# Speech Enhancement Algorithm based on Reassigned Spectrogram and Auditory Masking

Jie Wang*, Chengcheng Yang*, Manlu Huang*, Linhuang Yan*and Jinqiu Sang†
* School of Mechanical and Electric Engineering,Guangzhou University, Guangzhou, China
E-mail:2484888182@qq.com Tel: +86-20-39366923
† Communication Acoustics Laboratory, Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
E-mail:464722621@qq.com Tel: +86-10-82547851

***Abstract*— A single channel speech enhancement algorithm is proposed based on reassigned spectrogram and masking properties of the human auditory system. According to the strong correlation of speech harmonics, the correlation of adjacent frames and that of harmonics of the reassigned spectrogram are adopted to control the forgetting factor of the two-step-decision-directed method, and it can suppress the non-speech components better. Then, the estimated speech spectra is used to compute the noise masking threshold of a perceptual gain factor according to the masking properties of the human ear. Experimental results show that the proposed algorithm can improve the intelligibility of speech signals while removing the background noise.**

## I. INTRODUCTION

The main objective of speech enhancement is to withdraw the pure speech signal in noisy environments[1]. In recent years, the industry is paying increasing attention to Artificial Intelligence. Speech enhancement is an important problem in many speech processing applications, such as mobile communications, cars, medical treatments and home automation, as a human–machine interaction pretreatment module. Various approaches have been proposed to solve this problem, but most speech enhancement algorithms may generate annoying musical noise and speech distortion in a certain degree. Aiming at these problems, some novel algorithms based on human ear perception properties have been proposed [2]-[4]. When the noise is below the human auditory threshold, it can not be perceived by the human ear. Based on the human auditory system property, preserving noise below the auditory masking threshold can reduce the chances of speech distortion.

In many speech enhancement algorithms, spectral subtraction method is a well-known and widely used enhancement method for all types of speech. Based on the masking threshold of the key frequency ranges of each speech frame, YH Liu [5] dynamic calculate the smooth factor of spectral subtraction method, and apply it to traditional spectral subtraction method. At the same time, this algorithm has better performance low SNR. Other than, Hu and Loizou [6] firstly derived a new gain function in frequency domain by analyzing the relationship between speech distortion and residual noise. Combining the new gain function and noise masking threshold can reduce speech distortion, effectively. The above algorithms can effectively alleviate the influence of musical noise to a certain degree. But speech quality may degrade due to residual background noise. To solve this problem, Ching-Ta Lu [7] found that two-step-decision-directed method can get better speech spectral amplitude estimations. Combining noise masking threshold, of single-channel speech enhancement using perceptual-decision-directed approach was proposed. Experiments show that the algorithm can achieve residual noise removal and reducing speech distortion.

According to this, using the correlation of adjacent frames and that of harmonics, a reassigned spectrogram method is proposed to obtain enhanced speech by supervising the estimate of the *a priori* SNR. Then, applying better spectral estimate of speech is employed to compute the noise masking threshold of a perceptual gain factor. Experimental simulation shows that the proposed algorithm can better eliminate the residual noise and improve the speech quality.

## II. THE *A PRIORI* SNR ESTIMATOR BASED ON REASSIGNED SPECTROGRAM

### *A. Channel Instantaneous Frequency (CIF) and the Local*

*Group Delay (LGD)*

Reassigned spectrogram can be defined as the rearrangement of speech spectrum in certain rules [8]-[10]. It not only can embody the time-frequency characteristics of speech signal, but also can suppress the impact of cross-terms. The keys of reassigned spectrogram is to take advantage of both the Channel Instantaneous Frequency (CIF) and the Local Group Delay (LGD), where they can measure the correlation of adjacent frames and that of harmonics, respectively.

The reassigned spectrogram is not simply to distribute the signal energy from the geometric center to the energy center, but also closely related to the phase information of Fourier transformation. Next, the phase information of short time Fourier transform can be obtained from the angle of classical slide window reconstruction signal.

In the classical sliding window method with short time window function $h(t)$, the expression of short time Fourier transform of continuous time signal $x(t)$ can be expressed as follow:

$$\begin{aligned} X_h(\tau,\omega) &= \int x(\tau)h(t-\tau)e^{-i\omega(\tau-t)}\,d\tau \\ &= e^{i\omega t}\int x(\tau)h(t-\tau)e^{-i\omega\tau}\,d\tau \\ &= e^{i\omega t}\,X(t,\omega) \\ &= A(t,\omega)e^{i\varphi_t(\omega)} \end{aligned} \quad (1)$$

where $A(t,w)$is the amplitude of short-time Fourier transform, and $\varphi_t$(w) represents the phase.

So the CIF and the LGD can be defined as follows [10].

$$CIF(\tau,\omega) = \frac{\partial}{\partial\tau} arg\big(X_h(\tau,\omega)\big) \quad (2)$$

$$LGD(\tau,\omega) = -\frac{\partial}{\partial\omega} arg\big(X_h(\tau,\omega)\big) \quad (3)$$

According to Nelson [11], it is possible to know that the two partial derivatives of each short time Fourier transform phase can be expressed by the point estimated at two points by the correlation conversion surface, so the CIF and LGD can be exchanged as follows:

$$C(\tau,\omega,\Delta\tau) = X\left(\tau+\tfrac{\Delta\tau}{2},\omega\right)X^*\left(\tau-\tfrac{\Delta\tau}{2},\omega\right), \quad (4)$$

$$L(\tau,\omega,\Delta\tau) = X\left(\tau,\omega+\tfrac{\Delta\omega}{2}\right)X^*\left(\tau,\omega-\tfrac{\Delta\omega}{2}\right). \quad (5)$$

The formula (4) CIF of phase is derived as follows:

$$\begin{aligned} &\frac{1}{\triangle\tau}[C(\tau,\omega,\varepsilon)] = \frac{1}{\triangle\tau} arg\left(X\left(\tau+\frac{\Delta\tau}{2},\omega\right)X^*\left(\tau-\frac{\Delta\tau}{2},\omega\right)\right) \\ &= \frac{1}{\triangle\tau} arg\left(A\left(\tau+\frac{\Delta\tau}{2},\omega\right)e^{j\phi\left(\tau+\frac{\varepsilon}{2},\omega\right)}A\left(\tau-\frac{\Delta\tau}{2},\omega\right)e^{-j\phi\left(\tau-\frac{\Delta\tau}{2},\omega\right)}\right) \\ &= \frac{1}{\triangle\tau} arg\left(A\left(\tau+\frac{\Delta\tau}{2},\omega\right)A\left(\tau-\frac{\Delta\tau}{2},\omega\right)e^{j\left[\phi\left(\tau+\frac{\Delta\tau}{2},\omega\right)-\phi\left(\tau-\frac{\Delta\tau}{2},\omega\right)\right]}\right) \end{aligned} \quad (6)$$

As a result of A is a real number, formula (6) can be deduced:

$$\tfrac{1}{\triangle\tau} arg[C(\tau,\omega,\Delta\tau)] \approx \tfrac{1}{\triangle\tau}\left[\left(\tau+\tfrac{\Delta\tau}{2},\omega\right)-\left(\tau-\tfrac{\Delta\tau}{2},\omega\right)\right]. \quad (7)$$

According to the expression of Sean A Fulop and Kelly Fitz in the literature [10], formula (8) can be achieved:

$$CIF(\tau,\omega) \approx \tfrac{1}{\triangle\tau}\left[\left(\tau+\tfrac{\Delta\tau}{2},\omega\right)-\left(\tau-\tfrac{\Delta\tau}{2},\omega\right)\right]. \quad (8)$$

Finally, through formula (7) and formula (8), This is the derivative of the time domain,using the "≈",ignoring the phase independent variables.the phase expression of channel instantaneous frequency can be deduced, as shown:

$$CIF(\tau,\omega) \approx \tfrac{1}{\triangle\tau} arg[C(\tau,\omega,\Delta\tau)], \quad (9)$$

where the derivation method of LGD is similar to the derivation method of CIF, so LGD can be show as follows:

$$LGD(\tau,\omega) \approx -\tfrac{1}{\triangle\omega} arg[L(\tau,\omega,\Delta\omega)]. \quad (10)$$

In speech signal processing, because the input signal is discrete, the voice signal of the current frame can be expressed as:

$$x = [x(n-N+1), x(n-N+2), \dots, x(n)]^T, \quad (11)$$

where N is a frame length, and it is delayed by a bit to get $y$:

$$y = [x(n-N+1), x(n-N+2), \dots, x(n-1)]^T. \quad (12)$$

Therefore, according to the formula (9) to (12), the expression of CIF and LGD of the discrete speech signal can be obtained:

$$CIF(k) = \tfrac{Fs}{2\pi} * arg\big(X(k)Y^*(k)\big) \quad (13)$$

and

$$LGD(k) = \tfrac{-N}{2\pi Fs} * arg\big(X(k)X^*(k-1)\big), \quad (14)$$

where $F_s$ is the sampling frequency rate, superscript "$*$" means conjugate, and $arg(\cdot)$ represents the complex function

of principal value of radial angle. $X(k)$ and $Y(k)$ are Fourier transforms of input speech signals x and y, respectively.

*B. The tracking performance of CIF and LGD of noisy speech.*

In order to further illustrate reassigned spectrogram which can improve the estimate of the *a priori* SNR, noisy speech signals is used to analyze the tracking performance of reassigned spectrogram.

In Fig.1, the clean speech signal is added with stationary Gauss white noise with the SNR=10dB. The tracking performance of CIF and LGD for noisy speech are tested by CIF adjacent frame subtraction and LGD adjacent band subtraction. According to Fig.1, The instantaneous frequency CIF of the channel with inter-frame correlation between speech signals and the local group delay LGD with interfrequency correlation of speech signals have good tracking performance for the noisy speech signals. Therefore, a better *priori* SNR can be obtained by the modification of CIF and LGD, and the foundation of computing the noise masking threshold is laid.

*C. Reconstruction of the a priori SNR*

A noisy speech signal $y(n)$ can be modeled as the sum of clean speech $x(n)$and additive noise $d(n)$ in the frame of the time domain:

$$y(n) = x(n) + d(n). \quad (15)$$

In the spectral domain, the spectral estimate of a speech

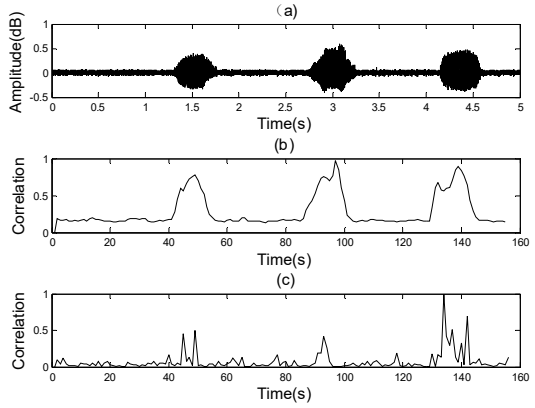


Fig. 1. The tracking performance of CIF and LGD of noisy speech. (a)noisy speech signal; (b)CIF tracking performance; (c) LGD tracking performance.

signal $\hat{X}(k,l)$ is obtained by multiplying a gain factor $\hat{G}(k,l)$ with the noisy spectrum $\hat{Y}(k,l)$ of a subband. The estimated speech spectra can be expressed by

$$\hat{X} = \hat{G}(k,l) \times Y(k,l). \quad (16)$$

The gain factor$\hat{G}(k,l)$ is decided by the estimated *a priori* SNR, given as

$$\hat{G}(k,l) = \hat{\xi}(k,l)/\left(1 + \hat{\xi}(k,l)\right), \quad (17)$$

where, according to [12], the *a priori* SNR of the decision-directed method can be described as follow:

$$\hat{\xi}_{\mathrm{DD}}(\mathrm{k,l}) = \alpha_{\mathrm{DD}} G^2_{\mathrm{DD}}(\mathrm{k,l-1})\hat{\gamma}(\mathrm{k,l-1}) + (1-\alpha_{\mathrm{DD}})\max\{\hat{\gamma}(\mathrm{k,l}) - 1, 0\}, \quad (18)$$

where $\alpha_{DD}$ is a constant, the typical value is from 0.92 to 0.98; $G_{DD}(k,l-1)$ is the gain function of the previous frame, $\hat{\gamma}(k,l-1)$ is the estimated value of the posterior SNR of the previous frame, $\hat{\gamma}(k,l)$ is the estimated value of the posterior SNR of the current frame.

So the *a priori* SNR $\hat{\xi}_{TSDD}(k,l)$of two-step-decision-directed method can be obtained by [13]:

$$\hat{\xi}_{TSDD}(k,l) = \max\{\beta_{TSDD}(k,l)\hat{\gamma}(k,l), \xi_{min}\}, \quad (19)$$

where $\beta_{TSDD}(k,l)$ is determined by *a priori* SNR of DD algorithm, as shown below:

$$\beta_{TSDD}(k,l) = \hat{\xi}_{DD}(k,l)/\left(1 + \hat{\xi}_{DD}(k,l)\right)^2. \quad (20)$$

According to above analysis, CIF and LGD will be used to control the forgetting factor of the prior SNR of the two-step-decision-directed method. The concrete steps can be divided into three steps.

First, update the *a priori* SNR estimators by using CIF. Compute the CIF difference between adjacent frames:

$$R_{CIF}(l) = CIF(l) - CIF(l-1), \quad (21)$$

where $l$ is the frame index, $R_{CIF}(l)$ can track the variation of the CIF. By comparing the mean value M of $R_{CIF}(l)$ of the current frame and the preset threshold th0, and $A\_CIF(l)$ can be obtained as follows:

$$A\_CIF(l) = \begin{cases} R_{CIF}(l), & \mathrm{M} \geq th0 \\ 0, & \mathrm{M} < th0 \end{cases}, \quad (22)$$

where $\mathrm{th0} = 1.0\mathrm{e}-12$, which can be obtained through multiple simulation experiments.

Find the maximum value of the current frame and the normalized CIF difference $B\_CIF(l)$ is defined by:

$$B\_CIF(l) = A\_CIF(l)/max\left(A\_CIF(l)\right). \quad (23)$$

In order to get a more accurate estimate of *a priori* SNR, the

forgetting factor usually ranges from 0.92 to 0.98. $C_{CIF(l)}$ is the mean value of $B\_CIF(l)$ in single frame and its modified value is $E\_CIF(l)$:

$$E_{CIF(l)} = C_{CIF(l)} + \beta, \quad (24)$$

where $\beta$ is a constant value ranging from 0.7 to 0.87.

So, the parameter $E\_CIF(l)$ can be obtained to update the *a priori* SNR estimators:

$$\hat{\xi}_{CIF}(k,l) = \max\{(1-E_{CIF})|G_{CIF}(k,l-1)\hat{\gamma}(k,l-1) + E_{CIF} * P[\hat{\gamma}(k,l-1)], \xi_{min}\}, \quad (25)$$

and

$$G_{CIF}(k,l-1) = \max(\hat{\gamma}_k(k,l) - 1, \xi_{min}). \quad (26)$$

Second, compute the LGD value for each subband to update the *a priori* SNR estimators and obtain$G_{LGD}(k,l)$, which is induced by:

$$G_{LGD}(k,l) = LGD(k,l)/Tseg \quad (27)$$

$$A_{LGD} = 1 - \min(G_{LGD}(k,l), \delta), \quad (28)$$

where $\delta$ is a constant 0.92 in the experiments, and $Tseg = 2$ is the amplitude correction for group delay. Experiments show that the greater $\delta$ corresponds better tracking performance of speech and more residual noise:

$$\hat{\xi}_{LGD}(k,l) = \max\{A\_LGD * \hat{\xi}_{LGD}(k-1,l) + (1-A\_LGD) * \hat{\xi}_{LGD}(k,l), \xi_{min}\}. \quad (29)$$

Third, combine the *a priori* SNR estimator with the CIF and that with the LGD to obtain the estimated SNR:

$$\hat{\xi}_{RS}(k,l) = k * \hat{\xi}_{CIF}(k,l) + \lambda * \hat{\xi}_{LGD}, \quad (30)$$

where $k + \lambda = 1$. If the sum of k and l is not equal to 1, it may cause the distortion of the input speech signal. $k$ is bigger than $l$ because big $\hat{\xi}_{LGD}(k,l)$ may lead to much residual noise.

Therefore, the enhanced speech spectrum can be expressed as:

$$\hat{X} = \hat{G}(k,l) \times Y(k,l). \quad (31)$$

The gain function $G(k,l)$ of the speech signal can be obtained as:

$$G(k,l) = \hat{\xi}_{RS}(k,l) / \left(1 + \hat{\xi}_{RS}(k,l)\right). \quad (32)$$

## III. Proposed Auditory Masking Algorithm Based on Reassigned Spectrogram

### *A. .Estimation of noise masking threshold*

As the subjects of receiving voice, human ears have different feelings for different intensities. The auditory masking effect is that when two (or more) sound acts on the human ear, the human ear is more sensitive to the louder sound. When the weak sound is lower than a certain threshold, it becomes unheard. And that threshold is the noise masking threshold (NMT), which is obtained through modeling the frequency selectivity of the human ear and its masking property [14]. The detailed procedure for estimating the NMT used herein is described as follows.

Initially, the estimated spectra of pre-processed speech can be accurately estimated by the RS algorithm. Hence, the critical-band energy is computed by:

$$B_i = \sum_{l=bl_i}^{bh_i} |\hat{X}(k,l)|^2, \quad (33)$$

where $bl_i$ and $bh_i$ represent the upper and the lower frequencies at the $i$ th critical band. And the sampling frequency determines the critical frequency band number.

Taking masking properties between different critical bands into account , an excitation pattern $B_i$ can be thought as an energy distribution along the basilar membrane. So the extended critical energy can be computed as follow:

Extended critical energy calculation

$$SF_{ij} = 15.81 + 7.5(\triangle + 0.474) - 17.5\sqrt{1 + (\triangle + 0.474)^2} \quad (34)$$

and

$$C_i = B_i * SF_{ij} = \sum_{j=1}^{i_{max}} B_i \times SF_{ij}, \quad (35)$$

where $\Delta = i - j$ is the difference between the two frequency bands and the maximum value is less than or equal to $i_{max}$.

A relative threshold offset which specifies whether a speech frame is tone-like or noise-like is imposed to adjust the log-critical-band energy.

Because pure tone and noise can produce different masking characteristics in the human ear, the scholars use the pure tone coefficient $\phi$ to compensate for the threshold, which can be obtained by the flatness of the Bark spectrum$SFM_{dB}$, and the flatness can be obtained by the geometric mean $Gm$ and the arithmetic mean $Am$ of each band power spectrum, so the

masking threshold offset $O_i$ is obtained as follows:

The threshold value compensation

$$O_i = \varphi(14.5 + i) + 5.5(1 - \varphi)\text{dB}, \quad (36)$$

$$\varphi = \min\{\frac{SFM_{dB}}{SFM_{dBmax}}, 1\}, \quad (37)$$

$$SFM_{dB} = 10lg\frac{Gm}{Am}. \quad (38)$$

From above, the noise masking threshold can be calculated as follow:

$$T(k,i) = 10^{lg[C_i - (O_i/10)]}. \quad (39)$$

$T'(k,l)$ can be obtained by extend the spectrum, so the final noise masking threshold can be expressed as:

$$T(k,l) = \max[T'(k,l), T_a(k,l)], \quad (40)$$

where the absolute threshold of human ear $T_a(k,l)$ is defined as follows:

$$T_a(k,l) = 3.64f^{-0.8} - 6.5exp(f - 3.3)^2 + 10^{-3}f^4, \quad (41)$$

where the units of $f$ and $T_a(k,l)$ are kHz and dB SPL (Sound Pressure Level), respectively.The above is the preliminary calculation of noise masking threshold, which is a very fixed solution method..

*B. Perceptual gain factor*

The spectral estimate of a speech signal is obtained by multiplying a perceptual gain factor with the noisy spectrum. Next, in order to demonstrate the performance of speech enhancement, spectral distortion measurement can be defined as the spectral difference before and after speech enhancement

$$\begin{aligned} E(k,l) &= \hat{X}(k,l) - X(k,l) \\ &= (1 - G(k,l))X(k,l) + G(k,l)D(k,l) \\ &= \varepsilon_s(k,l) + \varepsilon_d(k,l) \end{aligned} \quad (42)$$

where the spectra of speech distortion is $\varepsilon_s(k,l)$ and that of residual noise is $\varepsilon_d(k,l)$. It can be found that the trend of the above two changes is the opposite with the change of gain function $G(k,l)$. Therefore, how to select an optimal gain function to minimize speech distortion and maximize suppress noise is a key problem.

In order to solve the problem, it is assumed that the noise signal is additive and is uncorrelated with a speech signal. The gain factor can be optimized by minimizing the short-term spectral energy associated with the speech distortion, subject to a constraint on the short-term spectral energy related to residual noise below the noise masking threshold (NMT):

$$\min_{G(k,l)}\{\varepsilon_x^2(k,l)\} \quad (43)$$

subject to the constraint $\varepsilon_d^2(k,l) \le T(k,l)$,

where T(k,l) is the NMT corresponding to the frequency bin $k$. The values of NMT are all identical in a critical band.

So using the noise masking threshold, the gain function can be obtained as follow [15]:

$$G_1(k,l) = 1/\left(1 + max\sqrt{\frac{|D(k,l)|^2}{T(k,l)} - 1, 0}\right). \quad (44)$$

So the estimated spectra of processed speech can be obtained by:

$$\hat{X}_1 = \hat{G}_1(k,l) \times \hat{X}(k,l). \quad (45)$$

Finally, the enhanced speech signal can be obtained by the inverse-fast-Fourier transform (IFFT), given as:

$$\hat{x}_1(n) = IFFT|\hat{X}_1(k,l)| \cdot \exp(jargY(k,l)). \quad (46)$$

*B. The flow chart of the method*

According to the above analysis, the key of the algorithm is to update the gain function adaptively with the speech reassigned spectrogram and the auditory masking characteristics of the human ear, and then use the gain function to restore the speech from noisy environments. The reassigned spectrogram is introduced, which use the

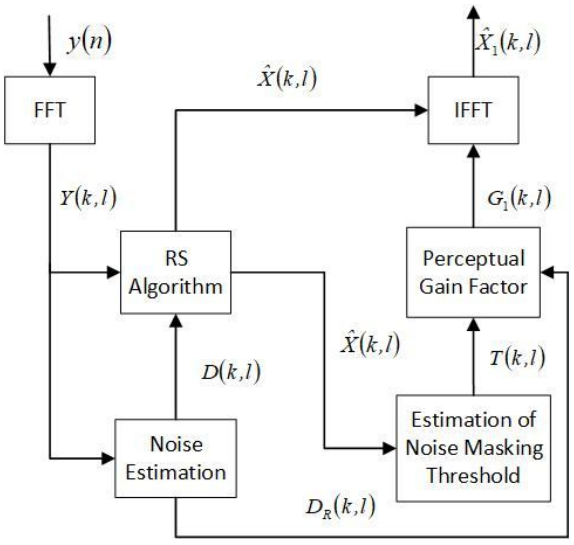


Fig. 2. Block diagram of the proposed method

correlation of adjacent frames and that of harmonics to control the forgetting factor of the two-step-decision-directed method, which can suppress the non-speech components better.

Then, the estimated speech spectra is employed to compute the noise masking threshold of a perceptual gain factor by the masking properties of the human ear. As a result of, speech distortion become less. Therefore, the algorithm block diagram can be described as Fig.2.D(k,l) is estimated noise.DR(k,l) is real noise which is generally used to obtain ideal masking thresholds.

## IV. Experimental Results

In the simulation experiments, head mounted microphone is used to record the clean speech signals. The distance between the microphone and the sound recorder is about 2cm, and we can put it as a test sample. The noise signals consist of white, babble, factory, stationary and f16 noise signals, which are selected from the NoiseX-92 database.

The following parameters are chosen in our simulation experiments: (1) all the signals are sampled by 8kHz; (2) each frame includes 512 samples with 50% overlap; (3) where the frequency bin index k = 50 is selected, and the other parameters are as follows: $\alpha = 0.93, \varphi = 0.92, k = 0.7,\ \lambda = 0.(4)$ the noise power spectrum is estimated by the MS algorithm.

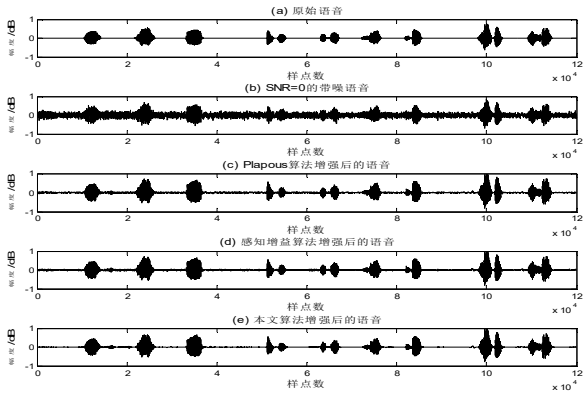


Fig. 3. Example of speech signal spoken in Mandarin by a male speaker (from top to bottom): (a) clean speech, (b) noisy speech (corrupted by factory noise with average segmental SNR = 0 dB), (c) enhanced speech using two-step-decision-directed method, (d) enhanced speech using perceptual two-step-decision-directed method, and (e) enhanced speech using proposed method

Next, algorithm enhancement experiments under different SNR are carried out.

Fig.3 and Fig.4 show waveform and spectrogram comparisons for various speech enhancement methods, respectively. In figure 3, the objective evaluation index is applied, and the distinction in the time domain is not obvious.Where speech signal is corrupted by factory noise signals with SNR = 0dB. The speech waveform processed by proposed method is more complete, and the background noise of the unvoiced segment is greatly eliminated. At the same time, it can be find that the speech stripe processed by proposed method is clearer in Fig.4, so the proposed method can improve the quality of the speech, effectively.

In order to further verify the effectiveness of the proposed algorithm, objective measures, including the average of segmental SNR improvement (SegSNRI) [16], Average logarithmic spectrum distance (LSD)[16] and the perceptual evaluation of speech quality (PESQ)[17] are conducted to evaluate the performance of a speech enhancement system. Where M1 is the two-step-decision-directed method[13], M2 is the perceptual-decision-directed approach[7] and M3 is the proposed algorithm.

Table 1 and Table 2 show the comparative results of SegSNRI and PESQ in the two-step-decision-directed method, the perceptual-decision-directed approach and the proposed algorithm, respectively. In the table 1, the proposed algorithm

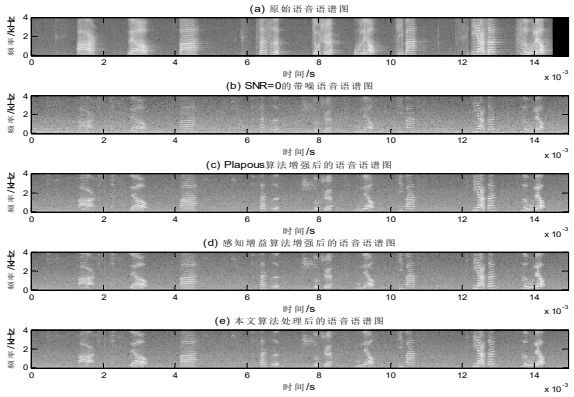


Fig. 4. Spectrograms of clean speech spoken by a female speaker: (a) clean speech, (b) noisy speech (corrupted by factory noise with average segmental SNR = 0 dB), (c) enhanced speech using two-step-decision-directed method, (d) enhanced speech using perceptual two-step-decision-directed method, and (e) enhanced speech using proposed method.

Table 1. Comparisons of segmental SNR improvement for enhanced speech in various noise corruptions.

| Noise type | SNR(dB) | Method | | |
|---|---|---|---|---|
| | | M1 | M2 | M3 |

| | | | | |
|---|---|---|---|---|
| White | 0 | 2.28 | 2.92 | 3.06 |
| | 5 | 2.53 | 2.33 | 2.54 |
| | 10 | 0.59 | 1.47 | 2.03 |
| Babble | 0 | 1.55 | 2.50 | 2.59 |
| | 5 | 0.76 | 1.66 | 1.84 |
| | 10 | 0.13 | 0.81 | 0.95 |
| Factory | 0 | 2.17 | 3.60 | 3.47 |
| | 5 | 1.39 | 2.39 | 2.87 |
| | 10 | 0.40 | 1.43 | 2.05 |
| F16 | 0 | 2.10 | 3.10 | 3.10 |
| | 5 | 1.35 | 2.47 | 2.64 |
| | 10 | 0.42 | 1.47 | 2.08 |

Table 2. Comparisons of PESQ for the enhanced speech in various noise corruptions.

| Noise type | SNR(dB) | PESQ | | |
|---|---|---|---|---|
| | | M1 | M2 | M3 |
| White | 0 | 0.46 | 0.47 | 0.52 |
| | 5 | 0.44 | 0.45 | 054 |
| | 10 | 0.33 | 0.36 | 0.53 |
| Babble | 0 | **0.43** | 0.45 | 0.39 |
| | 5 | 0.35 | **0.39** | 0.37 |
| | 10 | 0.23 | 0.28 | 0.29 |
| Factory | 0 | 0.39 | 0.42 | 0.44 |
| | 5 | 0.28 | 0.33 | 0.39 |
| | 10 | 0.21 | 0.28 | 0.40 |
| F16 | 0 | 0.56 | 0.57 | 0.57 |
| | 5 | 0.40 | 0.43 | 0.48 |
| | 10 | 0.34 | 0.39 | 0.43 |

all reaches a better SegSNRI than the other two algorithms for white, babble, factory and f16 noise. In the table 2, The PESQ of the proposed algorithm is better than those of the other two algorithms in white, factory and f16 noise. But for the babble noise, the PESQ of proposed algorithm is slightly lower than that of the two-step-decision-directed method and is close to and transcend the other two algorithms while the input SNR increases.

Table 3 shows the comparative results of LSD in three

Table 3. Comparisons of LSD for the enhanced speech in various noise corruptions.

| Noise type | SNR(dB) | LSD | | |
|---|---|---|---|---|
| | | M1 | M2 | M3 |
| White | 0 | 6.49 | **6.76** | 4.40 |
| | 5 | 3.91 | **4.02** | 3.63 |
| | 10 | 2.98 | 2.94 | 3.06 |
| Babble | 0 | 6.75 | 6.74 | 5.76 |
| | 5 | 5.05 | 5.03 | 4.50 |
| | 10 | 3.82 | 3.69 | 3.54 |
| Factory | 0 | 5.01 | 5.00 | 3.60 |
| | 5 | 3.85 | 3.71 | 2.90 |
| | 10 | 3.14 | 3.92 | 2.53 |
| F16 | 0 | 5.94 | 5.77 | 3.94 |
| | 5 | 4.18 | 3.96 | 3.07 |
| | 10 | 3.13 | 2.93 | 2.62 |

algorithms, which is the two-step-decision-directed method, the perceptual-decision-directed approach and the proposed algorithm. In the table 3, for the white noise, the LSD of perceptual-decision-directed approach is slightly higher than that of the two-step-decision-directed method in the low SNR, but it has a better PESQ. Which means some unseen noise is retained. This is in accordance with the principle of auditory masking effect. However, the proposed algorithm all reaches a better LSD than the other two algorithms for white, babble, factory and f16 noise. So according to above analysis, the proposed algorithm can suppress background noise and improve the intelligibility of speech signals.

## V. CONCLUSIONS

In this paper, a single channel speech enhancement algorithm based on reassigned spectrogram and masking properties of the human auditory system is proposed. According to the analysis of the correlation of adjacent frames and that of harmonics, we take advantage of CIF and LGD to control the forgetting factor of the two-step-decision-directed method, which can suppress the non-speech components and get a better estimate of speech spectral. Then, through study the auditory masking of human ear system, it is found that this estimate can be employed to compute the noise masking threshold of a perceptual gain factor. Experimental results show that the amounts of residual noise can be efficiently suppressed by proposed method.

## ACKNOWLEDGMENT

This work was supported by National Science Fund of

China (No.61302126), Special Innovation Project of Department of Education of Guangdong Province(No. 2017KTSCX141) , Key Lab of Information Processing & Transmission of Guangzhou(No.201605030014）and Modern Video &Audio Information Engineering Center of Guangdong Province.

## REFERENCES

[1] J. Benesty, S. Makino, J. Chen, “Speech enhancement,” [M]. Springer., 2005.

[2] N. Virag, “Single channel speech enhancement based on masking properties of the human auditory system,” IEEE Transactions on Speech and Audio Processing., vol. 7, no. 2 , pp. 126-137, 1999.

[3] MR. Schroeder, BS. Atal, JL. Hall, “Optimizing digital speech coders by exploiting masking properties of the human ear,” J. acoust. soc. am., vol. S1-66, pp. 139, 1979.

[4] O. Cappe, “Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor,” IEEE Transactions on Speech and Audio Processing., vol. 2, no. 2, pp. 345-349, 1994.

[5] YH. Liu, “Spectral subtraction speech enhancement based on the masking properties of human auditory system,” Information Technology., 2009.

[6] Y. Hu, PC. Loizou, “A perceptually motivated approach for speech enhancement,” Speech & Audio Processing IEEE Transactions on., vol. 11, no. 5, pp. 457-465, 2003.

[7] CT. Lu, “Enhancement of single channel speech using perceptual-decision-directed approach,” Elsevier Science Publishers B. V., vol. 53, no. 4, pp. 495-507, 2011.

[8] H. Wang, CJ. Huang, LP. Yao, Y. Qian, XC. Jiang, “Time-frequency Analysis of Partial Discharge for GIS Using the Theory of Reassignment Distribution,” High Voltage Engineering., vol. 36, no. 9, pp. 2236-2241, 2010.

[9] X. Wu, T. Liu, “Time-frequency analysis on Wigner-Ville distribution of seismic signal based on time-frequency rearrangement,” Oil Geophysical Prospecting., vol. a02, pp. 86-91, 2009.

[10] SA. Fulop, “Speech spectrum analysis,” Springer Berlin Heidelberg., pp. 41-68, 2011.

[11] DJ. Nelson, “Cross-spectral methods for processing speech,” Journal of the Acoustical Society of America., vol. 110, no. 1, pp. 2572-2592, 2001.

[12] [12]R. Martin, “Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors,” IEEE International Conference on Acoustics., vol. 1, pp. 253-256, 2002.

[13] C. Plapous, C. Marro, P. Scalart, “Improved signal-to-noise ratio estimation for speech enhancement,” IEEE Transactions on Audio Speech & Language Processing., vol. 14, no. 6, pp. 2098-2108, 2006.

[14] N. Ma, “Speech enhancement algorithms using Kalman filtering and masking properties of human auditory systems,” University of Ottawa., 2005.

[15] Y. Hu, PC. Loizou, “Incorporating a psychoacoustical model in frequency domain speech enhancement,” Signal Processing Letters IEEE., vol. 11, no. 2, pp. 270-273, 2004.

[16] Y. Hu, PC. Loizou, “Evaluation of objective quality measures for speech enhancement,” IEEE Trans on Audio, Speech and Language Processing., vol. 16, no. 1, pp. 229-238, 2008.

[17] AW. Rix, JG. Beerends, MP. Hollier, AP. Hekstra, “Perceptual evaluation of speech quality (pesq) a new method for speech quality assessment of telephone networks and codecs,” [C]. IEEE International Conference on Acoustics., vol. 2, pp. 749-752, 2002.