A CASA-Based System for Long-Term SNR Estimation

Arun Narayanan, *Student Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

***Abstract*—We present a system for robust signal-to-noise ratio (SNR) estimation based on computational auditory scene analysis (CASA). The proposed algorithm uses an estimate of the ideal binary mask to segregate a time–frequency representation of the noisy signal into speech dominated and noise dominated regions. Energy within each of these regions is summated to derive the *filtered* global SNR. An SNR transform is introduced to convert the estimated filtered SNR to the true broadband SNR of the noisy signal. The algorithm is further extended to estimate subband SNRs. Evaluations are done using the TIMIT speech corpus and the NOISEX92 noise database. Results indicate that both global and subband SNR estimates are superior to those of existing methods, especially at low SNR conditions.**

***Index Terms*—Computational auditory scene analysis (CASA), broadband SNR, ideal binary mask (IBM), signal-to-noise ratio (SNR), subband SNR.**

I. Introduction

ESTIMATION of the signal-to-noise ratio has been studied for decades, mostly in the context of noise estimation and speech enhancement. Typical algorithms estimate local or instantaneous SNR, i.e., the SNR at a particular time–frequency (T-F) unit (also referred to as short-time subband SNR) [23], which can then be directly used by speech enhancement algorithms [2]. Two assumptions made by most algorithms are: 1) the background noise is stationary, at least between speech pauses and during the time interval when the noise energy is estimated (or updated) and 2) regular speech pauses occur in speech. For the estimation to be effective, the interval size should be chosen wisely. Longer intervals are suited for tracking stationary background noises. When noise statistics change quickly, a shorter interval is preferred. But using a shorter interval reduces the chance of seeing noise-only frames. Recent techniques relax some of the above assumptions to deal with non-stationary noise types [21], [10]. In realistic noise conditions, such as the so called *cocktail-party* condition, most techniques falter [26].

Manuscript received November 10, 2011; revised March 12, 2012, May 15, 2012; accepted June 01, 2012. Date of publication June 19, 2012; date of current version August 24, 2012. This work was supported in part by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-08-1-0155. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sharon Gannot.

A. Narayanan is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: narayaar@cse.ohio-state.edu).

D. L. Wang is with the Department of Computer Science and Engineering and Center for Cognitive Science, The Ohio State University, Columbus, OH 43210-1277 USA (e-mail: dwang@cse.ohio-state.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2012.2205242

While most algorithms perform short-time subband SNR estimation, knowledge of the SNR at other levels is also useful. Global SNR of an utterance, for instance, can be used to devise SNR specific speech and speaker recognition strategies [8], [32]. In many applications, speech processing algorithms are optimized to function in certain specific SNR conditions. An SNR estimator can be used in such applications during the model selection process at runtime. Similarly, subband SNR estimates are useful in many speech processing tasks.

The main theme of this paper is to estimate broadband and subband global SNRs, i.e., SNRs at the utterance level. Typical utterance length is between 2–5 seconds (e.g., the utterances in the TIMIT core test set [9]). Traditional SNR estimation algorithms have difficulties dealing with such long intervals of speech when the underlying noise is non-stationary. Algorithms have been proposed for global broadband SNR estimation. They are based on identifying the noise and speech energy distributions [1], [5], or signal statistics [17].

We take a CASA-based approach for SNR estimation. A main goal of CASA is to estimate the ideal binary mask (IBM) [30], which identifies speech dominated and noise dominated units in a T-F representation of noisy speech. The IBM has been shown to be effective in improving speech intelligibility and robust automatic speech and speaker recognition in noise [31]. Motivated by this line of research, we investigate whether the IBM can be used to calculate broadband and subband SNRs. Although IBM estimation algorithms are commonly based on short-time SNR estimation [15], [22], few have used the IBM to estimate the global SNR of mixture signals. The proposed algorithm works under the assumption that at the utterance level, the total speech and noise energy can be well approximated using only the speech dominant and the noise dominant T-F units, respectively.

The remainder of the paper is organized as follows. In Section II we discuss existing SNR estimation strategies from the literature. A detailed description of our system is provided in Section III. Evaluation results are described in Section IV. We conclude with a discussion of our results in Section V.

II. Prior Work

We first discuss short-time subband SNR estimation algorithms. Herein, estimation of the noise level is an important subproblem and has been widely studied. Early methods include the spectral histogram based method of Hirsch [11], and the low-energy envelope tracking method of Martin [23]. Other strategies for SNR estimation include energy clustering to distinguish speech and noise portions of the mixture [28], [5], and explicit speech pause or voice-activity detection (VAD) [19]. Nemer *et al.* [25] make use of higher order statistics of speech and

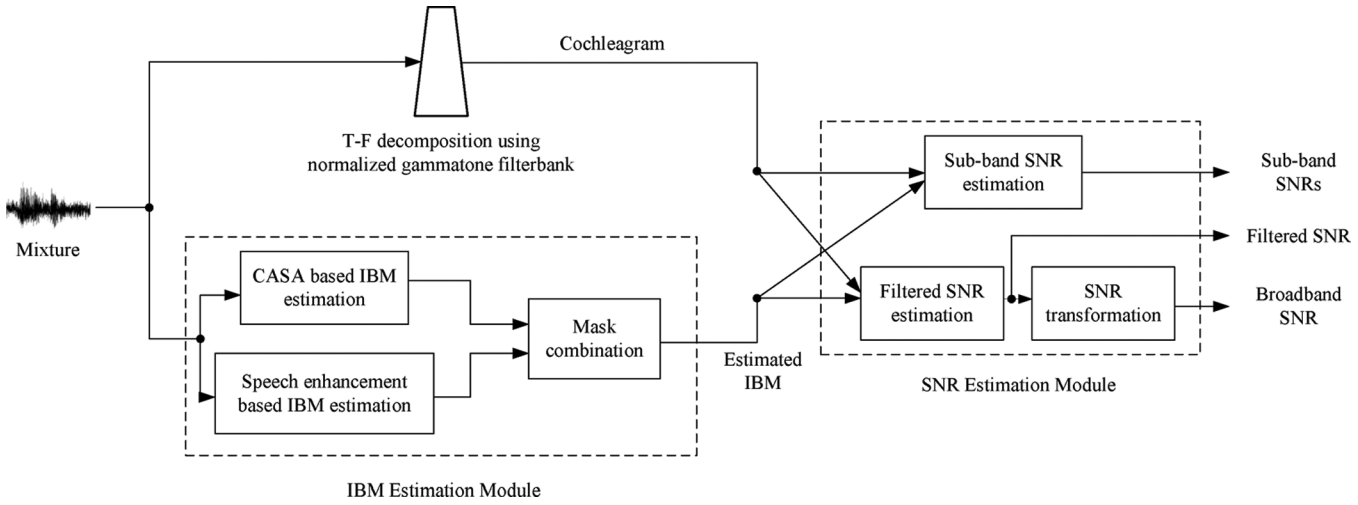

Fig. 1. Schematic diagram of the proposed system. The input to the system is a noisy mixture. The outputs are the broadband SNR, filtered SNR and subband SNRs. The system includes an IBM estimation module and an SNR estimation module.

noise, assuming a sinusoidal model for band restricted speech and a Gaussian model for noise. Supervised classification based methods have also been applied to this task. For example, features inspired from psychoacoustics and an MLP based classifier are used in [27], [18] to estimate broadband and subband SNRs in short intervals. The *a priori* SNR, which is the ratio of the speech and noise power, is widely used in speech enhancement algorithms and is typically estimated using the decision-directed approach of Ephraim and Malah [6]. Alternative techniques are based on GARCH models [4] and cepstro-temporal smoothing [3].

Global SNR estimation has also been studied, although not as widely as short-time subband SNR estimation. A commonly used algorithm from NIST [1] builds a histogram of short-time signal power using the noisy utterance, which is used to infer noise and noisy speech distributions. From these distributions, the peak signal-to-noise ratio is calculated rather than the mean SNR. The peak SNR is clearly an overestimate of the true SNR. Dat *et al.* [5] use a similar approach, but instead of fitting the histogram, they fit a 2-component Gaussian to the data using the expectation maximization (EM) algorithm. A similar approach was also used in [28] to model speech. Dat *et al.* extended the idea by using the learned Gaussians in a principled way to derive the SNR of the signal. Similar to [1], their approach would have problems when the bimodal Gaussian assumption fails. The method by Kim and Stern [17] is based on the waveform amplitude distribution. It assumes that clean and noisy speech have Gamma distributions, and noise a Gaussian distribution. It infers the global SNR based on the parameter of the distribution estimated from noisy speech. Their algorithm works well when these assumptions are met. Performance degradation occurs at low SNR conditions and when the background noise has non-Gaussian characteristics. An alternative, relatively straightforward approach would be to use speech enhancement algorithms to estimate the noise power spectral density (PSD) [10] and the squared-magnitude of speech in the DFT domain [7]. Assuming that the noise PSD approximates noise energy, which is reasonable, both global broadband and subband SNRs can be directly calculated by summating these estimates across time.

Long-term subband SNR estimation is not much studied, but global SNR estimation algorithms can be extended to perform subband SNR estimation. NIST [1] provides a subband SNR estimation algorithm based on the same principle as broadband SNR estimation. It is fairly easy to extend the methods in [17] and [5], and speech enhancement based strategies to perform subband SNR estimation. A supervised approach was proposed by Kleinschmidt and Hohmann [18]. Being supervised, the algorithm is likely dependent on training conditions.

A system related to ours is the one described in [14] (referred to as the Hu-Wang system). It estimates the SNR using a binary mask for only the voiced speech frames, by making the following assumptions: 1) the total voiced speech energy is approximately equal to the total noisy signal energy under the unmasked, speech dominant (1 s in the voiced IBM) T-F units, 2) the total signal energy can be inferred from the total voiced signal energy, and 3) the per-frame noise energy in both voiced and unvoiced frames remains unchanged. Their system produces reasonable results at SNRs close to 0 dB but biased estimates at other conditions. Since only the voiced IBM is used, estimating subband SNRs will be challenging, especially at high frequencies. In addition to providing a novel framework for SNR estimation, our algorithm differs from the Hu-Wang system since we use an estimate of the IBM in both voiced and unvoiced time frames.

III. System Description

The architecture of the proposed system is shown in Fig. 1. The input to the system is a noisy speech signal, which is first processed using a 128-channel gammatone filterbank to perform T-F decomposition. The center frequencies of the filterbank are uniformly spaced in the ERB (Equivalent Rectangular Bandwidth) rate scale from 50 Hz to 8000 Hz [31]. The signals are sampled at 16 kHz in our experiments, and the chosen frequency range ensures that almost all useful speech information is retained in the filtered signal. A typical gammatone filterbank performs loudness equalization across frequencies to match cochlear filtering. As a result, different frequency components are scaled differently. This may alter the SNR of the

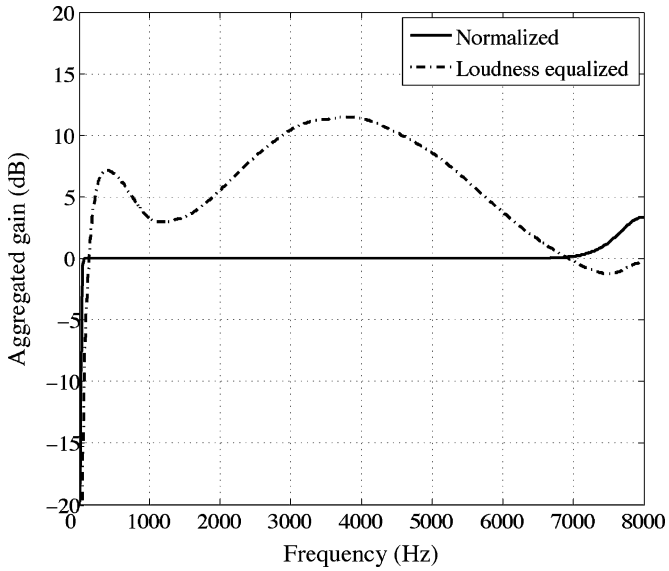

Fig. 2. Aggregated magnitude response of the normalized and loudness equalized gammatone filterbank. The gain for a specific frequency is calculated by aggregating the gains across the 128 filters of the filterbank. Notice that most frequency components undergo no attenuation/amplification when processed using the normalized gammatone filterbank.

filtered signal compared to the original signal in the time domain, even if the signal is band limited to 50–8000 Hz. In order to prevent this undesired effect, we normalize the gammatone filterbank. The normalized filterbank scales the frequency components covered by the filterbank so as to ensure that for speech signals, the filtered signal energy approximately equals its total time-domain energy. This may not be the case for noise and noisy speech signals if the underlying noise has significant energy in the low-frequency range (e.g., the car interior noise from the NOISEX92 corpus [29]). We will make use of the normalized filterbank in the subsequent SNR transformation step to estimate the true broadband SNR of a noisy signal, given its filtered SNR. Fig. 2 compares the aggregated magnitude responses of the conventional gammatone filterbank and the normalized gammatone filterbank.

After T-F decomposition, the filtered signal is windowed using a 20 msec rectangular frame with a 10 msec frame shift. A *cochleagram* [31] of the signal is then created by calculating the signal energy within each of these windows. Because of the 50% overlap between adjacent frames, the total energy within the cochleagram will roughly be twice the energy of the speech signal in the time domain.

Let $y(t)$, $x(t)$ and $n(t)$ represent the noisy, clean and noise signals, respectively, and $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{N}$ their corresponding cochleagrams. Noise is assumed to be additive in nature and independent of speech:

$$y(t) = x(t) + n(t).$$

Here, t denotes a time sample. We define the following SNRs:

$$\mathrm{SNR}_b = 10\log_{10}\left(\frac{\sum_t (x(t))^2}{\sum_t (n(t))^2}\right), \tag{1}$$

$$\mathrm{SNR}_f = 10\log_{10}\left(\frac{\sum_{m,c}\mathbf{X}(m,c)}{\sum_{m,c}\mathbf{N}(m,c)}\right), \tag{2}$$

$$\mathrm{SNR}_c = 10\log_{10}\left(\frac{\sum_{m}\mathbf{X}(m,c)}{\sum_{m}\mathbf{N}(m,c)}\right), \tag{3}$$

where SNR_b, SNR_f and SNR_c denote the broadband SNR, the filtered SNR and the subband SNR, respectively. m indexes a time frame and c a frequency channel. Each of the three SNRs can be useful depending on the target application. The goal of the proposed algorithm is to estimate these SNRs.

Since we only have access to $y(t)$ and $\mathbf{Y}$ in practice, to calculate these SNRs, we approximate the total target speech and noise energy using $\mathbf{Y}$ and an estimated IBM. The IBM is a two-dimensional binary matrix, with the same dimensionality as $\mathbf{Y}$. An element in the matrix takes the value 1 if the speech energy within the corresponding T-F unit is greater than the noise energy. Formally, the IBM is defined as:

$$\mathrm{IBM}(m,c) = \begin{cases} 1 & \text{if } \mathbf{X}(m,c) > \mathbf{N}(m,c) \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

Note that the IBM can also be defined in terms of a local SNR threshold at each T-F unit called the local criterion (LC). The above formulation implies an LC of 0 dB. Under certain conditions, the IBM obtained using an LC of 0 dB is the optimal binary mask in terms of SNR gain [20]. Given an estimated IBM and the cochleagram of the input signal, the SNRs are estimated by the SNR estimation module (Fig. 1) in the proposed system. This module is described in detail in the following subsection. IBM estimation itself is an important problem and is discussed in Section III.B.

A. SNR Estimation

For SNR estimation, we assume that the total target energy, both at the broadband and the subband level, can be estimated using *only* the speech dominant T-F units and the total filtered noise energy from the noise-dominant T-F units. As shown in the evaluations, this assumption is reasonable for long-term SNR estimation.

1) Global SNR Estimation: Given an estimated IBM ($\mathbf{M}$), the total speech and noise energy are estimated as follows:

$$\hat{E}_{\mathrm{speech}} = \sum_{m,c}\mathbf{Y}(m,c)\cdot\mathbf{M}(m,c), \tag{5}$$

$$\hat{E}_{\mathrm{noise}} = \sum_{m,c}\mathbf{Y}(m,c)\cdot\neg\mathbf{M}(m,c), \tag{6}$$

where '$\cdot$' and '$\neg$' denote the pointwise multiplication and 'NOT' operations, respectively. The filtered SNR ($\widehat{\mathrm{SNR}}_f$) is then estimated as shown below, using these estimates:

$$\widehat{\mathrm{SNR}}_f = 10\log_{10}\left(\frac{\hat{E}_{\mathrm{speech}}}{\hat{E}_{\mathrm{noise}}}\right). \tag{7}$$

The true broadband SNR is estimated by transforming $\widehat{\mathrm{SNR}}_f$ using an SNR transformation step. We transform the SNR based on the following observation. Recall that when speech signals are processed using the normalized gammatone filterbank, the total signal energy is not significantly altered since it applies a unit gain to most of the useful bands. Therefore, the difference between the energy of the noisy signal in the time domain and its energy after T-F decomposition using the normalized gammatone filterbank can mostly be attributed to noise. This is true

especially at low SNRs, where noise energy is comparable to or greater than the target energy. With this observation, the true broadband SNR can be calculated by compensating the noise energy with this difference during SNR estimation:

$$\triangle\hat{E} = 2\sum_{t}(y(t))^2 - \sum_{m,c}\mathbf{Y}(m,c), \tag{8}$$

$$\widehat{\mathrm{SNR}}_b = 10\log_{10}\left(\frac{\hat{E}_{\mathrm{speech}}}{\hat{E}_{\mathrm{noise}} + \max(0, \triangle\hat{E})}\right). \tag{9}$$

$\widehat{\mathrm{SNR}}_b$ is the estimated broadband SNR of the noisy signal. Note that, $\triangle\hat{E}$ compensates for the low frequency noise energy that gets *attenuated* by the filterbank. The implications of the approximation in (8) to compensate the total noise energy are described in Section IV.C.

2) Subband SNR Estimation: The subband SNRs are estimated similar to (7), but the energy values are summated only across time:

$$\widehat{\mathrm{SNR}}_c = 10\log_{10}\left(\frac{\sum_m \mathbf{Y}(m,c)\cdot\mathbf{M}(m,c)}{\sum_m \mathbf{Y}(m,c)\cdot\neg\mathbf{M}(m,c)}\right). \tag{10}$$

$\widehat{\mathrm{SNR}}_c$ denotes the estimated subband SNR for frequency channel c.

B. IBM Estimation

We consider three methods for IBM estimation. The first one is based on a recent CASA based IBM estimation algorithm described in [15]. The second one is based on the state-of-the-art speech enhancement algorithms in [7], [10]. With the goal of generalization to different test conditions, the final method combines the CASA and speech enhancement methods to estimate the IBM.

1) CASA Based IBM Estimation: The CASA algorithm in [15] uses the tandem algorithm [13] to estimate the voiced IBM (the IBM in voiced frames) and a spectral subtraction based method to estimate the unvoiced IBM. The tandem algorithm is an iterative procedure that estimates both the target pitch and the corresponding binary mask for up to two voiced sound sources in the signal. The algorithm does not link disjoint pitch contours, which is the task of sequential organization. Since we only deal with non-speech noise, multiple pitch points are typically detected only for a fraction of frames. In this work, sequential organization is performed based on: 1) plausible pitch range of speech, 2) length of a pitch contour, and 3) pitch continuity. The binary masks corresponding to the sequentially grouped pitch contours are then grouped to obtain an estimate of the voiced IBM. The algorithm estimates the unvoiced IBM by first removing periodic components from the mixture signal. It then forms a noise estimate for each unvoiced interval by averaging the energy within the noise dominant T-F units (0 s in the mask) of its neighboring voiced intervals. These estimates are finally used in spectral subtraction to obtain the estimated unvoiced IBM. Fig. 3(d) shows an estimated IBM obtained in this fashion. It captures most of the voiced segments (T-F regions) and a good number of unvoiced segments. Comparing with the IBM shown in Fig. 3(c), we can see that it still misses a few target-dominant segments.

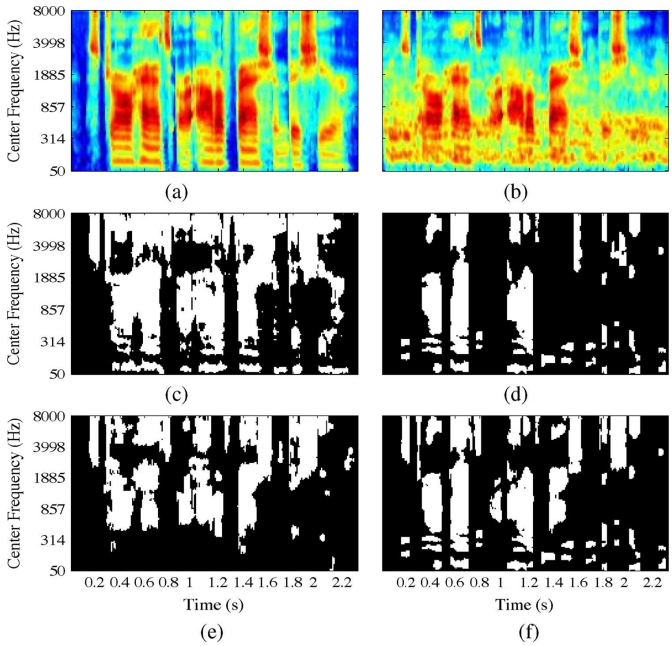

Fig. 3. IBM estimation. (a) Cochleagram of the utterance 'Straw hats are out of fashion this year' from the core test set of the TIMIT corpus. (b) Cochleagram of the same utterance mixed with babble noise; the filtered SNR is set to 5 dB. (c) The IBM. (d) The mask estimated by the CASA system described in Section III.B.1. (e) The mask estimated by the speech enhancement method described in Section III.B.2. (f) The mask obtained by combining the two methods.

2) Speech Enhancement Based IBM Estimation: The speech enhancement mask estimation is based on a state-of-the-art noise tracking algorithm described in [10]. The algorithm operates in the linear frequency domain, using the FFT to perform T-F decomposition. To estimate the noise PSD, it uses an MMSE estimator of noise magnitude-squared DFT coefficients assuming that both speech and noise DFT coefficients follow a complex-Gaussian distribution. The squared-magnitudes of the speech DFT coefficients are estimated using the algorithm in [7], which assumes that speech magnitude-DFT coefficients follow a generalized Gamma distribution with parameters $\gamma = 1$ and $\nu = 0.6$. The algorithms use the decision-directed method [6] to estimate the *a priori* SNR at each T-F unit. Given these estimates, the noise and speech energy within a T-F unit are approximated as the estimated noise power and estimated squared-magnitude of the speech DFT coefficient, respectively. These estimates are then transformed to the nonlinear frequency domain of the gammatone filterbank using the frequency response of the individual gammatone filters:

$$\hat{\mathbf{X}}(m,c) = (1/K)\sum_{k=0}^{K-1}(\hat{\mathbf{X}}_{\mathrm{FFT}}(m,k)\cdot|G_c(k)|^2). \tag{11}$$

Here, $\hat{\mathbf{X}}$ is an estimate of $\mathbf{X}$, $\hat{\mathbf{X}}_{\mathrm{FFT}}$ the estimated speech energy in the DFT domain and G_c the frequency response of the filter channel c. Index k denotes a DFT coefficient and K the number of DFT bins used for T-F analysis, which is set to 512 in our experiments. A similar equation is used to estimate $\mathbf{N}$. The IBM is finally estimated by substituting $\hat{\mathbf{X}}$ and $\hat{\mathbf{N}}$ in (4). Fig. 3(e) shows a binary mask estimated in this way.

3) Combining CASA and Speech Enhancement for IBM Estimation: The motivation behind combining the speech enhancement method and the CASA method is that the former works well when the SNR is high, whereas the latter algorithm is designed for low SNR conditions. Moreover, from Figs. 3(a)–(c), it can be seen that some target dominant units missed by one method are identified by the other.

The CASA mask in the combined system is obtained using the algorithm described in Section III.B.1 without any change. The goal of the speech enhancement method in the combined system is to identify units having high SNR. Therefore, the mask is estimated by calculating the local SNR at each T-F unit using $\hat{\mathbf{X}}$ and $\hat{\mathbf{N}}$ obtained as described in Section III.B.2, and comparing it to LC which is set to a value greater than 0 (unlike (4)). This also helps to reduce false alarms (0 s wrongly labeled as 1 s) in the final mask. A reasonable value for LC is chosen using a small development set of noisy mixtures (see Section IV.A for details).

To combine the two masks, we use the simple logical 'OR' operation. Fig. 3(f) shows the mask estimated by this algorithm. The final mask is more similar to the IBM than the masks estimated using CASA and speech enhancement based methods.

IV. Evaluation Results

We start by describing the experimental setup in Section IV.A. Results that highlight the role of the SNR transform are presented in Section IV.B. Since the idea of using *binary* masks for SNR estimation is relatively new, we provide an initial set of results using the IBM directly in Section IV.C. This is followed by a description of the results using the estimated IBMs and comparisons in Section IV.D. Finally, we compare an FFT based representation for SNR estimation using binary masks with the proposed method in Section IV.E.

A. Experimental Setup

All our experiments are conducted using the TIMIT speech corpus [9] and the NOISEX92 noise database [29]. Specifically, the experimental results are obtained on the core test set of the TIMIT database which consists of 192 clean speech utterances from 24 speakers recorded at 16 kHz. Four noises are chosen from the NOISEX92 database—white noise, car noise, babble noise and factory noise. The first two noises are stationary and the last two relatively non-stationary. Car noise is chosen as it has a considerable amount of low frequency energy as a result of which the broadband and the filtered SNRs are quite different, thereby enabling us to measure the performance of the proposed algorithm in estimating these SNRs more thoroughly. The noise signals are downsampled to 16 kHz to match the sampling rate of the speech signals.

Two test sets—Set A and Set B—are created for evaluating the performance of the proposed system in estimating the broadband SNR and the filtered SNR, respectively. Both test sets consist of the 4 noises mixed with clean speech at 6 SNR conditions ranging from -10 dB to 15 dB, in increments of 5 dB. To create a noisy signal, a randomly selected segment of the noise is scaled to the desired level and added to the speech signal. Test Set A is created by scaling the signals so as to set the broadband SNR (SNR_b) to the desired level. Similarly, Test Set B is created by controlling the filtered SNR (SNR_f). Test Set B is also used to evaluate subband SNR estimation performance.

TABLE I
Mean Absolute Error in Estimating the Broadband SNR With and Without the SNR Transformation Step With the True Filtered Speech and Noise Energies Assumed to be Available

Noise	SNR					
	-10 dB	-5 dB	0 dB	5 dB	10 dB	15 dB
Without SNR Transformation						
White	0.00	0.00	0.00	0.00	0.00	0.00
Car	7.66	7.78	7.64	7.68	7.77	7.78
Babble	0.12	0.07	0.05	0.12	0.09	0.06
Factory	0.86	0.89	0.89	0.95	0.89	0.91
With SNR Transformation						
White	0.00	0.00	0.00	0.00	0.00	0.00
Car	0.00	0.00	0.00	0.00	0.01	0.08
Babble	0.00	0.00	0.00	0.00	0.00	0.00
Factory	0.00	0.00	0.00	0.00	0.01	0.04

The broadband and filtered SNR estimation results are presented for the following systems. The first one is the SNR estimation algorithm (WADA) proposed in [17], which was shown to significantly outperform the algorithm from NIST [1]. The second system uses the noise power and speech squared-magnitude estimate obtained as described in Section III.B.2 using the speech enhancement algorithms [10], [7] directly to estimate the SNR (HND). The frame length and the frame shift are set to 20 msec and 10 msec, respectively, to match those used by our algorithm. We use 512 DFT bins for T-F analysis. The remaining parameters are set as suggested in [10], [7][1]. The SNR is estimated by summating the estimated noise power and the estimated squared-magnitudes of speech across time and frequency in the DFT domain. The remaining approaches are based on estimated IBMs. The Hu-Wang system [14] is the third, and is slightly modified so as to make use of the normalized filterbank and the SNR transform. These modifications improve the performance reported in [14]. The forth method uses the IBM estimated using the speech enhancement method described in Section III.B.2. We denote this method HND_MOD. The final method is based on the IBM estimated using the combined method described in Section III.B.3. The method is denoted as *Proposed*. Note that the only difference between HND_MOD and Proposed is in the way the IBM is estimated.

WADA and HND make use of all the frequencies of the signal to estimate the SNR. Therefore, before estimating the filtered SNR using these algorithms for Test Set B, the original mixture is processed using a filter that has a frequency response similar to the aggregated response of the gammatone filterbank (see Fig. 2). These algorithms then calculate the broadband SNR using the filtered signal, which is equivalent to estimating the filtered SNR of the signal.

A development set is created by randomly choosing 30 utterances from the training set of the TIMIT corpus to tune the LC value that is used to estimate the speech enhancement mask in the combined system (Section III.B.3). Values ranging from

[1]An implementation of this algorithm is available at http://siplab.tudelft.nl/content/mmse-based-noise-psd-tracking-algorithm, which was used to generate the results reported in this paper.

TABLE II
MEAN ABSOLUTE ERROR AND STANDARD DEVIATION OF THE ERROR (IN PARENTHESIS) IN ESTIMATING THE FILTERED SNR (SNR_f) AND THE BROADBAND SNR (SNR_b) USING THE IBM

Noise type	SNR type	SNR -10 dB	-5 dB	0 dB	5 dB	10 dB	15 dB	Mean
White	SNR_f	$0.16(\pm0.37)$	$0.00(\pm0.00)$	$0.00(\pm0.00)$	$0.09(\pm0.28)$	$0.55(\pm0.50)$	$0.99(\pm0.19)$	$0.30(\pm0.22)$
	SNR_b	$0.16(\pm0.37)$	$0.00(\pm0.00)$	$0.00(\pm0.00)$	$0.07(\pm0.26)$	$0.51(\pm0.50)$	$0.98(\pm0.25)$	$0.29(\pm0.23)$
Car	SNR_f	$0.00(\pm0.00)$	$0.00(\pm0.00)$	$0.00(\pm0.00)$	$0.00(\pm0.00)$	$0.00(\pm0.00)$	$0.03(\pm0.17)$	$0.01(\pm0.03)$
	SNR_b	$0.00(\pm0.00)$	$0.00(\pm0.00)$	$0.00(\pm0.00)$	$0.00(\pm0.00)$	$0.02(\pm0.12)$	$0.09(\pm0.42)$	$0.02(\pm0.09)$
Babble	SNR_f	$0.89(\pm0.56)$	$0.05(\pm0.21)$	$0.09(\pm0.29)$	$0.79(\pm0.41)$	$1.10(\pm0.31)$	$1.72(\pm0.53)$	$0.77(\pm0.39)$
	SNR_b	$0.89(\pm0.55)$	$0.03(\pm0.16)$	$0.12(\pm0.33)$	$0.74(\pm0.44)$	$1.09(\pm0.34)$	$1.79(\pm0.51)$	$0.78(\pm0.39)$
Factory	SNR_f	$0.77(\pm0.47)$	$0.01(\pm0.07)$	$0.07(\pm0.26)$	$0.71(\pm0.45)$	$1.02(\pm0.20)$	$1.53(\pm0.51)$	$0.68(\pm0.33)$
	SNR_b	$0.55(\pm0.50)$	$0.01(\pm0.07)$	$0.10(\pm0.31)$	$0.71(\pm0.45)$	$1.01(\pm0.22)$	$1.47(\pm0.54)$	$0.64(\pm0.35)$

0 dB to 10 dB in 1-dB steps are tested. Based on the SNR estimation performance on the development set across the 4 noise conditions, the final value is set to 8 dB.

Subband SNRs are estimated across the frequency bands of a 64-channel gammatone filterbank, which is a typical number of channels used in CASA systems. Among the algorithms described earlier, only modified versions of WADA and HND are compared with the proposed subband SNR estimation algorithm. As described in Section II, WADA assumes that speech is Gamma distributed with a fixed parameter $\alpha = 0.4$. Although this holds for broadband signals, we have noticed that this value does not hold for band-limited signals. Therefore, the 30-utterance development set is used to find an optimal α for each subband. This is done by fitting a Gamma distribution to the clean subband signal amplitudes (in the maximum-likelihood sense). The mean α for the 30 utterances for each channel is then chosen as the final parameter for that channel. HND is adapted to estimate subband SNRs in the domain defined by the gammatone filterbank by first transforming the energy estimates using (11) and then using (3). The IBM estimation module of the proposed algorithm estimates a 128-channel mask. Instead of re-estimating a 64-channel mask for the purpose of subband SNR estimation, we sub-sample this mask to 64 channels. This is reasonable because the center frequencies (c) of the 64-channel gammatone filterbank and those of the odd numbered channels $(2c-1)$ of the 128-channel gammatone filterbank are identical, since both of them are uniformly distributed in the ERB rate scale. Sub-sampling is done by additionally accounting for the wider bandwidths of filters in the 64-channel filterbank; a T-F unit, $\mathbf{M}_{64}(m,c)$, in the 64-channel mask is labeled 1 only if at least 2 out of the 3 corresponding T-F units, $\mathbf{M}_{128}(m,2c-2))$, $\mathbf{M}_{128}(m,2c-1)$ and $\mathbf{M}_{128}(m,2c)$, in the 128-channel mask are speech dominant. The subband SNRs are restricted to the range of -20 dB to 30 dB, i.e., any estimate not falling in this range is rounded to the boundary values.

In order to remove minor effects of windowing on the global SNR, the estimated values from each of these algorithms are rounded to the nearest integer before calculating error metrics[2]. In the case of broadband/filtered SNR estimation, the mean absolute errors and standard deviations are reported. In the case of subband SNR estimation, only the mean absolute errors are reported.

[2]In the default setting, the minimum step size in WADA is 1 dB.

B. SNR Transformation

In this section, we illustrate the effectiveness of the SNR transform by performing an oracle experiment assuming that the true filtered speech and noise energies are available to the system. Turning off the SNR transform implies that the broadband SNR is approximately equal to the filtered SNR. The mean absolute errors are shown in Table I for the 4 noise types at the tested SNR conditions. As can be seen, there are no differences in the results for white noise since the amount of low-frequency energy is negligible compared to the total noise energy that passes through the filterbank. In contrast, for car noise, without the transformation the errors are much larger. On average, SNR transformation improves performance by around 7.7 dB for this noise. The difference is less dramatic for babble noise, as it has only a small amount of energy in the low-frequency range. For factory noise, the transformation improves the average performance by around 0.9 dB. With the SNR transform the mean absolute error is near 0 dB for all 4 noises at the tested SNR conditions. The results corroborate our claim that the broadband and the filtered SNR can be different and the proposed SNR transform compensates for this difference for broadband SNR estimation. The transform plays an important role when the underlying noise type in a mixture has a considerable amount of low-frequency energy.

C. IBM Results

The mean absolute errors and the standard deviations of the errors in estimating the filtered SNR and the broadband SNR of the signal using the IBM are shown in Table II. The error trends in estimating these SNRs are quite similar. It can be clearly seen from the results that excellent performance is obtained using the IBM. When the noise is relatively stationary, the IBM based system is even able to perfectly estimate the SNR in a few test conditions. It is interesting to note that the errors are slightly larger in extreme SNR conditions (-10 dB and 15 dB). This is because at such high (low) SNRs masked (unmasked) T-F units are much fewer, leading to an underestimation (overestimation) of the total noise energy. This bias is noise dependent, which makes it difficult to compensate for without prior knowledge about the noise type. It should be pointed out that the advantage of SNR transformation persists even when the IBM is used to approximate the total speech and noise energy, especially for noises with significant low-frequency energy. Since noise is slightly overestimated

TABLE III
THE MEAN ABSOLUTE ERROR IN ESTIMATING THE FILTERED SNR(SNR_f) USING WADA, HND, HU-WANG, HND_MOD AND THE PROPOSED ALGORITHM. THE STANDARD DEVIATION OF THE ERROR IS SHOWN WITHIN PARENTHESIS. THE BEST RESULT IN EACH CONDITION IS MARKED IN **BOLD**. ALSO SHOWN ARE THE RESULTS AVERAGED ACROSS SNRS AND ACROSS DIFFERENT NOISE TYPES

Noise type	Method	SNR						Mean
		-10 dB	-5 dB	0 dB	5 dB	10 dB	15 dB	
White	WADA	3.32(±4.51)	1.17(±1.31)	0.90(±0.82)	0.88(±0.91)	1.02(±1.22)	1.55(±2.01)	1.47(±1.80)
	HND	**0.55**(±**0.63**)	0.81(±0.58)	0.69(±0.53)	0.64(±0.51)	0.85(±0.42)	1.07(±0.38)	**0.77**(±0.51)
	Hu-Wang	1.15(±0.94)	**0.45**(±0.78)	**0.39**(±0.71)	0.76(±1.81)	1.44(±1.86)	2.89(±3.20)	1.18(±1.55)
	HND_MOD	2.44(±0.97)	1.28(±**0.48**)	0.92(±**0.39**)	0.86(±**0.37**)	0.86(±**0.35**)	**0.90**(±**0.32**)	1.21(±**0.48**)
	Proposed	1.65(±0.78)	0.78(±0.57)	0.46(±0.51)	**0.54**(±0.54)	**0.71**(±0.63)	1.09(±0.75)	0.87(±0.63)
Car	WADA	6.29(±7.05)	3.11(±4.88)	1.09(±1.30)	0.86(±1.03)	0.93(±1.24)	1.56(±2.01)	2.31(±2.92)
	HND	2.24(±0.92)	0.74(±0.56)	0.20(±0.43)	0.15(±0.38)	0.28(±0.47)	0.70(±0.64)	0.72(±0.57)
	Hu-Wang	0.77(±1.00)	0.55(±0.88)	0.60(±1.54)	0.92(±2.08)	1.56(±2.15)	3.44(±3.77)	1.31(±1.90)
	HND_MOD	0.42(±0.78)	0.13(±0.37)	0.07(±0.26)	0.14(±0.36)	0.43(±0.52)	0.77(±0.59)	0.33(±0.48)
	Proposed	**0.35**(±**0.67**)	**0.07**(±**0.25**)	**0.01**(±**0.10**)	**0.01**(±**0.07**)	**0.06**(±**0.23**)	**0.24**(±**0.43**)	**0.12**(±**0.29**)
Babble	WADA	5.88(±2.59)	2.94(±1.86)	1.75(±1.09)	1.28(±1.06)	1.32(±1.39)	1.65(±2.09)	2.47(±1.68)
	HND	3.59(±**1.20**)	1.09(±**1.09**)	0.65(±0.86)	0.72(±0.83)	1.17(±0.96)	1.56(±1.03)	1.46(±1.00)
	Hu-Wang	2.32(±1.49)	1.06(±1.42)	1.12(±1.64)	1.61(±2.19)	2.29(±2.91)	3.65(±4.20)	2.01(±2.31)
	HND_MOD	2.80(±1.50)	1.00(±1.21)	0.59(±0.82)	0.61(±0.81)	0.72(±0.85)	**0.85**(±0.93)	1.10(±1.02)
	Proposed	**1.96**(±1.44)	**0.86**(±1.16)	**0.40**(±**0.65**)	**0.50**(±**0.69**)	**0.60**(±**0.78**)	0.93(±**0.91**)	**0.88**(±**0.94**)
Factory	WADA	6.39(±3.81)	2.86(±1.71)	1.84(±1.14)	1.43(±1.18)	1.30(±1.42)	1.74(±2.17)	2.60(±1.91)
	HND	2.15(±2.04)	**0.96**(±1.42)	0.93(±0.99)	1.09(±0.84)	1.41(±0.88)	1.92(±1.06)	1.41(±1.20)
	Hu-Wang	1.75(±2.09)	1.14(±1.72)	1.03(±1.53)	1.05(±1.45)	1.65(±2.57)	3.02(±3.40)	1.61(±2.13)
	HND_MOD	2.32(±3.16)	1.14(±1.52)	0.82(±0.95)	0.88(±0.81)	1.06(±0.80)	**1.11**(±**0.96**)	1.22(±1.37)
	Proposed	**1.30**(±**1.83**)	0.97(±**1.23**)	**0.54**(±**0.70**)	**0.50**(±**0.67**)	**0.72**(±**0.78**)	1.13(±0.98)	**0.86**(±**1.03**)
All	WADA	5.47(±5.73)	2.52(±3.17)	1.39(±1.23)	1.11(±1.10)	1.14(±1.35)	1.63(±2.08)	2.21(±2.45)
	HND	2.13(±1.20)	0.90(±0.91)	0.62(±0.70)	0.65(±0.64)	0.93(±0.68)	1.31(±0.78)	1.09(±0.82)
	Hu-Wang	1.50(±1.99)	0.80(±1.30)	0.79(±1.41)	1.08(±1.94)	1.74(±2.43)	3.25(±3.68)	1.53(±2.12)
	HND_MOD	1.99(±1.61)	0.89(±0.89)	0.60(±0.60)	0.62(±0.59)	0.77(±0.63)	0.91(±**0.70**)	0.96(±0.84)
	Proposed	**1.31**(±**1.18**)	**0.67**(±**0.80**)	**0.35**(±**0.49**)	**0.39**(±**0.49**)	**0.52**(±**0.61**)	**0.85**(±0.77)	**0.68**(±**0.72**)

TABLE IV
THE MEAN ABSOLUTE ERROR IN ESTIMATING THE BROADBAND SNR (SNR_b) USING WADA, HND, HU-WANG, HND_MOD AND THE PROPOSED ALGORITHM. THE STANDARD DEVIATION OF THE ERROR IS SHOWN WITHIN PARENTHESIS. THE BEST RESULT IN EACH CONDITION IS MARKED IN **BOLD**. ALSO SHOWN ARE THE RESULTS AVERAGED ACROSS SNRS AND ACROSS DIFFERENT NOISE TYPES

Noise type	Method	SNR						Mean
		-10 dB	-5 dB	0 dB	5 dB	10 dB	15 dB	
White	WADA	3.52(±4.69)	1.10(±1.25)	0.87(±0.77)	0.86(±0.84)	1.02(±1.13)	1.57(±1.88)	1.49(±1.76)
	HND	**0.62**(±**0.57**)	1.00(±0.55)	0.92(±0.42)	0.86(±0.44)	0.97(±**0.34**)	1.17(±0.37)	0.92(±**0.45**)
	Hu-Wang	1.21(±0.92)	**0.44**(±0.69)	**0.39**(±0.70)	0.67(±1.08)	1.51(±2.89)	3.03(±3.77)	1.21(±1.68)
	HND_MOD	2.40(±0.92)	1.30(±**0.50**)	0.96(±**0.31**)	0.86(±**0.38**)	0.86(±0.35)	**0.89**(±**0.36**)	1.21(±0.47)
	Proposed	1.68(±0.77)	0.80(±0.55)	0.52(±0.55)	**0.55**(±0.52)	**0.73**(±0.63)	1.10(±0.75)	**0.89**(±0.63)
Car	WADA	6.93(±7.61)	4.48(±6.23)	1.53(±1.78)	1.21(±1.29)	1.22(±1.34)	1.64(±1.93)	2.84(±3.36)
	HND	6.22(±1.69)	3.80(±1.25)	2.22(±0.86)	1.79(±0.90)	1.42(±0.70)	1.21(±0.78)	2.78(±1.03)
	Hu-Wang	0.35(±1.52)	0.31(±1.51)	0.46(±1.71)	0.93(±2.58)	1.62(±2.30)	3.57(±4.05)	1.21(±2.28)
	HND_MOD	0.03(±0.16)	0.01(±0.07)	**0.00**(±**0.00**)	0.01(±0.10)	0.03(±0.18)	0.20(±0.49)	0.05(±0.17)
	Proposed	**0.01**(±**0.07**)	**0.00**(±**0.00**)	**0.00**(±**0.00**)	**0.00**(±**0.00**)	**0.01**(±**0.07**)	**0.06**(±**0.32**)	**0.01**(±**0.08**)
Babble	WADA	5.72(±2.58)	2.94(±1.66)	1.70(±1.08)	1.30(±1.06)	1.28(±1.23)	1.70(±2.05)	2.44(±1.61)
	HND	4.03(±1.75)	1.24(±1.17)	0.60(±0.89)	0.77(±1.00)	1.15(±0.98)	1.53(±1.08)	1.55(±1.15)
	Hu-Wang	2.18(±1.66)	0.98(±1.39)	1.06(±1.44)	1.62(±2.49)	2.22(±2.80)	4.08(±4.67)	2.02(±2.41)
	HND_MOD	2.51(±**1.60**)	0.96(±1.18)	0.49(±0.76)	0.56(±0.74)	0.68(±**0.78**)	**0.73**(±**0.91**)	0.99(±0.99)
	Proposed	**1.82**(±**1.60**)	**0.72**(±**1.03**)	**0.41**(±**0.67**)	**0.41**(±**0.65**)	**0.57**(±**0.78**)	0.90(±0.93)	**0.81**(±**0.94**)
Factory	WADA	5.70(±3.92)	2.82(±1.66)	1.72(±1.14)	1.24(±1.03)	1.30(±1.22)	1.81(±2.02)	2.43(±1.83)
	HND	2.34(±**1.40**)	0.83(±1.23)	0.67(±0.84)	0.91(±0.75)	1.22(±0.82)	1.59(±0.92)	1.26(±1.00)
	Hu-Wang	1.53(±1.96)	0.89(±1.23)	0.82(±1.21)	0.93(±1.39)	1.40(±1.89)	2.95(±3.66)	1.42(±1.89)
	HND_MOD	1.65(±2.34)	1.00(±1.37)	0.66(±0.77)	0.78(±0.69)	0.85(±0.74)	**0.89**(±0.93)	0.97(±1.14)
	Proposed	**1.16**(±1.65)	**0.77**(±**0.95**)	**0.44**(±**0.68**)	**0.36**(±**0.54**)	**0.59**(±**0.73**)	1.01(±**0.90**)	**0.72**(±**0.91**)
All	WADA	5.47(±5.81)	2.83(±3.63)	1.46(±1.32)	1.15(±1.09)	1.21(±1.24)	1.68(±1.97)	2.30(±2.51)
	HND	3.30(±1.36)	1.72(±1.05)	1.10(±0.75)	1.08(±0.77)	1.19(±0.71)	1.38(±0.79)	1.63(±0.91)
	Hu-Wang	1.32(±2.00)	0.65(±1.27)	0.68(±1.33)	1.04(±2.03)	1.69(±2.53)	3.41(±4.08)	1.46(±2.21)
	HND_MOD	1.65(±1.26)	0.82(±0.78)	0.53(±**0.46**)	0.55(±0.48)	0.60(±**0.51**)	**0.68**(±**0.67**)	0.80(±0.69)
	Proposed	**1.17**(±**1.02**)	**0.57**(±**0.63**)	**0.34**(±0.47)	**0.33**(±**0.43**)	**0.47**(±0.55)	0.77(±0.73)	**0.61**(±**0.64**)

at extremely low SNRs, the transformation may worsen the performance for noises that do not have significant low-frequency energy, like babble and factory. The mean absolute errors without the transform for babble and factory noise at -10 dB are 0.72 dB and 0.28 dB, respectively, slightly better than the results with the transform. At all other SNRs, the transformation improves performance. The results point to the fact that the IBM, despite being binary, can indeed be used for SNR estimation.

D. Estimated IBM Results

1) Global SNR Estimation: Global SNR estimation results are tabulated in Tables III and IV. Each table consists of 5 sets

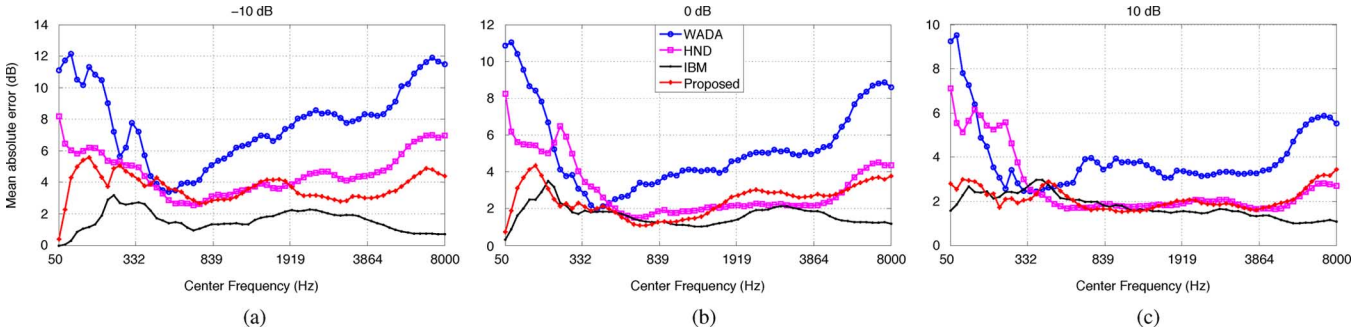

Fig. 4. Subband SNR estimation results using WADA, HND, the IBM, and the estimated IBM by the proposed algorithm, averaged across the four noises—white, car, babble and factory. Mean absolute errors across the 64 sub-bands are shown for the following filtered SNR conditions: (a) -10 dB. (b) 0 dB. (c) 10 dB.

of results—one for each noise and one for the average across the 4 noises.

The mean absolute errors in estimating the filtered SNR are shown in Table III. The proposed algorithm obtains the best average results across all noise types. It also obtains the best results in most of the individual test conditions. Similar to the IBM results, errors gradually increase at positive SNR levels but are still reasonably small. The second best performance is obtained using another binary masking method—HND_MOD. On average, it is around 0.2 dB worse than the proposed method. The proposed algorithm outperforms WADA and HND by about 1.5 dB and 0.4 dB, respectively. WADA performs reasonably when the SNR > 0 dB. But at lower SNRs, the noisy speech does not follow the Gamma distribution leading to poor estimation results. Not surprisingly, both WADA and HND perform the best in white noise conditions. WADA assumes that noise is Gaussian distributed, which holds better in white noise conditions compared to the other noises. Similarly, the distribution and statistical independence assumptions made by HND about the DFT coefficients of noise also hold well for white noise[3]. Hu-Wang outperforms the proposed system when the background noise is white and the SNR ≤ 0 dB. It is interesting to note that the proposed algorithm works better than the IBM in a few conditions when the SNR is high (e.g., for factory noise at 10 dB). This is possible because the IBM does make errors in SNR estimation, as can be seen from Table II. However, on average the IBM obtains better results than the proposed algorithm in every noise condition. The standard deviations of the errors are also shown in Table III. In terms of this error metric, the proposed algorithm also works the best in most test conditions.

The errors in estimating the broadband SNR are shown in Table IV. Again, the trends are very similar to Table III. Compared to HND_MOD, the average mean absolute error of the proposed algorithm is better by about 0.2 dB. Compared to WADA and HND, it is better by about 1.7 dB and 1 dB, respectively. The standard deviation profiles are similar to those for filtered SNR estimation.

These results clearly show that the proposed algorithm is able to obtain accurate estimates of global SNR—both broadband and filtered. Unlike WADA and HND, which work reasonably well at high SNRs, the proposed algorithm works well at all SNR/noise conditions.

[3]Note that these properties are unrelated to the color of the noise.

2) Subband SNR Estimation: Subband SNR estimation results are shown in Fig. 4. For simplicity, we only show the average performance across the 4 noises at 3 SNR conditions: -10 dB, 0 dB and 10 dB (see [24] for more detailed results). Unlike the global SNR estimation results, the errors are larger even when the IBM is used, where the best performance is typically obtained. For the proposed algorithm, better performance is usually obtained when the noise type is stationary. Barring a few conditions, the mean absolute error of the proposed algorithm is ≤ 5 dB.

Excluding the IBM results, the best performance in the low frequency channels (center frequency ≤ 350 Hz or the first 10–15 channels) is typically obtained by the proposed algorithm. The only noted exception is when the noise is babble and the SNR ≤ 0 dB, where the mean errors are greater than 5 dB for some channels. HND outperforms the proposed algorithm in such conditions. If we consider the average performance across all noise conditions, the mean absolute error of the proposed algorithm is well within 5 dB for these frequency channels, significantly better than both HND and WADA.

For the mid-frequency channels (center frequency between 300 Hz and 3800 Hz, or frequency channels 13–51), no one method works uniformly better than the rest. Both HND and the proposed algorithm work well in most conditions. WADA obtains results similar to HND and the proposed algorithm when the background noise is white; and performs better when the background noise is car and the SNR ≥ 10 dB. This is largely because the true subband SNRs in these conditions are well above 0 dB. At other conditions, performance of WADA is significantly worse than the other methods, as reflected in the average performance (see Fig. 4). When the background noise is non-stationary, the proposed algorithm is slightly better than HND at most SNRs. Under stationary conditions, the performance of the proposed algorithm is mostly comparable or better than HND. In a few cases, especially when the SNR is high, HND works slightly better. Similar mixed trends are observed for the high frequency channels (center frequency ≥ 3800 Hz, or the last 10–15 channels), with the proposed algorithm working slightly better than HND especially when the noise type is non-stationary. When the noise type is factory and the SNR ≤ 0 dB, the errors are greater than 5 dB but still better than both WADA and HND.

We can observe a few overall trends in estimation errors from Fig. 4. For example, we can see that as the filtered SNR of the

TABLE V
THE MEAN ABSOLUTE ERROR IN ESTIMATING THE BROADBAND SNR (SNR_b) USING IBM_DFT, IBM_GF, EBM_DFT AND THE PROPOSED ALGORITHM. THE RESULTS ARE AVERAGED ACROSS THE 4 NOISES

Method	SNR					
	-10 dB	-5 dB	0 dB	5 dB	10 dB	15 dB
IBM_DFT	0.18	0.00	0.00	0.10	0.49	0.77
IBM_GF	0.40	0.01	0.06	0.43	0.66	1.08
EBM_DFT	2.99	1.58	1.10	1.11	1.23	1.35
Proposed	1.17	0.57	0.34	0.33	0.47	0.77

signal increases, the performance of the proposed algorithm also improves. When the SNR is -10 dB, the mean absolute errors are about 4 dB. And when the SNR is 10 dB, the errors are about 2 dB. Also note that improvements in mask estimation can clearly improve the average performance of the proposed method, since the IBM results are significantly better especially at low SNR conditions. These results indicate that the proposed algorithm can additionally be used to estimate subband SNRs with considerable accuracy.

E. Comparison of FFT and Gammatone Filterbank Based Representations

The use of the cochleagram representation rather than the more commonly used FFT based representation to estimate the SNRs deserves some justification. A system that uses binary masks estimated in the DFT domain has an advantage—the SNR transformation step is not needed to estimate the broadband SNR. But IBM estimation in the DFT domain is less studied compared to that in the auditory domain. Furthermore, auditory cues typically used by CASA-based estimation algorithms, like pitch and amplitude modulation are less prominent in representations that use a linear frequency scale [12].

Nevertheless, we perform a comparison between the performance obtained using the ideal and estimated binary masks in these two domains. When using the IBM (or the estimated mask) in the DFT domain, (5)–(7) are used without any SNR transformation, after replacing the cochleagram with the spectrogram. The IBM in the DFT domain is defined using (4), by comparing the energy (squared-magnitude) of clean speech and noise at each T-F unit. When estimating the binary mask, a recently proposed MMSE-based mask estimator is used [16]. We use Type-II binary masks as defined in [16] that minimize the spectral 'squared-magnitude' MSE. It has the same form as the spectral magnitude MMSE mask derived in [16], except that the spectral squared-magnitude MMSE gain function is used in place of the gain function used in [16]. The results are summarized in Table V. The table shows results obtained using the IBM (IBM_DFT) and the MMSE-optimal binary mask (EBM_DFT) in the DFT domain. For comparison, we also show the results obtained using the IBM in the gammatone filterbank domain (IBM_GF), and those obtained using the IBM estimated using the algorithm described in Section III.B.3 (Proposed). Note that the SNR transform is used by both IBM_GF and Proposed. Clearly, the performance is better when the IBM defined in the DFT domain is used. This is expected because the DFT domain uses a better frequency resolution (512 vs. 128). On the other hand, when estimated binary masks are used, better performance is obtained in the auditory domain. To conclude, if effective algorithms exist to estimate the IBM in the DFT domain, one can choose such a representation. But since relatively accurate mask estimation algorithms operate in the auditory domain, it seems preferable to perform SNR estimation in this domain.

V. DISCUSSION

The results presented in this paper show that binary masks can be used for long-term SNR estimation—both at subband and broadband levels. The results further indicate that we only need a reasonable estimate of the IBM to obtain good SNR estimates. If an algorithm is able to correctly label the high energy regions as belonging to the target or the noise, the long-term SNR can be estimated with very good accuracy as the energy in these regions dominates the total energy. In most of the test conditions, the best performance is obtained when the masks estimated by CASA and speech enhancement algorithms are combined.

The proposed algorithm cannot be used to estimate short-time SNR of a signal, which would lead to a chicken-and-egg problem as the short-time SNR can directly be used to estimate the IBM. A disadvantage of the proposed algorithm is its computational complexity. The CASA component involves computation of autocorrelation and envelope extraction at each T-F unit during the feature extraction stage, both of which are computationally expensive. The feature extraction stage dominates the time complexity of the proposed algorithm. Autocorrelations can be efficiently calculated in $O(N \log N)$ time and since frequency channels are independent of each other, computations can be parallelized [13], [15]. Even so, the algorithm takes longer than WADA or HND. Nevertheless, the performance in SNR estimation obtained by the proposed system is significantly better than these approaches.

Binary masking described in this work is quite different from VAD based algorithms that have been proposed in the literature for SNR estimation [19], [26]. A VAD tries to identify *noise-only* frames to obtain an estimate of the noise energy by assuming stationarity. On the other hand, our approach identifies *noise-dominant* T-F units, which are used to approximate the total noise energy in the algorithm. The algorithm can easily be extended to estimate the SNR in speech-present frames, by simply dropping noise-only frames during estimation. In experiments not reported in the paper, we have confirmed that dropping noise-only frames does not have a significant impact on performance. As such, our algorithm can deal with situations when the target signal contains long pauses. Such pauses would appear as long sections of time frames with no unmasked units. In contrast, methods like WADA and the algorithm from NIST [1] will have greater difficulty dealing with such signals.

Note that, the mask estimation and the SNR estimation in the proposed system are two separate modules. The IBM estimation module used by the current system can be replaced with any other mask estimation algorithm. Therefore, the proposed algorithm can potentially be used in more challenging conditions like reverberant noisy environments and multi-talker conditions by replacing the existing mask estimation algorithm with those that work well in such conditions.

To summarize, we have proposed a novel CASA based SNR estimation algorithm. The algorithm estimates the filtered, broadband and subband SNRs with high accuracy. Results

show that the performance of the proposed system is better than existing long-term SNR estimation algorithms. The algorithm additionally estimates the IBM, which can be used for speech separation purposes. An insight from our work is that binary masks can be effectively used for SNR estimation.

ACKNOWLEDGMENT

The authors would like to thank G. Hu, K. Hu and K. Han for helpful discussions and providing software implementations; R. Hendriks, R. Heusdens, J. Jensen and J. Erkelens for sharing MATLAB codes for the work described in [10], [7]; and C. Kim for providing an implementation of WADA.

REFERENCES

[1] NIST Speech Quality Assurance (SPQA) Package v2.3, 1994 [Online]. Available: http://www.itl.nist.gov/iad/mig/tools/

[2] M. Berouti, R. Schwartz, and R. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE ICASSP*, 1979, pp. 208–211.

[3] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel *a priori* SNR estimation approach based on selective cepstro-temporal smoothing," in *Proc. IEEE ICASSP*, 2008, pp. 4897–4900.

[4] I. Cohen, "Speech spectral modeling and enhancement based on autoregressive conditional heteroscedasticity models ," *Signal Process.*, vol. 86, no. 4, pp. 698–709, 2005.

[5] T. H. Dat, K. Takeda, and F. Itakura, "On-line Gaussian mixture modeling in the log-power domain for signal-to-noise ratio estimation and speech enhancement," *Speech Commun.*, vol. 48, pp. 1515–1527, 2006.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[7] J. Erkelens, R. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 6, pp. 1741–1752, Dec. 2007.

[8] S. Furui*, Digital Speech Processing, Synthesis, and Recognition*, 2nd ed. New York: Marcel Dekker, 2000.

[9] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus, 1993, [Online]. Available: http://www.ldc.upenn.edu/Catalog/LDC93S1.html

[10] R. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE ICASSP*, 2010, pp. 4266–4269.

[11] H. G. Hirsch, "Estimation of noise spectrum and its applications to SNR-estimation and speech enhancement," Int. Comput. Sci. Inst., Berkeley, CA, Tech. Rep. TR-93-012, 1993.

[12] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[13] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2067–2079, Nov. 2010.

[14] G. Hu and D. L. Wang, "Segregation of unvoiced speech from nonspeech interference," *J. Acoust. Soc. Amer.*, vol. 124, pp. 1306–1319, 2008.

[15] K. Hu and D. L. Wang, "Unvoiced speech segregation from nonspeech interference via CASA and spectral subtraction," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 6, pp. 1600–1609, Aug. 2011.

[16] J. Jensen and R. Hendriks, "Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 92–102, Jan. 2012.

[17] C. Kim and R. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proc. Interspeech*, 2008, pp. 2598–2601.

[18] M. Kleinschmidt and V. Hohmann, "Sub-band SNR estimation using auditory feature processing," *Speech Commun.*, vol. 39, pp. 47–64, 2003.

[19] A. Korthauer, "Robust estimation of the SNR of noisy speech signals for the quality evaluation of speech databases," in *Proc. ROBUST'99 Workshop*, 1999, pp. 123–126.

[20] Y. Li and D. L. Wang, "On the optimality of ideal binary time–frequency masks," *Speech Commun.*, vol. 51, pp. 230–239, 2009.

[21] P. C. Loizou*, Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007.

[22] Y. Lu and P. Loizou, "Estimators of the magnitude-squared spectrum and methods for incorporating SNR uncertainty," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 19, no. 5, pp. 1123–1137, Jul. 2011.

[23] R. Martin, "An efficient algorithm to estimate the instantaneous SNR of speech signals," in *Proc. Eurospeech*, 1993, pp. 1093–1096.

[24] A. Narayanan and D. L. Wang, "A CASA based system for SNR estimation,' Dept. Comput. Sci. and Eng., The Ohio State Univ., Columbus, OH, Tech. Rep. OSU-CISRC-11/11-TR36, 2011 [Online]. Available: ftp://ftp.cse.ohio-state.edu/pub/tech-report/2011/

[25] E. Nemer, R. Goubran, and S. Mahmoud, "SNR estimation of speech signals using subbands and fourth-order statistics," *IEEE Signal Process. Lett.*, vol. 6, no. 7, pp. 504–512, Jul. 1999.

[26] C. Ris and S. Dupont, "Assessing local noise level estimation methods: Application to noise robust ASR," *Speech Commun.*, vol. 34, pp. 141–158, 2001.

[27] J. Tchorz and B. Kollmeier, "SNR estimation based on amplitude modulation analysis with applications to noise suppression," *IEEE Trans. Audio, Speech, Signal Process.*, vol. 11, no. 3, pp. 184–192, May 2003.

[28] D. van Compernolle, "Noise adaptation in a hidden Markov model speech recognition system," *Comput. Speech Lang.*, vol. 3, pp. 151–168, 1989.

[29] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, 1993.

[30] D. L. Wang, "On ideal binary masks as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Boston, MA: Kluwer, 2005, pp. 181–197.

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

[32] X. Zhao, Y. Shao, and D. L. Wang, "Robust speaker identification using a CASA front-end," in *Proc. IEEE ICASSP*, 2011, pp. 5468–5471.

Arun Narayanan (S'11) received the B.Tech. degree in computer science from the University of Kerala, Trivandrum, India, in 2005, and the M.S. degree in computer science from the Ohio State University, Columbus, USA, in 2012, where he is currently pursuing the Ph.D. degree. From November 2005 to June 2008, he was a System Engineer at IBM India.

His research interests include computational auditory scene analysis, robust automatic speech recognition, and machine learning.

DeLiang Wang, (F'04) photograph and biography not available at the time of publication.