Improvement Approaches of Ordered Spectra Warping Criteria for Noise Estimation

Dragomir D. Nikolov

Tampere University of Technology
drago@ieee.bg / dragomir.nikolov@gmail.com

Abstract

Still Ephraim&Malah method has an outstanding performance, but it needs a priori speech/noise information and complicated calculations. On the contrary, noise estimation, based on ordered spectra, is computationally simple and needs no voice activity detection. The known methods have either big statistical error or high memory requirements. They also suffer from the lack of adaptation. The proposed solution is based on a warping of the ordered spectra when a speech/music is present. This warping provides dual information about noise level and speech presence at the same time. The derived method shows very robust voice activity detection and a resistance to speech artefacts like lip "pocks" and breathing. It has reasonable memory requirements. The method can improve the word error ratio up to 18.6%, compared to quantile based noise estimation. Similar performance is achieved with real world noises. A simple approach to suppress the musical noise is proposed too.

1. Definitions

Standing on the fact, that each frequency component of a given audio signal is active only within short time intervals, we state the definitions:

N *frame window* length – 256 samples/frame (at Fs = 11 ÷ 16 kHz) with overlapping=50%,

Q *analysis window* length – varying between 35 and 85 frames; chosen as a compromise between good statistics and good time resolution

s[n] - discrete time speech signal
d[n] - discrete time noise; we assume additive stationary
c[n] =s[n]+d[n] - discrete time noisy speech signal

Based on the STFT, the corresponding spectra are:

$S_m(k)$, $D_m(k)$ and $C_m(k)$, where:
m - frame number, which increases within the time,
k - is the DFT frequency bin $\omega_k=2\pi k/N$,

Using the moving analysis window we cut the last **Q** frames of the corrupted spectra to form the matrix $\mathbf{C_m(k,m-a)}$, **a=0,1,…Q-1**. Then we form the sorted matrix as follows:

$$O_m(k,p)= \mathbf{P}\{ C_m(k,m), C_m(k,m-1), \ldots, C_m(k,m-Q+1)\}. \quad (1)$$

In eq. (1) the operator **P** sorts the values of each row (the magnitudes for given **k**) in ascending order. Therefore, (for each frequency **k**) the sorted matrix O_m follows the rule:

$$O_m(k,0) \le O_m(k,1) \le O_m(k,2) \le \ldots \le O_m(k,Q-1)$$

For normalisation purpose we substitute the index **p** by a quantile position **q**, following the next relation:

p=<q(Q-1)> , where **q** is within the range of $\mathbf{0 \le q \le 1}$.

Further, *mean* μ and *deviation* σ will be related to spectrum magnitude of the noise at a given frequency:

$$\mu(k) = \frac{1}{Q-1}\sum_{m=1}^{Q} D_m(k) \text{ and}$$

$$\sigma(k) = \sqrt{\frac{1}{Q-1}\sum_{m=1}^{Q}[D_m(k)-\mu(k)]^2}\ .$$

2. Introduction

As known, sorting in ascending order of any frequency component forms a vector of two sections. The first section, on the left of the graph (fig. 1, line 3), is gradual. The second section on the right part of the draft is much steeper. Similar shapes are observed at almost all speech signals degraded by more than thirty types of real noises. In a case of environmental noise, it is defined by the speaker ability to adjust the signal to noise ratio of about 15 dB (Lombard effect) [1,2]. However, the same shape is observed within a wide range of signal-to-noise ratios, even for noise added after the speaking process.

Martin develops the above idea for small time intervals based on minimum statistics [3], thus characterised by good dynamics and by big statistical errors. Further findings by Stahl et all [4] improve the noise estimation using a value close to the median of stationary noise ordered spectra. The complexity in this case appears because of the real speech sound where a lot of ordered consecutive spectra frames are needed. That means high memory consumption and low dynamics of the noise estimation. Furthermore, the suggested value close to the median (q = 0.5) of ordered spectra quantiles [4], carries information only for noise *mean* value without considering the *deviation*. As a result the information for high magnitude noise spectral bins found within the interval q=[0.6, .., 0.9] is lost.

Thus a smart approach is necessary if we aim:

- to reduce the memory usage;
- to improve the accuracy;
- to avoid errors when speech goes under the median;
- to capture the noise dynamics.

The methods based on psycho-acoustical noise suppression, such as [5,6], are not appropriate since the optimisation must be done to the aim of word error ratio minimisation.

An alternative approach for noise reduction is the use of neural networks [7,8]. This, in fact, is a direct recognition of noisy features and requires a huge variety of training samples under different SNR and noise types. Hence the speech recognition is without denoising. For those reasons the training is a very complex task while aiming a real-world application. In this approach some of the neural network methods could also introduce additional distortions and musical noise.

3. Motivation

To clarify some initial questions about the structure of degraded speech ordered spectra the following experiment is performed: a synthetic white Gaussian noise with known characteristics (μ, σ, etc.) is added to an utterance.

Several cases of the ordered spectra and the exact noise positions can be seen in fig.1. When the analysis window is short, the position of the noise mean is varying within a wide range. Besides, one can see the importance of the standard *deviation* of the noise (for example, the end of the flat section is close to *mean+deviation* value).

The curves are normalised according to their maximum values (at q=1) in order to be displayed comparably in the same plot. The variation of the *mean* (displayed in fig.1) is due to the normalisation (see the y-axis distance between *mean* and *mean+deviation*). The variation of the noise position, as a function of **q**, depends mainly on the speech duration in the current analysis window.

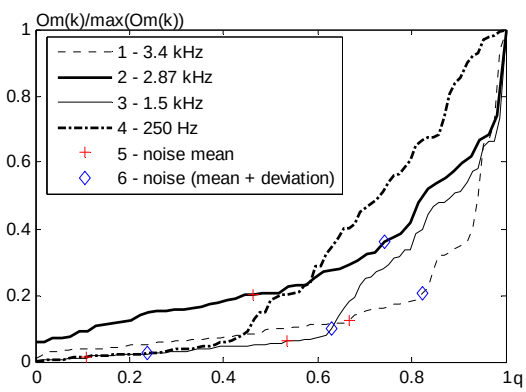

Fig.1

Fig.1 represents: (i) ordered spectra of noisy speech for 4 certain frequencies; and (ii) the positions of means of a known noise and their deviations.

Thus, we may define the following problems of the method:

- What are the specifics of the real noise ordered spectra?
- Can we experiment with artificial noise while developing the method?
- What kind of normalisation should we choose for comparison purpose?
- How could we estimate the dynamical position of the real noise (lines 1, 3 correspond to a good case, while lines 2, 4 – to a difficult one) ?

The first three from the above ones are discussed in depth [9], as they define the motivations and the experimental setup. The major problem is to find the *warping point* (between speech and noise), even in the dynamical situation caused by real speech. This will be in the center of the present discussion. It will define the Ordered Spectra Warping Criteria (OSWC) and the ways of its improvement.

4. Algorithm

The fundamental idea is demonstrated by plotting all the normalised ordered spectra in one picture (fig.2). There one can see the difference between theoretical noise (line 1) and continuous speech (line 5). If the plot for any real $O_m(k)$ is continuously redrawn when increasing the speech duration, a *warping* will be observed. In fig. 1 that can be seen by comparison of the two real cases denoted by line 2 and line 3. It is interesting to mention that this warping is a logarithmic relation between line 1 and line 5.

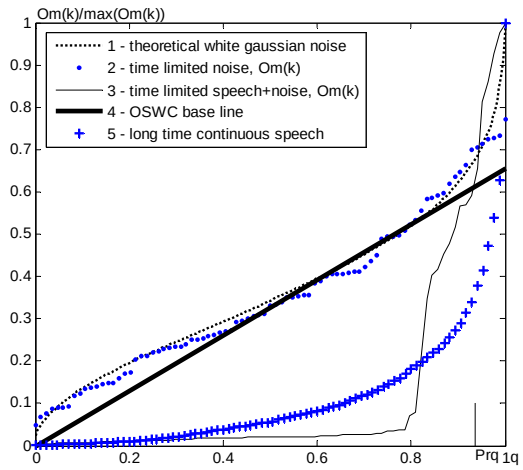

Fig. 2

The definition of the warping is related to the difference between any real ordered spectra and the theoretical one for noise (line 1). From practical viewpoint this method proposes comparison between real ordered spectra and a *base line*. This base line is an interpretation of the mid section of line 1. It is chosen in order to: (i) avoid the deviations from the right (above 3σ); (ii) avoid random pulse noises; and (iii) decrease the computational requirements. The base line is defined by the two points $O_m(0)$ and $O_m(<Q*Prq>)$.

Thus we form a matrix with base lines for each frequency **k**, based on the current sorted spectra, as follows:

$$L_m(k,p) = [O_m(k,p) - O_m(k,0)]\frac{p}{Q} + O_m(k,0) \qquad (2)$$

The base line matrix $L_m(k,p)$ has linearly increasing rows and each of them is based on the ordered spectra magnitude at the Protecting reference quantile (Prq).

The OSWC base line for given **k** is shown with line 4 (fig 1), constructed as a function of line 3.

The normalising Prq has an additional purpose in real signals. The appearance of a high magnitude random pulse noise will make the ordered spectra look warped like speech. Similar errors are avoided by displacing of Prq enough to the left from the maximum magnitudes. Its value is chosen experimentally to balance between noise and real speech. Hence, it prevents the positioning of the noise index within the speech quantiles. Prq is defined as a function of analysis window length. Additional specific of the protective quantile allows precise location of the beginning of the words.

Next step is to obtain the difference between base lines matrix and sorted values, thus producing the Ordered Spectra Warping matrix:

$$W_m(k,p) = L_m(k,p) - O_m(k,p) \tag{3}$$

Then the OSWC states that the noise estimation is got from the matrix O at indexes where the maximums of the warping lines are found.

$$\Theta_m(k) = \arg\max_{p=0}^{Q-1} W_m(k,p) \tag{4.1}$$

$$\hat{D}_m(k) = O_m(k, \Theta_m(k)) \tag{5}$$

Note that this criterion provides separate decisions for each frequency index **k** (while the solution [4] is for all **k**).

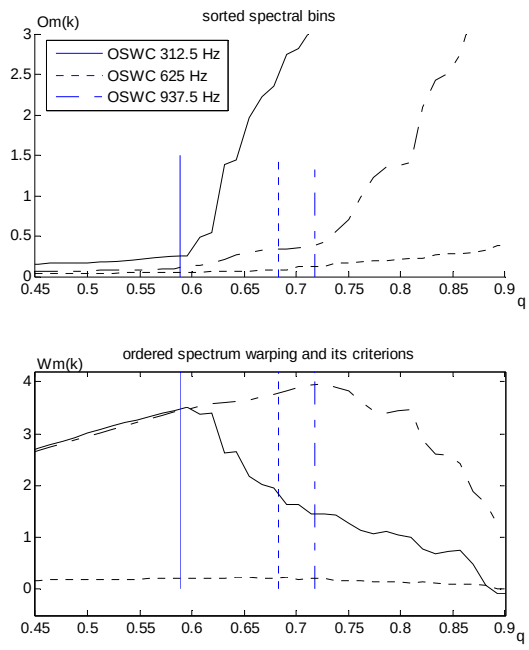

Fig. 3

In Fig.3 ordered spectra at three certain frequencies, their corresponding warpings and the decisions based on the maximums of the criterion are given. The noise level for given frequency **k** can be observed at quantiles with indexes below q=0.5 (median value).

The second application of the difference matrix W is used for VAD by defining the following criteria:

- The sum of the difference line vector $W_m(k,p)$ is used for VAD (voice activity detection/ word segmentation):

$$\Phi s_m = \sum_{k=0}^{N/2} \left[\sum_{p=1}^{Q} W_m(k,p) \right]; \tag{4.2}$$

- A product, based on OSWC, is similar and is calculated by the following way:

$$\Phi p_m = \prod_{k=0}^{Q} \left[\sum_{p=1}^{Q} W_m(k,p) \right] \tag{4.3}$$

5. Signal restoration

Sometimes there are errors caused by the maximum noise quantiles during the last Q frames. Since they result in a poor behaviour we modify the equations (5) and (5.1), as follows:

$$\hat{D}_m(k) = O_m(k, \langle \chi \Theta_m \rangle) \text{ , where} \tag{6}$$

χ is an experimentally chosen correction parameter.

Then a classical spectrum subtraction [10, 11] is performed as follows:

$$R_m(k) = \begin{cases} \left[C_m(k) - \lambda \hat{D}_m(k) \right] & \text{, if } \quad C_m(k) - \lambda \hat{D}_m(k) > 0 \\ 0 & \text{, otherwise} \end{cases} \tag{7}$$

where λ is the oversubtraction parameter.

Further the signal reconstruction is done by the overlap-add method:

$$r(t_m \ldots t_m + N\Delta t_s) = r(t_m \ldots t_m + N\Delta t_s) + \mathrm{Re}\left[FFT^{-1}(R_m(\omega_k) e^{j\varphi(k)}) \right] \tag{8}$$

where

$\Delta t_s = 1/F_s$ - sampling period

$e^{j\varphi(k)}$ - phase spectrum of the original signal frame

If the application of the method requires an audio synthesis, there are possibilities to suppress the "musical noise" – typical for spectral subtraction. If the argument of the maximum Θ is at large **q** value, this frequency band is noise and hence the "musical noise" suppression should be applied. The suppressor is a simple third order median filter (certain frequency through time) over the restored signal. This filter transforms the musical noise into more pleasant type of noise. This solution is simpler than [12].

6. Improvement approaches

As this paper targets the audio processing (especially speech recognition), it is more important to optimise the performance in order to improve the accuracy (the Word Error Ratio *WER*, respectively) than to minimise the error E_m:

accuracy = 100 correct_words/total_words ,*

WER=100 - accuracy

$$E_m = \sum_{k=0}^{N/2} \left(D(k,m) - \hat{D}(k,m) \right)^2$$

In the present study each word is a separate recording into a waveform file. Furthermore *correct_words* is the number of correctly recognised files from the neural network based Speech Recognition System (SRS) into a given experiment. While *total_words* is the number of the files used for testing.

Three types of artificial noise with pink, white and uniform distribution will be used in the experiments.

6.1. Oversubtraction

Widely used approach in most of the methods which improves the performance. Nevertheless, for OSWC (short analysis window) results with λ=1 (Eq. 7) are good and there is no need of such a technique.

The OSWC gives better results with oversubtraction of λ=1.5 for Gaussian noise, but the average results with different noises suggest λ=1.

noise	pink		gaussian		uniform	
λ	1	1.5	1	1.5	1	1.5
OSWC	6.83	8.66	7.83	8.5	10.9	10.5

Table 1. WER [in %] at different oversubtraction (λ=1, λ=1.5)

6.2. First order OSWC smoothing for noise estimation

The essence of Θ is to compress the information, but this supposes some random errors, especially at the non-speech segments. To avoid this, its equation can be modified as follows:

$$\Theta'_m = \nu\Theta_m + (1-\nu)\Theta'_{m-1} \qquad (4.1.1)$$

ν	pink	gaussian	uniform
1	7.5	10.91	7.24
0.1	7.08	10.58	7.16
0.13	7.24	10.66	7.16
0.16	7.41	10.82	7.08

Table 2. WER [in %] vs. the smoothing parameter

6.3. Small area ordered spectra averaging

During the experiments an improvement was found when (6) was modified as follows:

$$D_m(k) = \frac{1}{\beta-\alpha}\sum_{a=\alpha}^{\beta} O_m(k, a+\Theta_m(k)) \qquad (6.1)$$

noise	pink		gaussian		uniform	
eq.	(6)	(6.1)	(6)	(6.1)	(6)	(6.1)
Shtal's	7.9	7.8	10.9	10.6	6.6	6.5
OSWC	13.5	13.4	17.3	17.2	7.8	7.9

Table 3. WER [in %] at 4 frames averaging (β = 1 , α = 4)

6.4. First order reconstructed spectrum smoothing

If equation (7) is further modified as follow:

$$R'_m(k) = \gamma R_m(k) + (1-\gamma)R'_{m-1}(k), \qquad (7.1)$$

the performance can be slightly improved.

Table 4 shows different results of the accuracy vs. the smoothing coefficient γ. Interesting fact is that there are two peaks in the parameters relation. While the first peak is reasonable (γ=0.5), the second one (γ=0.1) probably works only with ANN and features independent to signal energy, such as MFCC or MFCC(M) [13].

γ	pink	gaussian	uniform
1	10.3	7.74	7.24
0.5	8.91	6.49	6.33
0.1	9.1	6.49	6.58

Table 4. WER [in %] when smoothing the reconstructed signal

6.5. Ephraim and Malah suppression rule and OSWC combination

The combination shows slightly better performance for noisy note identification than the other methods. Experiments with speech in this promising direction are not made yet. Such a combination could be an advance in the area of denoising because of their positive features:

- OSWC is adaptive and provides good noise estimation.
- Ephraim and Malah suppression rule [14] is able to improve even the accuracy at low SNR (0,-5db).

6.6. Combined performance

All the experiments are made under assumption of superposition, as the interdependence between parameters controlling different ways of improvements are not taken into account. Having this statement and choosing the best values for the parameters, the results are as follows:

noise	pink	gaussian	uniform
corrupted	11	18.4	6.74
Stahl's [4]	7.9	10.9	6.58
OSWC	5.16	7.24	3.75
ephr&malah [14]	3.33	6.66	2.75
clean	2.08		

Table 5. WER [in %] after final improvements

7. Real noise performance

A street noise recordings from the TUT database were added to the speech with a pre-defined SNR. The drawing shows the definite advantage in comparison to Stahl's method [4].

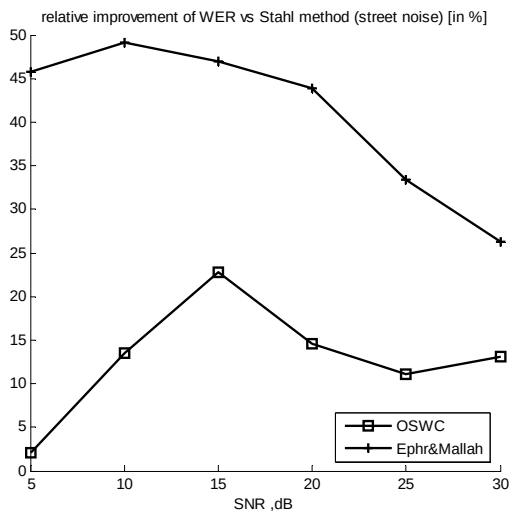

Fig. 4

The above pictures demonstrate the best improvement to be at 15 dB SNR. It is so as the method development and the improvements were performed at this ratio. Therefore, parameters adaptation in respect of SNR is a very important research direction.

8. Implementation

The experiments are based on the SpeeCon II database for Finnish language. A set of 2400 digits recorded with "*CLOSE_HEADSET*" microphone position was used. The set contains 50% female speakers. Equal number of samples per set was used for training and testing. All results presented in the paper are calculated over the test sample set with the same trained neural network. The speech recognition system is implemented as a four layer feed-forward network with a total of 1400 neurons, trained using conjugate gradient back-propagation with Powell-Beale restarts. The network has 10 output neurons corresponding to each of the words. The features used are Modified Mel-Frequency spaced Cepstral Coefficients [13].

9. Voice activity detection

The usability for voice activity detection is visualised in fig. 5.

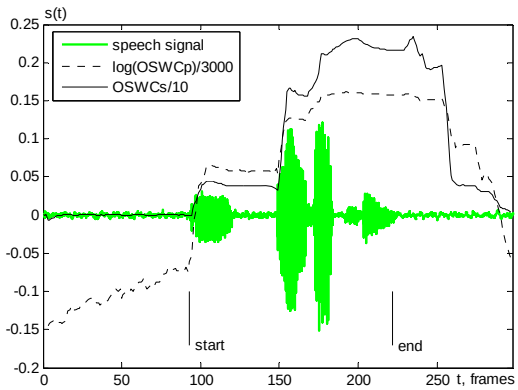

Fig. 5

The specific feature of all the criteria is based on the Protective Reference Quantile. The magnitudes of the first few speech frames are significantly higher than the noise. Thus the warping curve (for certain k) will have high negative values at its end. Therefore, a local minimum could be observed right before the beginning of the speech. Furthermore, only a level crossing is needed to confirm the speech activity. The robustness can be improved if the detection algorithm takes into account the changes in clean spectrum energy:

$$e_m = \frac{2}{N}\sum_{k=0}^{N/2}(D(k,m))^2$$

The only drawback, if VAD is not based on Φ+energy, is that the end of the word/segment will be detected Q frames later.

The VAD tests are conducted using four types of noise: car, street, crowd and train. The table presents the degree of accuracy when detecting the speech activity. The accuracy is calculated as an average value of four noise types: each one with seven SNRs from 0 to 30 dB with a step of 5 dB.

	A	B	C	D
$\log(\Phi_p)/300$	91.7%	3.37%	1.78%	3.16%

Table 6. VAD accuracy vs manual

The accuracies are presented as a histogram according to the methodology [15] and represent the percentage of the decisions falling into each group. The groups are formed in four regions for the difference in frames between the algorithm decision and the manual (human) segmentation:

A (excellent) error is ≤4 frames;
B (good) error is >4 frames and <10 frames;
C (average) error is ≥10 frames and ≤15frames;
D (poor) error is >15 frames.

10. Conclusions

OSWC is a new way to estimate the noise level at the price of slightly increased computations. The result is computed by simple straight-forward mathematics. As a final solution OSWC leads to an enhanced signal, containing only meaningful information with increased signal to noise ratio. The noise estimation is more precise within a short time statistics. It also provides applicable approach even if no a-priori noise information is present. Moreover, it avoids errors caused by wrong frame decision. At the same time it gives a good performance with real type noises as well as with synthesised ones. In addition it detects the speech activity.

11. References

[1] Mattila, V.-V. "Advanced Methods for Mobile Audio – Speech enhancement", *Nokia research centre, report, 2004*

[2] Sang-Mun Chi and Yung-Hwan Oh, "Lombard Effect Compensation and Noise Suppression for Noisy Lombard Speech Recognition", *ICSLP October 1996, pp. 2013-2016.*

[3] Martin R., "Spectral subtraction based on minimum statistics" *Proceedings of EUSIPCO-94, 1994, pp. 1182 1185.*

[4] Stahl, V. Fischer, A. Bippus, R. "Quantile based noise estimation for spectral subtraction and wiener filtering", *IEEE ICASSP 2000, VOL 3, pages III-1875-1878*

[5] Gustafsson, S. Jax, P. Vary, P. "A Novel Psychoacoustically Motivated Audio Enhancement Algorithm Preserving Background Characteristics", *IEEE ICASSP'98, 1998, VOL 1, pages I-397-I-400*

[6] Wolfe P., Godsill S., "Efficient Alternatives to the Ephraim and Malah Suppression Rule for Audio Signal Enhancement", *EURASIP Journal of Applied Signal Processing, 2003, NUMB 10, pages 1043-1051*

[7] Xie, F. Campernolle, D.V. "A family of 'MLP' based nonlinear spectral estimators for noise reduction", *IEEE ICASSP'94, pages 53-6, 1994.*

[8] Helge B.D. Sorensen, "A cepstral noise reduction multi-layer neural network" *ICASSP'91, pages 933-6. IEEE, 1991*

[9] Nikolov, D. "Statistical Properties of Ordered Spectra and Warping Criteria for Noise Estimation and Voice Activity Detection", *Conference Electronics 2006, Bulgaria, Sofia –accepted*

[10] Boll, S. "Suppression of acoustic noise in speech using spectral subtraction"; Acoustics, Speech, and Signal Processing, *IEEE Transactions on, Issue 2 1979 pages 113 – 120*

[11] Vaseghi S., "*Advanced Digital Signal Processing and noise reduction*", John Wiley & Sons.

[12] Goh, Z. Tan, K.-C. Tan, B. T. G., "Postprocessing Method for Suppressing Musical Noise Generated by Spectral Subtraction", *IEEE Transaction on Speech and Audio Processing, 1998, VOL 6; NUMBER 3, pages 287-292*

[13] Nikolov, D. "About speech feature extraction", *Scientific and Applied Conference Electronics ET1999, Bulgaria, Sozopol*

[14] Ephraim Y., Malah D., "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", *IEEE Transactions on Acoustics, Speech, And Signal Processing – No 6, 1984*

[15] Ivanov, R. "*Methods and algorithms for noise immunity improvements of distributed speech recognition systems*", PhD thesis, Bulgaria 2001,