Acoustic Scene-adaptive Speech Enhancement

Jianfeng Li
Signal Processing Systems, Dept. of Electrical Engineering,
Eindhoven University of Technology, The Netherlands
Email: j.li.1@student.tue.nl

***Abstract*—Conventional single-microphone speech enhancement (SE) systems strive to retrieve from the noisy observed signal an 'enhanced spectrum', which is as close to the clean speech spectrum as possible. In other words, these systems are optimized on the signal level. However, since the human auditory system parses speech at higher conceptual levels than the spectrum alone, minimization of signal spectrum error does not necessarily result in improved perception. In this paper, we propose an alternative optimization framework, which optimizes the performance at a perceptual level based on an on-line estimation of the acoustic scene. This optimization framework is closer to the end user, and thus the resulting optimal tuning parameters are more meaningful in practical application, *e.g.* a hearing aid device. In our system, an off-line training session first maximizes the PESQ score of the processed noisy speech for each acoustic scene. Next, during on-line operation of the speech enhancement algorithm, an optimal steering strategy is used to tune the parameters to the observed acoustic scenes.**

In order to support the proposed SE method, we develop a low-complexity Acoustic Scene Analyzer (ASA) with the following characteristics: 1) The computational complexity is very low, allowing real-time application (in MATLAB); 2) The valid analysis time period is tunable without retraining of the ASA model; 3) The ASA algorithm takes advantage of any existing noise spectral tracking algorithm (which is usually the case for hearing aid algorithms). In this work we also develop a novel noise spectral tracking algorithm, based on probabilistic modeling ideas.

The proposed methods and algorithms are evaluated on various non-stationary noise sources over a wide range of SNR's. Experimental results show that using the proposed method and noise estimation algorithm, the basic spectral subtraction enhancement system can result in an increment in PESQ about 0.3 point on average.

I. Introduction

It is well known that background noise reduces the intelligibility of the speech, especially for the people with hearing loss. Hence, modern Hearing Instrument (HI) has a form of speech enhancement (SE) system to reduce the additive noise.

Fig. 1 shows such a system. The uncolored part is a warped filterbank proposed in [1], which is suitable for HI. The other parts together realize a SE algorithm. For the given system, to have a better quality of the enhanced speech, besides of having a more accurate SNR estimation, an optimal steering strategy for the tuning parameter is also needed.

To achieve this object, in this paper, we develop an optimal acoustic scene-adaptive SE with three novel algorithms, OASA, SIC and PMCRA. Table I lists the mathematical notation used in this paper. The steering strategy for the tuning parameter (OASA, pink block) will be discussed in Section II. As its input, an acoustic scene analyzer (ASA) (SIC, blue block) will be developed in Section III. To further increase the accuracy of the analyzer and the SNR estimation, in Section IV, a novel noise spectral tracking algorithm (PMCRA, purple block) based on probabilistic modeling ideas will be proposed. Some state-of-the-art algorithms will be briefly reviewed in this section as well. Both the performance of the noise spectral tracking algorithm and the tuning strategy will be evaluated in Section V. The reference system is adopted as the one without the blocks in dash box, and the purple block will be substituted with the MCRA [2] algorithm. In the final section, a summary and concluding remarks will be drawn.

II. Optimal Perceptual Speech Enhancement

The SNR estimation can be optimized on the signal level by minimizing the estimation error. The impact of θ on the quality of the enhanced speech, however, has never been studied mathematically. In this section, we will study the performance of the SE system as a function of θ, and formulate the optimal steering strategy accordingly.

A. Mathematical Formulation

The given SE algorithm is denoted as

$$y(t) = H\left(x(t), \theta\right) \tag{1}$$

where θ is a tuning parameter ranging from 0 to 1.

To evaluate the performance of the SE algorithm, we define a utility model

$$u\left(y, s\right) \tag{2}$$

which quantitatively measures the quality of the enhanced speech y relative to the reference clean speech s. Note that, here we drop the variable t to indicate y and s are whole sentences rather than time-varying signals. For ease of notation, we simplify it as $u\left(y\right)$ when this does not cause confusion. It is pointed out in [3] that PESQ correlated well on predicting subjective speech quality ratings. Hence, we select PESQ as the utility model in our work.

With a selected utility model $u\left(y\right)$ and the SE algorithm $y(t) = H\left(x(t), \theta\right)$, we can re-formulate Eq. (2) as $u\left(x, \theta\right)$ to explicitly express the dependency of the utility model on the tuning parameter. It is calculated as first computing Eq. (1) followed by evaluation of Eq. (2).

Assume the performance is evaluated on a corpus

$$\mathcal{X} = \{x_1, \ldots, x_L\} \tag{3}$$

of L noisy speech sentences, with the relevancies $p(x_l)$ ($\sum_l p(x_l) = 1$), related to the importance weight of the

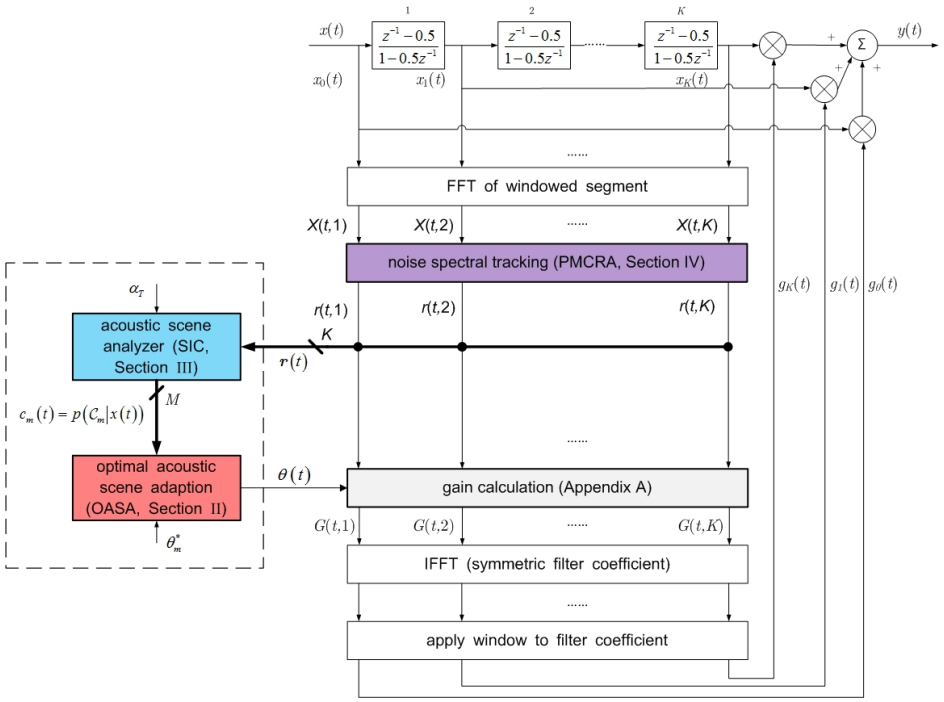


Fig. 1. Block diagram of the speech enhancement system discussed in this paper.

TABLE I
MATHEMATICAL NOTATION USED IN THIS PAPER

Notation	Affectation	Notation	Affectation
$x(t)$	noisy speech	$y(t)$	enhanced speech
$s(t)$	clean speech	$n(t)$	additive noise
$u\,(y)$	utility of sentence y	$c_m(t)$	out put of ASA on m^{th} channel at time t
$r(t,k)$	estimated SNR at time t on k^{th} frequency bin	$\theta(t)$	the value of tuning parameter at time t
$X(t,k)$	STFT coefficient of x at time t on the k^{th} frequency bin	$\lambda_x(t,k)$	random variable $\ln E\left[\lvert X(t,k)\rvert^2\right]$
$N(t,k)$	STFT coefficient of n at time t on the k^{th} frequency bin	$\lambda_n(t,k)$	random variable $\ln E\left[\lvert N(t,k)\rvert^2\right]$
$S(t,k)$	STFT coefficient of s at time t on the k^{th} frequency bin	$\lambda_s(t,k)$	random variable $\ln E\left[\lvert S(t,k)\rvert^2\right]$
$\bar{\lambda}_x(t,k)$	actual measurement of $\lambda_x(t,k)$, similar for $\bar{\lambda}_n(t,k)$ and $\bar{\lambda}_s(t,k)$	$\hat{\lambda}_x(t,k)$	estimated mean value of $\lambda_x(t,k)$, similar for $\hat{\lambda}_n(t,k)$ and $\hat{\lambda}_s(t,k)$
$\sigma_x^2(t,k)$	estimated variance of $\lambda_x(t,k)$, similar for $\sigma_n^2(t,k)$ and $\sigma_s^2(t,k)$	$p\left(\mathcal{C}_m\lvert x(t)\right)$	probability of $x(t)$ belonging to m^{th} class
$P^{(1)}(t,k)$	updating value for $\hat{\lambda}_n(t-1,k)$ at time t on the k^{th} frequency bin	$P^{(2)}(t,k)$	updating value for $\sigma_n(t-1,k)$ at time t on the k^{th} frequency bin
$\tilde{\alpha}_d^{(1)}(t,k)$	smoothing factor for $\hat{\lambda}_n(t-1,k)$ at time t on the k^{th} frequency bin	$\tilde{\alpha}_d^{(2)}(t,k)$	smoothing factor for $\sigma_n(t-1,k)$ at time t on the k^{th} frequency bin
$\alpha_s^{(1)}$	smoothing factor for $\hat{\lambda}_n(t-1,k)$ at time t on the k^{th} frequency bin when $p_s(t,k)=1$	$\alpha_s^{(2)}$	smoothing factor for $\sigma_n(t-1,k)$ at time t on the k^{th} frequency bin when $p_s(t,k)=1$
$\alpha_n^{(1)}$	smoothing factor for $\hat{\lambda}_n(t-1,k)$ at time t on the k^{th} frequency bin when $p_s(t,k)=0$	$\alpha_n^{(2)}$	smoothing factor for $\sigma_n(t-1,k)$ at time t on the k^{th} frequency bin when $p_s(t,k)=0$
$p_s(t,k)$	speech presence probability at time t on the k^{th} frequency bin	$\alpha_p(t,k)$	smoothing factor for $p_s(t-1,k)$ at time t on the k^{th} frequency bin
α_a	smoothing factor for $p_s(t-1,k)$ when speech attack	α_r	smoothing factor for $p_s(t-1,k)$ when speech release
$I(t,k)$	voice-activity-detector at time t on the k^{th} frequency bin	$T(t,k)$	decision threshold between $\lambda_x(t,k)$ and $\lambda_n(t,k)$
ϑ	separating point for applying the non-linear weight factor function when estimating $\hat{\lambda}_n(t,k)$	γ	inverse proportionality constant for weighting factor and the prior SNR

sentence x_l in the corpus. Then the overall performance can be defined as the *expected utility*

$$EU(\theta) = \sum_l p(x_l) u\,(x_l, \theta) \tag{4}$$

where the sum runs over the entire corpus. Note that, during evaluation, the SE algorithm is fixed.

The optimal θ should be the one that maximizes $EU(\theta)$, that is

$$\theta^* = \arg\max_\theta EU(\theta) = \arg\max_\theta \sum_l p(x_l) u\,(x_l, \theta) \tag{5}$$

A straightforward way to carry out this optimization is to calculate a global optimum. However, this optimum is only in the context of the given audio corpus $\mathcal{X}$, and thus lacks of flexibility. Next, we propose another approach to find optimal values θ^* that may vary as a function of acoustic scene.

B. Optimal Scene-adaptive Speech Enhancement

Assume that we have defined M different *acoustic scenes* $\mathcal{C}_m$, each of which represents a certain acoustic scenario or environment, such as a cock-tail party, a train station environment, a concert, or a low-SNR scenario. Then, mathematically, the expected utility can be broken down as

$$\begin{aligned} EU(\theta) &= \sum_l p(x_l) u\left(x_l, \theta\right) \\ &= \sum_l \left[\sum_m p\left(x_l|\mathcal{C}_m\right) p(\mathcal{C}_m) \right] u\left(x_l, \theta\right) \\ &= \sum_m p(\mathcal{C}_m) \left[\sum_l p\left(x_l|\mathcal{C}_m\right) u\left(x_l, \theta\right) \right] \end{aligned} \tag{6}$$

Here, $p(\mathcal{C}_m)$ can be regarded as our beliefs on the proportion of scene m occurring in real life, and we define the second factor as *scene-conditional expected utility*

$$EU_m(\theta) = \sum_l p\left(x_l|\mathcal{C}_m\right) u\left(x_l, \theta\right) \tag{7}$$

Similar to Eq. (5), the optimal tuning parameter for scene m can be computed as

$$\theta_m^* = \arg\max_\theta EU_m(\theta) \tag{8}$$

Note that, in Eq. (7) the term $p\left(x_l|\mathcal{C}_m\right)$ is the likelihood that we hear the sentence x_l in the scene m. This term, however, is unaccessible, since the sentences one could hear is unlimited, and it is hardly that the sentence would be repeated exactly the same in daily life. To make it computable, we rewrite it using Bayes rule

$$EU_m(\theta) = \frac{1}{p(\mathcal{C}_m)} \sum_l p\left(\mathcal{C}_m|x_l\right) p(x_l) u\left(x_l, \theta\right) \tag{9}$$

The term $p\left(\mathcal{C}_m|x_l\right)$ represents the probability that the sentence x_l occurs in the scene m, and we compute it as the average of a time-varying signal $c_m(t)$ over the whole sentence. $c_m(t)$ is the output of a *acoustic scene analyzer* (ASA)

$$c_m(t) = p\left(\mathcal{C}_m|x(t)\right) \tag{10}$$

Assume that we have the off-line data of ASA, Eq. (8) can be computed out using the global optimization approach, since we can rewrite it in a similar structure to Eq. (5)

$$\begin{aligned} \theta_m^* &= \arg\max_\theta \frac{1}{p(\mathcal{C}_m)} \sum_l p\left(\mathcal{C}_m|x_l\right) p(x_l) u\left(x_l, \theta\right) \\ &= \arg\max_\theta \sum_l p\left(\mathcal{C}_m|x_l\right) p(x_l) u\left(x_l, \theta\right) \end{aligned} \tag{11}$$

The term $1/p(\mathcal{C}_m)$ can be dropped because it does not affect the optimization result.

The optimal on-line steering strategy is consequently a fitting problem. We can store the optimal vectors $\theta_1^*, \ldots, \theta_M^*$ on the SE system. During executing of SE algorithm, the optimal tuning vector can be estimated as

$$\theta^*(t) = \sum_m p\left(\mathcal{C}_m|x(t)\right) \theta_m^* = \sum_m c_m(t) \theta_m^* \tag{12}$$

Eq. (10), (11) and (12) constitute the OASA algorithm.

III. Perceptual Acoustic Scene Analyzer

In the previous section, we proposed an optimal acoustic scene-adaptive (OASA) speech enhancement system. Both the optimization procedure and the adaption scheme are derived assuming $c_m(t)$ is known. In this section, we will develop a perceptual level ASA to support this framework, and we call it *speech intelligibility classifier* (SIC).

Traditional ASA uses various time and frequency domain features to estimate the probability that the given audio is a (noisy) speech, a pure noise, or a music [4] [5] [6]. This approach, however, lacks a perceptual interpretation in its output. Furthermore, it usually requires a completely independent computational module, which leads to a higher computational load.

In our work, instead of separating the acoustic scenes by its physical settings, we define the different classes according to the perceptual characteristics. Ideally, to cope with the OASA framework, the best classification feature for ASA should be some index reflecting the quality of the incoming signal. If the quality of the incoming signal is already quite high, the impact of our SE algorithm should be limited to a relatively low level to prevent the speech from being distorted. However, the speech quality is a multi-dimensional subjective term and hard to compute. For ease of computation, we separate different scenes according to the speech intelligibility (SI), since SI can be easily indicated by the Articulation Index (AI) [7], which is, generally speaking, a weighted sum of SNR's on different frequency bands. This computational scheme provides us an extra advantage that, the SIC could reuse the output of the noise tracking module, and thus be able to well integrated into any existing SE algorithm in HI.

The problem is that, however, the classic AI calculation asks for direct access to long-term speech and noise spectrum which are unaccessible in SE system. Next we develop a calculation scheme that estimates the short-time AI.

A. Short-time AI Estimation

The value of short-time AI is achieved by modifying the 20 equally contributing critical band calculation procedure in [7] to fit our SE system.

The first modification is in the AI analysis bands. The AI analysis bands are recomputed by comparing the frequency range of each analysis band in actual system with that of the equally contributing critical bands given in [7]. For the analysis filterbank shown in Fig. 1, its AI bands are given in Table II.

The second modification is the SNR calculation. In the original scheme, several corrections are made to the speech and noise spectra, such as compensation for speech peak and

TABLE II
AI BANDS FOR WARPED-31 COMPRESSOR [1], 16KHZ SAMPLING RATE

FFT bin index	Center frequency (Hz)	Limits (Hz)	AI band
1	0	0 to 83.5	
2	167	83.5 to 250.8	
3	337	250.8 to 423.6	1
4	513	423.6 to 602.8	2
5	699	602.8 to 794.7	3
6	898	794.7 to 1001.3	4
7	1116	1001.3 to 1230.9	5
8	1360	1230.9 to 1489.0	6
9	1639	1489.0 to 1788.4	7
10	1965	1788.4 to 2141.5	8
11	2357	2141.5 to 2572.0	9
12	2840	2572.0 to 3107.8	10
13	3451	3107.8 to 3794.4	11
14	4240	3794.4 to 4685.0	12
15	5260	4685.0 to 5834.8	13
16	6357	5834.8 to 7238.7	14
17	8000	7238.7 to 8000	

spread of masking. The SNR are calculated as the differences between the corrected spectra. In our system, however, the estimated SNR is directly obtained through noise tracking module. No correction is needed, since

- The spectrum is computed within a short time frame, so that the speech peak level can be well preserved.
- The noisy speech is pre-normalized before entering the system, so there is no chance that the speech level exceeds the sensation level.
- Neither the speech nor the noise spectrum is accessible, the SNR is estimated based on the noise power estimation. The estimation error has already incorporated the missing correction.

The short-time AI is thus calculated on every frame, proceeding in the following steps:

1) Obtain the estimated SNR $r(t,k)$ through the noise tracking module. The estimated values below 0 dB are set to 0, and the values above 30 dB are set to 30, which we will denoted as $\tilde{r}(t,k)$.
2) The result in step 1 is then normalized by a factor of 30.
3) The short-time AI can be seen as a linear combination of SNR's
$$AI(t) = \frac{1}{30|\mathcal{K}|}\sum_{k\in\mathcal{K}}\tilde{r}(t,k) \tag{13}$$
where $\mathcal{K}$ is the set of the FFT bin indices that correspond to AI bands, and $|\mathcal{K}|$ is the number of AI bands. For the example shown in Table II, $\mathcal{K} = 3,4,\ldots,16$ and $|\mathcal{K}| = 14$.

B. Speech Intelligibility Classifier

The short-time AI derived in the previous section can be regarded as the feature variable for the SIC, but only valid for the current analysis frame (typically a few milliseconds). For $c_m(t)$, we would like to make it flexible to represent the characteristics of a section $[x(t-T+1),\ldots,x(t)]$, where T is the analysis duration defined by user. According to [8], the AI for a long-time noisy speech can be calculated as the average of the short-time AI over time. Hence, we approximate this averaging procedure by a lowpass filter given in Appendix B, and thus getting the feature variable $v(t)$ at time t iteratively

$$\log v(t) = (1-\alpha_{_T})\log v(t-1) + \alpha_{_T}\log AI(t) \tag{14}$$

where $\alpha_{_T}$ is the smoothing factor calculated according to Eq. (B.40), by setting $\tau_{90} = T$.

With this feature variable $v(t)$ derived from $x(t)$, we define the following pdf for $\mathcal{C}_m$

$$p\left(v|\mathcal{C}_m\right) = \mathcal{N}\left(\mu_m, \Sigma_m\right) \tag{15a}$$
$$p\left(\mathcal{C}_m\right) = \pi_m \tag{15b}$$

where μ_m is the mean value of v for the m^{th} scene, and Σ_m is the corresponding variance.

The desired signal $c_m(t)$ can thus be regarded as the posterior scene probability, and be computed out using Bayes

$$c_m(t) = p\left(\mathcal{C}_m|x(t)\right) = \frac{p(v(t)|\mathcal{C}_m)p(\mathcal{C}_m)}{\sum_{j=1}^{M}p(v(t)|\mathcal{C}_j)p(\mathcal{C}_j)} \tag{16}$$

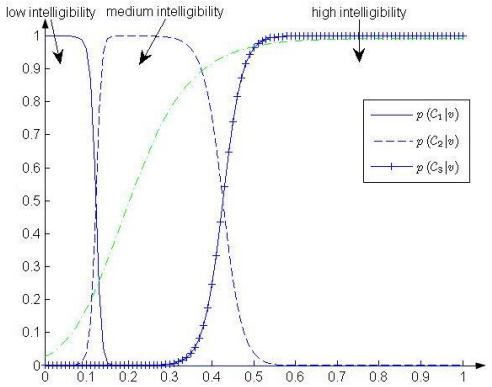


Fig. 2. SIC output $p\left(\mathcal{C}_m|x(t)\right)$ vs. feature variable v (blue lines) and the nonlinear relationship between AI and SI (green line). The nonlinear relationship is replotted from Fig. 15 in [7].

Fig. 2 shows the SIC we designed. The green dash line shows the nonlinear relationship between AI and SI, replotted from Fig. 15 in [7]. Consequently, we separate the acoustic environment into three scenes: low SI environment $\mathcal{C}_1$, medium SI environment $\mathcal{C}_2$, and high SI environment $\mathcal{C}_3$. The blue lines are the output of our SIC, versus the feature variable v. The corresponding scene-conditional feature variable distribution is defined as

$$p\left(v|\mathcal{C}_1\right) = \mathcal{N}\left(0.05, 0.05/2\right) \quad p\left(\mathcal{C}_1\right) = 0.3635 \tag{17a}$$
$$p\left(v|\mathcal{C}_2\right) = \mathcal{N}\left(0.3, 0.1/1.5\right) \quad p\left(\mathcal{C}_2\right) = 0.4912 \tag{17b}$$
$$p\left(v|\mathcal{C}_3\right) = \mathcal{N}\left(0.5, 0.07\right) \quad p\left(\mathcal{C}_3\right) = 0.1454 \tag{17c}$$

Eq. (13) to (17) constitute the SIC algorithm.

IV. NOISE TRACKING BASED ON PROBABILISTIC MODEL

It is clear that both the estimated SNR $r(t,k)$ and the tuning parameter $\theta(t)$ are relevant to the performance of the SE system given in Fig. 1. To support the framework for getting

the optimal $\theta^*(t)$, a SIC was developed in Section III, the feature variable of which is also based on $r(t,k)$. Hence, an accurate SNR estimation is compulsory for our optimization problem. Since the SNR estimation can be calculated as a transformation of the noise power estimation, in this section, we will discuss the noise spectral tracking problem. First, we will formulate the signal model in Section IV-A, and then briefly review some state-of-the-art noise spectral tracking algorithms. We develop a so-called PMCRA algorithm in Section IV-C and Section IV-D.

A. Spectral Modeling

We assume the simplest speech-independent additive-noise signal model of the form

$$X(t,k) = S(t,k) + N(t,k) \tag{18}$$

Following the standard assumption that $S(t,k)$ and $N(t,k)$ are statistically uncorrelated across both time and frequency, the expectations of the powers satisfy the following relationship:

$$\begin{aligned} E[|X(t,k)|^2] &= E[X(t,k)X^*(t,k)] \\ &= E[(S(t,k)+N(t,k))\,(S^*(t,k)+N^*(t,k))] \\ &= E[|S(t,k)|^2] + E[|N(t,k)|^2] \end{aligned} \tag{19}$$

In this paper, we work on the log domain, where $\lambda_x(t,k) = \ln E[|X(t,k)|^2]$, $\lambda_s(t,k) = \ln E[|S(t,k)|^2]$, and $\lambda_n(t,k) = \ln E[|N(t,k)|^2]$. The reason for doing this is that it is closer to the human perception, and thus the estimation accuracy is more consistent with the subjective effects.

The expectations cannot be observed, and thus an estimation approach is proposed. It is formulated as using an observation data $|X(t,k)|^2$ to estimate a maximum a posterior probability (MAP) for the noise power model $\lambda_n(t,k)$. The mathematical expression is

$$\lambda_n(t,k)^{\mathrm{MAP}} = \arg\max_{\lambda_n(t,k)} p\left(|X(t,k)|^2|\lambda_n(t,k)\right) f(\lambda_n(t,k)) \tag{20}$$

where f is the prior distribution for $\lambda_n(t,k)$.

B. Conventional Noise Spectrum Tracking Algorithms

The noise power estimation remains to be one of the difficult problem that draws a lot of attention on for the past decades. Several methods have been proposed in the literature.

The minimum statistics (MS) method proposed by Martin [9] is one of the most cited articles in noise estimation problem. It estimates the noise level as the minima of the smoothed noisy speech spectrogram on each frequency bin. The minima is update every 0.5 s to 1.5 s to prevent speech leakage. To implement such a long searching window, a large memory is required to store all the spectral information within the time length. This is generally not applicable in HI, due to the limited hardware sources.

To overcome the drawback of MS, Cohen *et. al.* [2] propose an alternative method which recursively computes the noise power. The estimated noise power in the current frame can be computed as the average of the noise power estimation in the previous frame and the noisy speech power in the current frame. MCRA algorithm is favorable for HI because of its low complexity and low demand of memory. The drawback of MCRA is that, the inaccuracy of the speech presence probability estimation causes the problem of speech leakage [10], but there is no indication of the estimation accuracy available in the algorithm. Moreover, the paper does not indicate the criteria for determining the values of the smoothing factors, which makes it hard to adjust or optimize those coefficients.

In [11], the noise power is modeled as a single Gaussian distribution. The speech is represented as a Gaussian mixture model. The evolution of the hyperparameters in the model is implemented in the form of Kalman filter: given all the possible speech model, the noise estimation is calculated recursively over time, using the incoming observation and the combined speech and noise model. This method has twice the degrees of freedom than MCRA, which predicts the uncertainty for noise estimation as well. But it needs the speech model that is pre-trained with clean speeches. Not only the complexity of this approach is quite high, but the characteristics of the natural speech signal is inherently too complex to be accurately described by conventional HMMs.

C. Noise Spectrum Tracking Based on Probabilistic Model

Being aware of the pros and cons of the conventional algorithms, we intend to develop a new noise estimation algorithm, in which $\lambda_n(t,k)$ is modeled as $\mathcal{N}\left(\hat{\lambda}_n(t,k), \sigma_n^2(t,k)\right)$. We model it as a Gaussian so that not only the estimated noise power will be given, but the estimation accuracy can be provided. To make the complexity low enough in favor for HI application, both the mean value and its variance will be updated recursively as in MCRA, without a specific model assumption for the speech signal.

Algorithm 1 summarizes our probabilistic MCRA (PMCRA) algorithm. The updating rule for both mean and variance are similar to MCRA, as shown in line 25 and 26. The smoothing factors are determined by the speech presence probability (line 17 and 18), which is a smoothed voice-activity-detector (VAD) (line 16).

The major difference of our algorithm from the others is that, since we model the noise power as a probabilistic distribution, the conventional VAD problem can be formulated as a classification problem: the incoming signal is classified as either noise or noisy speech. According to decision theory, the optimal decision threshold $T(t,k)$ for minimizing the misclassification rate is the point where the curves for priors $p\left(\lambda_n(t,k)\right)$ and $p\left(\lambda_x(t,k)\right)$ cross [12].

We can assume the smoothness in the posterior distribution of $\lambda_n(t,k)$, then the prior of $\lambda_n(t,k)$ is approximately the posterior $\lambda_n(t-1,k)$, the hyperparameters of which are determined at the preceding calculation. The prior distribution of $\lambda_x(t,k)$ is computed according to the prior distribution of $\lambda_n(t,k)$ and $\lambda_s(t,k)$.

Following the model assumption in Section IV-A, the posterior distribution of $\exp\left(\lambda_s(t,k)\right)$ is a shifted version of the posterior of $\exp\left(\lambda_n(t,k)\right)$

$$\lambda_s(t,k) = \ln\left[\exp(\bar{\lambda}_x(t,k)) - \exp(\lambda_n(t,k))\right] \tag{21}$$

$$\sigma_x^2(t,k) = \sigma_n^2(t-1,k) + g'^2\left(\hat{\lambda}_s(t-1,k) - \hat{\lambda}_n(t-1,k)\right)\sigma_n^2(t-1,k) + g'^2\left(\hat{\lambda}_s(t-1,k) - \hat{\lambda}_n(t-1,k)\right)\sigma_s^2(t-1,k)$$
$$= \left[1 + \left(1 - \frac{\exp(\hat{\lambda}_n(t-1,k))}{\exp(\bar{\lambda}_x(t-1,k))}\right)^2 + \left(\frac{\exp(\hat{\lambda}_n(t-1,k))}{\exp(\bar{\lambda}_x(t-1,k))}\right)^2\right]\sigma_n^2(t-1,k) \quad (22)$$

After applying the first-order Taylor series expansion around $\hat{\lambda}_n(t,k)$, we get the posterior distribution of $\lambda_s(t,k)$ also in a form of Gaussian

$$\lambda_s(t,k) \sim \mathcal{N}\left(\hat{\lambda}_s(t,k), \sigma_s^2(t,k)\right) \quad (23)$$

where

$$\hat{\lambda}_s(t,k) = \ln\left[\exp(\bar{\lambda}_x(t,k)) - \exp(\hat{\lambda}_n(t,k))\right]$$
$$\sigma_s^2(t,k) = \left(\frac{\exp(\hat{\lambda}_n(t,k))}{\exp(\bar{\lambda}_x(t,k)) - \exp(\hat{\lambda}_n(t,k))}\right)^2 \sigma_n^2(t,k)$$

The prior distribution of the noisy speech power expectation $\lambda_x(t,k)$ can then be computed as the summation of the posterior of $\lambda_n(t-1,k)$ and that of $\lambda_s(t-1,k)$

$$\lambda_x(t,k) = \lambda_n(t,k) + g\left(\lambda_s(t,k) - \lambda_n(t,k)\right) \quad (24)$$

where g is a nonlinear function

$$g(z) = \ln\left[1 + \exp(z)\right]$$

Applying the Taylor series expansion to Eq. (24)

$$\lambda_x(t,k) = \lambda_n(t-1,k) + g\left[\tilde{\lambda}_s(k) - \tilde{\lambda}_n(k)\right] - g'\left[\tilde{\lambda}_s(k) - \tilde{\lambda}_n(k)\right]$$
$$\left[(\lambda_n(t-1,k) - \tilde{\lambda}_n(k)) - (\lambda_s(t-1,k) - \tilde{\lambda}_s(k))\right] \quad (25)$$

where $(\tilde{\lambda}_n(k),\ \tilde{\lambda}_s(k))$ is the expansion point (equivalent to $(\hat{\lambda}_n(t-1,k),\ \hat{\lambda}_s(t-1,k))$ in our case), and g' is the first-order series expansion coefficient which can be easily computed as

$$g'\left(\hat{\lambda}_s(t,k) - \hat{\lambda}_n(t,k)\right) = \frac{\exp(\hat{\lambda}_s(t,k))}{\exp(\hat{\lambda}_s(t,k)) + \exp(\hat{\lambda}_n(t,k))}$$
$$= \frac{\exp(\bar{\lambda}_x(t,k)) - \exp(\hat{\lambda}_n(t,k))}{\exp(\bar{\lambda}_x(t,k))} \quad (26)$$

Applying the standard Gaussian manipulation to Eq. (25), the prior distribution of the noisy speech power expectation can thus be described as a Gaussian

$$\lambda_x(t,k) \sim \mathcal{N}\left(\bar{\lambda}_x(t-1,k), \sigma_x^2(t,k)\right) \quad (27)$$

where $\sigma_x^2(t,k)$ is given in Eq. (22) at the top of this page.

The time-varying threshold $T(t,k)$ determined by the prior distribution of $\lambda_x(t,k)$ given in Eq. (27) and the posterior distribution of $\lambda_n(t-1,k)$ performs intuitively well, since it takes the accuracy of the prediction model into consideration, together with the observed value. Fig 3 gives two illustrations. In Fig. 3(a), the prediction model for $\lambda_n(t,k)$ is fixed. With a higher observation value $\bar{\lambda}_x(t-1,k)$, $T(t,k)$ is higher, so that the influence of the abrupt changing caused by non-stationary noise can be eliminated. Fig. 3(b) demonstrates the influence of the prediction model on $T(t,k)$. The mean value

Algorithm 1 Pseudocode for the proposed algorithm

for all $k = 1$ to K **do** {initialization}
  $\hat{\lambda}_n(0,k) = \bar{\lambda}_x(0,k)$, $\sigma_n^2(0,k) = 0$, $p_s(0,k) = 0$, $T(1,k) = \text{Inf}$, $t \leftarrow 1$
end for

while $t \neq$ END **do** {seen new data}
  for all $k = 1$ to K **do**
    if $\bar{\lambda}_x(t,k) \leq T(t,k)$ **then**
      $I(t,k) = 1$
    else
      $I(t,k) = 0$
    end if

    if $I(t,k) > p_s(t-1,k)$ **then**
      $\alpha_p(t,k) = \alpha_a$
    else
      $\alpha_p(t,k) = \alpha_r$
    end if

    $p_s(t,k) = (1-\alpha_p(t,k))p_s(t-1,k) + \alpha_p(t,k)I(t,k)$

    $\tilde{\alpha}_d^{(1)}(t,k) = \alpha_n^{(1)} + p_s(t,k)(\alpha_s^{(1)} - \alpha_n^{(1)})$
    $\tilde{\alpha}_d^{(2)}(t,k) = \alpha_n^{(2)} + p_s(t,k)(\alpha_s^{(2)} - \alpha_n^{(2)})$

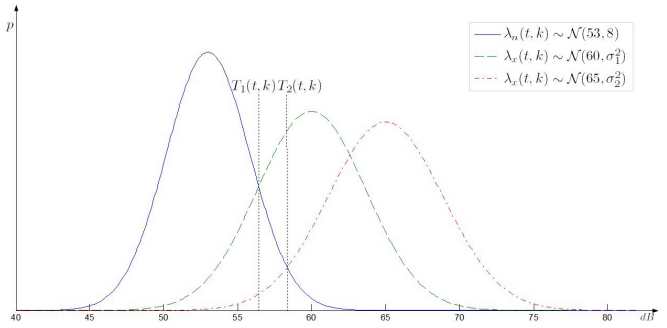
    if $\bar{\lambda}_x(t,k) - T(t,k) > \vartheta$ **then**
      $P^{(1)}(t,k) = \dfrac{T(t,k)}{\gamma(\bar{\lambda}_x(t,k) - T(t,k))}$
    else
      $P^{(1)}(t,k) = \min(\bar{\lambda}_x(t,k), T(t,k))$
    end if
    $P^{(2)}(t,k) = \left|\min(\bar{\lambda}_x(t,k), T(t,k)) - \hat{\lambda}_n(t,k)\right|$

    $\hat{\lambda}_n(t,k) = (1 - \tilde{\alpha}_d^{(1)}(t,k))\hat{\lambda}_n(t-1,k) + \tilde{\alpha}_d^{(1)}(t,k)P^{(1)}(t,k)$

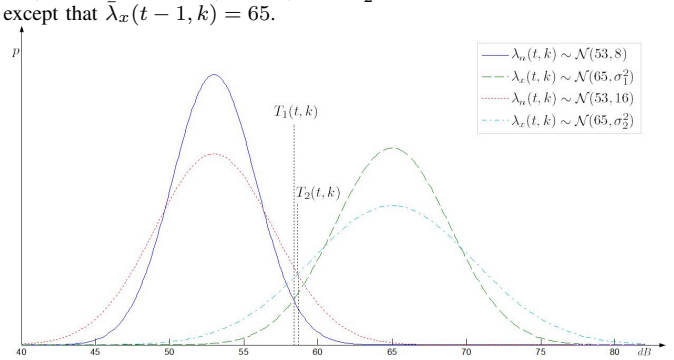
    $\sigma_n(t,k) = (1 - \tilde{\alpha}_d^{(2)}(t,k))\sigma_n(t-1,k) + \tilde{\alpha}_d^{(2)}(t,k)P^{(2)}(t,k)$

    if $\bar{\lambda}_x(t,k) \leq \hat{\lambda}_n(t,k)$ **then**
      $T(t+1,k) = T(t,k)$
    else
      Compute $\sigma_x^2(t+1,k)$ according to Eq. (22)
      Solve $\mathcal{N}\Big(T(t+1,k) \mid \hat{\lambda}_n(t,k), \sigma_n^2(t,k)\Big) = \mathcal{N}\Big(T(t+1,k) \mid \bar{\lambda}_x(t,k), \sigma_x^2(t+1,k)\Big)$
    end if
  end for
  $t \leftarrow t+1$
end while

of the posterior distribution of $\lambda_n(t-1,k)$ and the observed noisy speech power $\bar{\lambda}_x(t-1,k)$ are the same for both models. The prediction model with a larger variance indicates that its estimation $\hat{\lambda}_n(t-1,k)$ may not be so accurate. Consequently, the corresponding threshold $T(t,k)$ is set to be higher to compensate the uncertainty of the estimation.



(a) $T(t,k)$ is higher for a higher observation value. In the figure, σ_1^2 is computed according to equation1Eq. (22), with $\bar{\lambda}_x(t-1,k) = 60$, $\hat{\lambda}_n(t-1,k) = 53$, $\sigma_n^2(t-1,k) = 8$. σ_2^2 is computed in the identical way, except that $\bar{\lambda}_x(t-1,k) = 65$.



(b) $T(t,k)$ is higher for a prediction model with higher variance. In the figure, σ_1^2 is computed according to equation1Eq. (22), with $\bar{\lambda}_x(t-1,k) = 65$, $\hat{\lambda}_n(t-1,k) = 53$, $\sigma_n^2(t-1,k) = 8$. σ_2^2 is computed in the identical way, except that $\sigma_n^2(t-1,k) = 16$.

Fig. 3. The threshold $T(t,k)$ is determined by the observed value $\bar{\lambda}_x(t-1,k)$ together with the prior of $\lambda_n(t,k)$.

Another advantage of this Bayesian-based threshold $T(t,k)$ is, besides a hard binary decision for speech presence or absence, it also provides an extra quantitative information about the prior SNR (XNR), which can be computed as $\bar{\lambda}_x(t,k) - T(t,k)$. It is known that, for higher SNR, less information of $\lambda_n(t,k)$ can be achieved from $\bar{\lambda}_x(t,k)$. Hence, we formulate a weighting function shown in Fig. 4 for updating $\hat{\lambda}_n(t,k)$ using $T(t,k)$, which suppresses the contribution of noisy speech to the estimation in high XNR regions.

Mathematically, $P^{(1)}(t,k)$ is defined as the product of this nonlinear weighting function and $T(t,k)$, as given from line 19 till line 23. The updating value for $\sigma_n^2(t,k)$ is defined as the corresponding maximal estimation error shown in line 24.

D. Data-driven Coefficients Optimization

Noticed that the parameters α_a, α_r, $\alpha_s^{(1)}$, $\alpha_n^{(1)}$, $\alpha_s^{(2)}$, $\alpha_n^{(2)}$, ϑ and γ in our algorithm are undetermined coefficients. We will derive the values using a data-driven approach.

1) Optimization objective: Conventional data-driven approach for noise spectral tracking uses first-order-moment criteria as objective function, which tries to minimize the average

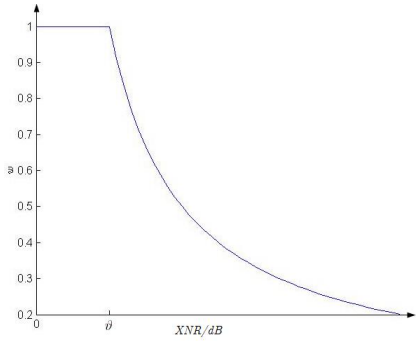


Fig. 4. Weighting function for updating $\hat{\lambda}_n(t,k)$ using $T(t,k)$.

of the term $\left|\bar{\lambda}_n(t,k) - \hat{\lambda}_n(t,k)\right|$ over all time-frequency bins. Our proposed algorithm, however, has one more degree of freedom, and thus, a more proper objective function should be used to incorporate $\hat{\lambda}_n(t,k)$ and $\sigma_n^2(t,k)$ into computation at the same time. Here we propose using the data *Evidence* [13] as objective function, and thus we try to maximizing the following term:

$$\frac{1}{TK}\sum_{\tau=1}^{T}\sum_{k=1}^{K}\mathcal{N}\left(\bar{\lambda}_n(\tau,k) \mid \hat{\lambda}_n(\tau,k), \sigma_n^2(\tau,k)\right) \qquad (28)$$

the average runs over all time-frequency bins.

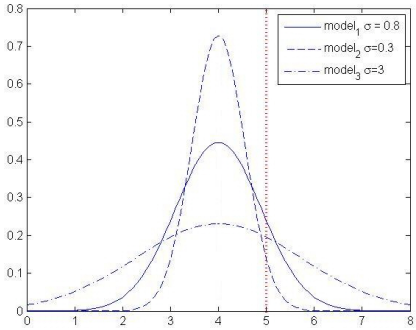


Fig. 5. Three different prediction models with the same prediction value (mean value of the Gaussian models). The real value is labeled by the dot line. Though $\left|\bar{\lambda}_n(t,k) - \hat{\lambda}_n(t,k)\right|$ are the same, we regard model 1 as the one having the best prediction capability.

Evidence is a more suitable measurement than the other objective measurements for the following reason. $\left|\bar{\lambda}_n(t,k) - \hat{\lambda}_n(t,k)\right|$ only measures the deviation of the mean value of $\lambda_n(t,k)$. The performance of a prediction model, however, is determined by both its mean and variance. This argument can be illustrated in Fig. 5. Three prediction models are presented, with the same prediction value 4 (mean value of the Gaussian models following our assumption). The true value is 5. Though the mean value deviation for three prediction models are the same, from the probabilistic point of view, model 1 has a superior prediction capability than the other 2. Compared with model 1, model 2 is too much confident in its prediction, while model 3 barely distinguishes different possible noise power levels. In contrast, the likelihood of the true value (evidence) under 3 different models, which can be read from the y coordinate of the cross point of the

dot line and the probability density curve, can well reflects the preference. In other words, *Evidence* can represent the prediction capability of each model better. Therefore, *Evidence* is a much proper measurement under probabilistic framework.

2) Training procedure: Our aim is finding the optimal coefficients that maximizing Eq. (28) for a wide range of SNR conditions. To cover the range of practical interest, we consider the range -5 dB to 20 dB of segmental SNR [14] in steps of 1 dB, which is defined as

$$SNR_{seg} = \frac{10}{M}\sum_{m=0}^{M-1}\log_{10}\frac{P_s(m)}{P_n(m)} \tag{29}$$

where $P_s(m)$ and $P_n(m)$ are the clean reference signal and the noise power at the m^{th} frame respectively. In our experiment, each frame is 256 ms.

A training set is constructed for training these parameters, and a validation set is used to prevent over-fitting problem.

The sound database we have used are constructed artificially adding noise to speech. The speech utterances are sentences taken from the TIMIT-TRAIN-DR1 database [15], the duration of which is between 1 s to 5 s. Noise signals are taken from database Noisex92 [16]. All the audio files are sampled at 16kHz sampling rate with 16 bits per sample.

Our training set consists of 100 noisy speech (1 to 5 s each). The generating procedure is as follows. The clean speech are 20 sentences spoken by 2 different people, one male and one female (FCJF0 and MDAC0 respectively). Chosen noise sources correspond to the following scene: flying jet with air breaks out (buccaneer1.wav), cockpit (f16.wav), plate-cutting and electrical welding factory (factory1.wav), HF radio channel (hfchannel.wav), and pink noise (pink.wav). In each audio class, 100 draws are performed. During each draw, one of the files will be selected with uniform probability. The corresponding SNR level is also determined by drawing one of the possible values under uniform distribution. The mixture audio is composed by the selected speech file and the selected noise file with the selected SNR level, and then the magnitude of which is normalized to 1.

The available audio materials for constructing the validation set is the same as that for the training set. 10 noisy speech are generated separately following the same procedure.

The optimization procedure is performed by the Global Optimization Toolbox provided by MATLAB. The chosen algorithm is pattern search which is suitable for non-smooth problems without available gradients. The optimal values are obtained before the validation set starting over-fitting.

E. Probabilistic SNR Estimation

With the estimated noise power, the SNR estimation can be obtained according to the standard power relation in Eq. (19). The estimated local SNR (dB) on each time-frequency bin is

$$\begin{aligned} r(t,k) &= 10\log_{10}\left(\frac{|S(t,k)|^2}{|N(t,k)|^2}\right) \\ &= \frac{10}{\ln 10}\ln\left(\frac{|X(t,k)|^2}{|N(t,k)|^2}-1\right) \\ &= \frac{10}{\ln 10}\ln\left(\exp\left(\bar{\lambda}_x(t,k)-\lambda_n(t,k)\right)-1\right) \\ &\triangleq \xi\ln\left(\exp\left(\bar{\lambda}_x(t,k)-\lambda_n(t,k)\right)-1\right) \end{aligned} \tag{30}$$

It is easy to show that the term $\exp\left(\bar{\lambda}_x(t,k)-\lambda_n(t,k)\right)$ is log-normal distributed, since

$$\bar{\lambda}_x(t,k)-\lambda_n(t,k) \sim \mathcal{N}\left(\bar{\lambda}_x(t,k)-\hat{\lambda}_n(t,k), \sigma_n^2(t,k)\right)$$

Applying the standard characteristics of log-normal random variables to Eq. (30), we can conclude that $r(t,k)$ is a Gaussian variable with the following distribution

$$r(t,k) \sim \mathcal{N}\left(\xi\mu(t,k), \xi^2\sigma^2(t,k)\right) \tag{31}$$

where

$$\sigma^2(t,k) = \ln\left(1+\frac{\left(e^{\sigma_n^2(t,k)}-1\right)e^{2\left(\bar{\lambda}_x(t,k)-\hat{\lambda}_n(t,k)\right)+\sigma_n^2(t,k)}}{\left(e^{\bar{\lambda}_x(t,k)-\hat{\lambda}_n(t,k)+\frac{1}{2}\sigma_n^2(t,k)}-1\right)^2}\right)$$

$$\mu(t,k) = \ln\left(e^{\bar{\lambda}_x(t,k)-\hat{\lambda}_n(t,k)+\frac{1}{2}\sigma_n^2(t,k)}-1\right)-\frac{1}{2}\sigma^2(t,k)$$

V. Experimental Results and Discussions

A. Experimental Setup

For constructing the test set, 20 more speech files spoken by other 2 speakers (FDAW0 and MEDR0 in TIMIT-TRAIN-DR1) are added into the clean speech database in Section IV-D2. Noise files recorded in cock-tail party (babble.wav), flying jet with different hight and speed (buccaneer2.wav), car production factory (factory2.wav), car interior (volvo.wav), and the white noise (white.wav) in Noisex92 are added into the noise database as well. 10 sentences are randomly selected from the speech database with uniform probability, and then concatenated with 500ms silent period at the beginning. All 10 different types of noise are then added to these speech files under 26 SNR levels (-5 dB to 20 dB in steps of 1 dB), resulting in 2600 noisy speech for testing.

The coefficients for PMCRA are obtained through the training procedure in Section IV-D2, and the values are $\alpha_s^{(1)}=0.0003$, $\alpha_n^{(1)}=0.7415$, $\alpha_s^{(2)}=0.0029$, $\alpha_n^{(2)}=0.2359$, $\alpha_a=0.0719$, $\alpha_r=0.0001$, $\vartheta=9.7178$ dB, $\gamma=0.1029$. For the reference system, MCRA are set to be the same as [2].

Using the same training set, scene-conditional optimal tuning parameters are derived as in Section II-B. The values are $\theta_1^*=0.0447$, $\theta_2^*=0.0944$, and $\theta_3^*=0.2239$, corresponding to 27 dB, 20.5 dB and 13 dB maximal attenuation respectively.

B. Evaluation Criteria

1) Direct Evaluation: In the direct evaluation method, the MAP estimate of $\lambda_n(t,k)$ will be compared to $\bar{\lambda}_n(t,k)$ using the system shown in Fig. 6. The results are demonstrated in Tabel III.

$LogErr$ computes the logarithmic estimation error which is defined in [10]:

$$LogErr = \frac{1}{|\mathcal{M}|K}\sum_{m\in\mathcal{M}}\sum_{k}\left|\bar{\lambda}_n(m,k)-\hat{\lambda}_n(m,k)\right| \tag{32}$$

where K is the number of frequency bins. The frames which do not contain noise are left out in the computation of $LogErr$, that is, frames with a noise energy more than 40 dB

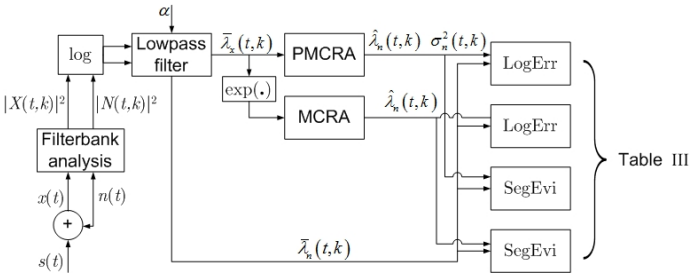


Fig. 6. System diagram for direct evaluation method. α is calculated by Eq. (B.40) setting $\tau_{90} = 32ms$ to approximate the 32 ms analysis window as required in [2].

below the noise energy of the frame with maximum noise energy were not included in the index set $\mathcal{M}$. $|\mathcal{M}|$ is the cardinality of $\mathcal{M}$. To exclude the influence of the transient adaption process, the output in first 500ms are ignored.

$SegEvi$ computes the segmental evidence defined as:

$$SegEvi = -\frac{1}{|\mathcal{M}|K} \sum_{m\in\mathcal{M}} \sum_{k} \log p\big(\bar{\lambda}_n(m,k) \mid \hat{\lambda}_n(m,k), \sigma_n^2(m,k)\big) \quad (33)$$

where the definitions of K, and $\mathcal{M}$ are the same as in Eq. (32). This criterion is equivalent to the mean value of the *Evidence* discussed in Section IV-D2 over one sentence. To make $SegEvi$ computable for MCRA, we assign a fixed value of 10 for it.

2) Indirect Evaluation: In the indirect assessment, the quality is evaluated as the PESQ score using the system shown in Fig. 7. The results are summarized in Table IV.

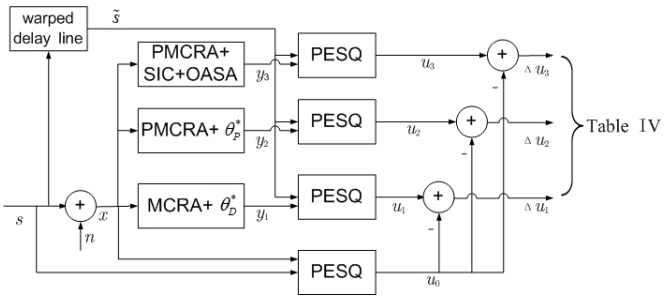


Fig. 7. System diagram for indirect evaluation method.

θ_P^* and θ_D^* are the global optimal tuning parameter defined in Eq. (5), where $\mathcal{X}$is the training set. The values are $\theta_P^* = 0.0398$ and $\theta_D^* = 0.0794$, corresponding to 28 dB and 22 dB maximal attenuation respectively. For PMCRA+OASA system, the averaging time for the feature variable α_τ is set to be 2.5 s.

C. Performance Analysis

We evaluate the performance on each sentence in the test set, and then compute the mean and the variance over the sentences. The values in the tables are averaged over all the sentences and conditions, and thus can be treated as expected $LogErr$, expected $SegEvi$ and expected utility respectively.

From Table III, we can see that, on average, PMCRA clearly outperforms MCRA in terms of $LogErr$ and $SegEvi$, especially in the cases of high input SNR. Fig. 8 and Fig. 9 give the details under separated noise conditions. These two figures further reveal that $Evidence$ is actually a more appropriate training/evaluation criterion since it is consistent over different conditions. In Fig. 8, under different background noises and SNR's, the $LogErr$ varies significantly. Consequently, if we use the expected $LogErr$ of these four noise types as optimization objective, the performance will be dominated by the performance under car interior noise, and thus over-fit to this specific noise type. On the contradictory, the differences of $Evidence$ among alternative conditions are small.

Table IV shows the results evaluated by the indirect criterion on our test set. Fig. 10 gives the details under individual noise. The PESQ score using PMCRA (either using SIC+OASA or not) is generally higher than using MCRA, which indicates a better quality in enhanced speech. The PMCRA+SIC+OASA system does not perform worse than PMCRA+θ_P^* system in general. In some certain cases, it even performs slightly better, though it cannot foresee the whole sentence as during off-line training. The reason that the improvement between PMCRA+θ_P^* and PMCRA+SIC+OASA is not significant is probably because the speech quality of the sentences in the test set are generally the same with those in the training set, so that the global optimization result on the training set still applicable on the test set. It is also worth to noticed that, though PMCRA performs better than MCRA in terms of both $LogErr$ and $SegEvi$ for the whole range of SNR's under HF radio channel noise, in Fig. 10(a), the enhanced speech with the SNR in the range of [0 dB, 10 dB] does not have a better quality than the other. This phenomenon supports our claim that a signal level optimization does not guarantee a better perceptual effect.

VI. Conclusion

Table V summarizes the major contributions in this paper.

TABLE V
CONTRIBUTIONS IN THIS PAPER.

	Algorithm	Results	Refer
1	PMCRA	$\Delta LogErr$ = -1.81 dB $\Delta SegEvi$ = -0.49 1/dB	Section IV-C Algorithm 1
2	SIC	ΔPESQ = +0.33	Section III Fig. 2 Eq. (13) Eq. (14) Eq. (16) Eq. (17)
3	OASA		Section II Eq. (10) Eq. (11) Eq. (12)

PMCRA algorithm models the noise power as a time-varying Gaussian. The advantage of adding one more degree of freedom to the model is that, the conventional VAD problem can be converted into the optimal classification problem, and the consequent decision threshold becomes computable and flexible. The coefficients in PMCRA algorithm are trained under a more proper optimization objective *Evidence*, since it is more consistent over different conditions. SIC is a perceptually

TABLE III

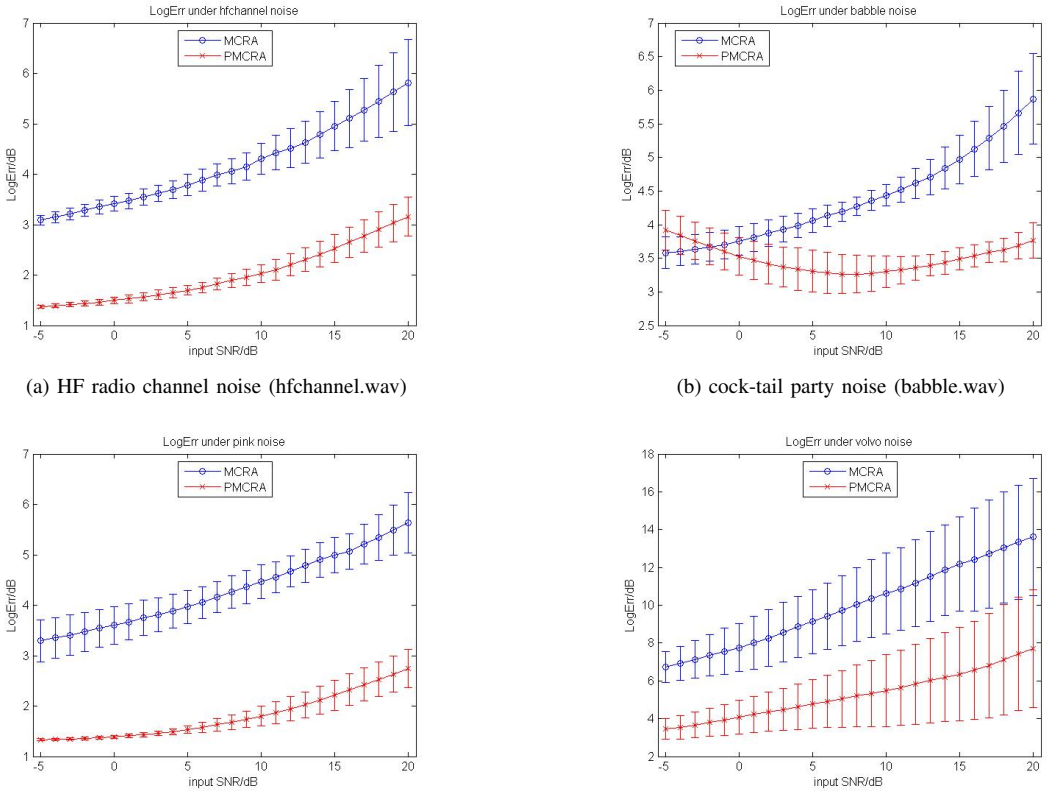
DIRECT EVALUATION RESULTS FOR VARIOUS NOISE TYPES AND LEVELS, OBTAINED USING MCRA AND PMCRA. THE RESULTS ARE THE AVERAGE VALUES OVER ALL THE NOISE CONDITIONS WITHIN THE CORRESPONDING SET.

SNR[dB]	$LogErr$ [dB]					$SegEvi$ [1/dB]				
	Training Set		Test Set		Δ	Training Set		Test Set		Δ
	MCRA	PMCRA	MCRA	PMCRA		MCRA	PMCRA	MCRA	PMCRA	
-5	2.87	1.59	3.52	2.89		1.23	1.14	1.40	1.29	
0	3.17	1.63	3.96	2.92		1.28	1.09	1.55	1.26	
5	3.54	1.76	4.54	3.01		1.34	1.08	1.78	1.26	
10	4.01	2.02	5.24	3.19		1.43	1.12	2.08	1.28	
15	4.59	2.43	6.09	3.57		1.58	1.18	2.44	1.34	
20	5.26	2.97	7.08	4.18		1.80	1.25	2.85	1.44	
AVG.	**3.88**	**2.03**	**5.03**	**3.25**	**-1.81**	**1.43**	**1.14**	**2.00**	**1.30**	**-0.49**

TABLE IV

THE PESQ SCORE FOR VARIOUS NOISE TYPES AND LEVELS, OBTAINED USING MCRA+θ_D^*, PMCRA+θ_P^*, AND PMCRA+SIC+OASA. THE RESULTS ARE THE AVERAGE VALUES OVER ALL THE NOISE CONDITIONS WITHIN THE CORRESPONDING SET.

PESQ evaluation on the test set							
SNR[dB]	Origin	MCRA+θ_D^*	Δ	PMCRA+θ_P^*	Δ	PMCRA+SIC+OASA	Δ
-5	1.89	2.13	+12.26%	2.23	+17.47%	2.23	+17.52%
0	2.25	2.52	+11.93%	2.61	+16.13%	2.62	+16.21%
5	2.61	2.87	+9.62%	2.98	+14.14%	2.99	+14.22%
10	2.97	3.18	+7.18%	3.32	+11.77%	3.32	+11.78%
15	3.30	3.46	+4.90%	3.59	+8.65%	3.59	+8.81%
20	3.61	3.69	+2.18%	3.82	+5.71%	3.83	+6.02%
AVG.	**2.78**	**2.99**	**+0.21**	**3.10**	**+0.32**	**3.11**	**+0.33**



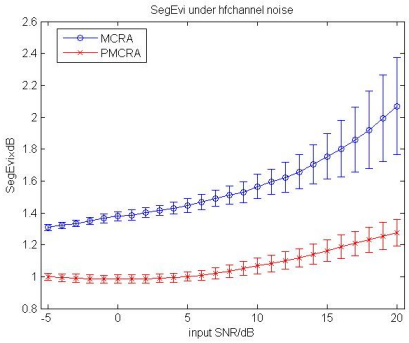
(a) HF radio channel noise (hfchannel.wav)

(b) cock-tail party noise (babble.wav)

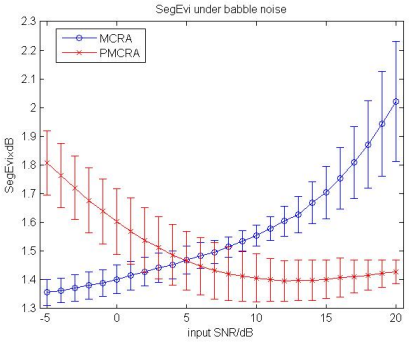
(c) pink noise (pink.wav)

(d) car interior noise (volvo.wav)

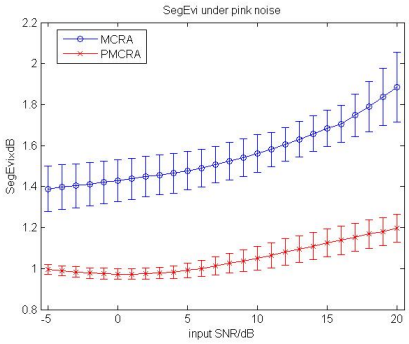
Fig. 8. Evaluation results of $LogErr$ under different noise conditions. The background noise in the figure listed in the left column are used in our training set, while that in the right column only appears in our test set. The x axis is the input SNR.



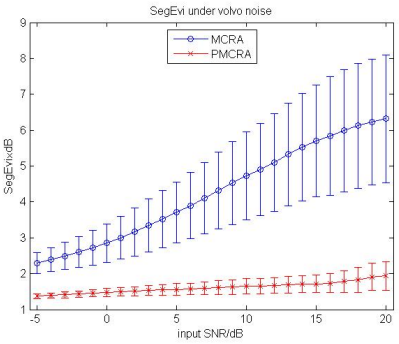
(a) HF radio channel noise (hfchannel.wav)



(b) cock-tail party noise (babble.wav)

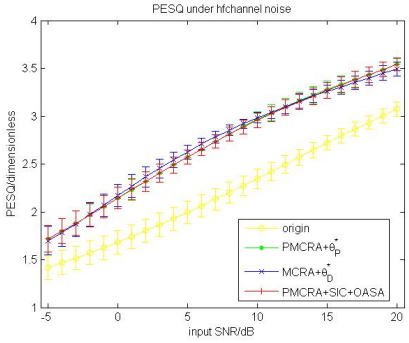


(c) pink noise (pink.wav)

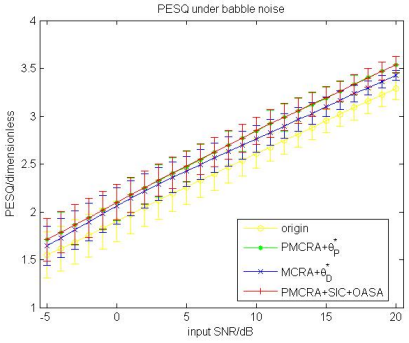


(d) car interior noise (volvo.wav)

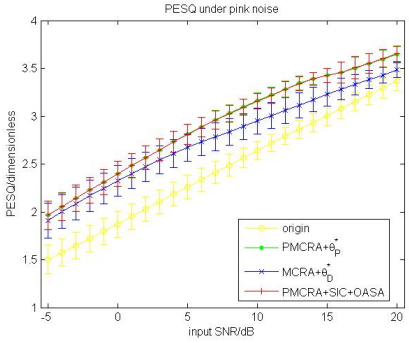
Fig. 9. Evaluation results of $SegEvi$ under different noise conditions. The background noise in the figure listed in the left column are used in our training set, while that in the right column only appears in our test set. The x axis is the input SNR.



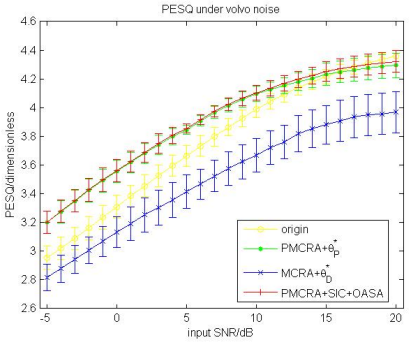
(a) HF radio channel noise (hfchannel.wav)



(b) cock-tail party noise (babble.wav)



(c) pink noise (pink.wav)



(d) car interior noise (volvo.wav)

Fig. 10. Evaluation results of PESQ under different noise conditions. The background noise in the figure listed in the left column are used in our training set, while that in the right column only appears in our test set. The x axis is the input SNR.

defined ASA, which can well integrated into any existing SE algorithm in HI. It is able to estimate the intelligibility of the incoming signal and label them as low, medium or high SI class accordingly. OASA is a novel optimization framework. Different from minimizing signal spectrum estimation error, it directly optimizes the perceptual performance of the SE algorithm. All three algorithms together result in a increment in PESQ about 0.33 on average, which is 0.12 higher than the reference.

Note that, in this work, though the input to SIC is an Gaussian, the variance is not used during classification. This could be one of the reason for the system adopting OASA does not clearly outperforms the global optimized system, and should be further investigated in future work.

APPENDIX A
NOISE REDUCTION ALGORITHM

As one of the most widely used method for reducing additive noise with extremely low complexity, the spectral subtraction method of speech enhancement has been proposed by Boll [17] in 1979. An estimation of the clean speech spectrum is obtained by subtracting an estimate of the noise spectrum from the noisy speech spectrum. The phase information is ignored based on the fact that human hearing perception are not sensitive to the phase distortion. The mathematical expression of this principle is

$$|S(t,k)| = |X(t,k)| - |\hat{N}(t,k)| \tag{A.34}$$

Consequently, the estimated speech signal $\hat{S}(t,k)$ can be computed as

$$\hat{S}(t,k) = G(t,k)X(t,k) \tag{A.35}$$

where $G(t,k)$ is a gain function independent of phase.

Based on the additive noise model given in Eq. (18), the gain function can be estimated as

$$\begin{aligned} G(t,k) &= 1 - \frac{|\hat{N}(t,k)|}{|X(t,k)|} = 1 - \sqrt{\frac{|\hat{N}(t,k)|^2}{|X(t,k)|^2}} \\ &= 1 - \sqrt{\exp\left(\hat{\lambda}_n(t,k) - \bar{\lambda}_x(t,k)\right)} \end{aligned} \tag{A.36}$$

In a hearing-aids setting, getting rid of all ambiance noise is not desirable. Complete suppressing the ambient environment leads to an unnatural sensory experience, and possibly even a dangerous situation. It is also desired that the residual noise spectrum be similar to the original noise spectrum. This leads to a minimal gain G_{min} when there is pure noise, *i.e.* when $\bar{\lambda}_x(t,k) = \hat{\lambda}_n(t,k)$. Consequently, the gain function in Eq. (A.36) is improved to be

$$G(t,k) = 1 - (1 - G_{min})\sqrt{\exp\left(\hat{\lambda}_n(t,k) - \bar{\lambda}_x(t,k)\right)} \tag{A.37}$$

The physical interpretation of G_{min} is the maximal attenuation in noise reduction, and the corresponding attenuation can be converted into dB unit as

$$G_{min}^{dB} = -20 \log G_{min} \tag{A.38}$$

APPENDIX B
TIME-AVERAGING LOWPASS FILTER

Time-averaging can be approximated by a lowpass filter in a form of leaky integrator

$$y(t) = (1-\alpha)y(t-1) + x(t) \tag{B.39}$$

where y is the approximated mean, x is the incoming signal, and α is the smoothing factor computed according to the length of time for average τ_{90}

$$\alpha = 1 - \exp\left(\frac{-2.3 \times T_s}{\tau_{90}}\right) \tag{B.40}$$

Here, T_s is sampling period. For the system in our work, with 16kHz audio sampling rate and 24 samples block processing manner, its value is 24/16000 s.

REFERENCES

[1] J. M. Kates and K. H. Arehart, "Multichannel dynamic-range compression using digital frequency warping," *EURASIP Journal on Applied Signal Processing*, vol. 18, pp. 3003–3014, 2005.

[2] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Processing Letters*, vol. 9, no. 1, pp. 12–15, January 2002.

[3] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *The Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387–3405, 2009.

[4] E. Alexandre, L. Cuadra, L. Álvarez, M. Rosa-Zurera, and F. López-Ferreras, "Two-layer automatic sound classification system for conversation enhancement in hearing aids," *Integr. Comput.-Aided Eng.*, vol. 15, no. 1, pp. 85–94, 2008.

[5] M. Büchler, S. Allegro, S. Launer, and N. Dillier, "Sound classification in hearing aids inspired by auditory scene analysis," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 2991–3002, 2005.

[6] S. Ravindran and D. Anderson, "Audio classification and scene recognition and for hearing aids," in *Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium*, vol. 2, 2005, pp. 860–863.

[7] K. D. Kryter, "Methods for the calculation and use of the articulation index," *The Journal of the Acoustical Society of America*, vol. 34, no. 11, pp. 1689–1697, 1962.

[8] K. S. Rhebergen and N. J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2181–2192, 2005.

[9] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Transactions on Speech and Audio Porcessing*, vol. 9, no. 5, pp. 504–512, July 2001.

[10] J. S. Erkelens and R. Heusdens, "Fast noise tracking based on recursive smoothing of MMSE noise power estimates," in *ICASSP*, 2008, pp. 4873–4876.

[11] L. Deng, J. Droppo, and A. Acero, "Incremental bayes learning with prior evolution for tracking nonstationary noise statistics from noisy speech data," in *ICASSP*, 2003, pp. 672–675.

[12] C. M. Bishop, *Pattern recognition and machine learning*. Springer New York, 2007.

[13] W. Penny, "Kalman filters." [Online]. Available: http://www.fil.ion.ucl.ac.uk/~wpenny/course/kalman.ps

[14] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Taylor & Francis Group, LLC, 2007.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Darpa timit acoustic phonetic continuous speech corpus cdrom," 1993.

[16] 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 Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[17] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, April 1979.