SPEECH NOISE ESTIMATION USING ENHANCED MINIMA CONTROLLED RECURSIVE AVERAGING

Ningping Fan, Justinian Rosca, Radu Balan

Siemens Corporate Research Inc.
{Ning.Fan, Justinian.Rosca, Radu.Balan}@siemens.com

ABSTRACT

Accurate noise power spectrum estimation in a noisy speech signal is a key challenge problem in speech enhancement. One state-of-the-art approach is the minima controlled recursive averaging (MCRA). This paper presents an enhanced MCRA algorithm (EMCRA), which demonstrates less speech signal leakage and faster response time to follow abrupt changes in the noise power spectrum. Experiments using real speech and noise recordings have validated the superiority of the proposed enhancements. EMCRA shows improvements both in intuitive subjective listening and objective quality measures in terms of higher output SNR and lower output distortion scores.

Index Terms— noise power spectrum estimation, noise control, speech enhancement, noise cancellation filter

1. INTRODUCTION

Noise power spectrum estimation is a key component in a practical speech enhancement system. An early approach is to average the noisy signal over non-speech section using a voice activity detector (VAD). However, its reliability severely deteriorates for weak speech components and low input SNR. Furthermore, there may not be sufficient non-speech sections detectable to track a varying noise spectrum in continuous speech. Alternatively, a minimum statistics [1] based noise estimator does not rely on VAD. The noise estimate is obtained from minima values of smoothed noisy signal power spectrum multiplied by a bias compensation factor, because even for continuous speech there are some time-frequency spectral discontinuities where the minima will converge to the noise spectral value. However using minima makes this approach sensitive to outliers and of high variance.

The MCRA [2] method uses spectral minima indirectly. It estimates noise by averaging past spectral power values with a smoothing parameter that is adjusted by the signal presence probability in sub-bands. That in turn is determined by the ratio between the local energy of the noisy speech and its spectral minima within a specified time window. The ratio is compared to a certain threshold, with smaller value indicating the absence of speech. Temporal smoothing is also performed to reduce fluctuations between speech and non-speech segments, which exploit the strong correlation of speech presence in neighboring frames. The MCRA method is computationally efficient, robust with respect to the input SNR and type of underlying additive noise.

However, there are also some problems. (1) The original MCRA algorithm uses an efficient local minima tracking technique, which reduces the complexity of searching the minima within a time window but it also doubles the delay to follow an abrupt noise spectral rise. (2) The speech leakage refers to the amount of speech spectral components being misclassified into the noise power spectrum, which is one of the causes of distortion in filtered speech. Generally speaking, the shorter the minima searching time window, the larger the speech leakage, because more likely the minima will hit on the weak speech components instead of noise.

This paper proposes two novel techniques to remedy these problems, which jointly reduce the delay and the speech leakage. These two steps are the basis for the EMCRA algorithm that has achieved better results than the original MCRA.

We first review the original MCRA method in section 2. Then the techniques for reducing time delay and speech leakage are presented in section 3 and section 4 respectively. These are followed by the experimental results and conclusions.

2. MINIMA CONTROLLED RECURSIVE AVERAGING

In a spectral filtering based speech enhancement approach, the input noisy speech samples are organized as overlapped frames and each frame is transformed into spectral domain with each frequency as a subband. Let H_0 and H_1 indicate the speech presence and absence hypotheses in the *i*-th frame of the *k*-th subband, then

$$\begin{aligned} H_0 &: X(k,i) = S(k,i) + N(k,i) \\ H_1 &: X(k,i) = N(k,i) \end{aligned} \qquad (1)$$

where $X(k,i)$, $S(k,i)$, and $N(k,i)$ represent the STFT of noisy, clean, and noise signals, respectively. The purpose of noise estimator is to obtain accurate value of noise spectral power magnitude as $\lambda_n(k,i) = E\{|N(k,i)|^2\}$. The MCRA method uses temporal recursive averaging as below:

$$\begin{aligned} H_0 &: \hat{\lambda}_n(k,i+1) = \alpha_n \hat{\lambda}_n(k,i) + (1-\alpha_n)|N(k,i)|^2 \\ H_1 &: \hat{\lambda}_n(k,i+1) = \hat{\lambda}_n(k,i) \end{aligned} \quad (2)$$

where $\alpha_n (0<\alpha_n<1)$ is a smoothing coefficient. Unlike the clear-cut decision such as from a VAD, a conditional speech presence probability $p(k,i) = P(H_0 \mid X(k,i))$ is used, so that (2) can be written as

$$\begin{aligned} \hat{\lambda}_n(k,i+1) &= \hat{\lambda}_n(k,i)p(k,i) \\ &\quad + (\alpha_n \hat{\lambda}_n(k,i) + (1-\alpha_n)|N(k,i)|^2)(1-p(k,i)) \\ &= \tilde{\alpha}_n(k,i)\hat{\lambda}_n(k,i) + (1-\tilde{\alpha}_n(k,i))|N(k,i)|^2 \end{aligned} \quad (3)$$

where

$$\tilde{\alpha}_n(k,i) = \alpha_n + (1-\alpha_n)p(k,i) \quad (4)$$

is a time-varying smoothing parameter that is adjusted by the signal presence probability estimated as follows

$$\hat{p}(k,i+1) = \begin{cases} \alpha_p \hat{p}(k,i) + (1-\alpha_p) & if \ \dfrac{S(k,i)}{S_{\min}(k,i)} > \delta \\ \alpha_p \hat{p}(k,i) & otherwise \end{cases} \quad (5)$$

where and δ is a threshold for speech presence,

$$S(k,i) = \alpha_S S(k,i) + (1-\alpha_S)|X(k,i)|^2 \quad (6)$$

and

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

which is calculated as follows:

$$S_{\min}(k,i) = \begin{cases} S(k,0) & if \ i=0 \\ \min\{S_{\min}(k,i-1), S(k,i)\} & if \ i\%L \neq 0 \\ \min\{S_{tmp}(k,i-1), S(k,i)\} & otherwise \end{cases}$$
$$S_{tmp}(k,i) = \begin{cases} S(k,0) & if \ i=0 \\ \min\{S_{tmp}(k,i-1), S(k,i)\} & if \ i\%L \neq 0 \\ S(k,i) & otherwise \end{cases} \quad (8)$$

3. TIME DELAY TO FOLLOW ABRUPT NOISE RISE

Because (8) updates the minima using 2 variables in sequel with each of length L time window, the original MCRA algorithm takes $2L$ time delay to response to abrupt noise rise, such as when turning on the air conditioner in a car. This can be improved by using the following new minima-tracking scheme.

$$S_{\min}(k,i) = \begin{cases} S(k,0) & if \ i=0 \\ \min\{S_{\min}(k,i-1), S(k,i)\} & \\ & if \ (i+\dfrac{l}{n})\%L \neq 0 \\ \min\{S_l(k,i-1), S(k,i)\} & otherwise \end{cases}$$

$$S_l(k,i) = \begin{cases} S(k,0) & if \ i=0 \\ \min\{S_l(k,i-1), S(k,i)\} & \\ & if \ (i+\dfrac{l}{n})\%L \neq 0 \\ S(k,i) & otherwise \end{cases} \quad (9)$$

$$l = 1,2,\cdots,n \quad and \quad n = 2^m \quad m \in I$$

The new scheme contains *n* pairs of (8) but with overlapped positions, the noise rise delay is reduced from $2L$ to $(1+\frac{1}{n})L$.

The effects of this technique are shown in Figure 1 ($n=4$). The initial noise is low as indicated by a narrow dark strip in the beginning of the noise power spectrum in Figure 1(b), and then it rises abruptly to a high level. The corresponding estimated noise spectra for three different algorithms are shown underneath in Figure 1(c-e). The right side color strip is the color-coding of spectral sample values from bottom to top indicating zero to maximum spectral value in the figure. The time window size L is set to 96 frames for the Rainer-Martine algorithm [1], and 64 frames for the MCRA and the EMCRA algorithms. The original MCRA takes $2L=128$ frames delay and the EMCRA only takes $(1+.25)L=80$ frames delay. The delay enhancement can be easily observed.

4. SPEECH LEAKAGE

Figure 1 also shows that even with a larger window size, the speech leakage is clearly visible for the Rainer-Martin method, while it is not observable for the MCRA and the EMCRA algorithms. The speech leakage is mostly visible only when the noise estimator is applied on clean speech signal with no noise. Because the noise spectrum should be always zero, any detected noise magnitude is erroneously produced from the speech component.

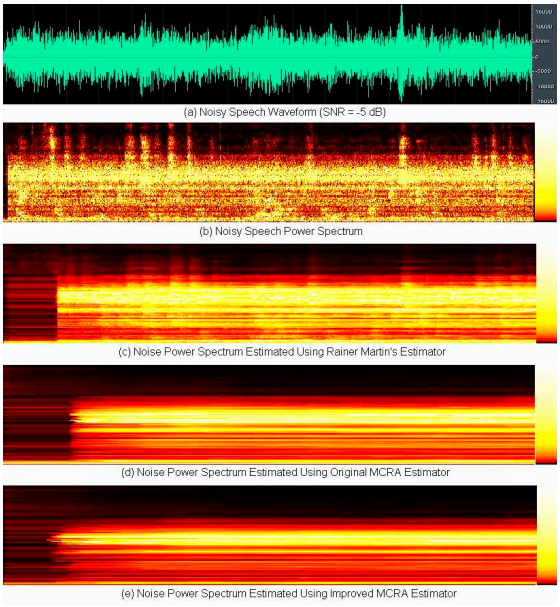

Figure 1– The comparison of (a) noisy waveform, (b) its power spectrum, and (c-e) various estimated noise power spectra. Note the speech leakage in (c), and smaller delay in (e).

This error certainly affects the speech quality when clean speech is filtered using that noise estimator. It also reflects the likelihood of misdetection of weak speech components as noise that will cause distortion in the filtered noisy speech signal. The technique that we propose to alleviate this problem is to introduce a control parameter at each time-frequency as follows.

$$C(k,i)=\begin{cases} S(k,i) & \text{if } C(k,i-1)>S(k,i) \\ \varepsilon C(k,i-1) & \text{if } 2C(k,i-1)<S(k,i) \\ \max\{\varepsilon C(k,i-1), S_{\min}(k,i-1)\} & \\ & \text{otherwise} \end{cases} \quad (10)$$

$$\text{where} \qquad 1<\varepsilon<2$$

The original equation (3) is then replaced by the following conditional updating formula

$$\hat{\lambda}_n(k,i+1)=\begin{cases} \tilde{\alpha}_n(k,i)\hat{\lambda}_n(k,i)+(1-\tilde{\alpha}_n(k,i))\left|N(k,i)\right|^2 & \\ \quad \text{if } 2C(k,i)>S(k,i) \text{ or } S(k,i)>\varsigma S_{\min} & \\ \hat{\lambda}_n(k,i+1) & \text{otherwise} \end{cases} \quad (11)$$

$$\text{where} \qquad \varsigma>1$$

The control parameter is compared with the smoothed noisy signal spectrum and acting like a loose time-frequency strong voice component mask. The power spectral values are updated only at spots that do not have strong speech components. Examples of speech leakage for the clean speech signal are shown in Figure 2. The EMCRA algorithm has achieved least speech leakage, even equipped with the time-delay reduction technique that actually results in a slight increase in speech leakage shown in Figure 2(f).

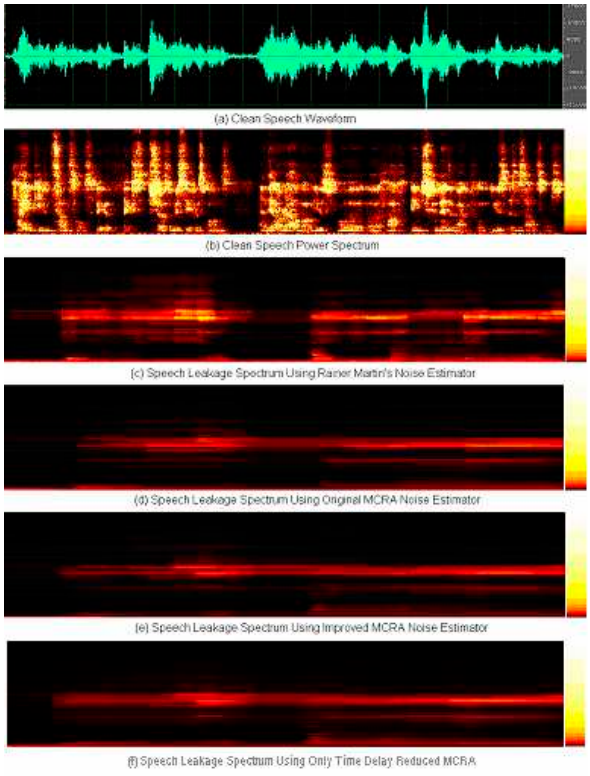

Figure 2- Speech leakage comparison for different noise power spectrum estimation techniques: (a) clean speech waveform, (b) its power spectrum, and (c-f) various estimated noise power spectra. Note the reduced speech leakage from (f) to (e).

5. TEST RESULTS

The EMCRA noise estimator has been tested in comparison with the original MCRA and several other popular noise estimators. In all the tests, an adaptive parametric Wiener filter [3] has been used to perform the noise removal. The objective quality measures used are standard from literature [4]. The SNR indicates the global signal to noise ratio, the larger the better. The Itakura-Saito distortion (It-Sa) and Weighted Spectral Slope (WSS) are distortion criterions, the smaller the better.

For one particular noisy speech file, results are summarized in the following table in terms of objective speech quality measures of filtered signal, as well as CPU computation time spent.

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

	SNR	It-Sa	WSS	Time (msec)
noisy speech	-5.17	0.4910	41.8865	
true noise	4.78	0.0767	23.2398	
Rainer Martin	-1.66	0.4089	41.9764	3976
MCRA	-1.55	0.4382	41.1159	3315
PSD	-2.09	0.4112	44.2091	3425
EMCRA	-0.86	0.4113	40.9234	3385

The "true noise" is the power spectrum of the actual noise used for mixing with the clean speech to form the noisy speech, which serves as the ground truth. It also demonstrates the achievable potential for this noise reduction filter, should noise estimation totally accurate. The results show that the proposed enhanced MCRA noise estimator achieved the best speech quality among all the estimators tested. It consumes more CPU time than the original MCRA algorithm, but much less than the Rainer Martin algorithm and less than the PSD noise estimator [5].

The same dataset of 112 different speech and noise mixtures used in [3] is tested for the average performance. The results are shown in Figures 3 in terms of the average filtered SNR scores, average Itakura-Saito distortions, and average Weighted Spectral Slope values. Our EMCRA algorithm achieved top rank in SNR improvement while maintaining the second rank in other two evaluation criteria. The first ranks in the Itakura-Saito distortion and the Weighted Spectral Slope measures are the PSD and the MCRA respectively. When considering the two distortions simultaneously, the EMCRA algorithm is clearly the winner. Subjective listening to the wave files also confirms the conclusion.

6. CONCLUSIONS

Novel techniques to enhance the MCRA noise estimation algorithm have been developed for speech enhancement in adverse environments. Our approach is to reduce the time delay for adapting to abrupt noise change while at the same time decreasing the speech leakage to avoid speech distortions. Experimental tests have demonstrated positive results. The need to reduce time delay to follow abrupt noise changes has also been addressed by others like [6] and [7]. However they did not consider the speech leakage problem, so that such alternative solutions may be enhanced in a rapid varying noisy environment at the cost of some decreased performance and distortions in a relatively stable noise environment.

7. REFERENCES

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

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

[3] N. Fan, "Low distortion speech denoising using an adaptive parametric Wiener filter", *In ICASSP2004*, Montreal, Canada, pp. 309-312, 2004.

[4] J.H.L. Hansen, and B.L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithm," *In Inter. Conf. on Spoken Language Processing,* vol.7, pp. 2819-2822, Sydney, Australia, December 1998.

[5] S. Aalburg, C. Beaugeant, S. Stan, T. Fingscheidt, R. Balan, and J. Rosca, "Single- and Two-Channel Noise Reduction for Robust Speech Recognition in Car," *In ISCA Tutorial and Research Workshop on Multi-Modal Dialogue in Mobile Environments*, Stuttgart, Germany, June 2002.

[6] Z. Lin, and R.A. Goubran, "Instant noise estimation using Fourier transform of AMDF and variable start minima search", *In ICASSP2005*, Philadelphia, USA, pp. 161-164, 2005.

[7] S. Rangachari, and P. Loizou, "A noise estimation algorithm for highly non-stationary environments," *Speech Communication,* 28, pp. 220-231, 2006.

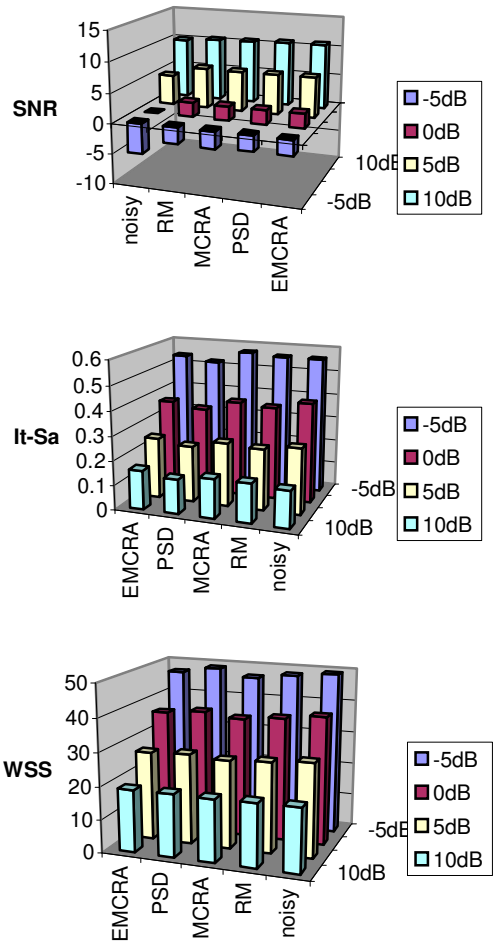

Figure 3 - Average filtered SNR (top), Itakura-Saito distortion (middle), and Weighted Spectral Slopes (bottom) scores at different input SNR for various noise estimators: RM [1], MCRA[2], PSD[5], and EMCRA.