ON THE OPTIMALITY OF IDEAL BINARY TIME-FREQUENCY MASKS

Yipeng Li and DeLiang Wang

Department of Computer Science and Engineering
The Ohio State University
{liyip, dwang}@cse.ohio-state.edu

ABSTRACT

Recently the concept of ideal binary time-frequency masks has received attention and their optimality in terms of signal-to-noise ratio has been presumed. However the optimality is not rigorously analyzed. In this paper we treat this issue formally and clarify the conditions for ideal binary masks to be optimal. We also experimentally compare the performance of ideal binary masks in terms of signal-to-noise ratio 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.

Index Terms— Ideal binary mask, ideal ratio mask, optimality, sound separation, Wiener filter

1. INTRODUCTION

The human auditory system has the ability to segregate an acoustic mixture into perceptual streams that correspond to different sound sources. Bregman [1] proposed an influential theory, called *auditory scene analysis* (ASA), to explain this ability. Inspired by ASA, *computational auditory scene analysis* (CASA) attempts to build monaural and binaural systems that possess the same functionality [2]. An important idea in CASA systems developed is binary time-frequency (T-F) masking that is used to extract a target sound [2]. After an input is transformed to a T-F representation such as a spectrogram or a cochleagram, an element of such a representation, called a T-F unit, is assigned 1 if its energy is considered as from the target or 0 otherwise. Hu and Wang [3, 4] proposed a binary mask where a T-F unit is assigned 1 if in that unit target energy is stronger than interference energy and 0 otherwise. They called such a mask the *ideal binary mask* (IBM) since it represents the computational objective of their system and its construction requires premixing target and interference. The IBM dramatically improves the intelligibility of speech corrupted by noise [5, 6, 7]. Several CASA algorithms that directly estimate the IBM [5, 4] have produced good results in speech separation.

Signal-to-noise ratio (SNR) has been widely used as a performance measure in sound separation. For sound separation, it is defined as

$$\text{SNR} = 10\log_{10}\frac{\sum_n x^2[n]}{\sum_n(\hat{x}[n]-x[n])^2}, \qquad (1)$$

where $x[n]$ is a target signal and $\hat{x}[n]$ is the estimated target signal. 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 lowers the SNR in that unit. It has also been assumed that the IBM is globally optimal, i.e., the IBM produces an output with the highest SNR gain among *all binary masks*. Two arguments exist for the global optimality of the IBM. The argument by Hu and Wang [4] is based on the local optimality of the IBM. At each T-F unit, the IBM either maximally retains target energy or removes interference energy. Therefore it minimizes the sum of missing target energy that is discarded and interference energy that is retained, i.e., the denominator in (1). As a result, the IBM would achieve the highest SNR. The other argument by Ellis [8] is based on 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. These two arguments are flawed because the SNR calculation is nonlinear: the local optimality may not lead to the global optimality.

The concept of the IBM with its assumed global optimality has received attention recently. Many computational systems have used the IBM as a measure of ceiling performance for sound separation [9, 10, 11, 12, 13]. In this paper, we give a rigorous treatment on the optimality of the IBM. In Section 2 we consider the optimality of the IBM at three different levels: the T-F unit level, the time frame level, and the global level. We show that, at each level, the IBM can be optimal under certain conditions. We also give a counterexample showing that the IBM is not optimal when these conditions are violated. In Section 3 we compare SNR gains of the IBM to those of ideal ratio masks that are closely related to the Wiener filter. Section 4 concludes the paper.

2. THE OPTIMALITY OF THE IDEAL BINARY MASK AT DIFFERENT LEVELS

2.1. T-F Unit Level

Given a T-F decomposition, we consider $\mathbf{X}_{c,m}$ and $\mathbf{Y}_{c,m}$, the spectral values of a target signal and an interference signal at T-F unit u_{cm}, respectively. c is the frequency index and m the

frame index. At the T-F unit level, the definition of SNR in (1) should be changed slightly when spectral values instead of time-domain signals are used:

$$\text{SNR} = 10\log_{10}\frac{|\mathbf{X}_{c,m}|^2}{|\hat{\mathbf{X}}_{c,m}-\mathbf{X}_{c,m}|^2}, \tag{2}$$

where $\hat{\mathbf{X}}_{c,m}$ is the estimated spectral value of the target. According to the definition of the IBM,

$$\hat{\mathbf{X}}_{c,m} = \begin{cases} \mathbf{X}_{c,m}+\mathbf{Y}_{c,m}, & \text{if } |\mathbf{X}_{c,m}|^2 > |\mathbf{Y}_{c,m}|^2, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

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

$$|\hat{\mathbf{X}}_{c,m}-\mathbf{X}_{c,m}|^2 = |\mathbf{X}_{c,m}+\mathbf{Y}_{c,m}-\mathbf{X}_{c,m}|^2 = |\mathbf{Y}_{c,m}|^2. \tag{4}$$

On the other hand, if u_{cm} is assigned 0, the denominator is

$$|\hat{\mathbf{X}}_{c,m}-\mathbf{X}_{c,m}|^2 = |0-\mathbf{X}_{c,m}|^2 = |\mathbf{X}_{c,m}|^2. \tag{5}$$

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

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

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

If u_{cm} is assigned 1, then

$$|\hat{\mathbf{X}}_{c,m}-\mathbf{X}_{c,m}|^2 = |\mathbf{X}_{c,m}+\mathbf{Y}_{c,m}-\mathbf{X}_{c,m}|^2 = |\mathbf{Y}_{c,m}|^2. \tag{7}$$

Since $|\mathbf{X}_{c,m}|^2 \leqslant |\mathbf{Y}_{c,m}|^2$, the IBM yields a denominator no greater than its alternative. Based on the above discussion, the IBM always minimizes the denominator among binary masks and consequently maximizes the SNR. Therefore we can conclude that the IBM is optimal at the T-F unit level.

2.2. Time Frame Level

Now consider $x_m[n]$, the one-frame time-domain target signal. Without loss of generality, we assume that the index of n is from 0 to $N-1$. Denote the time-domain estimate of $x_m[n]$ as $\hat{x}_m[n]$. The SNR of $\hat{x}_m[n]$ with respect to $x_m[n]$ can be calculated using (1) with summation of n from 0 to $N-1$. It is clear from (1) 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 [14], we have

$$\sum_{n=0}^{N-1}(\hat{x}_m[n]-x_m[n])^2 = \frac{1}{N}\sum_{c=0}^{N-1}|\hat{\mathbf{X}}_{c,m}-\mathbf{X}_{c,m}|^2, \tag{8}$$

where $\mathbf{X}_{c,m}$ is the spectral value of the target at frequency c and $\hat{\mathbf{X}}_{c,m}$ is the estimate of $\mathbf{X}_{c,m}$.

In Section 2.1, we have shown that the IBM minimizes $|\hat{\mathbf{X}}_{c,m}-\mathbf{X}_{c,m}|^2$ for each c. Therefore the IBM also minimizes the summation $\sum_{c=0}^{N-1}|\hat{\mathbf{X}}_{c,m}-\mathbf{X}_{c,m}|^2$. As a result, the IBM yields the highest SNR among all binary masks. Since the Parseval's theorem holds for any orthogonal decomposition, we can conclude that a sufficient condition for the IBM to be optimal at the time frame level is that frequency decomposition is orthogonal.

2.3. Global Level

Now consider the entire target signal $x[n]$, which is processed frame by frame. We can write $x[n]$ as

$$x[n] = \frac{1}{A[n]}\sum_{m=0}^{M-1}x_m[n], \tag{9}$$

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

$$\hat{x}[n] = \frac{1}{A[n]}\sum_{m=0}^{M-1}\hat{x}_m[n]. \tag{10}$$

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

After straightforward derivation, the energy of the entire error signal is

$$\begin{aligned} &\sum_n(\hat{x}[n]-x[n])^2 \\ =&\sum_n\frac{1}{A^2[n]}(\sum_m(\hat{x}_m[n]-x_m[n])^2+ \\ &2\sum_{m_1}\sum_{m_2>m_1}(\hat{x}_{m_1}[n]-x_{m_1}[n])(\hat{x}_{m_2}[n]-x_{m_2}[n])). \end{aligned} \tag{11}$$

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 (11) do not contribute to the overall error energy and (11) becomes

$$\sum_n(\hat{x}[n]-x[n])^2 = \sum_n\frac{1}{A^2[n]}\sum_m(\hat{x}_m[n]-x_m[n])^2. \tag{12}$$

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

$$\sum_n(\hat{x}[n]-x[n])^2 = \frac{1}{A^2}\sum_m\sum_n(\hat{x}_m[n]-x_m[n])^2. \tag{13}$$

Note that the order of summation is also switched in (13). Since the IBM minimizes $\sum_n(\hat{x}_m[n]-x_m[n])^2$ for each frame m as discussed in Section 2.2, it also minimizes the energy of the entire error signal. Consequently, the IBM is optimal

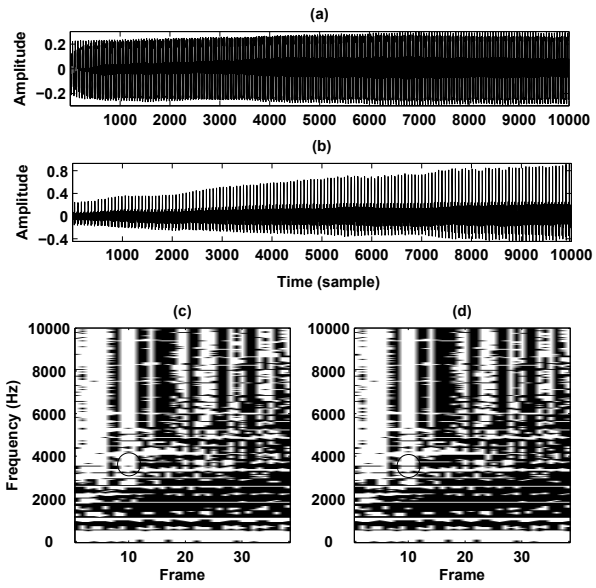

Fig. 1. An example showing that the IBM is not optimal when frames overlap even with a rectangular window. (a). The waveform of a target music signal. (b). The waveform of an interference music signal. (c). The IBM. (d). A mask generated with a local SNR threshold of 0.4 dB. In the masks white indicates 1 while black 0.

among binary masks. For $A[n]$ to be constant the window function w must be rectangular. Non-overlapping windowing can be considered as an orthogonal decomposition of a signal in the time domain. Therefore we conclude that *a sufficient condition for the IBM to be optimal at the global level is orthogonal T-F decomposition with a rectangular window*.

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.

2.4. A Counterexample

We use an example to illustrate that the IBM is not optimal when frames overlap. The top two panels in Fig. 1 plot the waveforms of two musical signals sampled at 20 kHz. The two signals are mixed to 0 dB and the first one is chosen as the target. In T-F decomposition, we use a frame length of 512 samples with 50% overlapping and at each frame we apply DFT. The windowing function is rectangular. The lower left plot in Fig. 1 is the IBM while the lower right is a mask generated with a local SNR threshold of 0.4 dB, i.e., u_{cm} is labeled 1 if and only if $10\log_{10}\frac{|\mathbf{X}_{c,m}|^2}{|\mathbf{Y}_{c,m}|^2} > 0.4$. The circle marks one noticeable difference between the two masks. The SNR gain of the IBM is 16.7 dB while the SNR gain for the other mask is 16.9 dB. Therefore the IBM is not optimal.

3. THE IDEAL BINARY MASK AND THE IDEAL RATIO MASK

Since most sound separation systems decompose a signal into overlapping frames to reduce boundary effects caused by windowing, the IBM may not be optimal. But we find that its SNR gains are actually close to those of ideal ratio masks (IRM). The IRM is defined as [15]

$$R_{c,m} = \frac{|\mathbf{X}_{c,m}|^2}{|\mathbf{X}_{c,m}|^2 + |\mathbf{Y}_{c,m}|^2} \tag{14}$$

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 [16]. If non-causality is allowed and a target signal is uncorrelated with an interference signal, the Wiener filter amounts to the same ratio as (14) with spectral values replaced by power spectral densities [17]. The conditions for the Wiener filter to be a ratio mask are satisfied in most cases since most sound separation systems operate offline and sound sources are generally independent.

Although one can show that the IRM always achieves a local SNR gain no smaller than the IBM, it is difficult to theoretically quantify the global difference between the two. We investigate this issue experimentally using a speech mixture database and a music database. The speech mixture database is collected by Cooke [18], which includes different types of interference that are commonly encountered in real environment. The music database is constructed from Bach's works, each of which has two lines and each line of music is synthesized using instrument samples from the RWC database [19]. Targets and interference are mixed to 0 dB.

Table 1 shows the SNR gain for two databases and two frequency decomposition methods. GF represents the gammatone filterbank which has been widely used in CASA [2]. The parameters used for DFT are mentioned in Section 2.4 while the parameters used for GF can be found in [4] except that the frame length is 512 samples and the frequency range is set to 50 to 8000 Hz. We can see from the table that in all cases the SNR gain of the IRM is higher. On the other hand, the SNR gain of the IBM is close to that of the IRM, particularly for gammatone decomposition. This indicates that, although the IBM is not optimal, it still gives a very reasonable performance metric for sound separation.

4. CONCLUSION

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. For the IBM to be globally optimal, the T-F decomposition should be orthogonal and the windowing function rectangular. Our experimental results have also shown that the performance of

Database	DFT		GF	
	IBM	IRM	IBM	IRM
Speech	14.5	15.2	14.4	14.8
Music	12.4	13.1	10.5	10.6

Table 1. SNR gains of IBM and IRM

the IBM is close to that of the IRM. Therefore the IBM is still a good objective for sound separation. Note that IBM estimation, unlike IRM estimation, requires only binary decisions, which makes applicable a host of classification and clustering methods.

In our discussion of the optimality of the IBM, we treat a signal deterministically, i.e., we do not consider the statistical properties of a signal such as stationarity. We believe that our treatment is more appropriate because it makes no assumption about signals. For example, if signals are treated statistically, the energy of the error signal has to be replaced by the expectation of the energy. In this case, we have $E(\sum_n (x_m[n] - \hat{x}_m[n])^2) = E(\sum_c |\mathbf{X}_{c,m} - \hat{\mathbf{X}}_{c,m}|^2)$. To proceed, i.e., to switch the expectation and summation, one has to assume that error terms of different c are statistically independent. However such an assumption is difficult to justify.

Acknowledgments. This research was supported in part by an AFOSR grant (F49620-04-1-0027) and an NSF grant (IIS-0534707).

5. REFERENCES

[1] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

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

[3] G. 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*, 2001.

[4] ——, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1135–1150, 2004.

[5] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *Journal of the Acoustical Society of America*, vol. 114, no. 4, pp. 2236–2252, 2003.

[6] M. C. Anzalone, L. Calandruccio, K. A. Doherty, and L. H. Carney, "Determination of the potential benefit of time-frequency gain manipulation," *Ear and Hearing*, vol. 27, no. 5, pp. 480–492, 2006.

[7] D. Brungart, P. S. Chang, B. D. Simpson, and D. L. Wang, "Isolating the energetic component of speech-on-speech masking with an ideal binary time-frequency mask," *Journal of the Acoustical Society of America*, vol. 120, pp. 4007–4018, 2006.

[8] D. P. W. Ellis, "Model-based scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Application*, D. L. Wang and G. J. Brown, Eds. Hoboken, NJ: Wiley/IEEE Press, 2006.

[9] S. Harding, J. Barker, and G. J. Brown, "Mask estimation for missing data speech recognition based on statistics of binaural interaction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 58–67, 2006.

[10] Y.-I. Kim, S. J. An, and R. M. Kil, "Zero-crossing based time-frequency masking for sound segregation," *Neural Information Processing - Letter & Review*, vol. 10, pp. 125–134, 2006.

[11] P. Li, Y. Guan, B. Xu, and W. Liu, "Monaural speech separation based on computational auditory scene analysis and objective quality assessment of speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2014–2023, 2006.

[12] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "A maximum likelihood estimation of vocal-tract-related filter characteristics for single channel speech separation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, pp. Article ID 84 186, 15 pages, 2007, doi:10.1155/2007/84186.

[13] A. M. Reddy and B. Raj, "Soft mask methods for single-channel speaker separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1766–1776, 2007.

[14] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Prentice Hall, 1999.

[15] S. Srinivasan, N. Roman, and D. L. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, pp. 1486–1501, 2006.

[16] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Cambridge, MA: MIT Press, 1949.

[17] H. L. van Trees, *Detection, Estimation, and Modulation Theory, Part I*. New York: Wiley, 1968.

[18] M. P. Cooke, *Modeling Auditory Processing and Organization*. Cambridge, U. K.: Cambridge University Press, 1993.

[19] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *International Conference on Music Information Retrieval*, 2003.