Spectral Subtraction with Full-Wave Rectification and Likelihood Controlled Instantaneous Noise Estimation for Robust Speech Recognition

Haitian Xu, Zheng-Hua Tan, Paul Dalsgaard and Børge Lindberg

SMC-Speech and Multimedia Communication, Department of Communication Technology
Aalborg University, Denmark
{hx,zt,pd,bli}@kom.aau.dk

Abstract

In standard Spectral Subtraction (SS), Half-Wave Rectification SS (HWR-SS) is normally applied to avoid negative values in the Power Spectral Density (PSD) that occur mainly due to inaccurate noise estimation caused by a Voice Activity Detector (VAD).

In this paper analyses show that, given accurate noise estimation, the phase relationship between speech and noise becomes the dominant cause of the negative values. Full-Wave Rectification based SS (FWR-SS) combined with Instantaneous Noise Estimation (INE) is therefore proposed to be applied instead of VAD based HWR-SS as it is better capable of maintaining the speech information in those negative values. It is also shown in the paper that FWR-SS provides optimum orthogonality between the estimated noise and speech signals.

The INE method proposed in this paper is Likelihood Controlled Instantaneous Noise Estimation (LCINE), which combines long-term statistical characteristics of noise resulting from a VAD with a method of short-term INE.

The combination of FWR-SS and LCINE is computationally efficient and shows a 51% error rate reduction on the Aurora 2 database in comparison to the basic Aurora front-end provided by ETSI [1].

1. Introduction

Automatic Speech Recognition (ASR) has reached a stage where ASR in moderate noisy environments has a very high level of performance – even for large vocabulary tasks. However, deploying an ASR system in environments in which the acoustic noise is less controlled – e.g. for speech operated hand-held terminals in connection with wireless communication – the speech signal may be severely contaminated resulting in dramatic degradation in recognition performance. It is therefore a challenging research task to develop effective and efficient methods for robust speech recognition.

SS is often used as an efficient front-end noise reduction method, mainly due to its simple implementation. As proposed in [2] and further improved in [3], SS to some extent effectively removes additive noise by subtracting noise estimated in non-speech segments. The drawback of this SS technique, however, is that it may generate "musical noise" due to inaccurate noise estimation. A further drawback is that it may also result in negative values occurring in the PSD. Forcing these to zeros or small positive floor values, as is extensively used by HWR-SS, results in the loss of speech information.

In this paper an initial analysis is conducted which illustrates that negative values occur not only due to inaccurate noise estimation but also to the phase relationship between noise and speech. Further analysis shows that if accurate noise estimation is available the phase relationship is the dominant cause of negative values and this research applies FWR-SS with the aim of maintaining the speech information in negative valued bins and proves that it can provide optimum orthogonality between the estimated noise and speech signals than HWR-SS.

Accurate noise estimation is vital for successful deployment of SS. INE has been used [4], [5], [6] to estimate the noise without explicitly utilising speech pause detection. It is shown that INE can provide SS with more accurate noise estimation than methods based on VAD and improved performance for both speech enhancement and robust ASR is obtained. However, based on short-term estimation only, the INE method inevitably causes large variations in the estimated noise. To counteract this, [7] introduced a method in which the probability of speech presence is taken into account in the noise estimation strategy.

In this paper a different strategy is proposed, which estimates the reliability of the INE: Based on the fact that non-speech segments can provide reliable statistical information about the noise, LCINE combines long-term noise statistical characteristics with the short-term INE resulting in more accurate noise estimation.

The paper is organised as follows. Section 2 presents the analysis of negative values and the basic rationale behind applying FWR-SS. Section 3 proposes LCINE technique to provide better noise estimation. Results from experiments on Aurora 2 are presented and discussed in Section 4. The conclusions are in Section 5.

2. Rationale Behind Full-Wave Rectification Based SS

Assuming that $\boldsymbol{X}(k)$, $\boldsymbol{Y}(k)$ and $\boldsymbol{N}(k)$ are the k*th* FFT bin of the spectrum for clean speech, noisy speech and noise in a frame, respectively, SS subtracts the estimated noise from the noisy speech signal in the PSD domain according to the following formula:

$$|\hat{\boldsymbol{X}}(k)|^2 = |\boldsymbol{Y}(k)|^2 - |\hat{\boldsymbol{N}}(k)|^2 \qquad (1)$$

where $\hat{\boldsymbol{N}}(k)$ and $\hat{\boldsymbol{X}}(k)$ are estimations of $\boldsymbol{N}(k)$ and $\boldsymbol{X}(k)$, respectively, and $\hat{\boldsymbol{N}}(k)$ is normally calculated from non-speech segments identified by a VAD.

2.1. Analysis of negative values

It is clear from Eq(1) that negative values of $|\hat{\boldsymbol{X}}(k)|^2$ may occur. Based on the fact that most negative values result from inaccurate noise estimation, the HWR-SS method replaces them with zeros or small positive floor values [3]. However, the negative values can also be caused by the phase relationship between the clean speech and the noise. Since $\boldsymbol{X}(k)$, $\boldsymbol{Y}(k)$ and $\boldsymbol{N}(k)$ are complex valued stochastic variables and can be treated as vectors in complex plane, the relationship among them can be illustrated as:

$$|\boldsymbol{Y}(k)|^2 = |\boldsymbol{X}(k)|^2 + |\boldsymbol{N}(k)|^2 + 2|\boldsymbol{N}(k)||\boldsymbol{X}(k)|\cos\theta_k \quad (2)$$

where θ_k is the random phase difference between $\boldsymbol{X}(k)$ and $\boldsymbol{N}(k)$. When $\cos\theta_k < -|\boldsymbol{X}(k)|/[2*|\boldsymbol{N}(k)|]$, $|\boldsymbol{Y}(k)|$ is smaller than $|\boldsymbol{N}(k)|$, and therefore a negative value for $|\hat{\boldsymbol{X}}(k)|^2$ occurs when applying Eq(1) even though the noise estimation is exact.

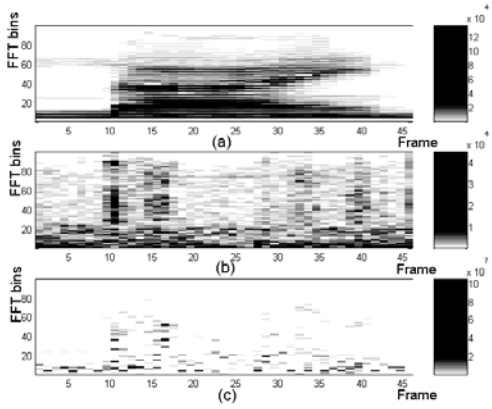

Figure 1: Example of negative valued bins during the SS caused by the phase relationship
(a) Amplitude spectrum of clean speech;
(b) Amplitude spectrum of "car" noise;
(c) Absolute value of negative $|\hat{\boldsymbol{X}}(k)|^2$

To illustrate the occurrence of negative values caused by the phase relationship, Fig.1 shows three spectra as an example. The noisy speech is calculated by adding "car" noise (Fig.1(b)) to clean speech (Fig. 1(a)) with a SNR equal to 0dB. $|\hat{\boldsymbol{X}}(k)|^2$ is calculated from Eq(1) using the added noise PSD as the noise estimate. Fig.1(c) shows the negative values of $|\hat{\boldsymbol{X}}(k)|^2$ where it is observed, that even with exact noise estimation, some negative values still occur. These negative values are solely caused by the random phase difference θ_k between $\boldsymbol{X}(k)$ and $\boldsymbol{N}(k)$.

From the analysis it is concluded that the more accurate the noise estimation is, the larger part of the negative values is caused by the phase relationship. In this case, setting them artificially to zeros or small positive floor values as in HWR-SS will inevitably result in the loss of speech amplitude information. Using the FWR-SS algorithm instead and reversing the negative values will exploit speech information remaining in the bin, which can be expressed as:

$$|\hat{\boldsymbol{X}}(k)|^2 = \left| |\boldsymbol{Y}(k)|^2 - |\hat{\boldsymbol{N}}(k)|^2 \right| \quad (3)$$

2.2. FWR-SS and orthogonality

Geometrical considerations are given below aimed at explaining and interpreting FWR-SS. SS – as given by Eq(1) – assumes that the vector $\boldsymbol{X}(k)$ and $\boldsymbol{N}(k)$ are uncorrelated and the mean value of the noise is equal to zero, namely they are orthogonal in a statistical sense [3]. Further assuming that $|\hat{\boldsymbol{N}}(k)| = |\boldsymbol{N}(k)|$, $\boldsymbol{Y}(k)$, $\hat{\boldsymbol{X}}(k)$ and $\hat{\boldsymbol{N}}(k)$ should statistically build a right angled triangle in which $\boldsymbol{Y}(k)$ acts as the hypotenuse. For $|\boldsymbol{Y}(k)| > |\hat{\boldsymbol{N}}(k)|$ and $|\boldsymbol{Y}(k)| < |\hat{\boldsymbol{N}}(k)|$, FWR-SS can be illustrated in Fig.2(a) and 2(b), respectively.

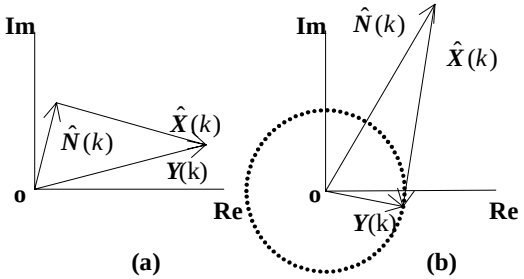

Figure 2: Geometrical illustration of FWR-SS
(a) for $|\boldsymbol{Y}(k)| > |\hat{\boldsymbol{N}}(k)|$ (b) for $|\boldsymbol{Y}(k)| < |\hat{\boldsymbol{N}}(k)|$

For $|\boldsymbol{Y}(k)| > |\hat{\boldsymbol{N}}(k)|$, FWR-SS exactly constructs such a right angled triangle mentioned above. However, for $|\boldsymbol{Y}(k)| < |\hat{\boldsymbol{N}}(k)|$, by reversing the negative values, FWR-SS actually constructs a right angled triangle with $\hat{\boldsymbol{N}}(k)$ the hypotenuse. This in fact maximally meets the requirement for orthogonal relationship as the angle between $\hat{\boldsymbol{X}}(k)$ and $\hat{\boldsymbol{N}}(k)$ is closest to 90^0 when $\hat{\boldsymbol{X}}(k)$ lies on the tangent line of the circle whose radius is $|\boldsymbol{Y}(k)|$.

Therefore, for exploiting speech information and keeping the orthogonal relationship, FWR-SS is better than HWR-SS and should be used in the precondition of accurate noise estimation.

3. Instantaneous Noise Estimation

Accurate noise estimation is of paramount importance for the SS algorithm, particularly the FWR-SS. The Minimum Statistics Noise Estimation (MSNE) presented in [4] and [6] is a method based on the assumption that speech cannot take up a frequency bin all the time. Therefore a window of for example 0.5 second is set, and the minimum value in the PSD domain in the window of each frequency bin is treated as the noise estimate within the current frame. The advantages of this method are that it does not need VAD and that it tracks noise changes even during speech which gives relatively more accurate and instantaneous estimation of the noise. The disadvantage is large variations in the estimate of noise [5]. To reduce these variations and enhance the accuracy of MSNE, the next sections introduce the use of noise estimation smoothing and LCINE.

3.1. Smoothing

First, a simple smoothing algorithm is applied to average the variations in the instantaneous estimation of noise. Assuming $\hat{P}_n(k)_M$ and $\hat{P}_n(k)_S$ are the PSDs of MSNE and the smoothed noise estimation for the k*th* FFT frequency bin in the n*th* frame respectively, the smoothing is performed as follows:

$$\hat{P}_n(k)_S = \xi \hat{P}_{n-1}(k)_S + (1-\xi)\hat{P}_n(k)_M \quad (4)$$

where ξ is the memory factor, chosen to be smaller than 0.5 as a compromise to enable tracking both stationary and non-stationary noise.

3.2. LCINE

Using this simple smoothing only has a limited improvement. To further enhance the performance of the INE, the effect of exploiting the long-term information available - from non-speech frames identified by the VAD - is analysed. Two facts have been observed. Firstly, the long-term characteristics of noise such as the Probability Density Function (PDF) are relatively stable, which may assist the INE in reducing variations. Secondly, INE is a short-term estimation method that introduces estimation errors. The likelihood of the INE in each bin calculated from long-term properties may be used to discriminate such errors. LCINE can be fulfilled as follows:

- Firstly, the statistical characteristic, the mean μ of the noise value in a PSD bin, is estimated from non-speech frames. Then the PDF of the PSD for Gaussian noise is defined by an exponential distribution [6]:

$$f(x) = \frac{U(x)}{\mu} e^{-x/\mu} \quad (5)$$

  where *U(x)* is the unit step function and *x* is the noise value in the PSD bin.
- Secondly, the normalized likelihood of MSNE $L_n(\hat{P}_n(k)_S)$ is calculated by Eq(6a) according to the noise PDF obtained in Eq(4), and smoothing is carried out to track the long term tendency by Eq(6b):

$$L_n(\hat{P}_n(k)_S) = \frac{f(\hat{P}_n(k)_S)}{f(0)} \quad (6a)$$

$$L_n(\hat{P}_n(k)_S) = \gamma L_n(\hat{P}_n(k)_S) + (1-\gamma) L_{n-1}(\hat{P}_{n-1}(k)_S) \quad (6b)$$

  where γ is a forgetting factor, normally chosen to be smaller than 0.5.
- Finally, the noise estimation is given as follows:

$$\hat{P}_n(k) = L_n(\hat{P}_n(k)_S)\hat{P}_n(k)_S \quad (7)$$

  By Eq(7), the final estimation of noise is proportional to its likelihood. Therefore, the larger the likelihood is, the more noise SS will subtract. In the extreme case, if the likelihood is zero indicating the MSNE is completely unreliable, by applying Eq(7), no noise will be removed in this frequency bin, which prevents the loss of speech information due to unreliable noise estimation.

4. Experiments and Discussions

To evaluate the performance of the proposed methods, a number of experiments on the basis of noise-free speech data training and multi-condition testing on the Aurora 2 database are conducted. The English digit utterances are artificially contaminated with different types of noise in Set A, Set B and Set C. The recognizer and the baseline results below are from the Aurora 2 CDs provided by ETSI[1].

4.1. FWR-SS

Table 1 summarises the results for the combination of the HWR-SS and FWR-SS methods with the noise estimation methods VAD and MSNE, where HWR-SS is the algorithm in [3] with β = 0.02. It is observed that the combination of FWR-SS and MSNE gives the largest improvement by 29% average error rate reduction. Such significant improvement verifies that FWR-SS is able to exploit the speech information inherent in negative valued bins on the basis of a relatively accurate noise estimation. When the noise estimation is provided by a VAD, the performance of FWR-SS is even worse than HWR-SS. The reason for this may be that the VAD-based noise estimation is non-instantaneous and makes the estimation inaccuracy the dominant cause for negative values.

Table 1: Recognition rates (%) and Relative Improvement (%) over baseline for FWR-SS, HWR-SS, VAD, MSNE

Algorithm	Set A	Set B	Set C	Overall	**Imprv.**
Baseline	61.34	55.75	66.14	60.06	**--**
HWR-SS+VAD	65.67	65.13	58.55	64.03	**9.9**
FWR-SS+VAD	65.36	61.29	62.29	63.12	**7.7**
HWR-SS+MSNE	68.59	65.42	72.20	68.04	**20.0**
FWR-SS+MSNE	71.86	69.56	75.53	71.67	**29.1**

4.2. LCINE combined with FWR-SS

Table 2: Recognition rates (%) and Relative Improvement (%) over baseline for smoothing and LCINE with FWR-SS

Algorithm	Set A	Set B	Set C	Overall	**Imprv.**
Smoothing	74.71	73.64	75.95	74.53	**36.2**
LCINE	81.27	79.15	81.07	80.38	**50.9**

Further experiments are conducted to evaluate the effect of smoothing and LCINE. The results are given in Table 2. All experiments are based on FWR-SS. It is noticed that the simple smoothing algorithm gives a 36% improvement. Significant improvement has been observed by applying LCINE, indicating that LCINE is an effective method for noise estimation.

Additionally, detailed comparisons between LCINE and MSNE, combined with FWR-SS are given in Fig.3 and Fig.4. The LCINE method shows better overall performance than

the MSNE method both for different SNRs and for different noise types. This indicates that the LCINE is a more general method.

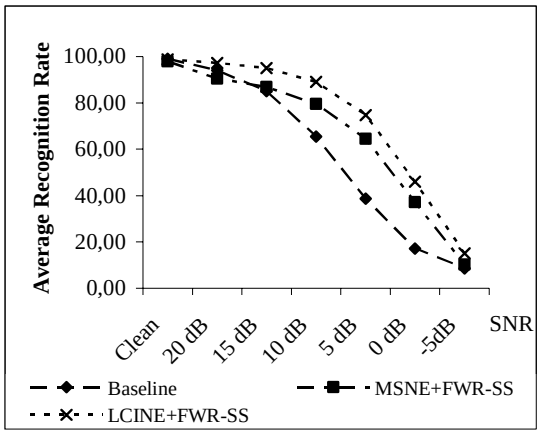

Figure 3: Comparison for LCINE and MSNE across a range of SNR values

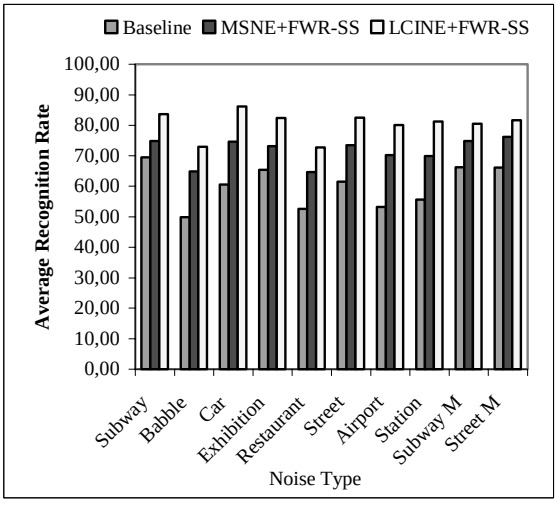

Figure 4: Comparison for LCINE and MSNE for different noise types ("M" represents Set C in the Aurora 2 database where convolutional noises are added)

5. Conclusions

This research focuses on methods aimed at improving speech recognition front-end processing to be deployed in noisy environments. Negative values during SS are not only caused by the inaccurate noise estimation but also by the phase relationship between the spectra of noise and clean speech. These negative values are usually substituted with zeros or small positive floor values to compensate for the annoying musical effects from inaccurate noise estimation. The substitutions, however, cause the loss of speech information within these negative PSD bins. Assuming that instantaneous noise can be estimated accurately, analyses show that FWR-SS is superior to HWR-SS by reversing the negative values and thus retaining potential speech information. Consequently, MSNE based INE is applied with the further improvement by using LCINE for enhancing the estimation accuracy by combining long-term statistical characteristics with instantaneous estimation.

The performance of the proposed methods is verified by experiments on the Aurora 2 database. It is shown that LCINE used in combination with FWR-SS improves the performance by 51% over the baseline front-end algorithm while FWR-SS alone achieves 29%. It is pointed out that the combined use of both FWR-SS and LCINE has low computational complexity and is rather robust to varying types of noise. These properties are of importance when applied in mobile devices.

6. Acknowledgements

This work is supported by a PhD grant from the CNTK (Centre for Network and Service Convergence) project that is partly granted by the Danish Ministry of Science, Technology and Development, partly the participating industrial partners.

7. References

[1] ETSI draft standard doc. "Speech Processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms", ETSI ES 202 108 V1.1.2 (2000-04), http://pda.etsi.org/pda/queryform.asp, April 2000,

[2] Boll, S.F, "Suppression of Acoustic Noise in Speech using Spectral Subtraction", *IEEE Trans. on Acoustic, Speech and Signal Proc*., pp. 113-120, 1979.

[3] Berouti, M., Schwartz, R., and Makhoul, J., "Enhancement of Speech Corrupted by Acoustic Noise", *Proc. of ICASSP, 1979*, pp.208-211.

[4] Martin, R., "An Efficient Algorithm to Estimate Instantaneous SNR of Speech Signals", *Proc. of Eurospeech*, vol. 3, pp. 1093-1096, 1993.

[5] Ris, C., Dupont, S., "Assessing Local Noise Level Estimation Methods: Application to Noise Robust ASR", *Speech Communication*, Vol. 34, Issues 1-2, pp. 141-158, April 2001.

[6] Martin, R., "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics", *IEEE Trans. on Speech and Audio Processing*, vol. 9, No. 5, pp. 504-512, July 2001.

[7] Cohen, I., Berdugo, B., "Noise estimation by minima controlled recursive averaging for robust speech enhancement", *Signal Processing Letters, IEEE*, Vol.9, Issue: 1, pp.12 - 15, Jan. 2002.