A HIGHLY NON-STATIONARY NOISE TRACKING AND COMPENSATION ALGORITHM, WITH APPLICATIONS TO SPEECH ENHANCEMENT AND ON-LINE ASR

Md Foezur Rahman Chowdhury[†] *Sid-Ahmed Selouani*[⋆] *Douglas O'Shaughnessy*[†]

[†] INRS-EMT, Université du Quebéc, Montréal, H5A 1K6, QC, Canada
[⋆] Université de Moncton, Campus de Shippagan, E8S 1P6, NB, Canada

ABSTRACT

This paper presents a noise tracking and estimation algorithm for highly non-stationary noises using the Bayesian on-line spectral change point detection (BOSCPD) technique. In BOSCPD, the local minima search window update technique of minima controlled recursive averaging (MCRA) algorithm is made a function of spectral change point detection. The novelty of this algorithm is that it can detect the rapid changes instantly and quickly update the non-stationary noise estimate compared to the MCRA-based algorithms. The BOSCPD algorithm shows improvement in objective quality measures in terms of higher SNR and lower output distortion scores for speech enhancement. It is also tested to track and compensate for rapidly varying noises in on-line automatic speech recognition (ASR) using the Aurora 2 speech database. The simulation results show significant improvement in recognition accuracy compared to the baseline MCRA technique.

Index Terms— Non-stationary noise tracking and estimation, Bayesian on-line change point detection, on-line automatic speech recognition, joint additive and channel noises compensation, MCRA

1. INTRODUCTION

The state-of-the-art MCRA [1] is a single channel noise tracking algorithm. It tracks the minimum of the noisy speech power spectrum within a finite search window. However, the major disadvantage of MCRA is that it takes more time than the minima search window duration to update the noise estimate when the noise floor increases abruptly and stays at that level [2]. Several improved versions of MCRA, e.g., MCRA2 [2], improved MCRA (IMCRA) [3], enhanced MCRA (EMCRA) [4] have been proposed for minimizing this delay problem. These algorithms reduce the delay in adaptation to new conditions to some extent compared to the MCRA. However, they fail to minimize this delay significantly and it still remains a challenging problem to speech researchers.

Recently, the Bayesian on-line change point detection (BOCPD) technique has found its application for real-world time series, e.g., finance, volatility in stock market, oil drilling, robotics, and satellite fault detection and tracking [5]. One salient feature of BOCPD is that it allows one to express uncertainty about the number and location of change points. This property makes it easy to be used as a model for a frame-based causal predictive filter, i.e., to generate an accurate predictive distribution of the next unseen spectral data of the speech frame, given only the statistical properties of the already observed speech frames.

In this paper, we propose a new method using the BOCPD technique to minimize the delay problem of the MCRA. The proposed method detects the abrupt changes in speech spectral properties, and becomes a function for the local minima search window process. The BOSCPD technique significantly reduces the delay problem in detecting abrupt changes in MCRA-based algorithms. And also it does not overestimate the noise spectrum. The proposed BOSCPD technique is tested for speech enhancement and compensation in the front-end of an on-line ASR and it shows better recognition results than the MCRA.

This paper is organized as follows. Section 2 describes the MCRA method. The BOCPD method is presented in section 3. Section 4 presents the proposed BOSCPD method. Front-end processing of on-line ASR is shown in section 5. These are followed by the experimental results and conclusions.

2. MCRA FOR SINGLE CHANNEL NOISE TRACKING

Let $y(n) = x(n) + d(n)$, where $y(n)$ is the noisy speech signal, $x(n)$ is the clean signal and $d(n)$ is the uncorrelated additive noise. The short-time Fourier transform (STFT) $Y(m,k)$ of the noisy signal can be calculated by first applying a window $w(n)$ to N samples of $y(n)$ and then computing the N-point FFT of the windowed signal:

$$Y(m,k) = \sum_{\lambda=0}^{N-1} y(\lambda + mM)w(\lambda)e^{-j\frac{2\pi}{N}\lambda k}, \tag{1}$$

where m is the frame index, $k(k = 0,1,2,...,N-1)$ is the frequency bin index, and M is the frame update step. The periodogram $P(m,k)$ of $Y(m,k)$ can be estimated using a first-order recursive formula as follows:

$$P(m,k){=}\alpha(m,k)P(m{-}1,k) + (1{-}\alpha(m,k))|Y(m,k)|^2, \tag{2}$$

where $\alpha(m,k)$ is a time and frequency dependent smoothing parameter.

In MCRA, the smoothing parameter $\alpha(m,k)$ is made dependent on the conditional speech presence probability based on the following hypothesis:

$$\begin{aligned} H_0^k &: \hat{\sigma}_d^2(m,k){=}\alpha_d\hat{\sigma}_d^2(m-1,k) + (1-\alpha_d)|Y(m,k)|^2, \\ H_1^k &: \hat{\sigma}_d^2(m,k){=}\hat{\sigma}_d^2(m-1,k), \end{aligned} \tag{3}$$

where α_d $(0 \leq \alpha_d \leq 1)$ is a smoothing parameter. H_0^k and H_1^k designate hypothetical speech absence and presence, respectively, in the mth frame of the kth frequency bin. $\sigma_d^2(m,k)$ denotes the variance of the noise in the kth frequency bin [1].

In Eq. (3), the noise estimate is updated whenever speech is absent, otherwise it is kept constant. The noise power spectral density (PSD) can be estimated in the mean-square sense as follows:

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

$$\hat{\sigma}_d^2(m,k) = \tilde{\alpha}_d(m,k)\hat{\sigma}_d^2(m-1,k) + [(1-\tilde{\alpha}_d(m,k))]|Y(m,k)|^2, \quad (4)$$

where $\tilde{\alpha}_d(m,k) = \alpha_d + (1-\alpha_d)p(m,k)$ is a time-varying smoothing parameter and it varies within the range $\alpha_d \leq \tilde{\alpha}_d(m,k) \leq 1$.

Accordingly, the noise spectrum can be estimated by averaging past spectral power values, using a smoothing parameter that is adjusted by the signal presence probability as follows:

$$\hat{p}(m,k) = \alpha_p \hat{p}(m-1,k) + (1-\alpha_p)I(m,k), \quad (5)$$

where α_p $(0 < \alpha_p < 1)$ is a smoothing parameter, $I(m,k) = 1$ if $S_r(m,k) > \delta$ and $I(m,k) = 0$ otherwise. Here $S_r(m,k) = S(m,k)/S_{min}(m,k)$ is the ratio of the local energy of the noisy speech and its derived minimum. δ is a threshold for speech presence [1],

$$S(m,k) = \alpha_s S(m-1,k) + (1-\alpha_s)|Y(k,l)|^2, \quad (6)$$

where α_s $(0 < \alpha_s < 1)$ is a smoothing parameter. $S_{min}(m,k)$ is defined as follows:

$$S_{min}(m,k) = \min\{S(j,k)\}; \text{ for } m-2L < j < m, \quad (7)$$

which is calculated [4] as

$$S_{min}(m,k) = \begin{cases} S(0,k) & \text{if } m = 0, \\ \min\left[S_{min}(m-1,k), S(m,k)\right] & \text{if } m\%L \neq 0, \\ \min\left[S_{tmp}(m-1,k), S(m,k)\right] & \text{otherwise,} \end{cases} \quad (8)$$

$$S_{tmp}(m,k) = \begin{cases} S(0,k) & \text{if } m = 0, \\ \min\left[S_{tmp}(m-1,k), S(m,k)\right] & \text{if } m\%L \neq 0, \\ S(m,k) & \text{otherwise,} \end{cases} \quad (9)$$

where $\%$ sign is used to indicate modulus after division [4].

The parameter L determines the resolution of the local minima search. The local minimum is based on a window of at least L frames, but not more than $2L$ frames. The length of the window controls the bias upwards during continuous speech and the bias downwards when the noise level increases [1].

3. BAYESIAN ON-LINE CHANGE POINT DETECTION

The Bayesian On-line CPD (BOCPD) algorithm has mainly focused on the time since the last change point, called the run length *r* [5], [6]. It used an underlying predictive model (UPM) of the time-series that changes at each change point. It also used a hazard function $H(r|\theta_h)$ which describes how likely a change point is given the run length *r*. The UPM is used to compute the posterior predictive $p(x_t|x_{(t-\tau)}, \theta_m)$ for any $\tau \in [1, ..., (t-1)]$, at time *t*. The parameters $\theta = \{\theta_m, \theta_h\}$ form the set of hyper-parameters for the UPM model, and are assumed to be fixed and known.

The posterior run length $p(r_t|x_{1:t})$ at time *t* is estimated sequentially to predict the on-line changes by marginalizing the run length variable as follows:

$$\begin{aligned} p(x_{t+1}|x_{1:t}) &= \sum_{r_t} p(x_{t+1}|x_{1:t}, r_t)p(r_t|x_{1:t}) \\ &= \sum_{r_t} p(x_{t+1}|x_t^{(r)})p(r_t|x_{1:t}), \end{aligned} \quad (10)$$

where $x_t^{(r)}$ refers to the last r_t observations of x, and $p(x_{t+1}|x_t^{(r)})$ is computed using the UPM. The run length posterior can be found by normalizing the joint likelihood:

$$p(r_t|x_{1:t}) = \frac{p(r_t, x_{1:t})}{\Sigma_{r_t} p(r_t, x_{1:t})}. \quad (11)$$

The joint likelihood can be updated on-line using a recursive message passing scheme

$$\begin{aligned} \gamma_t &:= p(r_t, x_{1:t}) \\ &= \sum_{r_{t-1}} p(r_t, r_{t-1}, x_{1:t})) \\ &= \sum_{r_{t-1}} \underbrace{p(r_t|r_{t-1})}_{hazard} \underbrace{p(x_t|r_{t-1}, x_t^{(r)})}_{UPM} \underbrace{p(r_{t-1}, x_{1:t-1})}_{\gamma_{t-1}}. \end{aligned} \quad (12)$$

This defines a forward message passing scheme to recursively calculate γ_t from γ_{t-1}. The conditional can be restated in terms of messages as $p(r_t|x_{1:t}) \propto \gamma_t$. All the distributions mentioned so far are implicitly conditioned on the set of hyper-parameters $\theta = \{\theta_m, \theta_h\}$. The detailed mathematical description of this BOCPD technique for on-line change point detection can be found in [5] and [6].

4. BOCPD TO DETECT ABRUPT NOISE FLOOR CHANGE

In real-world acoustic environments, both the background additive noise and the channel distortions are highly non-stationary in nature and are not known *a priori*. The statistical properties of the noise power spectrum density (PSD) change very rapidly with time. Under these circumstances, the actual model of the speech signal is highly non-linear and non-Gaussian as follows [7]:

$$\mathbf{y} = \mathbf{x} + \mathbf{q} + IDFT\{\ln(1 + e^{DFT[\mathbf{d}-\mathbf{q}-\mathbf{x}]})\}, \quad (13)$$

where $\mathbf{y}$ is the observed noisy speech signal in the cepstral domain, $\mathbf{x}$ is the uncorrupted speech in the cepstral domain, $\mathbf{q}$ is the channel bias in the cepstral domain, and $\mathbf{d}$ is the additive noise in the cepstral domain.

The changes in real-world acoustic conditions can easily be monitored by tracking the statistical properties of the noise PSD for each frame of the observed speech signal using the UPM model of the BOCPD technique [6]. The BOCPD technique uses this UPM model to detect a change point by predicting the changes in second order statistical properties of the time-series in on-line conditions.

In this paper, we apply the UPM model to detect rapid changes in the noise floor by tracking and monitoring the second order statistic of the noise PSD for each noisy speech frame. The UPM is modeled with an independent and identically distributed (iid) Gaussian observation with changing mean and variance of the kth DFT bin. A distribution called the Normal-Inverse-Gamma on the mean and variance, which is computationally advantageous, is used as follows:

$$|Y(m,k)| \sim \mathcal{N}(\mu, \sigma^2), \quad (14)$$

$$\mu \sim \mathcal{N}(\mu_0, \sigma^2/\kappa), \sigma^{-2} \sim Gamma(\alpha, \beta). \quad (15)$$

In this proposed noise PSD tracking model, the product partition model in [6] is replaced by the speech frames assuming the arrival of each frame is independent of other frames. A Hamming window is used for windowing the speech signal and the temporal correlation effects between overlapped adjacent speech frames is neglected in order to make the UPM model simple. A constant hazard function $H(r|\theta_h) := \theta_{h_{constant}}$ similar to [5] is used in the BOSCPD model. A constant hazard function means $p(r_t = 0|r_{t-1}, \theta_h)$ is independent of r_{t-1} and gives rise to geometric inter-arrival times for change points. Under these conditions, the model hyper-parameters for the proposed Bayesian on-line spectral change point detection (BOSCPD) algorithm are:

$$\theta = \{\mu_0, \kappa, \alpha, \beta, \theta_{h_{constant}}\}. \tag{16}$$

The detailed description of these model hyper-parameters for the BOCPD-based model can be found in [5] and [6]. The authors of this paper published part of this proposed algorithm in [8].

The result of BOSCPD algorithm for each noisy speech frame is a decision whether there is an abrupt change in the noise PSD or not. If there is a change point detected in a noisy speech frame, the algorithm raises a flag and the noise tracking algorithm uses this decision to update its noise estimation process as follows:

$$f(C_{m,k}) = \begin{cases} 1 & \text{if } \textit{change point is detected,} \\ 0 & \text{otherwise,} \end{cases} \tag{17}$$

where $f(C_{m,k})$ is a function of change point C detected by the BOSCPD algorithm for the kth frequency bin of the mth frame of the noisy speech signal. Finally, the noise estimation in Eqs. 8 and 9 can be updated in response to abrupt environmental change detection as shown in Algorithm 1.

Algorithm 1 Updating noise estimation based on the proposed BOSCPD algorithm

```
if mod$(m/L) == 0 \;||\; f(C_{m,k}) == 1$ then
    $S_{min}(m,k) \leftarrow \min\{S_{tmp}(m-1,k), S(m,k)\}$
    $S_{tmp}(m,k) \leftarrow S(m,k)$
else
    $S_{min}(m,k) \leftarrow \min\{S_{min}(m-1,k), S(m,k)\}$
    $S_{tmp}(m,k) \leftarrow \min\{S_{tmp}(m-1,k), S(m,k)\}$
end if
```

Figure 1 shows an example noise spectrum estimated with our algorithm and with MCRA [1], MCRA2 [2], and EMCRA [4] for a scenario in which the spoken utterance is degraded with highly non-stationary babble noise. Our algorithm is able to track non-stationarity in environments and adapt to the new environment without delay while MCRA-based algorithms required large delay to adapt.

Figure 2 compares the performance of the proposed BOSCPD algorithm with MCRA for denoising the noisy speech signal degraded by babble noise. The time window size L is set to 64 frames for both the proposed algorithm and the MCRA algorithm. The proposed algorithm performed better than the original MCRA, which can be easily observed from Figure 2(d-e).

Figure 3 shows an example noise spectrum estimated with our algorithm and with the MCRA [1], MCRA2 [2] and EMCRA [4] algorithms for a scenario in which the noise environment changes suddenly with an increased noise floor. Our algorithm is able to adapt to the new environment within 0.08 sec, while the MCRA and EMCRA algorithms required 1.1 secs, and the MCRA2 algorithm required 1.3 secs to adapt.

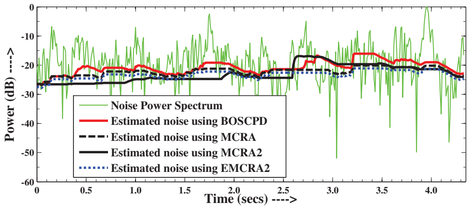
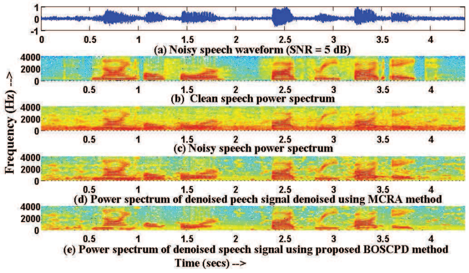


Fig. 1. Comparison between the noise spectrum (for f=1.5 kHz) estimated using the proposed algorithm and MCRA [1], MCRA2 [2] and EMCRA [4] algorithms for a sentence corrupted by babble noise at 5 dB SNR.



(a) Noisy speech waveform (SNR = 5 dB)

(b) Clean speech power spectrum

(c) Noisy speech power spectrum

(d) Power spectrum of denoised peech signal denoised using MCRA method

(e) Power spectrum of denoised speech signal using proposed BOSCPD method

Fig. 2. Comparison of speech enhancement performances using the proposed algorithm and MCRA algorithms for the test utterance corrupted by babble noise.

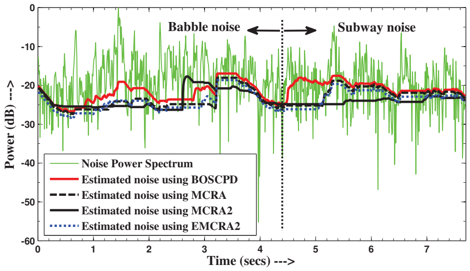


Fig. 3. Comparison between the noise spectrum (for f=1.5 kHz) estimated using the proposed algorithm and the MCRA [1], MCRA2 [2] and EMCRA [4] algorithms for a sentence corrupted by babble noise (t < 4.4 sec) followed by a sentence corrupted by subway noise (t > 4.4 sec).

5. FRONT-END PROCESSING OF ON-LINE ASR

In real-world environments, both the background noise and the channel distortion vary abruptly with a change of temporal-spatial conditions. A frame dynamic joint additive and channel distortions compensation (JAC) [7] method would be suitable to address this problem. In the JAC technique, first the noise tracking algorithm compensates the background noise in the linear spectral domain and then the channel bias needs to be compensated in the cepstral domain for each speech frame. A first order recursive filter with a time smoothing constant could be used to estimate the channel bias as follows:

$$\bar{\mathbf{x}}_t = \hat{\mathbf{y}}_t - \bar{\mathbf{b}}_{t-1}, \tag{18}$$

$$\bar{\mathbf{b}}_t = \alpha\bar{\mathbf{b}}_{t-1} + (1-\alpha)\hat{\mathbf{y}}_t, \tag{19}$$

where $\hat{\mathbf{y}}_t$ is the additive noise compensated observed cepstral feature for the current frame, $\bar{\mathbf{x}}_t$ is the bias compensated cepstral feature, $\bar{\mathbf{b}}_t$

is the bias estimate in the cepstral domain from the previous frame and the current observation using a first order recursive filter, and $\alpha\{\alpha = 0.995\}$ is a time smoothing constant. The on-line ASR uses a simultaneous bias compensation and recognition scheme and such an approach is very much suitable for real-world applications, where the end of each speech utterance is not known *a priori*.

6. EXPERIMENTAL RESULTS

The proposed noise tracker has been tested in comparison with the popular MCRA [1], MCRA2 [2] and EMCRA [4] for noisy speech enhancement. In all the tests, a standard spectral subtraction-type speech enhancement method has been used to perform the noise removal. Speech signals sampled at 8 kHz are segmented into 25-ms frames using a Hamming window with 60% overlap. We use several standard objective quality measures such as i) global SNR (GSNR), ii) segmental SNR (segSNR), iii) Itakura-Saito distortion (It-Sa), iv) weighted spectral slope (WSS), and v) perceptual evaluation of speech quality (PESQ). For one particular noisy speech file, results are summarized in Table 1.

The proposed BOSCPD algorithm has also been tested in comparison with MCRA for front-end speech processing of an on-line ASR. The on-line ASR is simulated using the ATK toolkit [9] for restaurant noise, street noise, airport noise and train station noise environments using the Aurora 2 speech database. We use a clean-training model for this simulation. The on-line ASR uses 39 MFCCs (13 static MFCC coefficients C_0, C_1, C_2, C_3,...,C_{12}, 13 ΔC coefficients, and 13 $\Delta\Delta C$ coefficients) for training HMMs and testing the utterances. It also uses whole word HMM models with 18 states per word including 2 dummy states at the beginning and the end. These HMM models are left-to-right models without skip-over states. They use mixtures of 6 Gaussians per state.

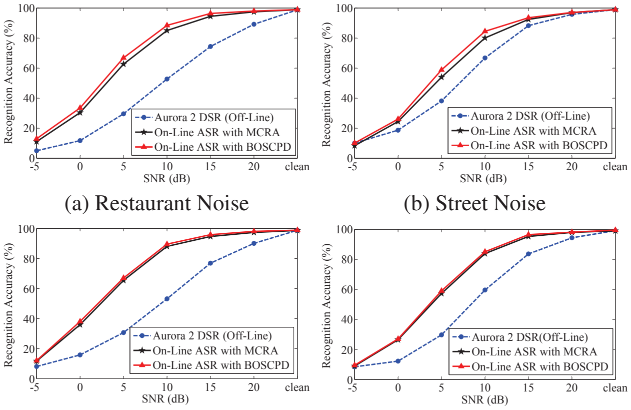
The proposed BOSCPD algorithm performs better than the MCRA technique for on-line ASR. The average increase in the word accuracy for test set 'b' is 23.69% for MCRA and 26.43% for the proposed method compared to the results of the Aurora 2 DSR [10]. Graphically these performance results are shown in Figure 4.

Table 1. Speech Enhancement Comparison of Different Noise Power Spectrum Estimation Techniques.

	GSNR	SegSNR	It-Sa	WSS	PESQ
Noisy Speech	5.264	-1.545	3.848	93.927	1.987
MCRA	9.304	0.623	3.061	85.681	2.357
MCRA2	8.672	0.166	2.612	88.046	2.316
EMCRA	9.352	0.524	3.507	86.488	2.373
BOSCPD	9.397	0.631	3.050	85.382	2.420

7. CONCLUSIONS

In this paper, we present a novel noise estimation algorithm algorithm to track and compensate rapidly varying acoustic noises for speech enhancement and on-line ASR in adverse conditions. The proposed noise tracking algorithm is based on the Bayesian on-line inference for change point detection (BOCPD) technique and the MCRA algorithm. The objective of this algorithm is to reduce the time delay for adapting to abrupt changes in background acoustic noises. Compared to the most popular MCRA-based algorithms, the BOSCPD tracking and noise estimate technique responds more quickly to noise variations. Experimental tests for both the speech enhancement and front-end processing of on-line ASR have demonstrated positive results.



(a) Restaurant Noise

(b) Street Noise

(c) Airport Noise

(d) Train Station Noise

Fig. 4. Performances of the proposed method and MCRA for on-line ASR compared to the results of Aurora 2 DSR(Batch mode) [10].

8. REFERENCES

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

[2] S. Rangachari, and P. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, vol. 48, no. 2, pp. 220–231, February 2006.

[3] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, September 2003.

[4] N. Fan, "Speech noise estimation using enhanced minima controlled recursive averaging," in *Proc., IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 4, Honolulu, Hawaii, USA, April 2007, p. 581 584.

[5] R. Turner, "Bayesian change point detection for satellite fault prediction," in *Proc., Interdisciplinary Graduate Conference (IGC)*, Cambridge, UK, June 2010, pp. 213–221.

[6] R. P. Adams, and D. J. C. MacKay, "Bayesian online changepoint detection," University of Cambridge, Tech. Rep., 2007, report.arXiv:0710.3742v1[stat.ML].

[7] A. Acero, *Acoustical and environmental robustness in automatic speech recognition*. Kluwer Academic Publisher, 1993.

[8] M. F. R. Chowdhury, S. -A. Selouani, and D. O'Shaughnessy, "A rapid adaptation algorithm for tracking highly non-stationary noises based on bayesian inference for on-line spectral change point detection," in *Proc., INTERSPEECH*, Florence, Italy, August 2011, pp. 1205–1208.

[9] S. Young, *ATK Real-Time API for HTK (ver. 1.6)*, Machine Intelligence Laboratory, University of Cambridge, University of Cambridge, UK, June 2007.

[10] H. G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc., ASR-2000*, Beijing, China, October 2000, pp. 181–188.