Impulsive Noise Detection and Location in Speech

Liang Chang, Kun Tang, and Huijuan Cui
National Laboratory for Information Science and Technology, Department of Electronics Engineering, Tsinghua University, Beijing, China
Email: {changliang, tangkun, cuihuijuan}@mail.tsinghua.edu.cn

***Abstract*—The most important part in speech applications is parameter extraction, but the existence of impulsive noise can influence the parameter extraction dramatically. To minimize the influence caused by the impulsive noise, this paper proposed a new impulsive noise detection and location method that incorporates the advantages of both model-based and statistics-based methods. First the Projection Statistics (PS) is calculated to detect and locate the impulsive noise as outliers from statistics view, and then the residual signals of speech is calculated and compared with a pre-defined threshold to refine the raw result. Experiments show that the proposed method is obviously better than the model-based and statistics-based methods in detection accuracy and location accuracy, especially in the false alarm rate.**

***Index Terms*—speech applications, impulsive noise, Projection Statistics, residual signals**

I. Introduction

Impulsive noise is very harmful in speech related applications. It usually has huge power, and can twist the spectrum of the speech tremendously. Therefore it can influence the speech parameter extraction greatly and further influence the speech related applications.

For Impulsive noise detection and location, the simplest way is the median filtering [1] and its various improvement versions [2]-[4]. But median filtering has two obvious drawback: first the procedure is the same to all kinds of signals, which means no characteristic of the signals is exploited; second the filtering window width must be larger than double of that of the impulsive noise, otherwise the effect is contrary. But the width of impulsive noise is not always very short, thus the width of filtering widow must follow, and this will limit the filtering effect. The classical method for detecting impulsive noise in speech is Linear Prediction (LP) model based method [5]. First the signal is filtered by Linear Prediction Coefficients (LPC) to obtain residuals. LP model is the most successful model for speech modeling, but not for noise modeling. Therefore the residual of speech is very small and like white noise while the residuals of speech that contains impulsive noise remain high. Therefore it is very easy to detect impulsive noise in residual signals since the impulsive noise is much more obvious in residual signals than in original signals. The drawback of the model-based method is that it can only detect but not locate. To be specific, it can only locate the very center of impulsive noise which is usually 1 or 2 points. Because of this, the double-threshold method [6] is proposed to solve this problem, one threshold for detection and the other for location. Model-based methods also include Warped Linear Prediction (WLP) method [7]. WLP can model the spectrum envelope better than LP, but is more complicated. Anti-noise LPC estimation method which is not limited to impulsive noise proposed sample-selective [8] and combined sample-selective method [9] to detect noise, and their principle is similar to the model-based methods. To summary, the location problem is the main problem of the model-based methods.

2005 Gandhi and Mili use statistic method to detect and locate impulsive noise [10]. They incorporated the Projection Statistics (PS) [11], and use the way of finding outliers to detect and locate impulsive noise in speech. The higher of PS value is, the more likely the point is an outlier. The PS method has high location accuracy because PS is calculated from original speech but not residuals. However PS method has another problem, which is relatively high false alarm rate.

By analyzing the defects of model-based and PS-based methods, this paper proposed a new impulsive noise detection and location method, which is PS-based but incorporates the advantage of model-based method. Experiments show that the proposed method has very low false alarm rate and at the same time has high location accuracy.

The paper is organized as following: Part II introduces model-based method, Part III introduces PS method and Part IV introduces the proposed method. Experiments are in Part V, and conclusions are made in Part VI.

II. Model-Based Method

The impulsive noisy speech can be modeled by the following formula:

$$y(n) = x(n) + I(n), n = 1, ..., N \quad (1)$$

where $y(n)$ is noisy speech, $x(n)$ is clean speech, $I(n)$ is impulsive noise and N is the number of the points in one frame.

Manuscript received July 24, 2015; revised May 3, 2016.

doi: 10.18178/ijsps.4.6.446-450

To detect $I(n)$ from $y(n)$, we have the Signal to Noise Ratio (SNR) $s1$ is:

$$s1 = \frac{\mathrm{E}(I(n)^2)}{\mathrm{E}(x(n)^2)} \tag{2}$$

where E() is expectation. Notice that $I(n)$ is the one to be detected, so in (2) $I(n)$ has been put on the position of signal.

According to LP model, $x(n)$ can be expressed as:

$$x(n) = \sum_{i=1}^{k} a_i x(n-i) + e(n) \tag{3}$$

where a_i are LPC, k is prediction order, $e(n)$ is residual signal of speech.

So (1) can be rewritten as:

$$y(n) = \sum_{i=1}^{k} a_i x(n-i) + e(n) + I(n) \tag{4}$$

Use the estimate of LPC $\hat{a}_i$ to filter $y(n)$, the residual of noisy speech $r(n)$ is:

$$r(n) = \sum_{i=1}^{k} (a_i - \hat{a}_i) x(n-i) + e(n) + I(n) - \sum_{i=1}^{k} \hat{a}_i I(n-i) \tag{5}$$

The difference of a_i and $\hat{a}_i$ can be ignored comparing to other terms, $r(n)$ can be rewritten as:

$$r(n) = e(n) + I(n) - \sum_{i=1}^{k} \hat{a}_i I(n-i) \tag{6}$$

In (6) $e(n)$ is the residual of clean speech, which is small and like white noise, so $I(n)$ is the dominant term. Detect $I(n)$ from $r(n)$ the SNR $s2$ is:

$$s2 = \frac{\mathrm{E}(I(n)^2)}{\mathrm{E}(e(n)^2)} \tag{7}$$

Compare detecting $I(n)$ from $r(n)$ and detecting $I(n)$ from $y(n)$ the improvement in SNR is $\mathrm{E}(x(n)^2)/\mathrm{E}(e(n)^2)$. The improvement is substantial because the amplitude of $x(n)$ is usually much higher than that of $e(n)$.

The threshold for detecting $I(n)$ from $r(n)$ is defined as $K \bullet \sigma_e$ [3], σ_e is a robust estimate of residual signal power. The robust estimate is obtained from the median of residual signals. K is an adjust factor which reflect a tradeoff between the hit rate and the false alarm rate. K remains unchanged once determined while σ_e is calculated for every speech frame.

Detecting impulsive noise from residual signal can improve the accuracy a lot, but at the same time it brings one fatal drawback. The method cannot locate the whole impulsive noise but only a few points in the center, even one point sometimes. The reason is that the LPC inverse filtering can blur the border of impulsive noise and makes the impulsive noise thinner as Fig. 1 shows. Thus the detection can only locate the center of impulsive noise, but not the real range. And Even the double-threshold method [6] does not work under such conditions.

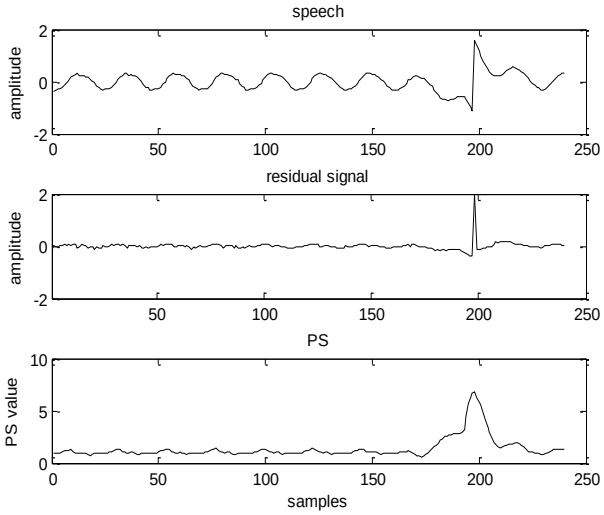

Figure 1. The residual signal and PS of impulsive noise

III. PS-Based Method

Put the noisy speech $y(n)$ into the following matrix form:

$$H = \begin{bmatrix} h_1^T \\ \cdots \\ h_n^T \\ \cdots \\ h_N^T \end{bmatrix} = \begin{bmatrix} y(0),\ldots,y(1-k) \\ \cdots \\ y(n-1),\ldots,y(n-k) \\ \cdots \\ y(N-1),\ldots,y(N-k) \end{bmatrix}$$

where $h_n = [y(n-1),\ldots,y(n-k)]^T$, k is the dimension of h_n.

The Project Statistic [11] is calculated as following:

$$PS_n = \max_{\|v\|=1} \left\{ \frac{\left| h_n^T v_l - \mathrm{med}_{j=1}^{N}(h_j^T v_l) \right|}{s \bullet \mathrm{med}_{i=1}^{N} \left| h_i^T v_l - \mathrm{med}_{j=1}^{N}(h_j^T v_l) \right|}, l = 1,\ldots,N \right\} \tag{8}$$

where $v_l = h_l - \mathrm{med}_{j=1}^{N}(h_j)$ is the direction vector of h_l to the center of points, $\mathrm{med}_{j=1}^{N}()$ is median operation, s is a constant.

For every point h_n, project it to N directions from v_1 to v_N. For one direction v_l calculate the distance of the projection of h_n to the center of all projections in the direction projected by all points. The distance is normalized using the median absolute deviation of all projections in the direction v_l. Finally find the maximum of all distances of h_n in all directions as the distance measure of h_n and the center of all points, and determine whether h_n is an outlier in the sample cluster. And the maximum of all distances of h_n is the PS of h_n.

If $y(n)$ is contaminated by impulsive noise, $y(n)$ must be an outlier in the signal sequences, and furthermore if some points from $y(n-k)$ to $y(n-1)$ are contaminated by impulsive noise, h_n must be an outlier in its own space. Thus if h_n is detected as an outlier, it means the points from $y(n-k)$ to $y(n-1)$ contain impulsive noise. The dimension k will not be too big, so the border will not be too wide.

PS is a statistics evolved from Mahalanobis Distance (MD) [12]. In 1982 Stahel and Donoho proved that MD can be calculated from the following formula [13]:

$$MD_n = \max_{\|v\|=1} \left\{ \frac{\left| h_n^T v_l - \frac{1}{N}\sum_{j=1}^{N}(h_j^T v_l) \right|}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(h_i^T v_l - \frac{1}{N}\sum_{j=1}^{N}(h_j^T v_l))^2}} , l = 1,...,N \right\} \quad (9)$$

where $v_l = h_l - \frac{1}{N}\sum_{j=1}^{N}(h_j)$.

PS use sample median to replace sample mean and median absolute deviation to replace deviation, which makes PS a very robust statistics comparing to MD. Because as long as the number of outliers doesn't exceed 50% of the total samples, the sample median and median absolute deviation will not be influenced.

Under the condition of h_n follow normal distribution, and number of samples is more than five times of sample dimension, which is easy to fulfill for speech, the squares of PS follow roughly chi-square distribution with degree of k [13]. So if the outlier detection threshold is 97.5%, the detection condition is:

$$PS_n > \sqrt{F^{-1}_{X^2,k}(0.975)} \quad (10)$$

where $F^{-1}_{X^2,k}()$ is the inverse of cumulative distribution function of chi-square distribution with degree of k.

Since PS is calculated by original speech signal, the PS method can locate the border of impulsive noise accurately. As Fig. 1 shows, the PS of impulsive noise is obviously wider than the residuals of impulsive noise.

IV. Improved PS Method

Since it is known that the PS method is much better than the model-based method in location accuracy, the improved method the paper proposed is based on PS method.

Experiments show that the PS method has a disadvantage of high false alarm. By observing the speech signals that are detected as impulsive noise by PS, it is found that the error usually occurs in the transition periods of voiced speech and unvoiced speech. The amplitude of voiced speech is usually higher than that of unvoiced speech, and if the amplitude difference is high enough, the PS method will misjudge the voiced as impulsive noise, as shown in Fig. 2.

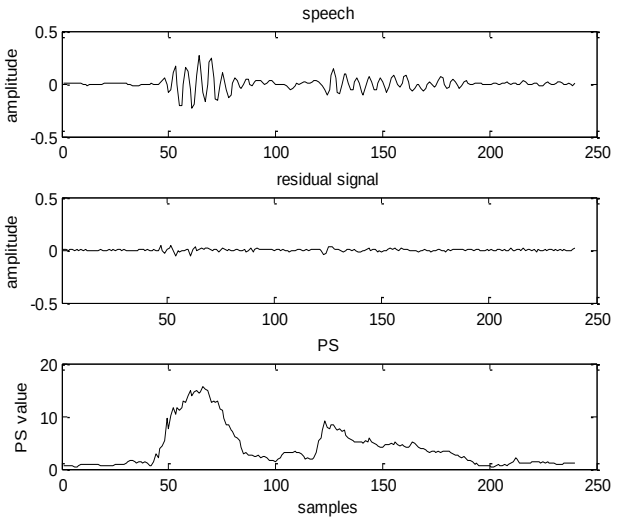

Figure 2. The residual signal and PS of speech transitions

But voiced speech and impulsive noise are different in nature. The most essential difference is that speech can be perfectly modeled by LP model, while impulsive noise cannot. The residual of speech is white-noise-like signals with low amplitude after LPC inverse filtering, while the residual of speech that contains impulsive noise, especially the center, can hardly be influenced by LPC inverse filtering, as shown in Fig. 1. In addition experiments results show that even the LPC is extracted from the noisy speech, the residual of speech is still white-noise-like. Based on these features, this paper proposed the improved PS method as following:

- Calculate the PS of the original noisy speech;
- Use PS method to detect impulsive noise from the noisy speech and locate the raw impulsive noise sections;
- Use traditional LPC extraction method to extract LPC from the noisy speech;
- Use the LPC to inverse filter the noisy speech and obtain the residuals;
- Compare the residuals with a pre-defined residual threshold;
- In the raw impulsive noise sections, if there are residual signals that are larger than the threshold, the section is confirmed as impulsive noise section, and if not, the section is determined as false alarm noise section and is removed.

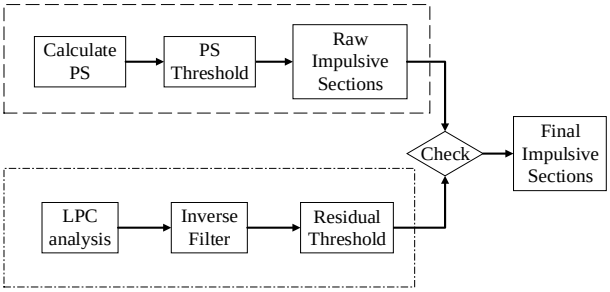

Figure 3. The flowchart of proposed method

The proposed method is based on PS method because of its high location accuracy. The residual signals are

used to refine the raw results because it is much easier to detect impulsive noise from them than the original signals. It takes the advantage of substantial improvement of SNR from model-based method and the advantage of location accuracy from PS method. As Fig. 3 shows, the improved PS method can be seen as "double-detection". The part in dotted line is PS method, and the part in dash dotted line is model-based method.

There are two ways to determine the residual threshold. In [5], $K \bullet \sigma_e$ is used. In this paper, we proposed another way to determine the threshold, which is $K \bullet y_{\max}$, where $y_{\max}$ is the maximum of the given speech frame $[y(1),...,y(n),...,y(N)]$. If speech is clean, the residual of each speech sample is $e(n)$, $y_{\max} = \max(x(n))$, the ratio of residual to $y_{\max}$ is $e(n)/\max(x(n))$, which is close to 0 since $x(n) \gg e(n)$. But if the speech frame is polluted by noise, the dominant term of residual signal is $I(n)$, $y_{\max} = \max(x(n) + I(n))$, the ratio of residual to $y_{\max}$ for clean speech sample is $e(n)/\max(x(n) + I(n))$ and for noisy speech sample is close to $I(n)/\max(x(n) + I(n))$. Due to the fact that $I(n) > x(n) \gg e(n)$, the former is close to 0, while the latter is close to 1. In one word, for speech sample, the ratio of residual to $y_{\max}$ is close to 0 while for noisy speech sample, this value is close to 1. Therefore $K \bullet y_{\max}$ method can detect impulsive noise in residual signal much easier. K is the factor that compromises between hit rate and false alarm rate.

V. Experiment Results

A. Test Sequences

The paper use the impulsive noise in [10] as the noise template, add it into five clean speech and generate five impulsive noisy speech, which are Back, Mix1, Mix2, Back_4, Mix1_4. Back is the test sentence in [10]. Back_4 is such a version of Back that the amplitude of clean speech is amplified four times, and the same operation is applied to Mix1_4. Mix1 is women's speech and Back and Mix2 are men's speech.

The five sentences are equal length with 8 KHz sampling. There are total 38 frames, and each frame contains 240 samples, and among them 12 frames contain impulsive noise.

B. Accuracy of Different Detection Methods

Since the location problem in model-based method is very serious, only PS method and the proposed improved PS method are compared.

First PS method is tested. The outlier detection threshold is set at 97.5%, and this parameter do not change. The dimension of h_n can influence the performance of PS method and it reflects the tradeoff between false alarm rate and miss rate. The results of PS method with different dimensions are shown in Table I.

It can be seen that for different dimensions from 3 to 10, the false alarm rates are all very high. Even for dimension 10, the smallest in false alarm rate, its false alarm rate is unbearable. These false alarms often happen in the transitions between voiced and unvoiced. If the amplitude of voiced is much higher than the unvoiced, the voiced will be detected as impulsive noise by PS method.

TABLE I. The Accuracy of PS Method

Dimension	Average number of misjudged frames	
	False alarm	Miss
3	17.6	0
4	15.2	0.2
5	13.8	0.6
6	12.2	0.8
7	10.6	1
8	9.8	1.4
9	9.8	1.4
10	8.6	1.6

Due to the huge power of impulsive noise and the bad consequence it will bring to speech, we tend to decrease the miss rate as small as possible. So the improved PS method set the dimension of h_n as 3. Using the same outlier detection threshold 97.5%, the improved PS method that uses $K \bullet y_{\max}$ residual determination method is tested with different K values. K values also influence the tradeoff between false alarm rate and miss rate. The results are shown in Table II.

TABLE II. The Accuracy of Improved PS Method ($K \bullet y_{\max}$)

K	Average number of misjudged frames	
	False alarm	Miss
0.4	8	0
0.5	6.2	0
0.6	3	0
0.7	2.4	0.2
0.8	1.2	0.2
0.9	0.8	0.2

It can be seen that the improved PS method decreased the false alarm rate greatly, especially when K is close to 1. And the miss rates are all very small too. Obviously the proposed method is far better than PS method.

Using the same outlier detection threshold and the same dimension, the improved PS method that uses $K \bullet \sigma_e$ residual determination method is then tested with different K values. As in $K \bullet y_{\max}$ method, K values in $K \bullet \sigma_e$ method also influence the tradeoff between false alarm rate and miss rate. The results are shown in Table III.

TABLE III. The Accuracy of Improved PS Method ($K \bullet \sigma_e$)

K	Average number of misjudged frames	
	False alarm	Miss
5	13.4	0
10	4.2	0
15	1.2	0.4
20	0.8	0.6
25	0.8	1.6
30	0.8	2.2
35	0.6	3.6

It can be seen that the improved PS method with $K \bullet \sigma_e$ threshold decreased the false alarm rate greatly too. Overall, the improved PS method, no matter which way the residual threshold is determined, is far better than the PS method.

Besides, the $K \bullet y_{\max}$ residual determination method is better than $K \bullet \sigma_e$ method for the improved PS method. Because for the same miss rate, the false alarm rate of $K \bullet \sigma_e$ method is higher than that of $K \bullet y_{\max}$ method and for the same false alarm rate, the miss rate of $K \bullet \sigma_e$ is also higher than that of $K \bullet y_{\max}$ method.

VI. Conclusions

The paper first analyzed the harm of impulsive noise to speech related applications. The power of impulsive noise is huge and it can twist the spectrum of speech tremendously, which will influence the parameter extraction greatly and further influence speech related applications. And second the paper analyzed the model-based and PS-based impulsive noise detection methods, and pointed out the defects of both methods. Based on this analysis, this paper proposed a new impulsive noise detection method that takes advantage of both model-based and PS-based methods. The PS method is used to calculate the raw detection results and the residuals are used to refine the results. Besides a new threshold determination method is proposed in this paper. Experiments show that the proposed method is obviously better than model-based method and PS method in location accuracy and detection accuracy, especially in the false alarm rate.

References

[1] S. V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction*, England: John Wiley and Sons, 1996.

[2] C. G. Zhao, L. Zhang, and F. B. Wu, "Application of improved median filtering algorithm to image de-noising," *Journal of Applied Optics*, vol. 32, pp. 678-682, 2011.

[3] X. G. Kang, M. C. Stamn, A. J. Peng, and K. J. R. Liu, "Robust median filtering forensics using an autoregressive model," *IEEE Transactions on Information Forensics and Security*, vol. 8, pp. 1456-1468, 2013.

[4] J. S. Chen, X. G. Kang, Y. Liu, and Z. J. Wang, "Median filtering forensics based on convolutional neural networks," *IEEE Signal Processing Letters*, vol. 22, pp. 1849-1853, 2015.

[5] S. V. Vaseghi and P. J. W. Rayner, "Detection and suppression of impulsive noise in speech communication systems," *Communications, Speech and Vision, IEE Proceedings*, vol. 137, pp. 38-46, 1990.

[6] P. A. A. Esquef, L. W. P. Biscainho, P. S. R. Diniz, and F. P. Freelanci, "A double-threshold-based approach to impulsive noise detection in audio signals," in *Proc. 10th European Signal Processing Conference*, 2000, pp. 1-4.

[7] P. A. A. Esquef, M. Karjalainen, and V. Valimaki, "Detection of clicks in audio signals using warped linear prediction," in *Proc. 14th International Conference on Digital Signal Processing*, 2002, pp. 1085-1088.

[8] M. Markovic, "The application of sample-selective LPC method in standard CELP 4800 bit/s speech coder," in *Proc. Third IEEE International Conference on Electronics, Circuits and Systems*, Rodos, Greece, 1996, pp. 506-509.

[9] M. Markovic, M. Milosavljevic, B. Kovacevic, and M. Veinovic, "Robust LPC parameter estimation in standard CELP 4800 bit/s speech coder," *Vision, Image and Signal Processing, IEE Proceedings*, vol. 145, pp. 19-22, 1998.

[10] M. A. Gandhi, C. Ledoux, and L. Mili, "Robust estimation methods for impulsive noise suppression in speech," in *Proc. IEEE International Symposium on Signal Processing and Information Technology*, Athens, Greece, 2005, pp. 755-760.

[11] M. Gasko and D. Donoho, "Influential observation in data analysis," *Proc. of the Business and Economic Statistics Section, American Stat. Assn.*, pp. 104-110, 1982.

[12] D. M. Roy, J. R. Delphine, and D. L. Massart, "The Mahalanobis distance," *Chemometrics and Intelligent Laboratory Systems*, vol. 50, pp. 1-18, 2000.

[13] L. Mili, M. G. Cheniae, N. S. Vichare, and P. J. Rousseeuw, "Robust state estimation based on projection statistics [of power systems]," *IEEE Transactions on Power Systems*, vol. 11, pp. 1118-1127, 1996.

Liang Chang was born in Hebei Province, China in 1985. He received his Bachelor's degree in Electronic Engineering from Harbin Institute of Technology in the year 2008. He received his Doctor's degree in Electronic Engineering from Tsinghua University in the year 2013. Now he is pursuing a postdoctoral degree in Tsinghua University. His research area includes speech coding and robust speech parameter extraction.

Kun Tang received his Bachelor's degree in Electronic Engineering from Tsinghua University in the year 1970. Then he worked at Tsinghua University until now. Now he is a professor in Department of Electronic Engineering in Tsinghua University. His research area includes signal processing, source coding and multimedia technology.

Huijuan Cui received her Bachelor's degree in Electronic Engineering from Tsinghua University in the year 1970. Then she worked at Tsinghua University until now. Now she is a professor in Department of Electronic Engineering in Tsinghua University. Her research area includes digital communication, source coding and multimedia technology.