applied sciences

MDPI

Article

Speckle Noise Detection and Removal for Laser Speech Measurement Systems

Yahui Wang [1,2], Wenxi Zhang [2], Zhou Wu [2], Xinxin Kong [2] and Hongxin Zhang [1,*]

[1] School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing 100876, China; wangyahui@aoe.ac.cn

[2] Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100081, China; zhangwenxi@aoe.ac.cn (W.Z.); wuzhou@aoe.ac.cn (Z.W.); xxkong@aoe.ac.cn (X.K.)

* Correspondence: hongxinzhang@bupt.edu.cn

Abstract: Laser speech measurement is a new sound capture technology based on Laser Doppler Vibrometry (LDV). It avoids the need for contact, is easily concealed and is ideal for remote speech acquisition, which has led to its wide-scale adoption for military and security applications. However, lasers are easily affected by complex detection environments. Thus, speckle noise often appears in the measured speech, seriously affecting its quality and intelligibility. This paper examines all of the characteristics of impulsive noise in laser measured speech and proposes a novel automatic impulsive noise detection and removal method. This method first foregrounds noise using decorrelation based on a linear prediction (LP) model that improves the noise-to-signal ratio (NSR) of the measured signal. This makes it possible to detect the position of noise through a combination of the average short-time energy and kurtosis. The method not only precisely locates small clicks (with a duration of just a few samples), but also finds the location of longer bursts and scratches (with a duration of up to a hundred samples). The located samples can then be replaced by more appropriate samples whose coding is based on the LP model. This strategy avoids unnecessary processing and obviates the need to compromise the quality of the relatively large fraction of samples that are unaffected by speckle noise. Experimental results show that the proposed automatic speckle noise detection and removal method outperforms other related methods across a wide range of degraded audio signals.

Keywords: laser speech measurement; speech enhancement; automatic speckle noise detection; average short-time energy; kurtosis

Citation: Wang, Y.; Zhang, W.; Wu, Z.; Kong, X.; Zhang, H. Speckle Noise Detection and Removal for Laser Speech Measurement Systems. *Appl. Sci.* **2021**, *11*, 9870. https://doi.org/10.3390/app11219870

Academic Editor: Doru Florin Chiper

Received: 2 October 2021
Accepted: 19 October 2021
Published: 22 October 2021

1. Introduction

Laser interferometry-based Laser Doppler Vibrometry (LDV) is widely used in the field of precision measurement. In addition to the vibration measurement, LDVs are also used in non-contact speech acquisition. Vibro-acoustic sensors based on LDV are a new type of voice acquisition equipment that are widely used in the context of anti−terrorism operations, national security, and other related fields.

An LDV-based laser speech measurement system uses phase Doppler measuring techniques to obtain a speech signal by measuring the phase change of the optical signal caused by sound vibration. It is usually composed of a laser transmitter, laser receiver, amplifier and equalizer, together with some other important components. As shown in Figure 1, the laser transmitter emits an invisible narrow-band laser, which is divided into a reference beam and a measurement beam through a polarized beam splitter (PBS1). The measurement beam then passes through a beam splitter (BS3), focusing lens (L), and quarter-wave plate (P), and is focused on the vibrating object. The reflected beam is directed via acousto-optical modulators (AOM), and is then merged and interfered with the reference beam by BS2. The laser receiver (photo-detector, D) receives the reference

signal and the coherent signal. After demodulation, filtering and some signal processing steps are performed to obtain the voice signal [1].

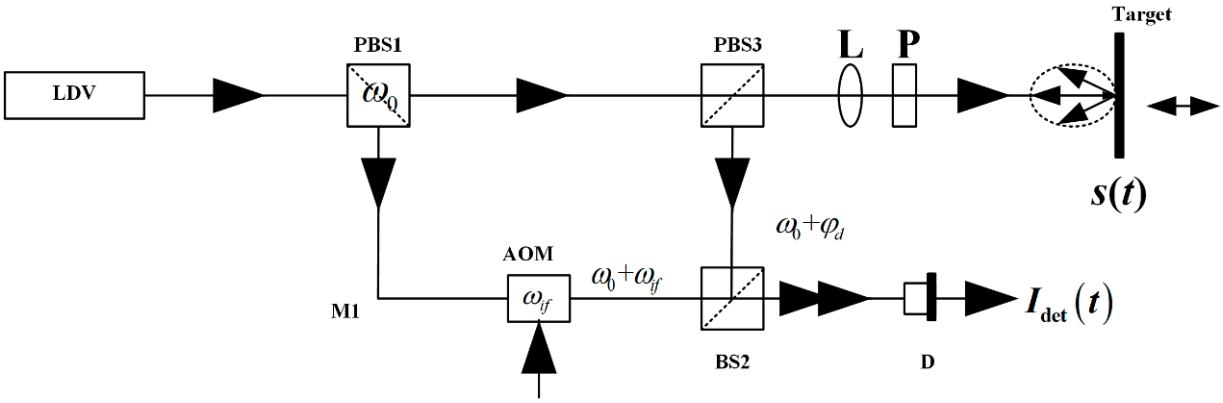

Figure 1. Internal structure of a laser speech measurement system.

As laser speech detection systems use lasers for voice measurement, they can detect speech in non-contact situations, undertake long−range measurements and are easy to conceal and operate [2].

Laser speech detection systems are usually set up in hidden places to detect conversations in conference rooms, cars, etc. Figure 2 shows a scene where a laser speech detection system is actually being used, with the objects near the speaker (such as the computer screens, tissue box, mineral water bottle, clothes, etc.) being used as sound sensors. As the sound signal is captured through the indirect measurement of vibrations, the selection of the objects for detection has a significant impact on the speech acquisition, as does the external environment. First of all, the surface of most detected objects is extremely rough in the laser band and scattered light is therefore emitted from numerous coherent points. These scattered sources propagate in different directions and can interfere with each other in space, resulting in a random distribution of light interference and generating what is known as speckle noise (in speech, it shows clicks and small burrs as shown in the black box in Figure 3). In addition, when the measured light is affected by interference (such as people walking, environmental occlusion, violent shaking, atmospheric turbulence, etc.), it not only makes it difficult for the measurement light to focus on the surface of the detection beacon, but also causes wavefront phase distortion in the reflected light, resulting in destructive interference. This makes the Doppler phase obtained by the detection system discontinuous between $-\pi$ and π [3]. These discontinuities lead to further kinds of speckle noise in the speech signal (such as bursts, outliers, crackles, scrapes, etc.), as shown in the red box in Figure 3.

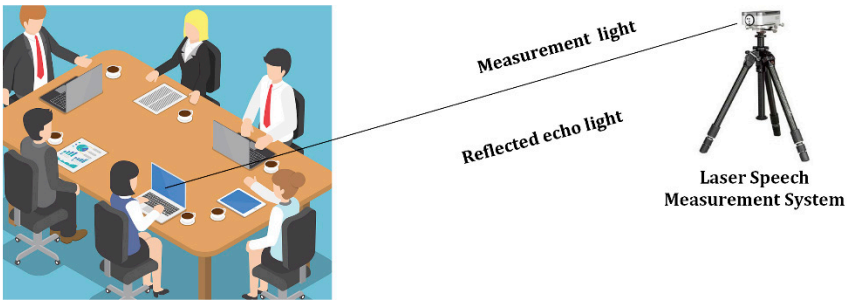

Figure 2. Real example of detection by a laser speech measurement system.

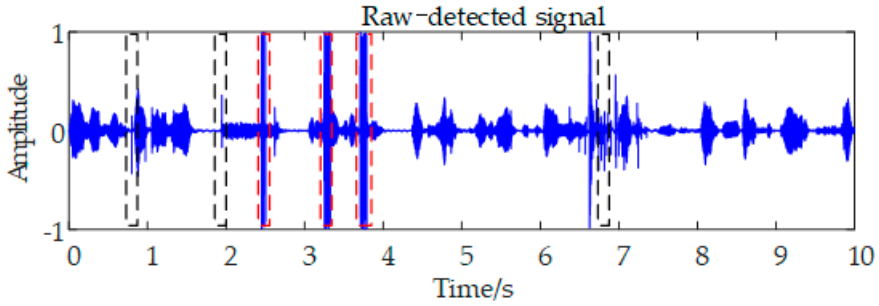

Figure 3. Raw-detected signal with speckle noise.

In speech, speckle noise is in the form of impulsive noise, which seriously reduces the quality and intelligibility of measured speech and has a significant impact on its viability for speech intelligence. Most of the people listening to laser detection signals agreed that due to the appearance of speckle noise, they were psychologically afraid to listen to the voice, and easily tired when listening to this voice, resulting in low speech recognition. In addition, because the background noise is usually removed based on the noise estimation, the irregular appearance of speckle noise affects the accuracy of this estimation. Therefore, removing the speckle noise first is also conducive to the subsequent background noise removal.

In previous works, we used a decorrelation method based on linear prediction (LP) model to detect the location of the speckle noise by improving the Noise-to-Signal Ratio (NSR) of the detection signal, and designed an interpolator to replace the speckle noise [4]. However, the previous method used the direct threshold to judge the noise position. For very weak signals, this method has limitations and the noise location accuracy is not high.

In this paper, we present a simple yet efficient technique that can restore laser measured speech signals that are corrupted by speckle noise. The speckle noise detection method, combining decorrelation preprocessing, average short−term energy and kurtosis to extract the signal and locate the noise according to the threshold which involves relatively little calculation, thereby increasing the computing speed. The decorrelation preprocessing and the double threshold criterion highly increase the noise positioning accuracy. The method of replacing contaminated samples with linear coded samples is also efficient in restoring the signal and reducing the distortion. The results show that the proposed automatic noise detection and removal method outperforms other related methods across a wide range of degraded audio signals.

2. Related Works

Restoring audio signals that are corrupted by targeted speckle noise is a tricky process. For information acquisition, any information loss is fatal; thus, for laser speech detection it is imperative to find the location of the speckle noise and remove it in a targeted fashion instead of denoising the whole speech signal.

The restoration process can be divided into two steps: detection (finding the locations of the degraded samples) and interpolation (replacing the degraded samples with more suitable values). On the other hand, the noise detection technology can be considered to be a Voice Activity Detection (VAD) [5,6] technology. VAD technology can be divided into frequency-domain methods and time-domain methods. Frequency-domain based methods assume that the energy of the noise is concentrated in the high frequency band, while the energy of the speech is mainly distributed in the low frequency band [7].

However, because laser speech detection focuses on the vibration of an object and there are numerous potential sources of interference, the frequency characteristics of the speckle noise are not completely consistent. Therefore, it is very difficult to distinguish speckle noise and speech using frequency-domain processing.

Time-domain methods include energy−based endpoint detectors [8,9], zero-crossing rate-based methods [10], Autocorrelation Function (ACF) based methods [11] and different

feature combination detection methods [12–15]. Energy-based noise detection methods use differences in energy to distinguish noise and speech. However, although the speckle noise energy in laser detected speech is relatively high, in some cases, speckle energy of a very short duration is close to the energy of the speech signal. Therefore, it is impossible to determine an appropriate threshold. Zero crossing rate-based methods represent the number of times a frame of speech signal waveform passes through a horizontal axis. This reflects, in outline, the frequency characteristics of the signal. It is generally thought that a speech segment will have a short-time zero crossing rate that is lower than a certain threshold, while the noise will be higher than the threshold [13]. However, the zero-crossing rate of noise in laser detected speech can be low or high, because the causes of the noise differ. Therefore, it is not possible to set an appropriate threshold to distinguish between speckle noise and speech using a zero-crossing rate. In view of the periodic nature of speech, its ACF is also periodic, with the period being equal to the pitch value. ACF shows peaks at various pitch and harmonics locations. Consequently, ACF-based algorithms are efficient in distinguishing between background noise with a small amplitude and speech. However, they are not so effective for speckle noise.

Outside of the above-mentioned, relatively straightforward time-domain and frequency-domain methods, Cristalli [16] and Lv [17] have developed a kurtosis-based approach for the detection of speckle noise in laser captured signals. As the kurtosis can measure the degree of deviation from a certain distribution [18], it can be applied to identify abnormal speckle samples. Their work introduced a kurtosis ratio (KR)—based method for the detection of speckle noise and the selection of undistorted regions within a signal. Their algorithm is composed of band-pass filtering, signal segmentation and computation of a scalar KR indicator for each signal segment, which can detect outlying samples that are caused by speckle noise. However, this method is not very robust for long−term speckle samples because the distribution of these impulsive samples becomes similar. Thus, the method is not effective for long-term speckle noise.

In speech, speckle noise is represented as impulsive noise in speech. Focusing directly upon impulsive noise, Oudre has proposed an Autoregressive (AR)—based impulsive noise detection [19] and interpolation [20] method that can be used in the detection phase. He transformed an original noisy signal into an excitation signal, while keeping the impulsive noise either unchanged or increased, by drawing on an AR model in order to increase the detection accuracy. After transforming the direct threshold by the estimated value of the excitation standard-deviation to locate the impulsive noise, the AR model can be used to generate samples to replace the noise samples and obtain an enhanced signal. This approach is very effective and can manage the targeted removal of impulsive noise. However, the impulsive noise detection accuracy is undermined by having to establish a direct threshold, especially if the transformed signal still contains a large amount of background noise.

Recently, some data driven methods have been proposed to suppress noise [21,22]. However, they all focus on background noise. For instance, Braun [23] has proposed a neural network-based architecture for VAD that works on a typical short audio frame basis. While the state-of-the-art neural network based VADs can achieve very good results, they often exceed computational budgets and cannot meet real-time operating requirements. Goyal [24] has presented a novel method to computationally determine when video data contains a person speaking through the recognition of full-lip facial closures within a given interval. However, the timing of video and sound detection is often not consistent and the problem of noise when processing the voice remains. Therefore, it is not feasible to use mouth movements to detect noise in laser-captured speech.

In the interpolation phase, several methods have been developed for the interpolation of missing samples in music or speech signals, which closely resembles the concern with noise removal based on accurately positioning the noise in speech detection. While some interpolation techniques, such as median filtering, are completely blind (no hypothesis regarding the signal is made) [25], they are often too crude to reconstruct gaps that are

larger than a few samples. The noise frame can be set to zero. However, this will destroy the periodicity of the speech signal, leading to frequency truncation, causing sudden changes in the frequency between the speech frame and the enhanced frame, and resulting in an audible "popping" sound. So, when a noise frame is located, more appropriate methods need to be selected to obtain the enhanced signal.

In summary, although several studies have been devoted to detecting and removing the noise in speech, most approaches focus on eliminating background noise. In the case of single judgment-based methods, the recognition rate for speckle noise is often low. Some novel data-driven methods have produced promising results. However, they are computationally expensive and time consuming. Therefore, they are not suitable for applications that depend on real time processing. Compared with these relevant methods, we fully analyze the characteristics of speckle noise in laser measured speech. On this basis, a novel automatic speckle noise detection and removal method is proposed. This method first foregrounds noise using decorrelation based on a linear prediction (LP) model that improves the NSR of the measured signal. This allows detection of the position of speckle noise through a combination of the average short-time energy and Kurtosis. The method not only precisely locates small clicks (with a duration of just few samples), but also finds the location of longer bursts and scratches (with a duration of up to a hundred samples). The located samples can then be replaced by more appropriate samples whose coding is based on the LP model. This strategy avoids unnecessary processing and obviates the need to compromise the quality of the relatively large fraction of samples that are unaffected by speckle noise. The proposed method has the advantages of high noise positioning accuracy, less distortion, small amount of computation, fast processing speed and low delay, which can meet the use scene of laser speech measurement.

3. Automatic Noise Detection and Removal

In order to avoid unnecessary processing and the distortion that wholesale processing can cause, the proposed speckle noise removal system consists of two subsystems: a detector and an interpolator (cf. Figure 4). The detector locates the position of each noise sample and the interpolator replaces it.

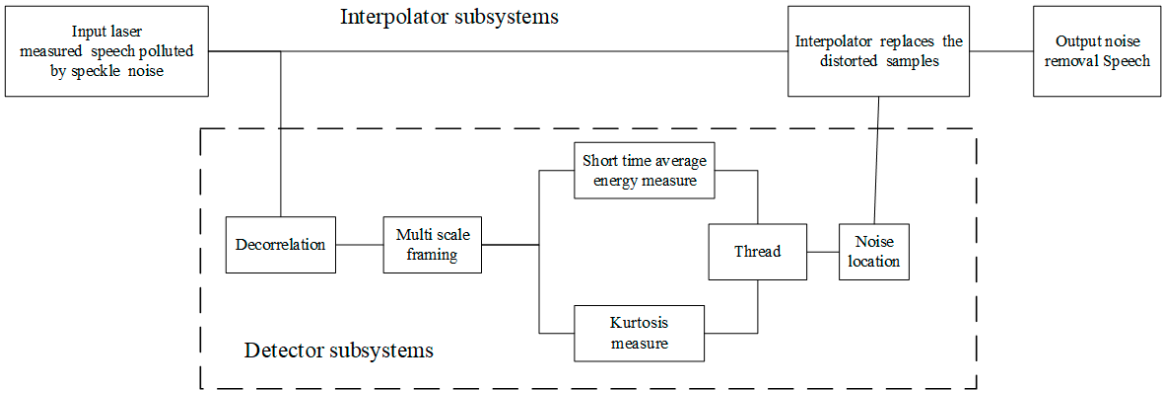

Figure 4. Automatic Detection and Removal of Speckle Noise.

The speckle noise detector plays a crucial role because accurate positioning is essential for effective noise removal. Speckle noise has certain characteristics, such as a large amplitude, random appearance, agitation, unpredictability, and indefinite duration. In view of these characteristics, a focus on energy and distribution can form the basis of distinguishing between speckle noise and speech. However, as previously mentioned, when there are only few occurrences of speckle noise or the amplitude of the speech is large, the energy of the speckle noise will not stand out. Similarly, if the speckle noise lasts for a long time, its distribution will not seem to be abnormal. Thus, relying on any single parameter will not produce a good result. We have also seen how the character of the

speech can itself challenge the detection accuracy. To solve this problem, we have assumed that the speech signal will be correlated while the noise signal will be uncorrelated. We can then begin by decorrelating the measured speech signal. Through decorrelation, the degraded audio signal can be transformed into an extracted signal and the influence of the speech signal and stationary background noise can be eliminated.

For a decorrelated signal, we make full use of the characteristics of the average short-term energy, which is sensitive to long-duration and high-amplitude speckle noise, and the kurtosis, which is sensitive to short-duration abnormal click speckle noise. These two noise detection methods complement each other and can accurately find the location of the speckle noise in a decorrelated signal.

After obtaining the position of the noise frames, we can then use a recursive LP model-based interpolator to replace the signals that are distorted by the speckle noise, one by one. Thus, we finally obtain an enhanced signal. The speckle noise removal process is shown in Figure 4 and the detailed steps involved in the process are given below.

3.1. Decorrelation

A LP model predicts the future value of a signal from a linear combination of its past values. LP models are used for several applications. The correlation structure of a signal can be modelled using a linear predictor by taking the amplitude of the signal at time m ($x(m)$), and then using a linearly-weighted combination of P past samples ($x(m-1)$, $x(m-2)$, ... ,$x(m-P)$):

$$\widetilde{x}(m) = \sum_{k=1}^{P} a_P(k) * x(m-k) \tag{1}$$

where the integer variable m is the discrete time index, $\widetilde{x}(m)$ is the prediction of $x(m)$ and $a_P(k)$ are predictor coefficients. The linear predictor coefficients can be calculated by using the Levinson-Durbin algorithm [26].

The linear prediction model for a signal with an error estimation, $e(m)$, can be expressed as:

$$e(m) = x(m) - \widetilde{x}(m) \tag{2}$$

Assuming that a clean speech signal (x) is corrupted by a random additive speckle noise (n). The detected signal is given by:

$$s = x + n \tag{3}$$

From Equations (1) and (2), we can rewrite the noisy signal model:

$$s(m) = \sum_{k=1}^{P} a_P(k) * x(m-k) + e(m) + n(m) \tag{4}$$

In practice, as there is no clean speech signal x, we use the noisy speech $s(m)$ to calculate an estimate $\hat{a}_N$ of the predictor coefficient vector a_N. This can then be used to invert and transform the noisy signal $s(m)$ to the noisy excitation signal $v(m)$ as:

$$\begin{aligned} v(m) = s(m) - \sum_{k=1}^{P} \hat{a}_P(k) * s(m-k) &= x(m) + n(m) - \sum_{k=1}^{P} [a_P(k) - \hat{a}_P(k)] * [x(m-k) + n(m-k)] \\ &= e(m) + n(m) + \sum_{k=1}^{N} \widetilde{a}_P(k) * x(m-k) + \sum_{k=1}^{P} \hat{a}_P(k) * n(m-k) \end{aligned} \tag{5}$$

where $\widetilde{a}_P(k)$ is the error in the predictor coefficient estimate. According to Saeed's [27] analysis of extracted signals, there are four basic terms that contribute to the noise in an excitation sequence:

(a) the error estimation $e(m)$;

(b) the speckle disturbance $n(m)$, which is usually the dominant term;

(c) the effect of the past noise samples, run over into the present time by the action of the inverse filtering;

(d) the increase in the variance of the excitation signal, caused by the error in the parameter vector estimate, and expressed as: $\sum_{k=1}^{P} \hat{a}_P(k) * x(m-k)$.

As $n(m)$ is usually the dominant term in a noisy excitation signal $v(m)$, when a detected speech signal is converted to an excitation signal, the relative energy of the noise in the signal is increased. In other words, the NSR is increased. Before decorrelation, the NSR of a noisy signal is given by:

$$\frac{power\ of\ speckle\ noise}{power\ of\ signal} = \frac{E\left[n^2(m)\right]}{E[x^2(m)]} \tag{6}$$

where E is the expectation operator. After applying the inverse filter, the NSR is expressed as:

$$\frac{power\ of\ speckle\ noise}{power\ of\ excited\ signal} = \frac{E\left[n^2(m)\right]}{E[v^2(m)]} \tag{7}$$

The overall gain of the NSR can be obtained by:

$$gain = \frac{E\left[x^2(m)\right]}{E[v^2(m)]} \tag{8}$$

This simple analysis demonstrates that an improvement in speckle noise detectability depends on the characteristics of the power amplification of the linear predictor model and the associated resonances. Figure 5 shows a comparison between a raw measured signal and a decorrelated signal. It can be seen that the speech signal is largely removed and the speckle noise in the signal is highlighted.

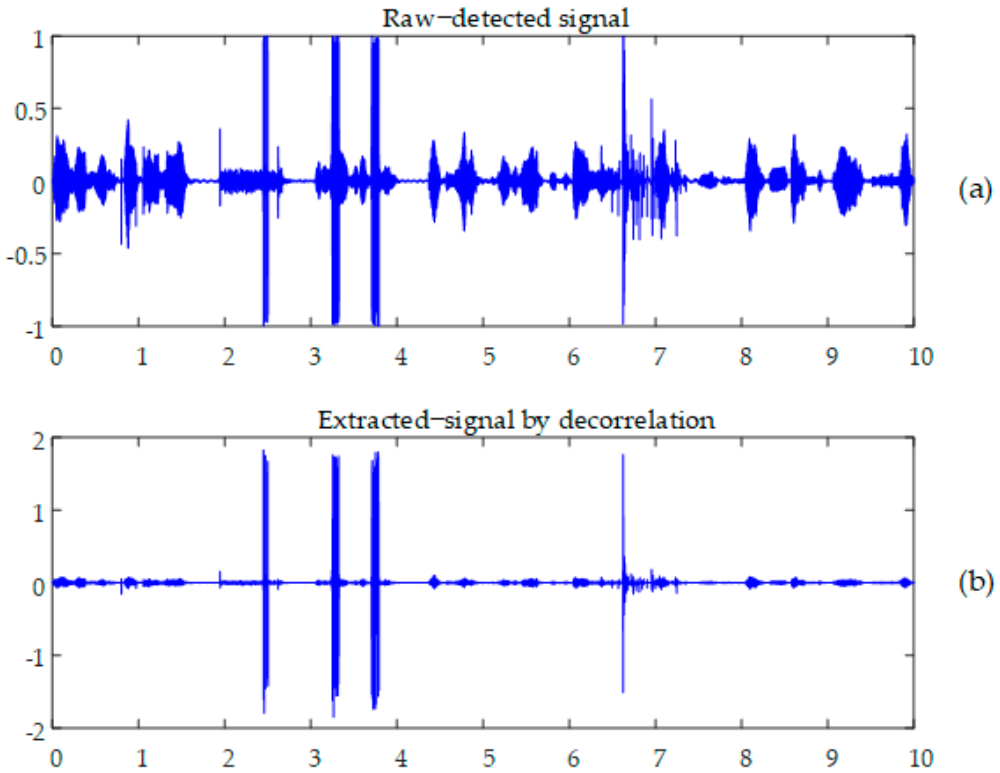

Figure 5. Comparison of (**a**) raw-detected signal and (**b**) extracted-signal by decorrelation.

3.2. Speckle Noise Detection

The average short-time energy and kurtosis are used as judgment indexes to distinguish the noise. The average short-time energy reflects the mean of the weighted sum of the squares of a frame sample values. The average short-term energy (E_n) of a speech signal at time n is expressed as:

$$E_n = \frac{1}{N} \sum_{m=n-(N-1)}^{n} [x(m)\omega(n-m)]^2 \tag{9}$$

where N is the window length and each window length represents one frame. As the extracted signal enhances the energy ratio of the speckle noise, the average short-term energy can be used to locate the speckle noise with a long duration and large amplitude.

On the contrary, the kurtosis is defined as the fourth central statistical moment normalized by the fourth power of the standard deviation. This describes the degree to which a sample deviates from the distribution:

$$K_i = \frac{E\left[(X_i - \mu)^4\right]}{\sigma^4} \tag{10}$$

where K_i is the kurtosis value of signal X, X contains samples of the ith frame $(n-(N-1))$, $x(n-N), \cdots x(n)$, μ and σ are respectively the mean value and standard deviation of x, and $E[\,]$ stands for the overall average.

Using the kurtosis as a method for locating statistical anomalies provides a strong sensitivity to sudden anomalies with a very short duration and a large numerical value.

From the given definitions of average short-term energy and the kurtosis, the average short time energy is sensitive to high-amplitude noise with a long duration. However, it can easily miss speckle noise with a very short duration and a slightly smaller amplitude. However, the kurtosis is sensitive to sudden anomalies with a very short duration. Thus, the two methods are complementary when applied to the detection of different types of noise in laser-detected speech.

Figure 6a,b show the short-time energy and kurtosis value of a raw detected speech signal, and the signal extracted by decorrelation, respectively. Looking at Figure 6a vertically and the long-duration speckle noise in the red dashed box, the average energy value is much greater than the voice. However, the average short-term energy value for the sharp speckle noise with a very short duration in the black box is very similar, or even lower than the voice signal. Simultaneously, the kurtosis value of the frame in the black box is very large. This confirms our proposition that the average short-term energy and kurtosis are complementary. If we then make a horizontal comparison between Figure 6a,b, it can be clearly seen that, by removing the influence of the correlated speech signal, the average short-term energy and kurtosis values of the speckle noise in the extracted signal are greater than they are in the raw signal, confirming that decorrelation can increase detectability of the noise.

Traditional methods usually use a fixed threshold, but as the distance, detection objects and speaker volume can all change, the threshold needs to be reinitialized on each occasion, which reduces the detection efficiency. Therefore, we use the ratio of the present moment to the past average to detect the speckle noise. If the energy E_i or the kurtosis K_i of the current frame is T_{E} times greater than the average energy E_{mean}, or T_{K} times greater than the average kurtosis K_{mean}, the current frame is judged to contain speckle noise.

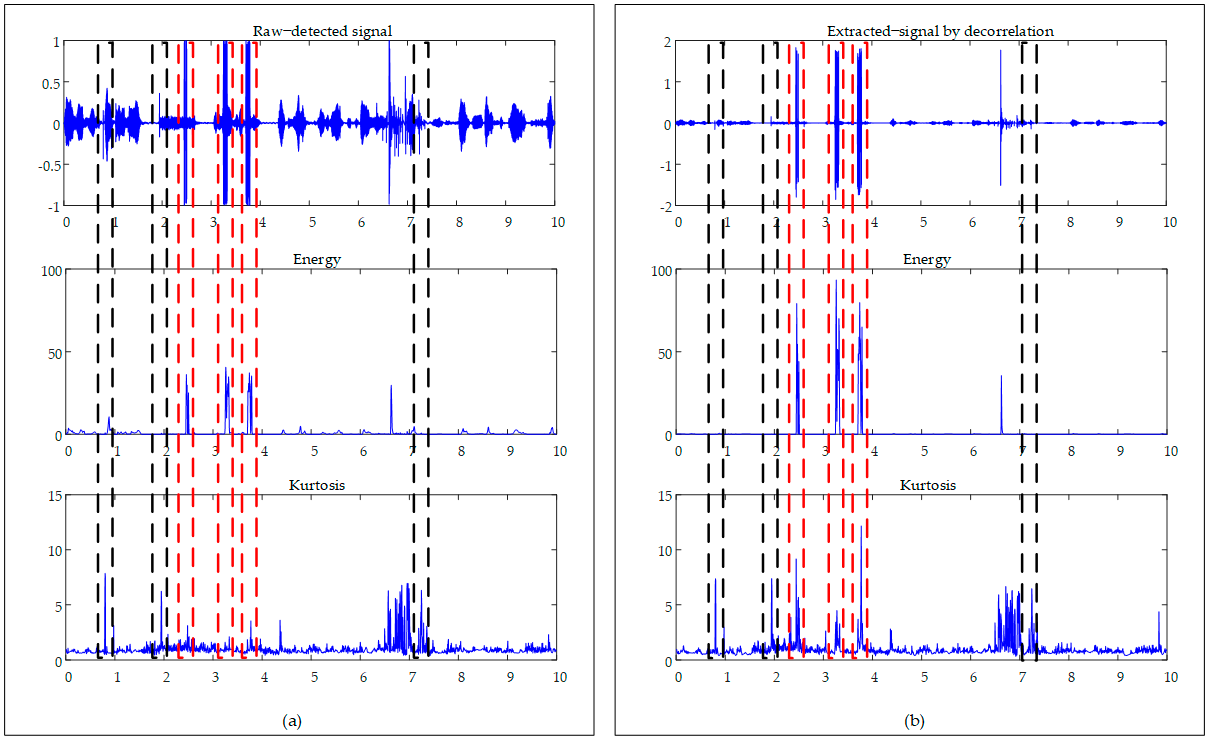

Figure 6. Laser captured speech signal, with the average short-time energy and kurtosis for (**a**) the raw-detected signal and (**b**) the extracted-signal by decorrelation.

3.3. Coding Samples to Replace Noisy Samples

We have focused on the method for locating every sample containing speckle noise. We then use a recursive LP model-based interpolator to replace the signals distorted by the noise, one by one. The basic idea of an LP model-based interpolator for noise removal can be expressed as follows: "The present value of a speech sample can be approximated by a weighted linear combination of past values of several speech samples" [27]. Samples irrevocably distorted by speckle noise are discarded and the gap to the left or the right is interpolated. Firstly, the available (clean) samples in the past of a noise pulse are used to estimate the linear predictor coefficients for the linear prediction model of the signal. Afterwards, the estimated model parameters and the samples on the left of the gap are used to interpolate the polluted sample.

For quasiperiodic signals, such as voiced speech, there are two types of correlation structures that can be utilized for an interpolation:

(1) the short-term correlation, which is the correlation of each sample with the P immediate past samples $x(m-1), x(m-2), \ldots, x(m-P)$.

(2) the long-term correlation, which is the correlation of a sample $x(m)$ with $2\mathrm{Q}+1$ similar samples, a pitch period T away $x(m-T+\mathrm{Q}), \ldots, x(m-T-\mathrm{Q})$.

Due to the disturbances of speckle noise that usually contaminates a relatively small fraction α of all the samples, the length of the samples to be interpolated is short, and the purpose is to remove the influence of speckle noise. Therefore, we do not use the periodic structure for interpolation, but only the interpolation of short-time correlation. That is the linear prediction of a sample $\hat{x}(m)$, based on P past samples. This can be defined as:

$$\hat{x}(m) = \sum_{k=1}^{P} a_P(k) \cdot \hat{x}(m-k) \tag{11}$$

where $\hat{x}(m)$ is the encoded sample at the location of the speckle noise, P is the coding order and $a_P(k)$ are predictor coefficients calculated using the Levinson–Durbin algorithm.

Note that each sample involved in the interpolation is the latest one after the code replacement. The advantage of using an LP model-based interpolator to replace contaminated samples is that it avoids truncation of the signal and keeps the signal consistent. It is not only effective in predicting the content of the signal, but also significantly improves the auditory character of the speech.

3.4. Parameter Setting

In the proposed automatic noise detection and removal method, the signal is divided into overlapping frames of length N with a hop size of N_h samples. In practice, we chose $N_h = N/4$, which corresponds to a 75% overlap. If the frames of length N are greater than the maximum length of the high-energy noise N_{max}, the average energy value will incorporate background noise or speech signals with a smaller amplitude. This results in the threshold rate T_{E}, being hard to set and the noise being missed or erroneously detected. Simultaneously, if the frames of length N are too small and they ignore the continuity of an impulse, only the point of the peak will be located, with the middle value of the peaks not being detected. To further complicate matters, in the case of the kurtosis, if the frames of length N are too short, applying a statistical process will be nonsensical. In terms of delay, overly-long frames of length N will lead to an excessive time delay.

In a laser speech measurement system, the duration of speckle noise is variable but generally ranges between 5 ms and 20 ms. Therefore, we set the frame length N to the largest potential value of 20 ms (i.e., 320 samples at a 16 khz sampling rate) in order to fulfill the requirements for the time delay, and to provide a segment length that is sufficient for reliable estimation of the average energy and kurtosis.

As for the coding order P, i.e., how many samples are used to predict and replace the contaminated samples, as the predictor order increases for a speech signal, the prediction error decreases. Saeed [27] stated that the interpolation error depends on the model order while usually a model order of two to three times the length of missing data sequence achieves good result. Janssen [28] suggests using $P = \{3N_{max} + 2\}$. However, the algorithmic complexity of calculating the prediction coefficients also needs to be considered with the increase of the coding order. Indeed, it seems fair to assume that P should be at least greater than N_{max}, so that only known samples are used for the reconstruction [20]. In order to reuse the parameters calculated during the decorrelation preprocessing, P is set equal to the frame length N, because we previously set the frame length as the maximum length of the speckle noise N_{max}. This is also advantageous since the linear predictor coefficients calculated during the decorrelation stage can be reused in the last step of the coding, which greatly reduces the computation complexity.

The threshold ratios T_{E} and T_{k} directly determine the detection accuracy. Several experiments demonstrated that, when $T_{\mathrm{E}} = 5$ and $T_{\mathrm{k}} = 2$, all the speckle noise is visibly removed without any false detection.

The value averages E_{mean} and K_{mean} were dynamically adjusted during the experiments in order to assess what would provide sufficient sensitivity with a minimum number of false alarms. Everyone has to take a break in order to breathe when speaking, and, on average, this occurs every 10 s. Therefore, the average energy value and kurtosis for over 10 s were taken as the basis of these calculations.

Figure 7 shows the result of locating and replacing the speckle noise in an example of laser measured speech. Figure 7a is the raw noisy signal. Figure 7b shows the noise locations given by the proposed method. The black box shows the noise points located by energy discrimination and the positions marked by an asterisk are the noise points identified by kurtosis. Figure 7c is the speech enhanced by replacing the noise points. For comparison, Figure 7d shows the pure voice.

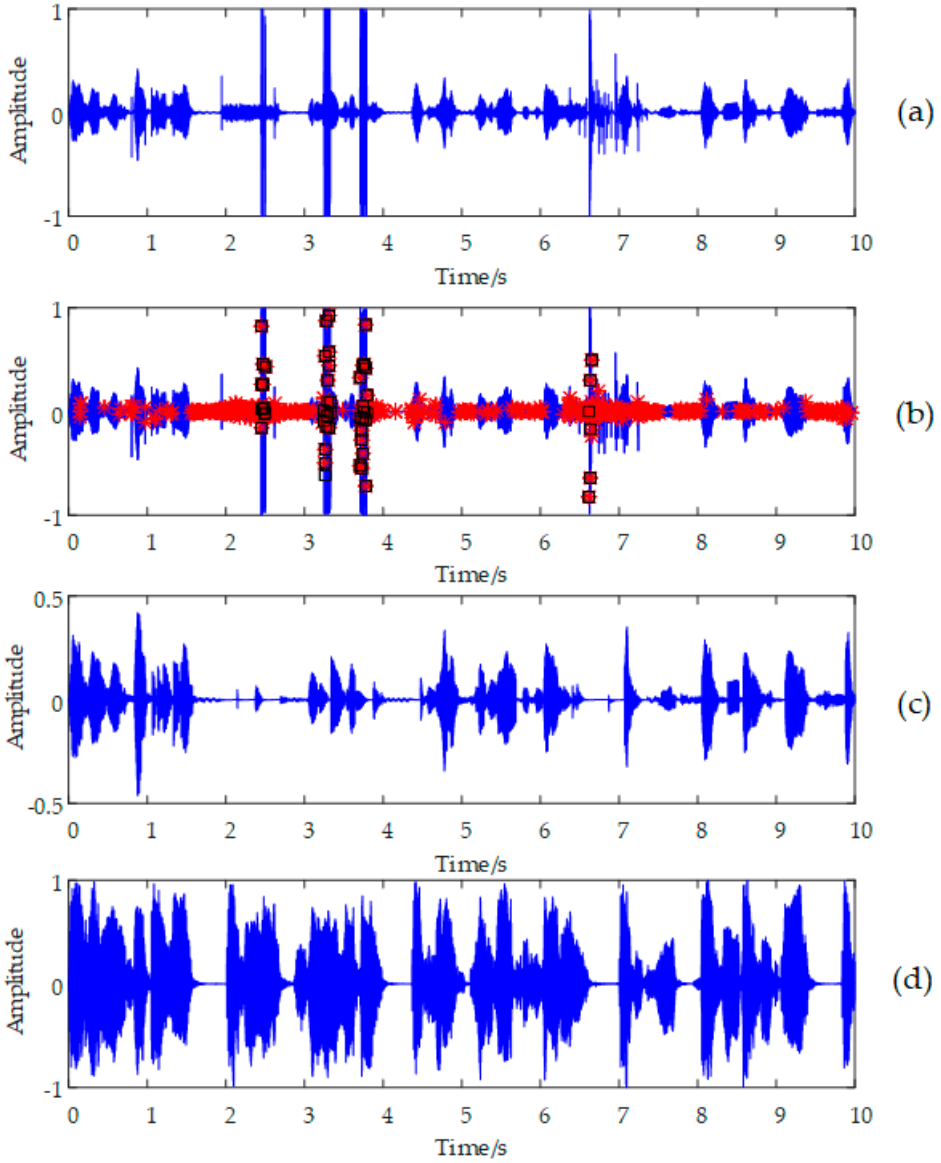

Figure 7. Laser captured speech signal showing (**a**) the raw-detected signal, (**b**) noise locations identified by the proposed method, (**c**) the enhanced-signal, and (**d**) clean signal.

4. Experiments and Discussion

In this section, the results of a performed experiment are shown to verify the applicability of the method and are compared with other related methods.

4.1. Related Methods

The methods we chose for comparison are Cristalli and Lv's kurtosis ratio (KR)-based method [16,17], and Oudre's Autoregressive (AR)—based speckle noise detection and interpolation method [19,20]. In order to ensure consistency, LPC coded samples were used for replacement by all of the methods to obtain the final enhanced speech.

4.2. Experimental Setting

The detection scene of the whole experiment is shown in Figure 8. We used a laser speech measurement system that we constructed by ourselves, as shown in Figure 8a, for the remote speech detection. The selected laser speech measurement system had a nominal working distance of 5–500 m and operated at a wavelength of 1550 nm. Figure 8b shows the target side scenario. The scenario was comprised of a loudspeaker playing clean speech,

an empty paper cup as a detection target, and an acoustic shell meter for detecting the decibel that reaches the surface of the paper cup. In the experiment, the loudspeaker was located 1 m away from the paper cup, and started playing pure voice. The volume of the loudspeaker was adjusted so that the sound intensity level on the surface of the paper cup did not exceed 80 dB. We focused the measuring light on the surface of the paper cup and then obtained a speech signal. By increasing the detection distance from 100 to 300 m and creating human interference, we obtained different intensities of speckle-noise for the noisy speech with various durations.

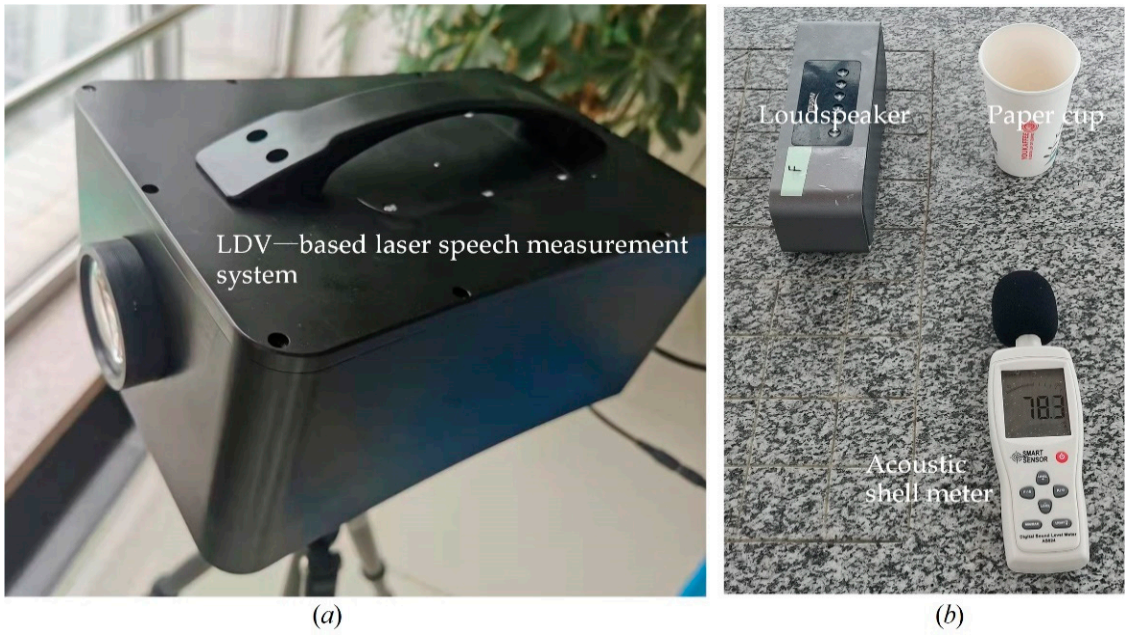

Figure 8. Real-world experiment: (**a**) The laser speech measurement system; (**b**) The loudspeaker, paper cup and acoustic level meter.

Clean speech was used for the speech played by the loudspeaker. This was taken from the Librispeech ASR corpus dataset [29]. LibriSpeech is a corpus of approximately 1000 h of 16 kHz read English speech. In this experiment, we randomly chose 50 groups of voice examples from one female reader and one male reader in the "dev-clean" subfolder, as the pure speech. In order to create speckle noise and therefore the noisy speech, we set interference with indefinite duration, such as occlusion on the measured optical path. We ended up with 50 groups of voice examples at each detection distance, giving a total of 150 groups of speech with speckle noise. The pure speech and noisy speech measured in the experiment are available on Gitee [30].

4.3. Evaluation Indexes

As it is difficult to count the number of points related to speckle noise in real-world detected speech, we first compared and evaluated the different methods by visually inspecting the resulting waveforms. Afterwards, as the accuracy of noise localization can also be indirectly judged by looking at the signal quality, we used the SNR [31] and PESQ [32] as objective metrics.

The SNR and PESQ indicators are calculated by comparing the examples where the speckle noise is removed, and the pure speech. To ensure the experimental results are accurate and to minimize the number of errors, all of the metrics are averaged across each group of 50 speech signals.

4.4. Results and Analysis

Figure 9 shows the raw detected signal at 300 m and the corresponding signal enhancements by the different methods. Figure 9a is the raw detected signal. Figure 9b shows the noise localization and interpolated signal when using AR. Figure 9c is the results for decorrelation of the signal using KR. Figure 9d shows the results obtained by the proposed method. Figure 9e is the clean signal. It can be seen that the missed detection rate for the KR and AR methods is very high. As the noise length increases, the probability of missed detection becomes higher. On the contrary, the proposed method is able to identify the full range of noise with different amplitudes and lengths. Its noise positioning accuracy is also greater.

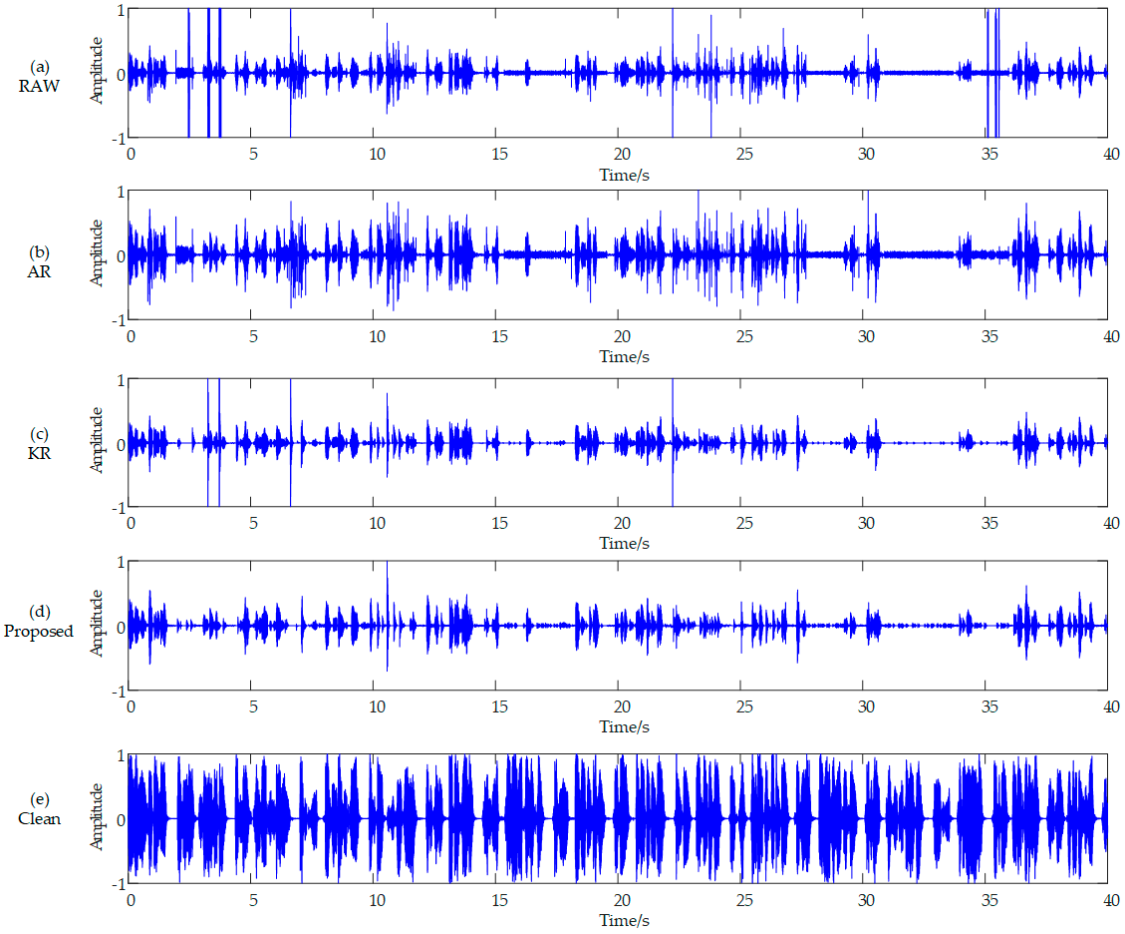

Figure 9. Laser captured speech signal and signal enhanced by different methods. (**a**) Raw signal; (**b**) AR results; (**c**) KR results; (**d**) proposed method results; (**e**) clean signal.

Table 1 shows the SNR and PESQ results for the signals enhanced by the different methods.

It can be seen from Table 1 that the signal enhanced by the proposed method has significantly higher scores than the AR and KR methods. When the detection distance is 100 m, the SNR of the speech signal enhanced by the proposed method is 2.064 dB. This is 0.54 dB and 0.42 dB higher than the AR and KR methods, respectively. When the distance is 200 m and 300 m, the SNR of the speech signal enhanced by the proposed method is 0.77 dB and 0.88 dB higher, respectively. This is a more significant improvement than that achieved by the other methods. The PESQ score of the proposed method at 100 m (2.064) is 0.39 points and 0.43 points higher than the score of the AR method (1.768) and KR method (1.726), respectively. This is even more pronounced when the distance is 300 m, with the PESQ for our method being 1.560 points, which was 0.78 points and 0.73 points higher than the AR and KR results, respectively. As speckle noise usually contaminates a relatively

small fraction of each overall sample, these experimental results do not seem surprising and are perfectly satisfactory for the removal of speckle noise.

Table 1. SNR and PESQ-based objective evaluation results.

Measurement Distance	Objective Metrics	Noisy Speech	AR	KR	Proposed Method
100 m	SNR	1.054	1.520	1.642	2.064
	PESQ	1.635	1.768	1.726	2.160
200 m	SNR	−0.835	−0.718	−0.606	−0.062
	PESQ	1.036	1.225	1.054	1.900
300 m	SNR	−2.159	−1.677	−1.628	−1.272
	PESQ	0.635	0.778	0.827	1.560

5. Key Outcomes and Conclusions

5.1. Key Outcomes

As demonstrated by the experimental results, the proposed algorithm is a simple yet efficient technique that can restore audio signals that are corrupted by speckle noise. The proposed combination of decorrelation preprocessing, average short-term energy, and kurtosis to detect the speckle noise in detected signals involves relatively little calculation, thereby increasing the computing speed. The decorrelation preprocessing and the double threshold criterion greatly improves the noise positioning accuracy. The method of replacing contaminated samples with linear coded samples is also efficient in restoring the signal and reducing the distortion. The advantages of the proposed method can be summarized as follows:

(a) Through decorrelation preprocessing, the influence of the captured speech on speckle noise detection is eliminated and the noise signal ratio is increased. This technique renders the speckle noise prominent, making it easy to detect.
(b) Combining the energy and distribution in the noise detection significantly increases the noise location accuracy because the two complement each other. The results show that our method can deal with speckle noise with different durations (clicks, bursts, outliers, crackles, scrapes, etc.) and different degradation types.
(c) Having a limited range of parameters and basing both the decorrelation and interpolation phases on LPC, keeps the calculation of the coefficients to a minimum, making it possible to process quickly and automatically.
(d) Replacing the degraded samples with more appropriate values greatly improves the voice quality and intelligibility, and avoids the generation of distortion.
(e) The method involves less calculation and less delay. Therefore, it is able to fit the practical needs of laser voice detection applications. The algorithm itself is not restricted to use in laser speech measurement, but also can be used for traditional microphone communication.

5.2. Conclusions

In this paper, we highlighted the problem of speckle noise for the laser speech measurement system, which can limit the efficiency and accuracy of speech information acquisition. We then proposed a novel automatic speckle noise detection and removal method. This method first foregrounds noise using decorrelation, based on a linear prediction (LP) model that improves the signal-to-noise ratio of the measured signal. This allows detection of the position of noise through a combination of the average short-time energy and kurtosis. The method not only accurately locates small clicks with a duration of few samples, but also finds the location of longer bursts and scratches with a duration of up to hundred samples. The located samples can then be replaced by more appropriate samples whose coding is based on the LP model. The method is not only able to improve the speech quality of

laser speech measurement systems, but also provides reference for the noise removal for ordinary electronic microphones.

5.3. Limitations and Future Work

Although the proposed method offers good speckle noise localization and removal, it does not currently consider the potential impact of background noise. In the case of background noise with a varying distribution, the decorrelation would not automatically eliminate its influence. If there is an excessive amount of background noise, this will also have an impact on the selection of the detection threshold and other parameters. In our future work, we will examine how to deal with the possible presence of background noise and will further develop a theoretical model of speckle noise to maximize its mathematical rigor. The parameter selection also requires more theoretical support, rather than relying on just experimental verification.

Author Contributions: Data curation, X.K.; Funding acquisition, H.Z.; Investigation, Z.W.; Methodology, Y.W.; Project administration, W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No: 62071057).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Long, C.V.; Flint, J.A.; Lepper, P.A. Wind turbines and bat mortality: Doppler shift profiles and ultrasonic bat-like pulse reflection from moving turbine blades. *J. Acoust. Soc. Am.* **2010**, *128*, 2238–2245. [CrossRef]
2. Bauer, M.; Ritter, F.; Siegmund, G. High-precision laser vibrometers based on digital Doppler signal processing. *Proc. Spie* **2002**, *4827*, 50–61.
3. Halkon, B.J.; Rothberg, S.J. Vibration measurements using continuous scanning laser Doppler vibrometry: Theoretical velocity sensitivity analysis with applications. *Meas. Sci. Technol.* **2003**, *14*, 382–393. [CrossRef]
4. Wang, Y.H.; Zhang, H.X.; Kong, X.; Wang, Y.; Zhang, H. Two-sided LPC-based speckle noise removal for Laser Speech Detection Systems. *Ieice Trans. Inf. Syst.* **2021**, *104*, 850–862. [CrossRef]
5. Muralishankar, R.; Ghosh, D.; Gurugopinath, S. A Novel Modified Mel-DCT Filter Bank Structure with Application to Voice Activity Detection. *IEEE Signal. Process. Lett.* **2020**, *27*, 1240–1244. [CrossRef]
6. Tan, Z.-H.; Sarkar, A.K.; Dehak, N. An unsupervised segment-based robust voice activity detection method. *Comput. Speech Lang.* **2020**, *59*, 1–21. [CrossRef]
7. Hu, D. Study on speech endpoint detection based on cepstrum distance and short-time energy. *Comput. Technol. Dev.* **2014**, *24*, 77–79.
8. LI, L.; Wang, Y.; Li, X. An Improved Wavelet Energy Entropy Algorithm for Speech Endpoint Detection. *Comput. Eng.* **2017**, *43*, 268–274.
9. Ganapathiraju, A.; Webster, L.; Trimble, J.; Bush, K.; Kornman, P. Comparison of energy-based endpoint detectors for speech signal processing. In Proceedings of the Southeastcon 96 Bringing Together Education, Science & Technology, Tampa, FL, USA, 11–14 April 1996.
10. Li, G.; Nan, Y.; Wang, B.-X. Study of Robust VAD Algorithm in Speech Operation. *Audio Eng.* **2005**, *9*, 41–45.
11. Ghaemmaghami, H.; Baker, B.; Vogt, R.; Sridharan, S. Noise robust voice activity detection using features extracted from the time-domain autocorrelation function. In Proceedings of the INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Chiba, Japan, 26–30 September 2010.
12. Haghani, S.K.; Ahadi, S.M. Robust voice activity detection using feature combination. In Proceedings of the Electrical Engineering (ICEE), 2013 21st Iranian Conference on, Mashhad, Iran, 14–16 May 2013.
13. Zaw, T.H.; War, N. The combination of spectral entropy, zero crossing rate, short time energy and linear prediction error for voice activity detection. In Proceedings of the 2017 20th International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 22–24 December 2017.
14. Korkmaz, Y.; Boyac, A. Unsupervised and supervised VAD systems using combination of time and frequency domain features. *Biomed. Signal. Process. Control* **2020**, *61*, 102044. [CrossRef]

15. Rahman, M.M.; Bhuiyan, M.A.A. Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches. *Int. J. Adv. Comput. Sci. Appl.* **2012**, *3*, 131–138.
16. Cristalli, C.; Torcianti, B.; Vass, J. A new method for filtering speckle noise in vibration signals measured by laser Doppler vibrometry for on-line quality control. In Proceedings of the SPIE-The International Society for Optical Engineering, Bellingham, WA, USA, 25 October 2006; Volume 6345.
17. Lv, T.; Han, X.; Wu, S.; Li, Y. The effect of speckles noise on the Laser Doppler Vibrometry for remote speech detection. *Optics Commun.* **2019**, *440*, 117–125. [CrossRef]
18. Mardia, K.V. Measures of multivariate skewness and kurtosis with applications. *Biometrika* **1970**, *57*, 519–530. [CrossRef]
19. Oudre, L. Automatic Detection and Removal of Impulsive Noise in Audio Signals. *IPOL J.* **2015**, *5*, 267–281. [CrossRef]
20. Oudre, L. Interpolation of missing samples in sound signals based on autoregressive modeling. *Line (Ipol)* **2018**, *8*, 329–344. [CrossRef]
21. Strake, M.; Defraene, B.; Fluyt, K.; Tirry, W.; Fingscheidt, T. Separated Noise Suppression and Speech Restoration: LSTM-Based Speech Enhancement in Two Stages. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 20–23 October 2019.
22. Sun, Z.; Li, Y.; Jiang, H.; Wang, Z. An RNN-based Speech Enhancement Method for a Binaural Hearing Aid System. In Proceedings of the 17th IEEE International New Circuits and Systems Conference (NEWCAS), Munich, Germany, 23–26 June 2020.
23. Braun, S.; Tashev, I. On training targets for noise-robust voice activity detection. In Proceedings of the 29th European Signal Processing Conference (EUSIPCO), Ithaca, NY, USA, 19 May 2021.
24. Goyal, A. Using Spasmodic Closure Patterns to Simplify Visual Voice Activity Detection. *Sn Comput. Sci.* **2021**, *2*, 1–8. [CrossRef]
25. Ko, S.J.; Lee, Y.H. Center weighted median filters and their applications to image enhancement. *IEEE Trans. Circuits Syst.* **1991**, *38*, 984–993. [CrossRef]
26. Boshnakov, G.N.; Lambert-Lacroix, S. A periodic Levinson-Durbin algorithm for entropy maximization. *Comput. Stat. Data Anal.* **2012**, *56*, 15–24. [CrossRef]
27. Vaseghi, S.V. *Advanced Digital Signal. Processing and Noise Reduction*, 2nd ed.; John Wiley: Hoboken, NJ, USA, 2000.
28. Janssen, A.; Veldhuis, R.; Vries, L. Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes. *IEEE Trans. Acoust. Speech Signal Process.* **2017**, *34*, 317–330. [CrossRef]
29. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An ASR corpus based on public domain audio books. In Proceedings of the ICASSP 2015—2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Piscataway, NJ, USA, 19–24 April 2015.
30. Real-Time Speckle Noise Detection and Removal for Laser Speech Measurement Systems. Available online: https://gitee.com/studylearning/real-time-speckle-noise-detection-and-removal-for-laser-speech-measurement-systems (accessed on 21 September 2021).
31. Hansen, J.; Pellom, B. An effective quality evaluation protocol for speech enhancement algorithms. *Inter. Conf. Spok. Lang. Process.* **1998**, *7*, 2822.
32. ITU. Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Itu T Recomm.* **2001**, 862.