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.