INSTANT NOISE ESTIMATION USING FOURIER TRANSFORM OF AMDF AND VARIABLE START MINIMA SEARCH

Zhong Lin and Rafik Goubran

Department of Systems and Computer Engineering, Carleton University, Canada

ABSTRACT

This paper proposes a new algorithm to estimate and suppress highly non-stationary background noise from speech. The algorithm consists of two spectral detectors. The first one uses strict criteria and is based on Fourier transform of AMDF (Average Magnitude Difference Function). The second one uses loose criteria and is based on variable start minima search. By combining the two detectors, the algorithm detects and tracks the sudden change of noise energy level instantaneously. The proposed algorithm is then merged to conventional MMSE-STSA to suppress non-stationary noise in speech. Simulation results are given to show the superiority of our proposed algorithm.

1. INTRODUCTION

Real-time noise power spectrum estimation is a crucial part in many speech quality enhancement algorithms based on frequency-domain. It is even more difficult to estimate those noises with nonstationary characteristics. Existing algorithms for nonstationary noise estimation can be divided into several categories: based on Minima Tracking [1]-[3]; based on Recursive Averaging [4]-[6], [10]; or based on Wavelet Threshold [7], etc.

In real environments, some types of noise are highly nonstationary and may arise and disappear suddenly. For example, a car passes by at a speed of 80km/h, a refrigerator starts working suddenly, or a car engine is just started. Few noise estimation algorithms work well for these types of noise. Particularly, most of them have several seconds' lag to track the noise energy change.

In this paper, we propose an instant noise spectral estimation algorithm to track noises whose energy change sharply. We use two speech spectral presence detectors with strict/loose criteria respectively to detect the rising edge of noise power spectrum and therefore follow it. Simulations show that compared to most of the other methods, the proposed method significantly reduces the tracking lag.

2. ALGORITHM DESCRIPTION

Let $x(n)$ and $d(n)$ denote speech and uncorrelated additive noise signals, respectively, where n is a discrete-time index. The observed signal $y(n)$, given by $y(n) = x(n) + d(n)$, is divided into overlapping frames by the application of a window function and analyzed using the short-time Fourier transform (STFT). Specifically,

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

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 frame update step in time.

The whole algorithm consists of four parts: Speech spectral presence detector with strict criteria by Fourier transform of AMDF; Speech spectral presence detector with loose criteria by variable start minima search; Noise sudden-rising detection; Recursive averaging.

2.1. Speech spectral presence detectors

2.1.1. Strict detector based on Fourier transform of AMDF

AMDF (Average Magnitude Difference Function) is a well-known algorithm for extracting pitches from a speech [8]. It is expressed by the following equation:

$$d(n,l) = \sum_{j=0}^{N-1} \left| y(j+lM) - y(j+lM-n) \right| \quad (2)$$

where n is the time-shift index, N is the length of window, l and M are the same as (1).

We propose that by using the Fourier Transform of AMDF (FT-AMDF), the dominant spectral component can be detected steadily from noisy spectrums, even when signals are severely corrupted by non-stationary noises.

Define $D(k, l)$ as the Fourier Transform of AMDF:

$$D(k,l) \stackrel{\Delta}{=} \sum_{n=0}^{N-1} b(n)d(n,l)e^{-j(2\pi N/nk)} \quad (3)$$

where l and k are the same as (1), b is an optional windowing function.

A linear smoother is then applied along the frequency axis and a recursive averaging is applied along the frame axis,

$$D_f(k,l) = \frac{1}{2Q+1} \sum_{q=-Q}^{Q} D(k+q,l) \quad (4)$$

$$D_s(k,l) = (1-\alpha_s)D_f(k,l) + \alpha_s D_s(k,l) \quad (5)$$

where α_s is a smoothing factor, Q is the length of the smoother.

Each frame is then normalized by subtracting its median value,

$$D_n(k,l) = D_s(k,l) - \underset{u=0}{\overset{N-1}{\text{Med}}}\{D_s(u,l)\} \quad (6)$$

where $\underset{u=0}{\overset{N-1}{\text{Med}}}\{D_s(u,l)\}$ refers to the median value of $\{D_s(0, l), D_s(1, l), \ldots, D_s(N\text{-}1, l)\}$.

Thereafter, we use the following equations to identify the speech spectral area,

$$\begin{cases} I_s(k,l) = 1, & D_n(k,l) > \delta_s \\ I_s(k,l) = 0, & D_n(k,l) \le \delta_s \end{cases} \quad (7)$$

δ_s is the threshold to determine if it is a speech spectral component.

2.1.2. Loose detector based on variable start minima search

A fixed start minima search algorithm has been described by Martin [1] and Cohen [4] as speech spectral detectors. In our algorithm, we use a variable start minima search algorithm to detect speech spectrum with loose criteria.

In a given frame, speech presence within a frequency band is determined by the ratio between the local energy $S(k,l)$ of the noisy speech and its minima $S_{\min}(k,l)$ within a specified time window.

$S(k,l)$ is calculated by smoothing the magnitude squared of the spectrum of noisy speech in time and frequency, as shown in the following equation,

$$S(k,l) = \alpha_l S(k,l-1) + (1-\alpha_l)\sum_{i=-W}^{W}\left|Y(k-i,l)\right|^2 \quad (8)$$

where α_l is a smoothing factor, W is the length of the smoother.

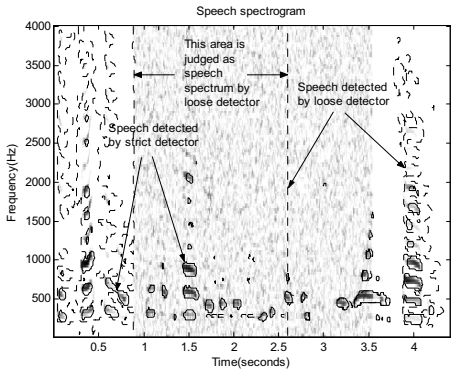
$S_{\min}(k,l)$ is calculated by the following variable start minima search algorithm:

In a specific subband k, we search minima of $S(k,l)$ within an L-sample window from nL+l_0+1 to nL+ l_0+L. l_0 indicates the start of search window, where $0 \le l_0 < L$. Initially l_0 is set to zero when the algorithm just starts from the first frame (l = 1). Then we have the following recursive equations with the increasing of l,

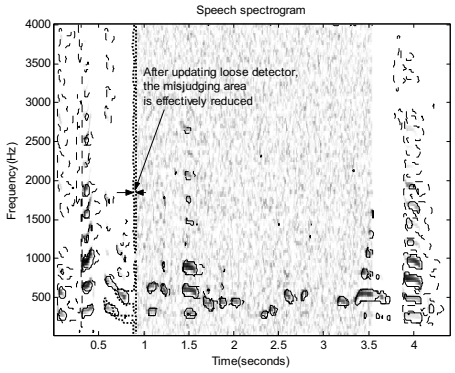
$$\begin{cases} S_{\min}(k,l) = \min\{S_{\min}(k,l-1), S(k,l)\}; \\ S_{tmp}(k,l) = \min\{S_{tmp}(k,l-1), S(k,l)\} \end{cases}$$
if l- l_0 cannot be divided by L
$$\begin{cases} S_{\min}(k,l) = \min\{S_{tmp}(k,l-1), S(k,l)\}; \\ S_{tmp}(k,l) = S(k,l) \end{cases}$$
if l- l_0 can be divided by L (9)

Let $S_r(k,l) = S(k,l) / S_{\min}(k,l)$, then the speech presence by loose criteria is calculated by:

$$\begin{cases} I_l(k,l) = 1, & S_r(k,l) > \delta_l \\ I_l(k,l) = 0, & S_r(k,l) \le \delta_l \end{cases} \quad (10)$$



a). Before using *loose detector update*



b). After using *loose detector update*

Figure 1. Speech spectral detection (solid contour is from the strict detector, dash contour is from the loose detector)

δ_l is the threshold to determine if it is a speech spectral component by loose criteria.

To the loose detector, if l_0 is fixed, usually the detector is apt to misjudge the rising noise spectrum as speech, as the areas within the dash contours illustrated in Figure 1a). In the figure, a piece of speech with 8kHz sampling rate is used. The signal is noise-free from 0s to 0.88s and from 3.52s to 4.4s. From 0.88s to 3.52s, it is corrupted by Gaussian white noise with -10dB SNR. Solid contour is the speech spectrum detected by the strict detector. Dash contour is the speech spectrum detected by the loose detector. From 0.88s to 2.6s, the noise spectrum is misjudged as speech spectrum by the loose detector. This is because in the loose detector, a speech is divided into many 1~2 seconds' long segments before processing. When the noise energy rises sharply within the pre-set time slot, the loose detector is not able to distinguish this from a speech rise.

On the other hand, the figure also shows that the strict detector can still steadily find speech components in this situation. This is because FT-AMDF is an effective harmonic spectral structure enhancer. It is particularly suitable to be used to detect voiced speech components from noisy spectrums. At the same time, unvoiced speech with a strong and broad spectrum will also be enhanced and detected by FT-AMDF. The drawback of the strict detector is it may lose some weak voiced speech when the energy of the speech is too low.

2.2 Noise energy rise detection and loose detector update

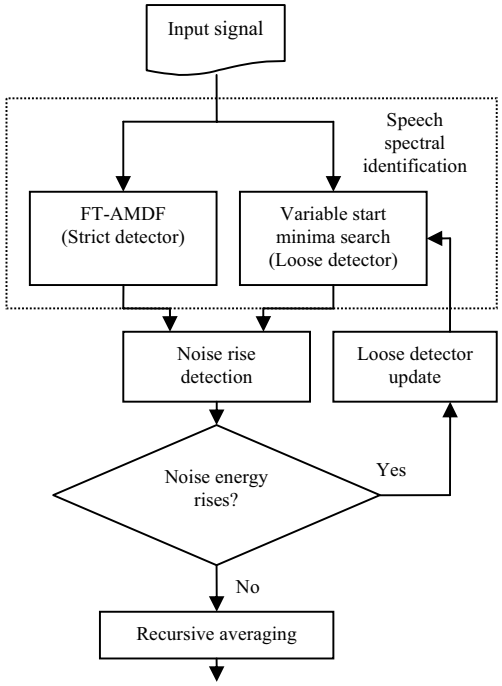


Figure 2. Flowchart of the proposed algorithm

We have the following rule to judge if there is a sudden-rise of noise energy: if at a specific frame, the algorithm does not find dominant speech components by the strict detector while the number of speech components found by the loose detector is more than a specified threshold, we conclude that the loose detector makes a wrong judgment. This is expressed by the following equation:

$$\begin{cases} N_r(l)=1, if\left[(\sum_k I_s(k,l)<\gamma_s) and (\sum_k I_l(k,l)>\gamma_l)\right] \\ N_r(l)=0, \qquad else \end{cases} \tag{11}$$

where $N_r(l)$=1 indicates that the noise energy changing sharply in frame l.

When the algorithm detects there is a sudden rise of noise energy, we recalculate the starting position of search window and the local energy $S(k,l)$:

$$\begin{cases} S(k,l)=\dfrac{1}{2P+1}\sum_{p=-P}^{P}\left|Y(k-i,l)\right|^2; \qquad if[Nr(l)]=1 \\ \quad and \quad l_0=l(\bmod L) \\ S(k,l) \quad Eq.(8); l_0 \quad no \quad change; \quad if[Nr(l)]=0 \end{cases} \tag{12}$$

We call equations (12) as *loose detector update*. Its effect is to recalculate signal energy S (k, l) and start a new round of minima search. Therefore the loose detector can update its parameters to keep up with the change of energy.

The outcome after using *loose detector update* is illustrated by the contours in Figure 1 b).

Obviously, by using *loose detector update*, the misjudging area in b) is significantly reduced, compared to the one in a).

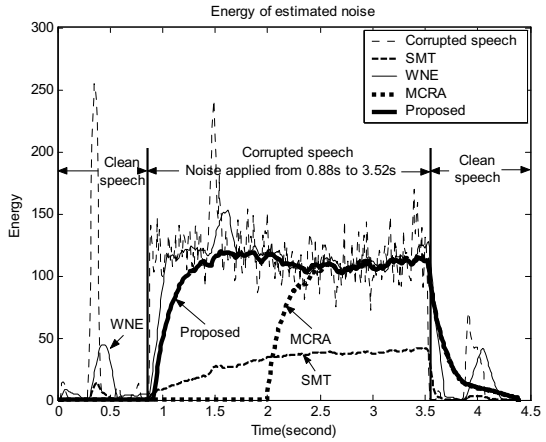


Figure 3. Noise energy estimated by various algorithms

2.3 Recursive averaging

The final decision of speech spectral presence is made by the combination of $I_l(k,l)$ and $I_s(k,l)$. Let $I(k,l)$ denote if the spectrum component at (k, l) is a speech spectrum. $I(k,l)$ is evaluated by the logical OR of $I_l(k,l)$ and $I_s(k,l)$:

$$I(k,l)=I_l(k,l)\,|\,I_s(k,l) \tag{13}$$

Finally the noise spectrum is estimated by the following equation:

$$\begin{cases} \hat{\lambda}_d(k,l+1)=\alpha_d\hat{\lambda}_d(k,l)+[1-\alpha_d]\left|Y(k,l)\right|^2; & if \quad I(k,l)=0 \\ \hat{\lambda}_d(k,l+1)=\hat{\lambda}_d(k,l); & if \quad I(k,l)=1 \end{cases} \tag{14}$$

where α_d ($0<\alpha_d<1$) is a smoothing parameter.

The flowchart is illustrated in Figure 2.

3. SIMULATION RESULT

3.1 Spectrum observation

The speech used in Figure 1 is used again to verify our proposed scheme. The following parameters are used in the proposed algorithm: N = 256; M = 64 (75% overlap); b = h = Hanning window; W = Q = 1; P = 2; α_s = 0.6; α_d = 0.95; α_l = 0.8; L = 125; δ_l = 8; δ_s = 5; γ_s = 3; γ_l = 150.

Three algorithms, including Spectral Minima Tracking (SMT) [2], Weighted Noise Estimation (WNE) [10], and Minima Controlled Recursive Averaging (MCRA) [4], are selected to compare with the proposed algorithm.

The energy of the estimated noise is illustrated in Figure 3. It shows that the estimation curve by SMT rises slowly and needs 2~3 seconds to get a stable output. Occasionally, it overestimates noise when there are speech spectrums. WNE follows the expectation of noise promptly; however it obviously overestimates noise when there is speech. MCRA works very well when it gets steady, but it reacts 1~2 seconds slower than the noise rises. Comparing with them, our method reacts to noise

rise within 100ms and tracks noise more quickly than SMT and MCRA, and more accurately than WNE.
The graph proves that our proposed method can instantly react and track sudden changes of noise quicker than other methods. Also it minimizes the possibility of mistakenly following speech spectral components.

3.2 Performance in speech enhancement

To evaluate the performance of the proposed algorithm in application of speech quality enhancement, we use it in MMSE-STSA (Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator), a conventional speech quality enhancement algorithm. The detail of MMSE-STSA is described in Ephraim's work [9]. For comparison, MCRA is used as the competitive algorithm. Both the output of the proposed algorithm and MCRA are merged into MMSE-STSA as its noise estimation part. Then we evaluate the quality of their enhanced speech.
Thirty pieces of speech, including fifteen males and fifteen females, are randomly chosen from the TIMIT database. Two pieces of noise recorded from real environments are selected. The first one is recorded from where a car with a high speed passes by a microphone. The other one is recorded close to where a man starts his car's engine. The car-pass-by noise is added in the middle of each speech. The engine-start noise is added into speech in three ways: Engine starts 1 second earlier than speech starts; engine and speech start at the same time; engine starts 1 second later than speech starts.
We use PESQ (Perceptual Evaluation of Speech Quality) [11], the current ITU objective speech quality evaluation standard to evaluate the performance of the proposed speech quality enhancement scheme. The PESQ improvement values are averaged over all of the thirty pieces of speech.
Table 1 shows the simulation results. From the table, we can find our method performs steadily better than MCRA.
Subjective listening tests confirm the result of objective evaluations. Especially at the time when the noise just rises, the output of MCRA+MMSE-STSA always has 1~2 second's disturbance of noise, while the proposed scheme reduces this disturbance period to less than 100ms.

Table 1. Average PESQ improvement of proposed scheme and MCRA+MMSE-STSA scheme

SNR (dB) \ Noise		Cars pass by	Car engine and speech start		
			Speech is 1s later	At the same time	Speech is 1s earlier
10	MCRA	0.312	0.355	0.303	0.133
	Proposed	0.346	0.374	0.317	0.137
5	MCRA	0.340	0.388	0.315	0.136
	Proposed	0.419	0.461	0.379	0.176
0	MCRA	0.221	0.290	0.192	0.017
	Proposed	0.367	0.428	0.304	0.095
-5	MCRA	0.061	0.158	0.050	-0.060
	Proposed	0.204	0.245	0.101	0.018

4. CONCLUSIONS

We propose an instant noise spectral estimation algorithm to estimate noise power spectrum with highly nonstationary energy in speech. An FT-AMDF method is proposed to detect dominant speech components in noisy spectrums with strict criteria. A variable start minima search method is introduced to detect speech components with loose criteria. By comparing the outputs of the two detectors, rise edges of noise energy are identified. We show that the proposed method significantly reduces the tracking lag of noise change, compared to other methods. And we show that the new algorithm is superior to other competitive methods in the performance of suppressing highly nonstationary noise.

5. ACKNOWLEDGEMENT

The author would like to acknowledge the funding from Canada Customs and Revenue Agency, OGS, CITO, and NSERC.

6. REFERENCES

[1] Rainer Martin, "Spectral subtraction based on minimum statistics", *EUSIPCO-94*, Sept. 1994, pp. 1182-1185.
[2] Rainer Martin, "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.
[3] G. Doblinger, "Computationally efficient speech enhancement by spectral minima tracking in subbands", in *Proc. EUROSPEECH'95*, vol. 2, 1995, pp. 1513-1516.
[4] I. Cohen, B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement", *IEEE Signal Processing Letters*, pp. 12-15, Jan. 2002.
[5] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging", *IEEE Trans. on Speech and Audio Processing*, pp. 466- 475, vol. 11, Sept. 2003.
[6] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments" in *Proc. IEEE ICASSP'99*, Mar. 1999, pp. 789-792.
[7] Sungwook Chang, Y. Kwon, Sung-il Yang, I-jae Kim, "Speech enhancement for non-stationary noise environment by adaptive wavelet packet", in *Proc. IEEE ICASSP'02*, May 2002, pp. 561-564.
[8] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, , H. Manley, "Average magnitude difference function pitch extractor", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 22, issue: 5, pp. 353-362, Oct. 1974.
[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator", *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109-1121, Dec. 1984.
[10] M. Kato, S. Akihiko, S. Masahiro, "Noise suppression with high speech quality based on weighted noise estimation and MMSE STSA", in *Proc. IWAENC2001*, Sept. 2001, pp. 183-186
[11] ITU-T P.862, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, 2001.