and the frame overlap is 50%. The sampling frequency is 20 kHz. In the gammatone filterbank analysis, the filterbank has 64 filters with center frequencies equally spaced on the ERB-rate scale from 50 to 8000 Hz.

Table 1 shows the SNR gains of the IBM and the IRM in dB for the Cooke database with both DFT and GF. In the database, there are ten voiced utterances spoken by male and female speakers. There are 10 different types of interference: 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. The length of the signals is about 1–2 s. In the experiment, each utterance is mixed with each interference so that the overall SNR is 0 dB. Each number in Table 1 represents an average over 10 utterances for one type of interference. The average SNR gain over the whole database is listed in the bottom row of the table. It can be seen that the IRM gives a higher SNR gain in all cases. On the other hand, the SNR gains of the IBM are close to those of the IRM. When DFT is used for frequency decomposition, the SNR gain of the IBM is 0.7 dB lower than that of the IRM on average. With the gammatone filterbank, the difference is only 0.4 dB. The variance of the SNR gain difference is also small—the largest difference is 0.8 dB when the interference is female speech (N7) and DFT is used.

The SNR gains for the music database are shown in Table 2. In this case, we group the SNR gains according to instrument combinations. Four instruments, a clarinet

Table 1
SNR gain (in dB) of IBM and IRM for a speech mixture database

Interference	DFT		GF	
	IBM	IRM	IBM	IRM
N0	18.3	19.0	21.4	21.7
N1	12.2	12.9	11.3	12.0
N2	17.6	18.3	17.6	18.1
N3	7.8	8.5	7.6	7.8
N4	12.4	13.0	11.2	11.6
N5	18.7	19.4	19.7	19.9
N6	21.0	21.7	20.6	20.9
N7	13.9	14.7	12.4	12.8
N8	13.2	13.8	12.2	12.7
N9	9.7	10.4	9.9	10.1
Average	14.5	15.2	14.4	14.8

Table 2
SNR gain (in dB) of IBM and IRM for a music database

Instruments	DFT		GF	
	IBM	IRM	IBM	IRM
CL + FL	12.9	13.5	12.3	12.0
CL + VN	13.2	13.9	12.3	12.1
CL + TR	11.3	12.2	9.0	9.3
FL + VN	13.7	14.8	11.9	11.8
FL + TR	11.1	12.1	8.8	9.4
VN + TR	12.1	12.8	8.9	9.2
Average	12.4	13.2	10.5	10.6

(CL), a flute (FL), a violin (VN), and a trumpet (TR) are used to synthesize different music lines in the music database and there are six different combinations. It can be seen that the IRM gives higher SNR gains for all instrument combinations when DFT is used for frequency decomposition. For the gammatone filterbank, the IBM actually performs better in several instrument combinations. One possible reason is that the uncorrelatedness assumption does not hold well in music. In Western music, pitches in harmonic relation—pitches form a simple integer ratio (Hubbard and Datteri, 2001)—are favored. As a result, harmonics of different notes may collide. This is more likely with the gammatone filterbank since the bandwidth of the filters are wider in the high frequency range. Nonetheless, on average, the IRM gives a better SNR gain than that of the IBM. Similar to speech, the SNR gains between the IBM and the IRM are small. With DFT, the IBM is 0.8 dB worse while with the gammatone filterbank, the difference is only 0.1 dB.

In summary, the IRM achieves higher SNR gains compared to the IBM. However, despite the fact that the IBM is binary and the IRM is not, the SNR gain of the IBM is surprisingly close to that of the IRM. This shows that the IBM is a very reasonable performance metric for sound separation. Indeed, there are reasons to prefer the IBM over the IRM as the computational goal of a separation system. The estimation of the IBM is considerably simpler than that of the IRM: the former requires only binary decisions, while the latter requires estimating the energy ratio of the two signals. Binary estimation is facilitated by the existence of numerous classification and clustering methods.

4. Concluding remarks

In this paper we have addressed the optimality of the IBM in terms of SNR gain at three different levels and clarified the conditions at each level for the IBM to be optimal. At the T–F unit level, the IBM is optimal. At the time frame level, the IBM is optimal when the frequency decomposition is orthonormal. At the global level, IBM is optimal when the T–F decomposition is orthonormal. We give counterexamples where the IBM is not optimal when the stated conditions are not satisfied. In most practical applications, frames overlap, and as a result the IBM is not expected to be optimal. However we have shown experimentally that the performance of the IBM is close to that of the IRM and therefore the IBM is still a good objective for sound separation systems.

The analysis in this paper is given in terms of SNR because the purported optimality of the IBM has been expressed in SNR. It is worth pointing out that the SNR metric does not correlate directly with speech intelligibility or quality when the signal is speech (Lim and Oppenheim, 1979). Attempts have been made to use several metrics together to evaluate the performance of a separation system. For example, Wang and Brown (1999) use signal-to-interference ratio (SIR) along with the percentage of recovered signal energy. Vincent et al. (2006) suggest the use of SIR, SDR (signal-to-distortion ratio), and SAR (signal-to-artifact ratio) jointly. These three measures are calculated by projecting an estimated signal onto the subspaces expanded by the target and the interference. Although multiple metrics may provide a fuller picture, it is often difficult to compare two separation systems if one performs better in some metric and worse in others. The SNR measure produces a single number making it easy to gauge the performance of a separation system relative to others, and this is probably a main reason why SNR remains the most widely used performance metric despite its shortcomings.

Acknowledgements

We wish to thank the three anonymous reviewers for their constructive suggestions/criticisms. This research was supported in part by an AFOSR Grant (F49620-04-1-0027) and an AFRL Grant (FA8750-04-1-0093).

References

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

Brown, G.J., Cooke, M.P., 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 Press, Cambridge, UK.

Deshmukh, O.M., Espy-Wilson, C.Y., Carney, L.H., 2007. Speech enhancement using the modified phase-opponency model. J. Acoust. Soc. Amer. 121 (6), 3886–3898.

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. 115–146.

Goto, M., Hashiguchi, H., Nishimura, T., Oka, R., 2003. RWC music database: music genre database and musical instrument sound database. In: Internat. Conf. on Music Information Retrieval.

Harding, S., Barker, J., Brown, G.J., 2006. Mask estimation for missing data speech recognition based on statistics of binaural interaction. IEEE Trans. Audio, Speech, Lang. Process. 14 (1), 58–67.

Hu, G., 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.

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

Hubbard, T.L., Datteri, D.L., 2001. Recognizing the component tones of a major chord. Amer. J. Psychol. 114 (4), 569–589.

Kim, Y.-I., An, S.J., Kil, R.M., 2006. Zero-crossing based time–frequency masking for sound segregation. Neural Inform. Process. – Lett. Rev. 10, 125–134.

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., Wang, D.L., 2007. Pitch detection in polyphonic music using instrument tone models. In: IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing, pp. II.481–484.

Li, P., Guan, Y., Xu, B., Liu, W., 2006. Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech. IEEE Trans. Audio, Speech, Lang. Process. 14 (6), 2014–2023.

Lim, J.S., Oppenheim, A.V., 1979. Enhancement and bandwidth compression of noisy speech. Proc. IEEE 67 (12), 1586–1604.

Oppenheim, A.V., Schafer, R.W., Buck, J.R., 1999. Discrete-Time Signal Processing, second ed. Prentice-Hall.

Princen, J.P., Bradley, A.B., 1986. Analysis/synthesis filter bank design based on time domain aliasing cancellation. IEEE Transactions on Acoustics, Speech, and Signal Processing 34 (5), 1153–1161.

Radfar, M.H., Dansereau, R.M., Sayadiyan, A., 2007. A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation. EURASIP Journal on Audio, Speech, and Music Processing 2007, Article ID 84186, p. 15.

Reddy, A.M., Raj, B., 2007. Soft mask methods for single-channel speaker separation. IEEE Trans. Audio, Speech, Lang. Process. 25 (6), 1766–1776.

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

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

Strang, G., Nguyen, T., 1996. Wavelets and Filter Banks. SIAM, Philadelphia, PA.

van Trees, H.L., 1968. Detection, Estimation, and Modulation Theory, Part I. Wiley, New York.

Vincent, E., Gribonval, R., Fevotte, C., 2006. Performance measurement in blind audio source separation. IEEE Trans. Audio, Speech, Lang. Process. 14 (4), 1462–1469.

Vincent, E., Gribonval, R., Plumbley, M.D., 2007. Oracle estimators for the benchmarking of source separation algorithms. Signal Process. 87, 1933–1950.

Wang, D.L., 2005. On ideal binary masks as the computational goal of auditory scene analysis. In: Divenyi, P. (Ed.), Speech Separation by Humans and Machines. Kluwer Academic, Boston, MA, pp. 181–197.

Wang, D.L., Brown, G.J., 1999. Separation of speech from interfering sounds based on oscillatory correlation. IEEE Trans. Neural Networks 10 (3), 684–697.

Wang, D.L., Brown, G.J. (Eds.), 2006. Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Wiley/IEEE Press, Hoboken, NJ.

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

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