Speech Enhancement of Noisy Speech Using Log-Spectral Amplitude Estimator and Harmonic Tunneling

Hyoung-Gook Kim, Markus Schwab, Nicolas Moreau and Thomas Sikora

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

Abstract

In this paper we present a two stage noise reduction algorithm for speech enhancement. The speech noise removal algorithm is based on a two stage noise filtering LSAHT by log spectral amplitude speech estimator (LSA) and harmonic tunneling (HT) with spectral subtraction. The performance of the system is measured by the segmental signal-to-noise ratio, mean opinion score (MOS) tests, and the recognition accuracy of an Automatic Speech Recognition (ASR) in comparison to other noise reduction methods.

1. Introduction

In many speech communication applications including mobile voice communications and speech recognition, the recorded and transmitted speech signals contain a considerable amount of acoustic noise. The background noise causes a signal degradation, which can lead to total unintelligibility of the speech and decreases the performance of speech coding and speech recognition systems. In speech enhancement, on of the main objectives is to maximize noise reduction while minimizing speech distortion. To attain such an objective, many approaches based on short-time spectral amplitude estimators have been developed. Such spectrum attenuation technique consists of two basic steps: (i) estimation of noise spectrum and (ii) the estimation of speech.

Concerning noise estimation, Martin [1] has proposed an efficient noise estimator based on minimum statistics to track non-stationary noise. An alternative solution is proposed by Cohen [2] who uses the minimum statistics as a voice activity detector and estimates the noise by a recursive averaging. The noise can also be tracked while speech periods by exploiting the harmonic structure of voiced speech. For this, Ealey et al. [3] proposed to estimate the noise between the harmonic components of the voiced speech and in the harmonic spectral peaks of the speech the noise estimation is achieved by tunneling.

Regarding speech estimation, spectral substraction based on the modified Wiener rule is a commonly applied method. In this, a good trade-off between overestimation factor and a spectral floor enables successful reducing musical tones.

Another efficient speech estimator such as log-spectral amplitude (LSA) [4] spectral gain function based on a Gaussian statistical model has been proposed by Ephraim and Malah.

In this paper we present a speech enhancement algorithm based on a two stage noise reduction method called LSAHT. A first noise reduction stage uses a modified minimum controlled recursive averaging noise estimation and LSA speech estimator. A second noise reduction stage is achieved by harmonic tunneling (HT) and spectral substraction. Especially, the speech enhancement based on LSAHT in combination with noise robust front-end improves both speech recognition performance and the quality of speech reconstruction at back-end of distributed speech recognition system under noisy conditions.

2. Structure of speech enhancement algorithm

Usually, the speech enhancement problem is addressed from the estimation point of view in which the clean speech is estimated under the uncertainty of speech presence [5] in noisy observations. The idea of utilizing the uncertainty of speech presence in the noisy spectrum has been applied by many authors to improve the performance of speech enhancement systems. In this paper, we present a simple modified log-spectral amplitude (LSA) speech estimator [4] and harmonic tunneling (HT) [3].

A simplified block diagram of a two stage noise filtering system LSAHT based on LSA speech estimator and HT is shown in figure 1.

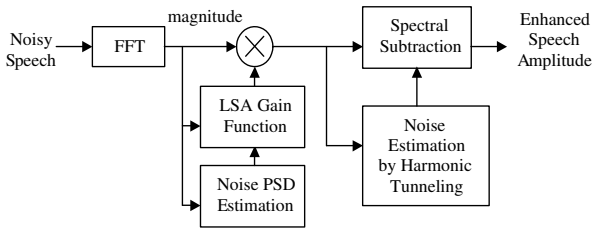

Figure 1: *Block diagram of the speech enhancement*

Let $x(n)$ denote the 8 kHz sampled input speech, which is assumed to be clean speech $s(n)$ and additional noise $d(n)$. The observed noisy signal $x(n)$ is divided into overlapping frames. A pre-emphasis filter is then used to emphasize the higher frequency components. In the frequency domain the short-time frequency components can be expressed with their magnitude $A(k,i)$ and their phase $\phi(k,i)$, where i denotes the time frame index and k the frequency index:

$$X(k,i) = A(k,i)e^{j\phi(k,i)} \tag{1}$$

The phase $\phi(k,i)$ remains unchanged during the noise reduction algorithm. We will only act on the short-time magnitude spectrum $A(k,i)$.

2.1. Modified log spectral amplitude estimator

First, a modified MCRA noise estimation is implemented. Therefore, an average of the short time spectrum is performed over B frames by:

$$E(k,i) = \frac{1}{B}\sum_{i=0}^{B-1} A(k,i). \quad (2)$$

The minimum values $M(k,i)$ of the averaged short-time magnitude $E(k,i)$ spectrum are calculated within windows of S frames. The minimum value for the current frame is found by a comparison with the stored minimum value:

$$M(k,i) = \min_{s=0..S}\{M(k,i-s), E(k,i)\}, \quad (3)$$

This minimum is used as a threshold and controls voice activity detectors in each subbands. The indicator function for the voice activity detectors $I(k,i)$ is defined by:

$$I(k,i) = \begin{cases} 1 & \text{if } E(k,i) < M(k,i)\tilde{T}(k,i) \\ & \quad A(k,i) < M(k,i)T(k,i) \\ 0 & \text{otherwise} \end{cases}, \quad (4)$$

where $T(k,i)$ and $\tilde{T}(k,i)$ are threshold functions:

$$T(k,i) = 1 + 4\exp\Big[-G_{LSA}(k,i-1)\Big] \quad (5)$$

and

$$\tilde{T}(k,i) = 1 + 0.5\exp\Big[-G_{LSA}(k,i-1)\Big]. \quad (6)$$

$T(k,i)$ and $\tilde{T}(k,i)$ depend on the spectral gain factor from the previous block $G_{LSA}(k,i-1)$. If $G_{LSA}(k,i-1)$ is high speech is more probable and the threshold is decreased. The first constrained $E(k,i) < M(k,i)\tilde{T}(k,i)$ will detect all speech pauses since the threshold function is low in this case and the second constraint $A(k,i) < M(k,i)T(k,i)$ ensures that there is only speech present.

The a priori probability for speech absence is then obtained by a smoothing equation using the indicator function $I(k,i)$:

$$q(k,i) = \alpha_q q(k,i-1) + (1-\alpha_q)I(k,i), \quad (7)$$

where $\alpha_q \in [0,1]$ is the time-smoothing factor and $q(k,i-1)$ denotes the speech absence probability from the previous frame. The noise estimation $\lambda_d(k,i)$ is then obtained by a recursive smoothing over the time:

$$\lambda_d(k,i) = \alpha_d(k,i)\lambda_d(k,i-1) + (1-\alpha_d(k,i))A(k,i) \quad (8)$$

using a smoothing parameter controlled by the speech presence probability $q(k,i)$:

$$\alpha_d(k,i) = 1 - F_d|\gamma(k,i-1)|q(k,i) \quad (9)$$

with $F_d \in [0,1]$ constant.

Now, we can define the a posteriori signal-to-noise-ratio SNR $\gamma(k,i)$, the a priori SNR $\xi(k,i)$, and $\nu(k,i)$:

$$\gamma(k,i) = \frac{A(k,i)}{\lambda_d(k,i)}, \quad (10)$$

$$\nu(k,i) = \frac{\xi(k,i)}{1+\xi(k,i)}\gamma(k,i), \text{ and} \quad (11)$$

$$\xi(k,i) = \beta G_{LSA}(k,i-1)\frac{\gamma(k,i)}{1-q(k,i)} + (1-\beta)P\{\gamma(k,i)-1\} \quad (12)$$

using

$$P\{\gamma(k,i)-1\} = \begin{cases} \gamma(k,i)-1 & \text{if } \gamma(k,i) \le 1 \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

where $\beta \in [0,1]$ is the SNR smoothing factor. The amplitude gain function $G_{LSA}(k,i)$ is then calculated with these parameters and the log spectral amplitude rule:

$$G_{LSA}(k,i) = \frac{\xi(k,i)}{1+\xi(k,i)}\exp\Big(0.5\int_{t=\nu(k,i)}^{\infty}\frac{e^{-t}}{t}dt\Big), \quad (14)$$

The output of the first noise reduction stage is then estimated by:

$$O(k,i) = G_{LSA}(k,i)A(k,i). \quad (15)$$

2.2. Harmonic tunneling and spectral substraction

Although this LSA estimator proved very efficient in reducing musical noise phenomena, there is still some remaining noise which lowers the speech quality. Some applications are very sensitive to this remaining residual background noise after LSA estimation. Such applications are low bit rate speech coders and speech recognition systems. Therefore, a second noise reduction stage is employed. From the magnitude spectrum $O(k,i)$, output of the first noise reduction stage, the voicing level is obtained by normalizing spectral autocorrelation at a lag equal to a pitch period in frequency domain. At the next stage, the peak detector is used to find the number of peaks and the frequency bin of the peak corresponding to the highest harmonic within the auto-correlation. Each of these candidate peaks is analyzed to categorize it as a peak coming from either a voiced speech harmonic or noise. To determine the harmonic amplitude $O(h,i)$ and harmonic frequency in the frame h, we proceed as follows:

$$O(h,i) = max_{m\in[a,b]}(|O(m,i)|), \quad (16)$$

where $a = floor((harmo-c)(f_0/S_r/N))$ using the sampling rate S_r and the estimated fundamental frequency f_0, and $b = ceil((harmo+c)(f_0/S_r/N))$. $c \in [0,0.5]$ determines the tolerated non-harmonicity. The estimate $\lambda_{HT}(k,i)$ of the noise is then obtained by sampling the noise spectrum in the tunnels between the harmonic spectral peaks and by interpolation of the frequency and time from the adjacent noise spectra in the surrounding tunnels. Finally, the enhanced spectral amplitude $\tilde{S}(k,i)$ is achieved by spectral subtraction:

$$\tilde{S}(k,i) = O(k,i) - \lambda_{HT}(k,i). \quad (17)$$

In figure 2 the results of the speech enhancement algorithm are shown. Figure 2 (a) denotes spectral magnitude of the clean speech. In figure 2 (b) f16 cockpit noise was artificially added to the clean speech at SNR of 7 dB. Its spectral magnitude at frequency bin $f = 9$ is illustrated. The valleys of figure 2 present a noise estimation as bold line while peaks correspond to the spectral amplitude of the noisy speech signal (thin line). Finally, 2 (d) presents the magnitude of the enhanced speech.

3. Noise Robust Front-End Speech Recognition

Our LSAHT speech enhancement algorithm can be used for robust feature extraction at an extended front-end for Distributed

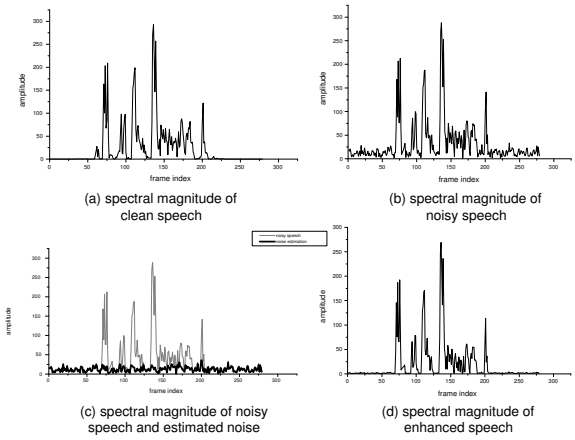

Figure 2: *Spectral magnitude of clean, noisy and enhanced speech and noise estimation.*

Speech Recognition (DSR) of tonal languages as well as speech reconstruction at the back-end. In such applications noise reduced features are extracted and transmitted with some parameters, i. e., pitch period, the number of harmonic peaks.

We used 13 MFCC coefficients as the noise reduced features for back-end speech recognition and reconstruction. The used frequency range is between 64 Hz and 4000 Hz and the noise-reduced power spectrum $\tilde{S}(k,i)$ is first mel-filtered into 23 half-overlapping bands. The 23 spectral values of mel-filtering are then logarithmised by a natural logarithm function. With a 23-point Discrete Cosine Transformation, the 13 MFCC coefficients are calculated.

For a low bit rate speech compression and decompression we use the method in [6]. The pitch period value is quantized using 7 bits. The number of harmonic peaks is quantized using 3 bits for the past, current and future frames. The MFCC feature vectors are quantized using a 4-split VQ with 37 bits. The streams of the compressed MFCC feature vectors, the compressed pitch period value, and the compressed number of harmonic peaks are multiplexed together to form the output bit stream for storage or transmission.

3.1. Reconstruction of speech at back-end server

The transmitted bit stream to server is fed into a stream of compressed MFCC feature vectors, a stream of compressed pitch, and a stream of number of harmonic peaks. The decompressed MFCC feature vectors may be used by the speech recognition back-end. For the speech reconstruction, the MFCC feature vectors are transformed back into the Mel-frequency domain by inverse DCT and the spectral magnitude $\breve{S}(i,l)$ is computed by exponentiation from the log-spectra [6].

Speech is synthesized using a harmonic sinusoidal model from the decompressed MFCC feature vectors, the decoded pitch values, and the number of harmonic peaks for voicing decision by

$$\tilde{s}_i(j) = \sum_{l=0}^{L-1} \bar{S}_l(j) \cos\Big(\tilde{\phi}_l(j)\Big), \qquad (18)$$

where the speech sample $\tilde{s}_i(j)$ is synthetisized as the sum of a number of harmonically related sinusoids with amplitude $\bar{S}_l(j)$ at multiples of the fundamental frequency and synthetic phase $\tilde{\phi}$. For voiced speech, the model is based on the assumption that the perceptually important information resides mainly in the harmonic samples of the pitch frequency. Because of the relatively slow variation in the amplitude between successive frames and the insensitivity of the human auditory system to slight inconsistencies in the speech amplitude, a straight forward linear interpolation is given by

$$\bar{S}_l^i(j) = \breve{S}_0^i \cdot j + \Big(\breve{S}_l^{i+1} - \breve{S}_l^i\Big)\Big(\frac{j}{L}\Big). \qquad (19)$$

The phase is reconstructed from the decoded pitch values using a quadratic model which assumes linear pitch variations:

$$\tilde{\phi}_l^i(j) = l f_0^{i-1} j + \frac{l(f_0^i - f_0^{i-1})}{2N} j^2 + \varphi_l, \qquad (20)$$

where f_0^{i-1}, f_0^i are the pitch frequency values for the $(i-1)^{th}$ frame and the i^{th} frame respectively, N is the frame size in samples, and φ_l is zero for harmonics below a threshold frequency called voicing and a random variable uniformly distributed in $[-\pi, \pi]$ for harmonics above the voicing frequency. For unvoiced speech, the magnitude spectrum is sampled at 100 Hz and a uniformly distributed random phase is applied to each frequency component.

4. Experimental Results

The performance of the proposed algorithm is measured using segmental SNR improvement in speech segments, recognition accuracy improvement, subjective study of speech spectrograms, and listening test.

4.1. Segmental SNR improvement

To measure the performance of the proposed algorithm in comparison to other one-channel noise reduction methods, the segmental signal-to-noise ratio ($segSNR$) at back-end of DSR is computed by $SNR_{improve} = segSNR_{out} - segSNR_{in}$ for the enhanced speech signals at back-end of DSR. Three types of background noise - white noise, car noise and factory noise - were artificially added to different portions of the data at SNR of 5 dB and -5 dB. Table 1 shows that LSAHT algorithm gives best results for input SNR 5 dB and -5 dB compared to the results of PSS, MS, DLSA and NSMR.

methods	Input SNR [dB]					
	white		car		factory	
	5	-5	5	-5	5	-5
PSS	4.3	7.3	5.3	8.1	4.1	7.3
MS	7.8	12.3	8.4	13.5	7.4	11.9
DLSA	7.9	12.6	8.6	13.2	7.2	12.1
NSMR	8.9	13.6	9.1	13.3	8.5	12.7
LSAHT	9.1	14.9	11.3	15.7	10.0	14.3

Table 1: *Comparison of segmental SNR improvement of different one-channel noise estimation methods. PSS: Power Spectral Subtraction, MS: spectral subtraction based on minimum statistics* [1]*, DLSA: log-spectral amplitude speech estimator by spectral minimum tracking* [7]*, NSMR: the ratio of the spectral amplitude of the noisy speech to its minimum* [8] *and LSAHT: the proposed noise reduction method using two stage noise filtering.*

4.2. Recognition accuracy in a DSR system

For evaluation of the improvement of speech recognition with the presented noise reduction algorithm, the Aurora 2 database together with a hybrid HMM/MLP ASR system (351 inputs, 420 hidden units and 24 outputs) using forward-backward training algorithm [9] have been chosen and two training modes are used: training on clean data and multi-condition training on noisy data. The feature vector consists of 39 parameters: 13 mel frequency cepstral coefficients plus delta and acceleration calculations. The mel-cepstrum coefficients are fed to the MLP (multi-layer perceptron) for the non-linear transformation consisted of 9 frames. The proposed LASHT-filtering front-end was compared to a NSMR front-end, LSA front-end, and MS front-end. For the noisy speech results, we averaged the word accuracies between 0 dB and 20 dB SNR. In the table 2, set A, B, and C refer to matched noise condition, mismatched noise condition, and mismatched noise and channel condition, respectively. Table 2 describes the results of the recognition accuracy.

Training Mode	Set A	Set B	Set C	Overall
Multicondition	86.91	86.61	86.66	86.73
Clean only	72.34	72.70	86.62	77.22
Average	79.63	79.65	86.64	81.97

(a) Word accurancy of DSLA front-end

Training Mode	Set A	Set B	Set C	Overall
Multicondition	89.92	88.41	86.86	88.40
Clean only	74.16	73.01	82.13	76.43
Average	82.04	80.01	84.50	82.42

(b) Word accurancy of MS front-end

Training Mode	Set A	Set B	Set C	Overall
Multicondition	89.65	88.35	86.88	88.29
Clean only	79.28	78.82	82.13	80.08
Average	84.47	83.59	84.51	84.19

(c) Word accurancy of NSMR front-end

Training Mode	Set A	Set B	Set C	Overall
Multicondition	91.45	90.21	89.13	90.26
Clean only	84.32	82.41	82.78	83.17
Average	87.89	86.31	85.96	86.69

(d) Word accurancy of LSAHT front-end

Table 2: *Comparisons of word accuracies (%) between four noise reduction algorithms (DSLA, MS, NSMR and LSAHT) on the Aurora 2 database*

As seen in the results of table 2, LSAHT provides much better performance than DLSA algorithm, MS algorithm, and NSMR algorithm.

4.3. Speech spectrograms and listening test

In order to visualize the effect of the noise reduction algorithm based on LSAHT, the spectrograms of noisy speech and the reconstructed speech at back-end server are shown in figure 3. The noisy spectrograms in the upper image of figure 3 was recorded in a busy street with a SNR of about 5 dB. The spectrogram of the reconstructed speech at back-end server is depicted in the lower parts of figure 3. Dark gray areas correspond to the speech components while background noise is light gray. The picture clearly indicates that only speech portions pass the system whereas the noise is suppressed. To evaluate the quality of four (MS,DLSA, NSMR, LSAHT) speech enhancement methods of DSR back-end speech synthesizers, a subjective Mean-Opinion-Score (MOS) was performed with noisy speech corrupted by car noise at SNR 10 dB. The noisy uncoded speech scored 2.16. The MS, the DLSA and, NSMR and LSAHT back-end synthesizer scored 2.53, 2.43, 2.65 and 2.83 respectively.

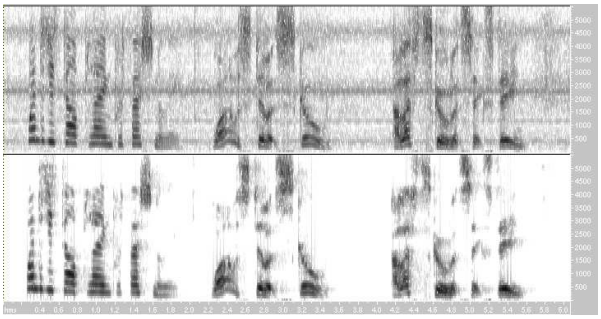

Figure 3: *Spectrograms of noisy speech, reconstructed speech at back-end of DSR system.*

5. References

[1] R. Martin, "Spectral subtraction based on minimum statistics", Proceedings of the Seventh European Signal Processing Conference, EUSIPCO-94, Edinburgh, Scotland, 13 16 September 1994, pp. 1182 1185.

[2] Israel Cohen, "Noise Estimation by Minima Controlled Recursive Averaging for Robust Speech Enhancement", IEEE Signal Processing Letters, vol. 9, no. 1, January 2002.

[3] Ealey, D., Kelleher, H., Pearce, D., "Harmonic tunneling: tracking non-stationary noises during speech", in EUROSPEECH, Aalborg, pp. 437–410, Sep. 1999.

[4] Y. Ephraim, D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator", IEEE Trans. Acoust. Speech Signal Process. ASSP-33 (2) (April 1985) 443–445

[5] Malah, D., Cox, R. V., AccardiLee, A. J., "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments", in Proc. ICASSP, Phonix, AZ, vol. 1, pp. 201–204, March 1999.

[6] Ramabadran, T., Meunier, J., Jasiuk, M., and Kusher B., "Enhancing Distributed Speech Recognition with back-end speech reconstruction", in EUROSPEECH, pp. 1859–1862, Sep. 2001.

[7] Doblinger, G., "Computationally efficient speech enhancement by spektral minimum tracking in subbands", in EUSPICO, pp. 1513–1516, Sep. 1995.

[8] Kim, H.-G., Ruwisch, D., "Speech enhancement in non-stationary noise environment", in ICLSP, pp. 1829–1832, Sep. 2002.

[9] Hennbert, J., Ris, C., Bourlard, H., Renals, S., Morgan, N., "Estimation of global posteriors and forward-backward training of hybrid HMM/ANN Systems", in EUROSPEECH, pp. 1951–1954, Sep. 1997