Robust Noise Estimation Applied to Different Speech Estimators

Markus Schwab, Hyoung-Gook Kim, Wiryadi and Peter Noll

Department of Communication Systems
Technical University of Berlin, Germany
{schwab|kim|wiryadi|noll}@nue.tu-berlin.de

Abstract

In this paper we present a robust noise estimation for speech enhancement algorithms. The robust noise estimation based on a modified minima controlled recursive averaging noise estimator was applied to different speech estimators. The investigated speech estimators were spectral substraction (SS), log spectral amplitude speech estimator (LSA) and optimally modified log spectral amplitude estimator (OM-LSA). The performance of the different algorithms were measured both by the signal-to-noise ratio (SNR) and recognition accuracy of an Automatic Speech Recognition (ASR).

1. Introduction

In recent years, the performance of automatic speech recognition has been improved drastically by applying statistical approaches. However, most speech recognizers still have serious problems in noisy environments. The recognition rate can significally degrade in severe conditions.

Noise reduction algorithms can improve the recognition rates in noisy environments. In general, noise reduction algorithms consist of two major components. The first component estimates the noise and the second one estimates the speech. Traditional noise estimator are based on voice activity detectors (VAD) which are difficult to tune and their application to low SNR speech results often in clipped speech. Israel Cohen proposed a noise estimator based on minimum statistics and recursive averaging, called minima controlled recursive averaging (MCRA) algorithm [1]. This algorithm is very robust and achieves also good results even in the case of non-stationary noise. In this paper we use a modified minima controlled recursive averaging where the threshold function for the speech probability depends also on the SNR calculations of the speech estimator.

A large number of speech estimators have been proposed in the past. Most of them are use a gainfunction to modify the spectral amplitude whereas the phase is normally unchanged. Traditional gain functions like spectral substraction [2] depend only on the measured signal level of the current frame and the estimated noise level. These methods cause musical tones which degrade the quality of the audio signal. A better solution was proposed by Ephraim and Malah [3]. They use the decision directed method to estimate an a priori SNR and the gain function minimizes the mean-square error of the log-spectra, based on a Gaussian statistical model. This estimator proved very efficient in reducing the musical residual noise phenomena [4]. In recent research speech presence probability has been used for further improvements in the performance of the algorithms [5], [6]. Whereas in [5] a multiplicative modification has been proposed, Cohen derived an optimally modified gainfunction in [6]. The optimal spectral gain function was obtained as a weighted geometric mean of the hypothetical gains associated with the speech presence uncertainty.

In this paper we use the modified MCRA noise estimator and compare the performances of three different speech estimators, namely spectral substraction (SS), log-spectral amplitude estimator (LSA), and optimally modified log-spectral amplitude estimator (OM-LSA). The different speech enhancement systems are evaluated with a SNR analysis and recognition rates of an automatic speech recognition (ASR) system. The database for the speech recognition task has been taken from AURORA 2. The training mode for the ASR system was multiconditional [7] in order to achieve best recognition results. But this lowers the improvement of speech enhancement systems because the ASR system is also trained to noisy conditions.

2. System Overview

In our signal model, clean speech $s(n)$ is corrupted by additive noise $d(n)$. The observed noisy signal $x(n)$ can then be expressed as:

$$x(n) = s(n) + d(n) \ . \tag{1}$$

First, the audio signal is segmented into overlapping blocks length N and an overlapping rate of 75%. Then a Hanning window is applied to the signal blocks. Afterwards, the signal is transformed into frequency domain with the fast Fourier transform (FFT)(Figure 1). With the power spectrum $|X(k,l)|^2$, k and l denote the frequency index and time index respectively, a modified minima controlled recursive averaging (MCRA) noise estimation is realized. Three different gain function algorithms are used, namely, spectral substraction (SS), log spectral amplitude estimator (LSA), and optimally modified LSA (OM-LSA), to estimate the speech signal. Finally, the signal is transformed back into time domain and reconstructed.

2.1. Noise Estimation Algorithm

The first step of the noise estimation algorithm is to average the power spectrum in frequency with a window function $w(i)$ and then in time domain with a recursive equation of first order:

$$A(k,l) = \sum_{i=-1}^{i=1} w(i)|X(k-i,l)|^2 \ , \tag{2}$$

$$E(k,l) = \beta E(k,l-1) + (1-\beta)A(k,l) \ . \tag{3}$$

A local minimum noise tracker is used to determine the noise floor.

$$M(k,l) = \min_{i=0..M}(E(k,l-i)) \tag{4}$$

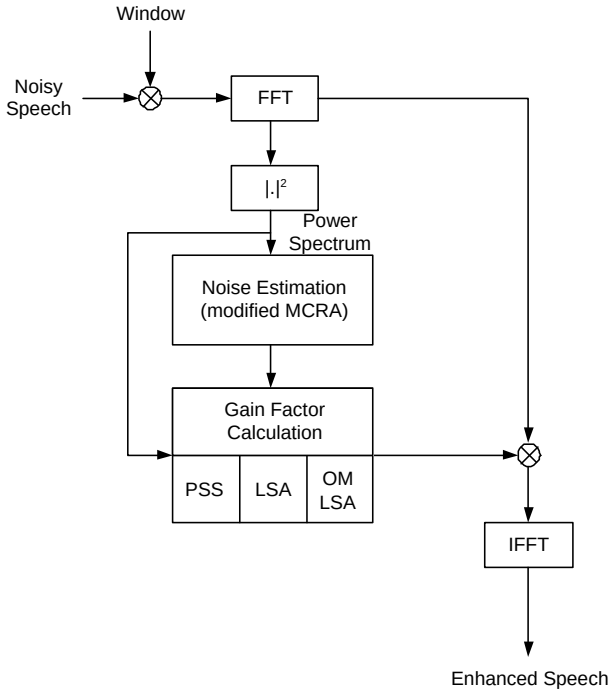

Figure 1: System Overview

With this noise floor a voice activity detector $I(k,l)$ (VAD) is employed in each frequency bin:

$$I(k,l) = \begin{cases} 1 & \text{if } E(k,l) > (1 + Ke^{-G(k,l-1)})M(k,l) \\ 0 & \text{otherwise} \end{cases} . \quad (5)$$

K is a constant value.

This VAD controls the update of the noise estimation. Therefore, a speech presence probability estimator $p(k,l)$ is computed to obtain an optimal smoothing parameter $\alpha(k,l)$ for the noise estimation:

if $I(k,l) = 1$
  $p(k,l) = \alpha_p + (1-\alpha_p) \cdot p(k,l-1)$
  $\alpha(k,l) = 1$
else
  $p(k,l) = (1-\alpha_p) \cdot p(k,l-1)$
  $\alpha_N(k,l) = \alpha_\alpha + (1-\alpha_\alpha) \cdot p(k,l)$

Whenever speech is assumed ($I(k,l) = 1$) the smoothing factor for the noise estimation $\alpha(k,l)$ is set to one and the noise estimation is immediately stopped. This avoids false noise estimation when speech is present. If the indicator function $I(k,l)$ is equal to zero the speech presence probability is recursively decreased with a high smoothing factor so that the noise estimation is very slow at the beginning of a noise only period. Weak speech components at the end of a speech period will be preserved.

Finally, the noise estimation $N(k,l)$ is done by a recursive update formula first order:

$$N(k,l) = \alpha_N N(k,l-1) + (1-\alpha_N)|X(k,l)|^2 \quad (6)$$

Figure 2 shows the noise estimation in the frequency bin $k = 39$. In speech periods the noise tracking is stopped and in noise only periods, there is a good noise tracking.

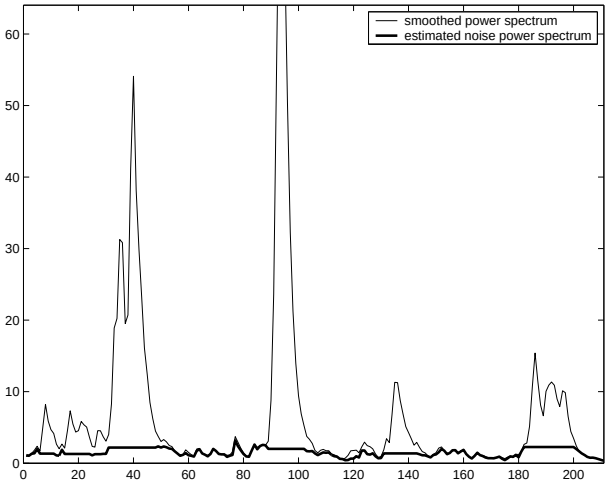

Figure 2: Power Spectrum of noisy speech and noise estimation

2.2. Speech estimation

The first implemented speech estimator is based on spectral substraction:

$$|\tilde{S}(k,l)| = |X(k,l)| - \sqrt{N(k,l)} \quad (7)$$

or written with a gain function:

$$\tilde{S}(k,l) = G_{SS}(k,l) \cdot X(k,l) \quad (8)$$

with

$$G_{SS}(k,l) = 1 - \sqrt{\frac{N(k,l)}{|X(k,l)|^2}} . \quad (9)$$

The second speech estimator is the log spectral amplitude estimator (LSA) as proposed by Ephraim and Malah.

$$\tilde{S}(k,l) = G_{LSA}(k,l) \cdot X(k,l) \quad (10)$$

For the calculation of $G_{LSA}(k,l)$, the *a priori SNR* $\xi(k,l)$ is estimated by

$$\xi(k,l) = \alpha G^2(k,l-1)\gamma(k,l) + (1-\alpha)\xi(k,l-1) \quad (11)$$

where

$$\gamma(k,l) = \begin{cases} \frac{|X(k,l)|^2}{N(k,l)} & \text{if } |X(k,l)|^2 > N(k,l) \\ 1 & \text{otherwise} \end{cases} , \quad (12)$$

and $\alpha \in [0,1]$. With these two parameters the log spectral amplitude estimator is given by:

$$G_{LSA}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \exp\Big(0.5 \int_{t=\nu(k,l)}^{\infty} \frac{e^{-t}}{t} dt\Big) \quad (13)$$

with

$$\nu(k,l) = \Big(\frac{\xi(k,l)}{1+\xi(k,l)}\Big)\gamma(k,l) . \quad (14)$$

The third speech estimator is an extension of the LSA algorithm. A speech presence probability is used to modify the

spectral gain function. Malah proposed a multiplicative modified LSA (MM-LSA) and Cohen derived an optimally modified LSA (OM-LSA) gain function. We have implemented the algorithm from Cohen where he proposed to use the *a priori SNR* to compute a speech absence probability $q(k,l)$ and then a speech presence probability $p(k,l)$. Therefore, the *a priori SNR* will be smoothed over time:

$$\zeta(k,l) = 0.7\zeta(k,l-1) + 0.3\xi(k,l) \tag{15}$$

and then smoothed in frequency domain

$$\zeta_{local}(k,l) = \sum_{i=-1}^{i=1} b(i)\zeta(k-i,l)\ , \tag{16}$$

where $b(i)$ is a window function.

A likelihood of speech presence can then be defined as follows:

$$p(k,l) = \begin{cases} 0 & \text{if } \zeta_{local}(k,l) \leq -10dB \\ 1 & \text{if } \zeta_{local}(k,l) \geq -5dB \\ \sin^2\left(2\pi \frac{\zeta_{local}(k,l)-\zeta_{min}}{\zeta_{max}-\zeta_{min}}\right) & \text{otherwise} \end{cases} \tag{17}$$

The estimate of the a priori probability for speech absence $q(k,l)$ is then given :

$$q(k,l) = 1 - p(k,l)\ , \tag{18}$$

with the restriction $q(k,l) \leq 0.94$. The conditional speech presence probability estimation, derived by Cohen, $p(k,l)$ is then given by

$$p(k,l) = \frac{1}{1 + \frac{q(k,l)}{1-q(k,l)}\left(1+\xi(k,l)\exp(-\nu(k,l))\right)} \tag{19}$$

In practice, we have taken equation 17 instead of recalculation $p(k,l)$. The difference to the derived equation is not very high. Finally, the OM-LSA gain function, derived by Cohen, becomes:

$$G_{OM-LSA}(k,l) = \Big(G_{LSA}(k,l)\Big)^{p(k,l)} \Big(G_{min}\Big)^{q(k,l)}\ , \tag{20}$$

where G_{min} is a spectral floor constant. The estimated speech is then obtained by:

$$\tilde{S}(k,l) = G_{OM-LSA}(k,l) \cdot X(k,l)\ . \tag{21}$$

3. Experimental Results

The parameters for the modified MCRA noise estimator have been chosen as follows:

$\beta = 0.6$
$K = 4$
$\alpha_p = 0.1$
$\alpha_\alpha = 0.6$

and the parameters for the speech estimator have been:

$\alpha = 0.92$
$\zeta_{min} = -10dB$
$\zeta_{max} = -5dB$
$G_{min} = 0.01$

3.1. SNR improvement and spectrogram

To measure the performance of the presented algorithms, the signal-to-noise ratios (SNR) have been computed with the knowledge of the clean speech signal $s(n)$:

$$SNR = \frac{\sum(x(n) - s(n)^2)}{\sum s(n)^2} \tag{22}$$

$x(n)$ is the signal from which the SNR is calculated. The SNR-improvement is then given by

$$SNR_{improve} = SNR_{out} - SNR_{in}. \tag{23}$$

In table 1 the SNR-improvement of each speech enhancement algorithm is shown. The examples are taken from [8]. Three types of background noises - white noise, factory noise and f16 cockpit noise - are artificially added to clean speech.

	white (6 dB)	factory (5 dB)	f16 (4.6 dB)
SS	5.35	3.52	4.80
LSA	5.86	4.08	4.96
OM-LSA	6.34	4.18	5.56

Table 1: *Comparison of SNR improvement of the different one-channel speech enhancement algorithms.*

Figure 3 shows the time signals, left side, and spectrograms, right side, of the F16 cockpit noise example. Above the unprocessed noisy signal can be seen and thereunder the outputs of the three implemented speech enhancement algorithms are shown.

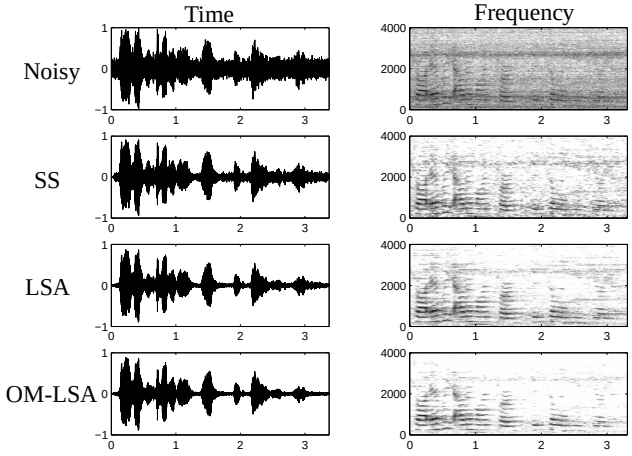

Figure 3: *Time signal and spectrogram of the noisy and the three enhanced signal of the F16 cockpit noise example*

3.2. Recognition accuracy in a ASR system

For evaluation of the improvement of speech recognition with the presented noise reduction algorithm, the Aurora 2 database

together with a HMM ASR system has been chosen and the used training mode was multi-condition training on noisy data. The feature vector consists of 39 parameters: 13 mel frequency cepstral coefficients plus delta and acceleration calculations.

	Set A	Set B	Set C
without noise reduction	87.81	86.27	83.77
SS	88.51	87.57	85.56
LSA	90.90	89.60	88.35
OM-LSA	90.93	89.48	88.92

Table 2: *Comparisons of word correctness (%) of the three noise reduction algorithms (SS, LSA, and OM-LSA) and the unprocessed signal based on the Aurora 2 database.*

4. Conclusions

In this paper we have presented a robust noise estimator for non-stationary noise environments. This robust noise estimator has been combined with three different speech estimators, spectral substraction, log-spectral-amplitude estimator, and optimally modified log-spectral-amplitude estimator, to give three different speech enhancement algorithms. The performances of the all presented speech enhancement algorithms has been evaluated on the basis of SNR-improvements and speech recognition rates. In both cases, SNR-improvements and speech recognition rates, an improvement can be stated for the three speech enhancement algorithms. The OM-LSA algorithm performs best in respect to the SNR-improvement, whereas for the speech recognition accuracy the LSA and the OM-LSA algorithms show almost similar recognition rates. The absolute improvement is about 4 per cent with multiconditional training.

5. References

[1] Israel Cohen and Baruch Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," in *IEEE Signal Processing Letters*, 2002.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," in *IEEE Trans. Acoustics, Speech and Signal Proc.*, 1979, vol. 27, pp. 113–120.

[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," in *IEEE Trans. Acoust., Speech, Signal Processing*, April 1985, vol. ASSP-33, pp. 443–445.

[4] O. Cappe, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," in *Eurospeech*, September 1993, pp. 1093–1096.

[5] R.V. Cox D. Malah and A.J. Accardi, "Tracking speech presence uncertainty to improve speech enhancement in non-stationary noise environments," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 1999, pp. 789–792.

[6] Israel Cohen, "On speech enhancement under signal presence uncertainty," in *International Conference on Acoustic and Speech Signal Processing*, May 2001, pp. 167–170.

[7] D. Pearce H. G. Hirsch, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000*, 2000.

[8] CSLU NSEL, "Neural speech enhancement," ULR: http://cslu.cse.ogi.edu/nsel/demos/.