INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Speech Enhancement based on Smoothing of Spectral Noise Floor

Hyoung-Gook Kim, Thomas Sikora

Communication Systems Group
Technical University of Berlin, Germany
kim@nue.tu-berlin.de

Abstract

This paper presents robust speech enhancement using noise estimation based on smoothing of spectral noise floor (SNF) for nonstationary noise environments. The spectral gain function is obtained by well-known log-spectral amplitude (LSA) estimation criterion associated with the speech presence uncertainty. The noise estimate is given by averaging actual spectral power values, using a smoothing parameter that depends on smoothing of spectral noise floor. The noise estimator is very simple but achieves a good tracking capability for a nonstationary noise. Its enhanced speech is free of musical tones and reverberation artifacts and sounds very natural compared to methods using other short-time spectrum attenuation techniques. The performance is measured by the segmental signal-to-noise ratio (SNR), the speech/ speaker recognition accuracy and the speaker change detection rate for the audio segmentation using MFCC-features (Mel-scale Frequency Cepstral Coefficients) in comparison to other single microphone noise reduction methods.

1. Introduction

In many speech communication applications, like audio-conferencing, hands-free mobile telephony and speech/speaker recognition devices, the recorded speech signals contain a considerable amount of acoustic noise, which not only degrades the subjective speech quality, but also hinders performance of recognition systems. Therefore efficient noise reduction algorithms are called for.

A variety of noise reduction algorithms have been suggested in the literature to achieve good speech quality and avoid musical tones. Most of the noise reduction algorithms are based on short time spectral amplitude (STSA) analysis. Among these, MMSE (minimum mean square error)-LSA (log-spectral amplitude) proposed by Ephraim and Malah [1] is the most popular STSA based algorithm. It minimizes the mean squared error of the log-spectra, based on a Gaussian statistical model. This method proved very efficient in reducing musical residual noise phenomena. A noise estimation based on minimum statistics by Martin [2] is a useful method. The spectral noise floor estimated by tracking spectral minima of the smoothed power estimate of the noisy signal within a finite time window length is used as the estimated noise. In this method, a larger window for the minimum tracking provides a good noise estimation for a stationary noise. However, the tracking capability for a nonstationary noise is degraded. Although a short window achieves better tracking capability it may introduce overestimation which results in poor speech quality for high SNRs. It is not easy to select an appropriate window length for good tracking capability without overestimation.

This paper proposes a noise reduction algorithm with good speech quality for a wide range of SNRs. To achieve the tracking capability for nonstationary noise the algorithm uses low computational costs due to a low number of turning parameters. The proposed algorithm continuously updates the noise estimate by averaging actual power spectral density balanced between the degree of smoothing and the noise tracking rate.

2. Description of Algorithm

Assuming an additive uncorrelated noise the proposed speech enhancement method mainly consists of two independent main parts, i.e. estimation of the noise spectrum and filtering of the noisy speech to obtain clean speech. Figure 1 depicts a block diagram of the speech enhancement configuration.

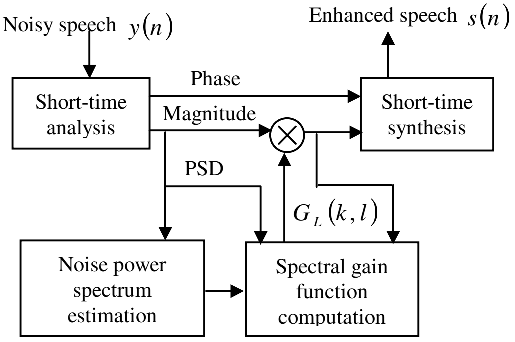
Noisy speech $y(n)$
Enhanced speech $\tilde{s}(n)$
Short-time analysis
Phase
Magnitude
Short-time synthesis
PSD
$G_L(k,l)$
Noise power spectrum estimation
Spectral gain function computation

Figure 1: Block diagram of speech enhancement

Let $s(n)$ and $d(n)$ denote clean speech and uncorrelated additive noise, where n is a time index. The noisy speech signal $y(n)$, given by $y(n)=s(n)+d(n)$, is divided into overlapping frame by a window function and analyzed using the short-time Fourier transform (STFT):

$$Y(k,l)=\sum_{n=0}^{N-1} y(n+lM)h(n)e^{-j(2\pi/N)nk}, \qquad (1)$$

where k is the frequency bin index, l is the time frame index, h is an analysis window of size N, and M is the framing step.

For a given frame l we have $Y(k,l)=S(k,l)+D(k,l)$, where $S(k,l)$ and $Y(k,l)$ are characterized by their amplitude $A(k,l)$ and $R(k,l)$ and their phases $\varphi(k,l)$ and $\theta(k,l)$, respectively, $S(k,l)=A(k,l)\exp(j\varphi(k,l))$, $Y(k,l)= R(k,l)\exp(j\theta(k,l))$.

The noise spectrum is estimated by a novel noise spectrum tracker using averaging actual spectral power values depending on smoothing of spectral noise floor. The spectral components of corresponding clean speech is obtained by multiplying the frequency dependent gain function, G_L, by the noisy magnitude spectrum, $Y(k,l)$, and it is defined as

$$\tilde{S}(k,l) = G_L(k,l)Y(k,l). \tag{2}$$

The filtered spectral values are transformed back into the time domain by applying inverse Fourier transformation in order to obtain the enhanced speech $\tilde{s}(n)$.

2.1. Noise power spectrum estimation

The noise power spectrum estimation consists of six steps; short-term spectral averaging, local minimum tracking, voice activity detection using smoothing of spectral noise floor, speech presence probability estimation, smoothing parameter computation and update noise spectrum estimation.

In the first step, the frequency smoothing of the noisy power spectrum in each frame is defined by

$$Y_F(k,l) = \frac{1}{2w+1}\sum_{i=-w}^{w} \left|Y(k-i,l)\right|^2. \tag{3}$$

Subsequently, the averaging of $Y_F(k,l)$ in time frame is performed over B frames by

$$Y_T(k,l) = \frac{1}{B}\sum_{l=0}^{B-1} Y_F(k,l). \tag{4}$$

In the second step, the spectral local noise floor values $M(k,l)$ of $Y_T(k,l)$ are estimated by tracking spectral minima within windows of C frames, for each frequency bin. The minimum value for the current frame is found by a comparison with the stored minimum value:

$$M(k,l) = \min_{c=0\ldots C}\{M(k,l-c), Y_T(k,l)\}. \tag{5}$$

In the third step, the frequency smoothing of the minima values in each frame is obtained by

$$M_F(k,l) = \frac{1}{2w+1}\sum_{i=-w}^{w} \left|M(k-i,l)\right|. \tag{6}$$

Afterwards, the time smoothing of $M_F(k,l)$ over B frames is computed by

$$M_T(k,l) = \sum_{l=0}^{B-1} M_F(k,l). \tag{7}$$

This new step is applied to more accurate voice activity detection (VAD), which decides whether speech is present in the kth bin. Using the ratio between noisy smoothed spectral amplitude and its derived smoothed spectral noise floor, the indicator function $I(k,l)$ for VAD is defined by

$$I(k,l) = \begin{cases} 1 & if \quad Y_T(k,l)/M_T(k,l) \geq \psi \\ 0 & else \end{cases}, \tag{8}$$

where $I(k,l)$ =1 denotes the speech present in kth bin while $I(k,l)$ =0 the speech absent, respectively.

The accurate VAD reduces the musical residual noise phenomena in high nonstationary noise environments and introduces the enhanced speech with good speech quality.

In the fourth step, the conditional speech presence probability is estimated by a first-order recursive averaging

$$p(k,l) = \alpha_p p(k,l-1) + (1-\alpha_p)I(k,l), \tag{9}$$

with time-varying smoothing parameter α_p ($0<\alpha_p<1$) to obtain an optimal smoothing parameter

$$\alpha_d(k,l) = \alpha + (1-\alpha)p(k,l), \tag{10}$$

where the noise can be more reduced if the noise threshold parameter α ($0<\alpha<1$) decreases.

A common noise estimation technique is to recursively average past spectral power values of the noisy measurement during periods of speech absence and hold the estimate during speech presence. Its main drawbacks are the very slow update rate of the noise estimate in case of a sudden rise in the noise energy level. The noise estimation depends on the window length of C frames. The noise estimate using a larger window is too low that the tracking capability for a nonstationary noise is degraded. A short window, on the other hand, may introduce overestimation which results in poor speech quality for high SNRs.

To achieve good tracking capability the noise estimation must track the change of the noise characteristics in both speech and non-speech periods. For this, as the new step, the noise estimate is obtained by averaging actual spectral power values using a time-varying frequency-dependent smoothing parameter $\alpha_d(k,l)$ in both speech and non-speech periods and it is defined as

If $\dfrac{Y_T(k,l)}{M_T(k,l)} \geq \psi$

$$\lambda_d(k,l) = \alpha_d(k,l)M(k,l) + (1-\alpha_d(k,l))\left|Y(k,l)\right|^2$$

else

$$\lambda_d(k,l) = \alpha_d(k,l) + (1-\alpha_d(k,l))\left|Y(k,l)\right|^2 \tag{11}$$

2.2. Estimation of speech

The estimation of clean speech is obtained by applying a log-spectral amplitude gain function G_L to each spectral component of the noisy speech signal:

$$\tilde{S}(k,l) = [G_{LSA}(k,l)]^{G_M(k,l)}Y(k,l) = G_L(k,l)Y(k,l) \tag{12}$$

where G_M is the gain modification function and G_{LSA} is the LSA gain function.

The G_{LSA} is derived by

$$G_{LSA}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)}\exp\left(0.5\int_{t=v(k,l)}^{\infty}\frac{e^{-t}}{t}dt\right), \tag{13}$$

where $\gamma(k,l)$ is the a posteriori SNR, $\eta(k,l)$ represents the a priori SNRs, $q(k,l)$ is an estimate of speech absence priori probability, β ($0<\beta<1$) is the SNR smoothing factor, and $\lambda_d(k,l)$ is a noise spectrum estimate:

$$\gamma(k,l) \cong \frac{R^2(k,l)}{\lambda_d(k,l)}, \quad \xi(k,l) \cong \frac{\eta(k,l)}{1-q(k,l)},$$

$$v(k,l) \cong \frac{\xi(k,l)}{1+\xi(k,l)}\gamma(k,l),$$

$$\eta(k,l) = \beta G_L^{\,2}(k,l-1)\cdot\frac{\gamma(k,l)}{1-q(k,l)} + (1-\beta)\cdot\max\left\{\gamma(k,l)-1\right\} \tag{14}$$

The amount of noise reduction can be reduced by overestimation $\eta(k,l)$ and increased by underestimating $\eta(k,l)$.

The G_M is applied to take into account the probability of the speech presence in the frequency *k*, and it is referred to as soft-decision modification of the optimal estimation. The G_M is given by

$$G_M = \frac{\Lambda(k,l)}{1+\Lambda(k,l)}, \tag{15}$$

where $\Lambda(k,l)$ is a likelihood ratio between speech presence and speech absence in frequency *k* and defined by

$$\Lambda(k,l) = \frac{1-q(k,l)}{q(k,l)}\frac{\exp(v(k,l))}{1+\xi(k,l)}\bigg|_{\xi(k,l)=\frac{\eta(k,l)}{1-q(k,l)}}. \tag{16}$$

The a priori speech absence probability $q(k,l)$ is estimated by $q(k,l)=bq(k,l-1)+(1-b)\mathrm{U}(k,l)$, where b $(0<b<1)$ is a time-smoothing factor and the likelihood ratio $\mathrm{U}(k,l)=1$ if $\gamma(k,l)>\zeta_1$, $\mathrm{U}(k,l)=(\zeta_1-\gamma(k,l))/(\zeta_1-\zeta_2)$ if $\zeta_2<\gamma(k,l)<\zeta_1$, and $\mathrm{U}(k,l)=0$ if otherwise.

3. Experimental results

For the performance of the proposed speech enhancement method, we measured segmental SNR improvement in speech segments, speech/speaker recognition rate and the speaker change detection rate for the audio segmentation in comparison to other single microphone noise reduction methods.

3.1. Segmental signal-to-noise ratio

The segmental signal-to-noise ratio (seg.SNR) is computed by *improve*SNR=seg.SNR$_{out}$-seg.SNR$_{in}$ for the enhanced speech signals.

Three types of background noise - white noise, car noise and factory noise - were artificially added to different portions of the data at SNR of 10 dB and 5 dB. The speech data used for the segmental SNR improvement were digitized at 22.05 kHz using 16 bits per sample.

Table 1: Comparison of segmental SNR improvement of different one-channel noise estimation methods.

Methods	Input SNR [dB]					
	White noise		Car noise		Factory noise	
	10	5	10	5	10	5
MM	7.3	8.4	8.2	9.7	6.2	7.7
OM	7.9	9.9	9	10.6	6.9	8.3
SNF	8.8	11.2	9.7	11.4	7.6	10.6

Table 1 shows that our SNF algorithm gives best improvement results compared to the results of MM (multiplicatively modified log-spectral amplitude speech estimator) [3] and OM (optimally modified LSA speech estimator and minima controlled recursive averaging noise estimation) [4].

3.2. Speech recognition

For evaluation of the improvement of speech recognition with the noise reduction algorithms, the Aurora 2 database together with a HTK software tools has been chosen. Two training modes are used: training on clean data and multi-condition training on noisy data.

The feature vector from the speech database with sampling rate 8 kHz consists of 39 parameters: 13 mel frequency cepstral coefficients plus delta and acceleration calculations. The mel-cepstrum coefficients were modeled by a simple left-to-right 16-state three-mixture whole word HMM. For the noisy speech results, we averaged the word accuracies between 0 dB and 20 dB SNR.

In the Table 2 und Table 3, set A, B, and C refer to matched noise condition, mismatched noise condition, and mismatched noise and channel condition, respectively. Table 2 and Table 3 describe the results of the recognition accuracy by training on clean data and multi-condition training on noisy data, respectively.

Table 2: Comparisons of word accuracies (%)between several front-ends on the Aurora 2 database under clean condition training. NR: noise reduction.

Feature extraction	Set A	Set B	Set C	Overall
Without NR	61.37	56.20	66.58	61.38
MM	79.28	78.82	81.13	79.74
OM	80.34	79.03	81.23	80.20
SNF	84.32	82.37	82.54	83.07

Table 3: Comparisons of word accuracies (%)between several front-ends on the Aurora 2 database under multi-condition training.

Feature extraction	Set A	Set B	Set C	Overall
Without NR	87.81	86.27	83.77	85.95
MM	89.68	88.43	86.81	88.30
OM	90.93	89.48	88.91	89.77
SNF	91.37	91.75	92.13	91.75

As seen in the results of Table 2 und Table 3, the proposed SNF provides better performance than MM front-end and OM front-end. The SNF method is very simple because it needs lower turning parameters compared to OM.

3.3. Speaker recognition

For speaker recognition we performed experiments where 25 speakers (11 male and 14 female) were used. Each speaker was instructed to read 15 different sentences. After recording of the sentences spoken by each speaker, we cut the recordings into smaller clips: 21 training clips (about 3 minutes long), and 10 test clips (50s.) per speaker. The speech data was digitized at 22.05 kHz using 16 bits per sample. The non-stationary Gaussian white noise is artificially added to the speech database at the SNR ratios ranging from clean over 20 dB to 5 dB in steps of 5 dB.

The training phase is done only on clean speech signals, while test phase is performed by using the noise reduction preprocessing on noisy speech signals (mismatched conditions).

Table 4 shows the results of speaker recognition. For decreasing SNR ratios, the speaker recognition rate without noise reduction is seriously decreased. However, the concatenation of MFCC with noise reduction using SNF yields a higher recognition rate than the front-end with other noise reduction methods.

Table 4: Speaker recognition accuracies (%). NR: noise reduction.

Feature extraction	Speech material				
	Clean speech	20 dB	15 dB	10 dB	5 dB
Without NR	95.6	66.2	37.8	24	19.6
MM	94.7	91.7	87.2	79.2	53.7
OM	95.1	92.5	88.9	79.8	55.3
SNF	95.7	92.8	90.3	82.2	57.8

3.4. Speaker change detection

Detecting speaker change in a given audio stream has received a great deal of interest in recent years. This is mainly due to its various potential applications ranging from retrieving information from audio materials to improving the accuracy of speech/speaker recognition systems.

In this paper, the speaker change detection step in absence of priori information about speakers is performed by the covariance of two sub-segments called divergence shape distance (DSD) [5] and splits the conversation into smaller segments that are assumed to contain only one speaker.

To evaluate the performance of the change detection with our proposed speech enhancement approach, we used audio tracks from television talk show programs. It is approximately 60 minutes long and contains 13 speakers.

The speech data were digitized at 22.05 kHz using 16 bits per sample. The non-stationary Gaussian white noise is artificially added to the speech database at the SNR ratios ranging from clean over 20 dB to 0 dB in steps of 5 dB. And the speech signal is parametrized with 30 MFCCs without addition of the delta- and acceleration coefficients.

For the measure of the performance we distinguish two types of errors related to speaker change detection. A false alarm (*FA*) occurs when a speaker change is detected although it does not exist. A missed detection (*MD*) occurs when the process does not an existing speaker change. Indeed, a missed detection may damage the grouping step. However, false alarm may be resolved during the grouping step. We can then define the false alarm rate (*FAR*):

$$FAR = \frac{N_{FA}}{NASC + N_{FA}} \tag{17}$$

and the missed detection rate (*MD*R):

$$MDR = \frac{N_{MD}}{NASC}, \tag{18}$$

where N_{FA} is a number of *FA*, *NASC* is a number of actual speaker changes and N_{MD} is the number of *MD*.

A good segmentation is then characterized by low values of *FAR* and *MDR*.

Table 5 describes the results of the change detection accuracy.

Table 5: *FAR* and *MDR* applied on the conversational speech of TV talk show program.

TV talk show	DSD		DSD+SNF	
	FAR (%)	*MDR* (%)	*FAR* (%)	*MDR* (%)
Clean	7.54	10.67	7.54	10.67
20 dB	13.76	20.24	10.75	15.14
15 dB	23.32	36.56	14.31	21.67
10 dB	43.51	63.82	24.58	34.33
5 dB	67.83	81.93	35.65	49.89
0 dB	87.68	96.35	43.70	66.79

Our experiments show that the proposed SNF speech enhancement algorithm significantly improves the performance of the speaker change detection.

4. Conclusions

In this paper, we have presented a speech enhancement system based on smoothing of spectral noise floor (SNF) and log-spectral amplitude (LSA) speech estimator for non-stationary noise environments. This algorithm applied on speech/speaker recognition and speaker change detection leads to significantly improved performance.

Our future work will apply the proposed SNF algorithm to automatic speaker segmentation algorithms integrated with speech recognition and speaker identification to enable indexing, quick browsing, and searching of audio documents.

5. References

[1] Ephraim, Y. and Malah, D., "Speech enhancement using a minimum mean-square error log spectral amplitude estimator", *IEEE Trans. Speech and Signal Proc. Vol. 32, pp. 443-445*, April 1985.

[2] Martin, R., "Spectral subtraction based on minimum statistics", *Proc. EUSIPCO, pp. 1182-1185*, April 1994.

[3] Malah, D., Cox, R., and Accardi, A., "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments", *Proc. ICASSP, Vol. 2, pp. 789-792*, 1999.

[4] Cohen, and Berdugo, B., "Speech enhancement for non-stationary environments", *Signal Processing, Elsevier, Vol. 81, pp. 2403-2418*, 2001.

[5] Lu, L., and Zhang, H.-J., " Speaker change detection and tracking in real-time news broadcasting analysis", *Proc.* of 10th ACM international conference on multimedia, *pp.602-610*, December, 2002.