Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-09849-8

Real-time speech enhancement algorithm for transient noise suppression

Ruiyu Liang[1] · Yue Xie[1] · Jiaming Cheng[2] · Guichen Tang[1] · Shinuo Sun[2]

Received: 18 August 2019 / Revised: 16 August 2020 / Accepted: 9 September 2020
Published online: 23 September 2020

Abstract
To effectively restrain stationary noise and transient noise, a real-time single-channel speech enhancement algorithm is proposed. First, to evaluate stationary noise, the quantile noise estimation method is used to obtain the spectrum of stationary noise. Then, based on the normalized variance and gravity center of the signal, the transient noise detection method is proposed to modify the spectrum of stationary noise. Next, the speech presence probability is estimated based on the speech features and harmonic analysis. Finally, the optimized-modified log-spectral amplitude (OM-LSA) estimator is adopted for speech enhancement. The experimental noise contains 115 environmental sounds with the SNR of −10 to 10 dB. The experimental results show that the performance of the proposed algorithm is comparable to the OM-LSA algorithm which has good denoising performance, but the real-time performance of the former is much better. Compared with the Webrtc real-time algorithm, under the overall performance of stationary noise and transient noise, the overall speech quality indicators of the improved algorithm increased by 7.5%, 7.8% and 5.0%, respectively. And the short-time objective intelligibility increased by 2.4%, 2.4% and 2.0%, respectively. Even compared with the recurrent neural network(RNN) algorithm, the suppression performance of the transient noise is better. Besides, the real-time experiment base on the hardware platform shows that the runtime of processing a 10 ms frame is 4.3 ms.

Keywords Speech enhancement · Transient noise suppression · The quantile noise estimation · Harmonic analysis

✉ Ruiyu Liang
liangry@njit.edu.cn

[1] School of Information and Communication Engineering, Nanjing Institute of Technology, Nanjing 211167, China

[2] School of Information Science and Engineering, Southeast University, Nanjing 210096, China

1 Introduction

Although speech enhancement algorithms have been studied for decades, they are still a hot topic in the field of speech processing. The early single-channel speech enhancement algorithm mainly studied how to effectively estimate the noise spectrum from noisy speech and thus suppress it. In recent years, with the concept of deep learning [13] and its successful application in the field of speech recognition [7], the speech enhancement algorithm based on supervised learning began to demonstrate its worth [40]. The deep neural network (DNN) [24, 45], convolutional neural network (CNN) [10], long short-term memory(LSTM) network [41, 42], generative adversarial network(GAN) [26, 30] are all used to implement speech enhancement. These supervised learning models demonstrate superior performance over traditional enhancement methods in the case of adequate training.

Tan et al. [35] develop a novel convolutional recurrent network(CRN) architecture for speech enhancement in real-time. The CRN incorporates a convolutional encoder-decoder and long short-term memory. However, this algorithm uses multi-frame 161-dimensional short-time Fourier transform (STFT) magnitude spectrum of noisy speech as input features and has a 12-layer network. Pandey et al. [28, 29] propose two fully convolutional neural networks for real-time speech enhancement. From the real-time performance, the input of the network is the frame length and the number of parameters is larger than 4.82 million. Takeuchi et al. [34] propose a speech enhancement method using a causal DNN for real-time applications. The equilibrated recurrent neural network (ERNN) is used for avoiding the vanishing/exploding gradient problem. This algorithm uses 512 points STFT magnitude spectrum as input features. Tan et al. [36] propose a novel deep-earning based framework for real-time speech enhancement on dual-microphone mobile phones in a close-talk scenario. However, the real-time performance of the algorithm is not evaluated. Valin et al. [37] propose a deep-learning approach to realize real-time full-band SE. The algorithm used an RNN with four hidden layers to estimate ideal critical band gains. And good experimental results are obtained. Although the above research shows excellent performance in speech enhancement, their real-time performance is not good except for reference [37]. These algorithms have too many network layers, usually more than 10 layers. And, the input parameters are usually STFT magnitude spectrum or time-domain signals with high dimension. So, the parameters of these algorithms are counted in millions, which limit their real-time performances. Also, these algorithms do not give the suppression effect of transient noise.

To improve the ability of transient noise suppression based on the supervised learning algorithm, it is necessary to collect these transient noise samples for training. However, due to the complexity of reality, it is unrealistic that a data set includes all situations. Both stationary and non-stationary noises may interfere with the speech. Even if the data set can contain all the conditions, supposing the types of noises can reach 10,000 [39], this will be a huge burden for model training. At this time, to make full use of these data, the deep neural networks algorithm has to increase the complexity of the network, such as using more hidden layers [11] and multi-segment networks [19]. If there are several problems with the labels of the data set, the result of the training cannot be guaranteed. Besides, generalization is an important indicator of any supervised learning algorithm. For speech enhancement, noise [39], speaker [4], and SNR [44] are the three main influencing factors. When the model is applied to an untrained noisy environment, it is difficult to obtain ample data in a short time. In other words, there are only a few noisy signals, and in most cases, it is difficult to obtain a clean label corresponding to these noisy signals.

Due to the difficulties in data annotation and acquisition, and the high complexity of the model, the applications for speech enhancement based on supervised learning algorithms are limited. Nowadays, real-time speech enhancement algorithms are still dominated by classical algorithms. Among these algorithms, spectral subtraction [1] is the earliest denoising algorithm. It generally detects the speech activity of noisy speech, estimated the power spectrum of the noise segment, and then performs spectral subtraction processing. However, if the power spectrum of the noise is underestimated, the new music noise will be introduced. On the other hand, overestimation will result in the loss of speech information and cause distortion. The residual noise caused by the Wiener filtering method is similar to Gaussian white noise, which is better than music noise for hearing, but the distortion still exists. In the 1980s, Ephraim et al. proposed the Minimum Mean Square Error (MMSE) [8] estimator based on Bayesian criteria. This algorithm has an optimal estimation of the amplitude spectrum, which improves the enhanced speech quality. Later, for the log-spectral amplitude is proportional to the ear's perception toward the acoustical loudness, the MMSE method based on Log-Spectral Amplitude (LSA) [9] is proposed. Additionally, Chen and Loizou proposed the Minima Controlled Recursive Averaging (MCRA) [5] noise estimation algorithm [44] and the optimized-modified log-spectral amplitude [6] (OM-LSA) estimator. Kumar [20] propose a generic model of minimum mean square error (MMSE) based speech enhancement technique and realize it on a TI DSP chip with a 1 GHz Processor. However, and the frame length of the algorithm is not given to judge the real-time performance. These algorithms focus on additive background noises and are designed based on complex statistical properties between noise and speech. Besides, they often assume that the noise is relatively stationary or slowly changing.

To this end, this paper proposes an improved real-time single-channel speech enhancement algorithm for transient noises with a quick change. The algorithm first estimates the stationary noise based on the quantile noise estimation algorithm. Then a method based on the combination of normalized energy variance and gravity center of the signal is proposed to detect the transient noise and modify the noise spectrum. Secondly, to reduce the misjudgment of the speech presence probability, the algorithm combines speech features and harmonic analysis to estimate the speech presence probability. Finally, the algorithm uses the OM-LSA estimator for speech denoising. The results of 115 kinds of environmental noises experiments show that the performance of the improved algorithm is comparable to that of the OM-LSA algorithm which has good denoising performance, but the real-time performance of the proposed algorithm is far superior to the OM-LSA algorithm. Compared with the Webrtc denoising [17] algorithm and the RNN-based algorithm, the proposed algorithm has better performance on the suppression of the transient noise.

The paper is organized as follows. Firstly, the algorithm model is briefly introduced in Section 2. Subsequently, the noise estimation method, speech probability estimation method and speech enhancement method, which the three core parts of the algorithm, are introduced in detail in Section 3, Section 4 and Section5. The experimental results are shown and discussed in Section 6. Finally, the conclusions are presented in Section 7.

2 Algorithm model

To effectively restrain stationary noise and transient noise, a real-time single-channel speech enhancement algorithm without introducing the deep learning network is proposed. Because the traditional noise estimation algorithms tend to estimate the minimum power, even the

MCRA algorithm has no time to respond to the transient noise. Therefore, the transient noise detection algorithm is adopted in the process of noise estimation. The specific idea is shown in Fig. 1. Firstly, the algorithm frames the noisy speech, adds the window, and computes the power spectrum of the signal. The parameters of the STFT were the 256 points Hann window, 128 points time-shifting, and 256 points FFT length. Then, the algorithm estimates the stationary noise and transient noise to obtain a priori and posterior signal-to-noise ratio (SNR) of the signal through the joint estimation. To reduce the effect of the voiced sounds on the discrimination transient noises, a combination discrimination method based on normalized energy variance and gravity center of the signal is proposed. Next, to effectively estimate the noises, the joint strategy of feature-based and harmonic-based speech presence probability estimation method is proposed. Finally, the filter coefficients are generated based on the estimated SNR and the signals are filtered.

Suppose that $s(n)$ represents the clean time signal and $d(n)$ represents the additive noise signal, then the noisy signal $x(n)$ is obtained as:

$$x(n) = s(n) + d(n) \tag{1}$$

The frequency-domain representation of the noisy signal $x(n)$ is

$$X(k,l) = S(k,l) + D(k,l) \tag{2}$$

where k represents the frequency points of the signal and l represents the frame label of the signal.

3 Noise estimation

3.1 Stationary noise estimation

In a given frequency band, about 80% to 90% of speech frames have very low signal energy, and only 10% to 20% of the frames contain high energy (voiced segment). To this end, a quantile noise estimation algorithm is proposed by Stahl et al. [32]. The algorithm computes

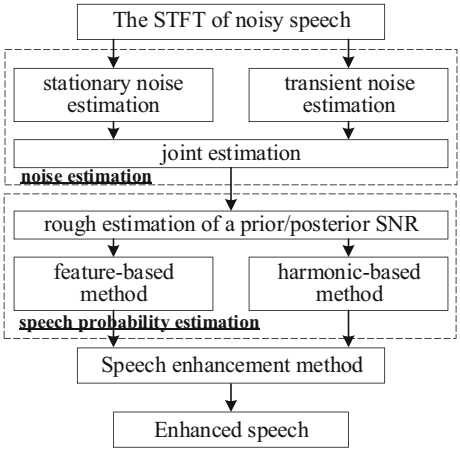


Fig. 1 The framework of the proposed real-time speech enhancement algorithm

the q-quantile point of each frequency band and uses it as a parameter to estimate the noise power spectral density from the noisy speech.

The basic steps are:

1) The noise inhibitory factor λ is computed based on the estimated q-quantile point.

$$\lambda = \begin{cases} \alpha/q(k), q(k) > 1 \\ \alpha, others \end{cases} \tag{3}$$

Here, the quantile q represents the noise probability. α is a constant which represents the maximum inhibitory value and set to 40.

2) Update quantile noise amplitude

$$D_q(k,l) = \begin{cases} D_q(k,l) + \beta\lambda/l, |X(k,l)| > D_q(k,l) \\ D_q(k,l) - (1-\beta)\lambda/l, others \end{cases} \tag{4}$$

Here, $D_q(k, l)$ represents the quantile noise amplitude at the frequency point k of the frame l. To make the noise value less than the actual amplitude value, the tradeoff factor β is less than 0.5 and set to 0.25.

Then, the estimated quantile noise is:

$$N_q(k,l) = e^{D_q(k,l)} \tag{5}$$

3) Update the quantile q

$$q(k) = \begin{cases} q(k) + \dfrac{1}{2\omega l}, \ \ abs\big(|X(k,l)| - D_q(k,l)\big) < \omega \\ q(k), others \end{cases} \tag{6}$$

Here, ω is a constant and set to 0.01. It measures the difference between the estimated noise and the actual spectral amplitude. When the difference is less than ω, the current frame is close to noise, and the quantile q is updated to regain the boundary between noise and speech. Since the noise is often unstable, the algorithm will set l to 1 every two seconds to reactivate the quantile noise estimation algorithm.

3.2 Transient noise estimation

The quantile noise estimation algorithm can effectively estimate the stationary noise, but it does not estimate transient noise in time. Because the presence of the transient noise is very short, its short-term energy is much larger than that of the speech signal. For transient noise, the common detection methods include energy ratio method [14], variance method [25] and wavelet analysis method [27]. Traditionally, spectral correlation coefficients [47] and the characteristics of harmonic [46] can be used to distinguish unvoiced sounds and transient noise. However, when both voiced sounds and transient noises are present, the former can greatly affect the characteristics of the latter to reduce the discrimination effect. So, a combination of normalized energy variance and gravity center of the signal [23] is proposed to detect the transient noise.

The normalized variance of the current frame is:

$$\overline{\mathrm{var}(k,l)} = \frac{\sum_{k=f_l}^{f_h}\left(|X(k,l)| - \overline{X(k,l)}\right)^2}{\sum_{k=f_l}^{f_h}|X(k,l)|^2} \tag{7}$$

Here, f_l and f_h represent low frequency and high-frequency points, corresponding to 100 Hz and 5000 Hz, respectively. The larger the normalized variance is, the more harmonic components of the signal are. Here, the signal is closer to the speech and the threshold is set to 0.3.

The algorithm distinguishes speech and noise according to the fact that the main energy of the speech concentrates on the middle and low-frequency band, the variance of speech is relatively large and various harmonics exist. In contrast, the noise is relatively stable. Although the detection algorithm based on the normalized energy variance can detect some transient noise, there are still some exceptions. For example, the overall amplitude of some noises deviates from the expectation. Therefore, the variance is large and not be used to judge the transient noise by the above formula. To solve this problem, the center of gravity of the frequency domain signal is adopted. Traversing from the center of gravity to both sides, the algorithm calculates the ratio of the traversed energy to the total energy. When the traversal distance is short, the signal energy is concentrated in a very short frequency band. At this time, the signal can be judged as transient noise.

The algorithm flow is:

1) Pre-whitening signal

To enhance the difference between transient noise and speech, the current signal can be estimated by linear prediction analysis. The formula is as follows:

$$x'(n,l) = x(n,l) - \sum_{p=1}^{P} a_p x(n-p,l) \tag{8}$$

where $x'(n)$ represents the pre-whitening speech signal. a_p is the autoregressive factor, which is obtained by the Levinson-Durbin algorithm [2].

2) Solving the center of gravity

Based on the characteristic of the pre-whitening signal, the transient noise detection method based on the center of gravity is adopted. The gravity center of the l-th frame can be expressed as:

$$C(l) = \frac{\sum_{n=0}^{N-1} nw(n)|x'(n,l)|}{\sum_{n=0}^{N-1} w(n)|x'(n,l)|} \tag{9}$$

Here, $w(n)$ represents the Hanning window and $C(l)$ represents the gravity center of the l-th frame.

3) Estimate the minimum length of time $B(l)$ to meet the following conditions:

$$\min_{B(l)}\left\{\frac{\sum_{n=C(l)-B(l)}^{C(l)+B(l)} x'(n,l)}{\sum_{n=0}^{N-1} x'(n,l)} \geq E\%\right\} \tag{10}$$

where E represents the energy ratio which is set to 90. In general, because the energy of transient noise is concentrated in a certain frequency band, $B(l)$ tends to be small. In contrast, because the speech has a wide energy distribution, $B(l)$ tends to be large. It can be considered that transient noise exists when $B(l) < C_{th}$. C_{th} is associated with the length of the frame. When the frame length is 256, C_{th} takes 75.

3.3 Joint estimation

The estimated noise $N_q(k, l)$ is the smallest power spectrum that has been counted over the past period and is the stationary noise of the current frame. To prevent overestimation, the power spectra of the current frame containing transient noise is superimposed to $N_q(k, l)$ with an attenuation factor κ (0.4) to reduce the effects of distortion. Therefore, the estimated noise $N_a(k, l)$ of the current frame can be expressed as:

$$N_a(k,l) = N_q(k,l) + \kappa|X(k,l)| \tag{11}$$

In summary, the estimated noise of the current frame is obtained by weighing different noises according to different conditions.

4 Speech probability estimation

Although there are many ways to estimate the probability of speech presence, the accuracy is not very high. To this end, the joint strategy of feature-based and harmonic-based speech presence probability estimation method is proposed. The feature-based method has better performance in judging the speech presence probability in each band. Besides, to guarantee speech quality, the algorithm enhances the probability of the speech harmonic band to improve the estimation accuracy. The harmonic-based estimation method mainly uses the speech harmonics simulated based on the estimated pitch period as the envelope of the harmonic frequency band. When the noise suppression is performed in these bands, the simulated signal is attenuated to retain important speech components.

4.1 Rough estimation of a prior/posterior SNR

A priori and posteriori SNR are calculated based on the estimated noise $N_a(k, l)$. The posteriori SNR refers to the ratio of the power spectrum of the observed signal to that of the estimated noise at frame l:

$$\delta_l(k) = \frac{|X(k,l)|^2}{|N_a(k,l)|^2} \tag{12}$$

A priori SNR $\rho_l(k)$ refers to the ratio of the power spectrum of clean signal to that of the estimated noise at frame l:

$$\rho_l(k) = \frac{|S(k,l)|^2}{|N_a(k,l)|^2} \tag{13}$$

Since the clean signal $S(k, l)$ is unknown, a priori SNR of the current frame is the weighted value of the estimated priori SNR of the previous frame and the posteriori SNR of the current frame [3]:

$$\rho_l(k) = r_{dd}\rho_{l\text{-}1}(k) + (1-r_{dd})\max(\delta_l(k)-1, 0) \tag{14}$$

Here, r_{dd} is the corresponding time smoothing parameter and its value is 0.98. The larger the value is, the higher the fluency is, and the greater the delay is.

4.2 Feature-based method

Feature selection has a great effect on audio and image signal processing [21, 22]. To estimate the speech probability of the current frame, various speech features [31] can be used to estimate the similarity between the current frame and the speech signal. Here, the likelihood ratio mean, the spectral flatness [18] and spectral difference are selected.

1) Likelihood ratio means

The likelihood ratio $\Delta(k, l)$ is defined as

$$\Delta(k,l) = \frac{P(X(k,l)|H_1)}{P(X(k,l)|H_0)} = \frac{\exp\left\{\frac{\rho_l(k)\delta_l(k)}{(1+\rho_l(k))}\right\}}{(1+\rho_l(k))} \tag{15}$$

Here, H_1 and H_0 represents the speech state and the noise state, respectively. Since the likelihood ratio factor $\Delta(k, l)$ between frames and frames will fluctuate greatly, the smoothed $\widetilde{\Delta}(k, l)$ can be expressed as

$$\log\left(\widetilde{\Delta}(k,l)\right) = r_{lrt}\log\left(\widetilde{\Delta}(k,l-1)\right) + (1-r_{lrt})\log(\Delta(k,l)) \tag{16}$$

Here, r_{lrt} represents a smoothing factor and is set to 0.5. The geometric mean of the smoothed $\widetilde{\Delta}(k, l)$ is a reliable indicator of the speech/noise state.

$$F_1 = \log\left(\prod_k \widetilde{\Delta}(k,l)\right)^{1/N} = \frac{1}{N}\sum_{k=1}^{N}\log\left(\widetilde{\Delta}_k(k,l)\right) \tag{17}$$

2) Spectral flatness

The spectral flatness reflects the number of peaks and troughs in the envelope curve of the frequency domain. When the sound waves resonate through the vocal tract, formants exist one after another. So, the number of harmonics is large and more peaks and troughs appear

correspondingly. Therefore, the speech spectrum is not smooth but the noise is the opposite. So, spectral flatness can be used as a means of distinguishing between speech and noise. The equation is:

$$F_2 = \frac{\sqrt[N]{\prod_{k=0}^{N-1}|X(k,l)|}}{\frac{1}{N}\sum_{k=0}^{N-1}|X(k,l)|} = \frac{\exp\left(\frac{1}{N}\sum_{k=0}^{N-1}\ln|X(k,l)|\right)}{\frac{1}{N}\sum_{k=0}^{N-1}|X(k,l)|} \tag{18}$$

Here, F_2 represents spectral flatness feature and its general interval is between [0, 1]. When F_2 is close to 1, it indicates the frequency spectrum of the current frame is stationary and can be approximated as noise. When F_2 is close to 0, it means that the frequency spectrum of the current frame fluctuates frequently, which is very similar to the harmonic characteristics of speech, so it can be considered as speech.

3) Spectral difference

The spectral difference can be obtained by calculating the variance of the current frame, the variance of the template and their covariance. The template spectrum is computed base on the segments which are most likely to be noise or speech pauses. The covariance is statistically used to measure the error between the two. If the trend of their changes is consistent, the covariance is positive. The variance represents the dispersion degree of variables. The specific equation is

$$F_3 = \mathrm{var}(X(k,l)) - \frac{\mathrm{cov}(X(k,l), N_m(k,l))^2}{\mathrm{var}(N_m(k,l))} \tag{19}$$

where var() and cov() represent the variance function and the covariance function, respectively. This feature reflects the degree of deviation from the current frame to the template.

Feature-based probability updates can use the following model:

$$q_l = \gamma_q q_{l-1} + \left(1-\gamma_q\right) f(F-T_F, w) \tag{20}$$

where γ_q is a smoothing factor. $f(F-T_F, w)$ is a mapping function of time and frequency (between 0 and 1) and the *tanh* function is used. Here, F is the measured feature and T_F is the threshold. The parameter w represents the shape/width feature of the mapping function and its value is 4. The mapping function divides the time-frequency slot into speech (f is close to 1) or noise (f is close to 0) based on the measured features, the threshold and width parameters. The specific expression of the mapping function is

$$f(F-T_F, w) = 0.5*(\tanh w(F-T_F) + 1) \tag{21}$$

To effectively track the change of the signal, the thresholds of the three features need to be updated dynamically. The algorithm uses the histogram to count the frequencies at which different features appear to compute the threshold. The likelihood feature needs to estimate the volatility of the feature while the other two features count two peaks and the threshold is calculated according to the distance between the two peaks.

1) Likelihood ratio threshold T_{F_1}

The equation of the likelihood ratio volatility is:

$$u_{F_1} = \sum_{l=1}^{L} (H_{F_1}(l))^2 - \frac{\sum_{l=1}^{L} H_{F_1}(l)}{L} * \frac{\sum_{l=1}^{M} H_{F_1}(l)}{M} \tag{22}$$

where L represents the number of frames and M represents the number of frames whose volatility is less than 1. The volatility is calculated to get the variance of the L frames. Since the speech is fluctuating and the noise is stationary, if the local expectation is not much different from the global expectation, the likelihood ratio is stationary in accordance with the noise characteristic and the obtained variance is relatively small.

At this time, the threshold of the likelihood ratio can be expressed as

$$T_{F_1} = \begin{cases} 1, & u_{F_1} < \delta \\ \xi \dfrac{\sum_{l=1}^{M} H_{F_1}(l)}{M}, & others \end{cases} \tag{23}$$

where δ is the threshold and its value is 0.05. When the volatility is less than δ, it indicates that the signal is most likely noise. At this time, the likelihood ratio threshold needs to be set higher so that the probability of the speech calculated by the eq. (21) can be smaller. On the contrary, if the signal is most likely speech, the likelihood ratio will be higher. To be conservative, the algorithm can set ξ a little higher to avoid affecting the noise reduction effect. Here, ξ is set to 1.2.

2) The threshold of spectral flatness and spectral difference

The thresholds of spectral flatness and spectral difference are computed in the same way. To make the distribution of the mapping function wider, the algorithm traverses the histogram to obtain the most frequent feature f_1, the second most frequent feature f_2, their occurrence count P_1 and P_2. If the difference between these two features is less than the threshold 0.1 and the occurrence count of the smaller feature is higher than that of the other feature, it is considered that the appearance probability of these two features is equal and their mean is

$$\begin{cases} f = \dfrac{f_1 + f_2}{2} \\ P = P_2 + P_1 \end{cases} \tag{24}$$

Otherwise, the occurrence count is P_1 and its corresponding feature is f_1.

Here, the threshold can be expressed as

$$T_{F_2} = \begin{cases} T_{F_2}, P < \varepsilon \\ \varsigma f, others \end{cases} \tag{25}$$

The above equation indicates that if the appearance count of the feature is too low (less than ε, its value is 150), the statistical samples are insufficient and the feature is not correctly estimated, then the previously estimated threshold will be used. Otherwise, the threshold is updated by eq. (24) and multiplied by the weight ς. Here, ς is 0.4.

Because different features contain different information, the joint processing provides a more stable and adaptive speech/noise updating probability. Therefore, three features are used to update the template of the speech/noise probability simultaneously, that is

$$q_l = \gamma_q q_{l-1} + \left(1-\gamma_q\right)[M(F_1-T_{F_1}) + M(F_2-T_{F_2}) + M(F_3-T_{F_3})] \tag{26}$$

The smoothed prior probability $\widetilde{q}_l$ is

$$\widetilde{q}_l = (1-\xi)*\widetilde{q}_{l-1} + \xi*q_l \tag{27}$$

Finally, the latest probability is calculated based on the likelihood ratio $\Delta(k, l)$ and a prior probability $\widetilde{q}_l$.

$$q_F = \frac{\widetilde{q}_l \Delta(k,l)}{\widetilde{q}_l \Delta(k,l) + 1-\widetilde{q}_l} \tag{28}$$

4.3 Harmonic-based method

According to the principle of speech generation, harmonics frequency is a mixture of multiple fundamental frequencies. In general, the noise is irregular and does not have obvious harmonic characteristics. The speech segment can be enhanced by increasing the probability of speech presence of the harmonic band. For other frequency bands, it is considered to contain less speech or is corrupted by noise and can be used to evaluate noise. To estimate the speech harmonics, a multi-band excitation (MBE) [12] fitting algorithm is adopted.

First, the excitation spectrum is computed.

$$\chi(k,l) = \sum_{q=1}^{Q} \varpi(k-qt_k) \tag{29}$$

where $\varpi(k)$ is the STFT of the window function, t_k is the frequency point corresponding to the pitch period, q is the index of the harmonic band, and Q is the number of harmonic sub-bands. Here, the speech frequency is set to 8000 Hz, so $Q = 8000 * N/fs$.

Then, the original signal is fitted by multiple harmonics. The fitting error is calculated as:

$$\varepsilon(l) = \sum_{q=1}^{Q} \sum_{k=a_q}^{b_q} \left| X\left(k,l\right) - \eta_q(l)\chi\left(k,l\right) \right|^2 \tag{30}$$

Here, a_q and b_q represent the intervals of the harmonic band and $a_q = (q - 0.5)t_k$, $b_q = (q + 0.5)t_k$. Equating the derivative of (30) with respect to $\eta_q(l)$ to zero, we have

$$\eta_q(l) = \frac{\sum_{k=a_q}^{b_q} X(k,l)\chi^*(k,l)}{\sum_{k=a_q}^{b_q} |\chi(k,l)|^2} \tag{31}$$

Then, $\eta_q(l)$ is substituted into the eq. (30), and the fitting error between all the fitted harmonics and the actual harmonics is obtained.

Based on the smallest error $\eta_q(l)$ and the excitation spectrum $\chi(k, l)$, the harmonics of the current frame can be computed. The equation is expressed as:

$$\widetilde{X}(k,l) = \eta_q(l)\chi(k,l) \tag{32}$$

Because the fitted signal is distorted, it cannot be directly used as an enhanced signal. However, it can be used to estimate the probability of speech presence. Considering that the fitted harmonics with high energy have a high probability of speech presence, the log power spectrum of the fitted harmonics signal is calculated. Then, the linear mapping is performed to obtain the probability of the speech presence. The equation is:

$$p_{mbe}(k,l) = \frac{\ell|\widetilde{X}(k,l)| - \min\limits_{k}\left(\ell|\widetilde{X}(k,l)|\right)}{\max\limits_{k}\left(\ell|\widetilde{X}(k,l)|\right) - \min\limits_{k}\left(\ell|\widetilde{X}(k,l)|\right)} \tag{33}$$

Here, $\ell|\widetilde{X}(k,l)| = 20\log_{10}\left(\widetilde{X}(k,l)\right)$. Combined with the likelihood ratio, the harmonic-based speech presence probability is:

$$p_M(k,l) = \frac{p_{mbe}(k,l)\Delta(k,l)}{p_{mbe}(k,l)\Delta(k,l) + 1 - p_{mbe}(k,l)} \tag{34}$$

5 Speech enhancement method

Based on the probability of speech presence calculated by the above two algorithms, the joint probability is:

$$p(k,l) = \tau p_F(k,l) + (1-\tau)p_M(k,l) \tag{35}$$

Here, τ is a weighting factor and takes a value of 0.3. It means that the noise probability is more determined by the feature-based method. The noise spectrum of the current frame is updated in combination with the joint probability, the spectrum of the current frame and the noise spectrum of the previous frame. The equation is:

$$\widetilde{N}_F(k,l) = p(k,l)X(k,l) + (1-p(k,l))N_F(k,l-1) \tag{36}$$

Smoothing with the noise spectrum of the previous frame, the updated noise spectrum is:

$$N_F(k,l) = \begin{cases} \xi_N N_F(k,l-1) + (1-\xi_N)\widetilde{N}_F(k,l), p(k,l) < T_N \\ \xi_X N_F(k,l-1) + (1-\xi_X)\widetilde{N}_F(k,l), others \end{cases} \tag{37}$$

Here, T_N is the probability threshold and its value is 0.3. When the probability is less than the threshold, the probability that the current frame is noise is large. Then, the noise smoothing coefficient ξ_N (0.9) is relatively smaller than the speech smoothing coefficient ξ_X (0.99). After updating the noise, a posteriori SNR $\widehat{\delta}_l(k)$ and a priori SNR $\widehat{\rho}_l(k)$ can be updated according to the eq. (12) and (14).

Based on the OM-LSA estimator, the enhanced signal $\widehat{X}(k,l)$ is obtained as:

$$\widehat{X}(k,l) = X(k,l)(G_{H_1}(k,l))^{p(k,l)}(G_{\min}(k,l))^{(1-p(k,l))} \tag{38}$$

where $G_{\min}(k,l)$ represents the maximum coefficient of noise suppression and is 0.1. $G_{H_1}(k,l)$ is the gain function and its definition [5] is:

$$G_{H_1}(k,l) = \frac{\widehat{\rho}_l(k)}{\widehat{\rho}_l(k)+1} \exp\left\{ \frac{1}{2} \int_{\frac{\widehat{\delta}_l(k)\widehat{\rho}_l(k)}{\widehat{\rho}_l(k)+1}}^{\infty} \frac{e^{-t}}{t} dt \right\} \tag{39}$$

6 Experimental simulation and analysis

6.1 Experiment setup

To evaluate the performance of the algorithm, the Webrtc noise reduction algorithm [17], OM-LSA algorithm [6], RNN-based algorithm [37] and the proposed algorithm are compared. The reasons are: 1) OM-LSA is considered to be the best algorithm in traditional single-channel speech enhancement; 2) Webrtc algorithm is an open-source algorithm of Google, which has excellent noise reduction performance and real-time performance, which makes it widely used in many audio and video products; 3) Compared with others algorithm based on the deep learning network, the RNN-based algorithm [37] has better real-time performance and can be transplanted to the low-power DSP chip. Most of the parameters in the proposed algorithm are empirical values and tested by experiments.

Performance test indicators include segmented SNR(segSNR), the perceptual evaluation of speech quality(PESQ) [15], the short-time objective intelligibility (STOI) [33] and overall quality assessment index (OQAI) [15]. Here, the PESQ [15], STOI [33] and segSNR [15] are used to evaluate the quality, intelligibility and SNR of the enhanced speech, respectively. Among these indexes, the segSNR is calculated based on the geometric average of SNR of all speech frames. It is also an objective indicator used to measure speech quality. The PESQ is the subjective voice quality assessment indicator recommended by the ITU-T (International Telecommunication Union Telecommunication Standardization Department), and its score range is −0.5 to 4.5. It compares the original (input) signal with the aligned degraded signal using a perceptual model. Then, two error parameters are computed in this model and combined to give an objective listening quality. The STOI is used to measure the intelligibility of speech and its score range is 0 ~ 1. It is computed based on a correlation coefficient between the temporal envelopes of the clean and degraded speech in short-time and overlapping segments. For these three indicators, the higher the value is, the better the speech quality. The OQAI is based on the comprehensive calculation of speech distortion and noise distortion, which uses the scale of the mean opinion score (MOS).

The experimental noise is selected from 15 noises of the NoiseX-92 [38] database and 100 environmental noises [16]. The selected speech files are from the TIMIT database, including 500 utterances. All speech samples and noises are resampled to 16 kHz. The noise signal is added on the speech with different SNR and the range is from −10 to 10 dB.

The test set of the RNN-based algorithm is the same as that of the other three algorithms. However, to test the generalization of the RNN-based algorithm, the speech and the noise samples of the train set are not completely consistent with those of the test set. The clean speech comes from the TSP speech database [37] and the Implementation Summary of

Mandarin Chinese Test [43]. Half of the samples (722 sentences) in the TSP speech database are used to generate training data; for the test CD, 30 paragraphs are selected, and 651 speech segments are generated. All noises in the NOISEX-92 database and half of the environmental noises [16] are used to build a training set. Here, six kinds of transient noise are not included. They are MachineGun in Noise92X database and applause, keyboard sound, dialing sound, footsteps and laughter in the environmental noise database. The noise SNR is from - 10 dB to 10 dB.

6.2 The suppression of the stationary noise

The experiment uses pink noise as a stationary noise to verify the suppression of stationary noise. The SNR of the signal is 0 dB. Figure 2 (a) is a clean speech, Fig. 2 (b) is a speech with 0 dB pink noise, Fig. 2 (c) is enhanced speech without harmonic estimation and Fig. 2 (d) is enhanced speech of the proposed algorithm. As shown in Fig. 2 (b), in the case of low SNR, the spectrum of clean speech has been almost submerged by pink noise. Comparing Fig. 2 (c) with Fig. 2 (d), both algorithms are better for stationary noise reduction. However, from the figure, the proposed algorithm has a lighter spectrum color, indicating that the effect of noise suppression is better. From the performance indicators of the proposed algorithm, segSNR increased from −3.05 to −2.92, PESQ increased from 2.18 to 2.30, OQAI increased from 2.11 to 2.29 and STOI increased from 0.5208 to 0.5243.

6.3 The suppression of the transient noise

To evaluate the ability to suppress transient noise, the experiment sample is the noisy speech with 0 dB keyboard noise. The enhancement signal is shown in Fig. 3. It can be seen from the figure that the proposed algorithm eliminates the most noise components and its suppression effect is significantly higher than that without harmonic estimation. From the performance indicators of the proposed algorithm, segSNR increased from −5.35 to −2.91, PESQ increased from 1.55 to 1.73, OQAI increased from 1.61 to 1.95 and STOI increased from 0.6163 to 0.6382.

6.4 Overall performance evaluation

To comprehensively evaluate the performance of the algorithm, the experiment comprehensively compared the ability of the four algorithms on 115 kinds of noise. The experimental results are shown in Table 1, in which 6 kinds of transient noise include MachineGun in Noise92X database and applause, keyboard sound, dialing sound, footsteps and laughter in the environmental noise database. The selected transient noise is characterized by a short duration of the time domain. For example, the 89th laughter sample of the 100 kinds of environmental noise is selected instead of the 90th laughter sample.

As can be seen from the table, among the four algorithms, the RNN-based algorithm has the best performance for stationary noise suppression and the proposed algorithm has the best performance for the transient noise suppression. It is shown that the algorithm based on deep learning can obtain superior performance through effective training, but the generalization needs to be further improved. Compared with the other three algorithms, the mean OQAI and STOI of the proposed algorithm are better at high SNR, and the mean segSNR of the OM-LSA algorithm is better. At low SNR, the mean PESQ of the proposed algorithm is better. It can be

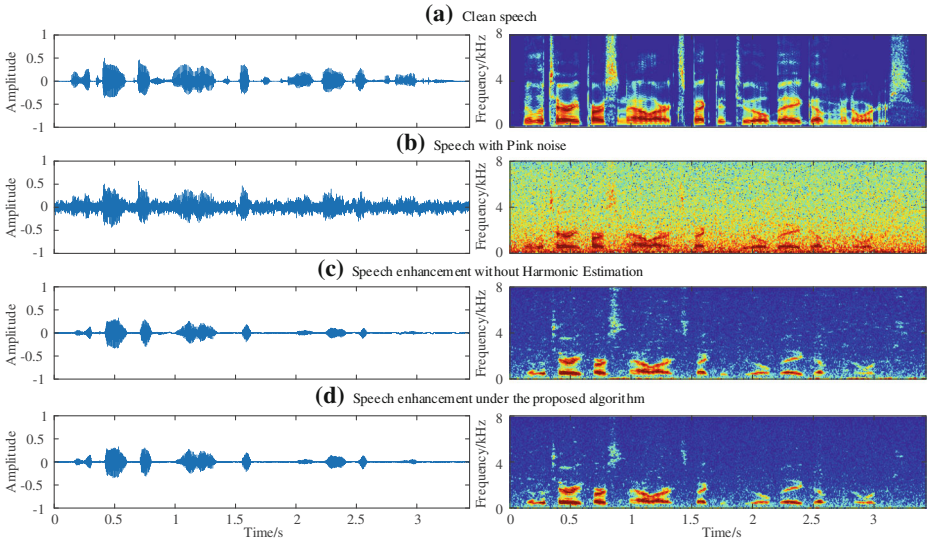


Fig. 2 The suppression of the stationary noise

seen that although OM-LSA has the best segSNR, it also has some damage to the speech quality, so other indicators are not good. Relatively speaking, the mean segSNR of the proposed algorithm is not good, but the speech quality is better.

Besides, the configuration of the experimental computer has Intel I7–7700 CPU and 8G memory. Processing the one second speech, the Webrtc algorithm, the OM-LSA [6] algorithm, the RNN-based algorithm and the proposed algorithm need 20 ms, 1.4 s, 47 ms and 45 ms, respectively. It can be seen that OM-LSA has the worst real-time performance.

Comparing the three real-time algorithms, the proposed algorithm is the best while the RNN-based algorithm is the worst for the transient noise suppression. The reason is maybe that

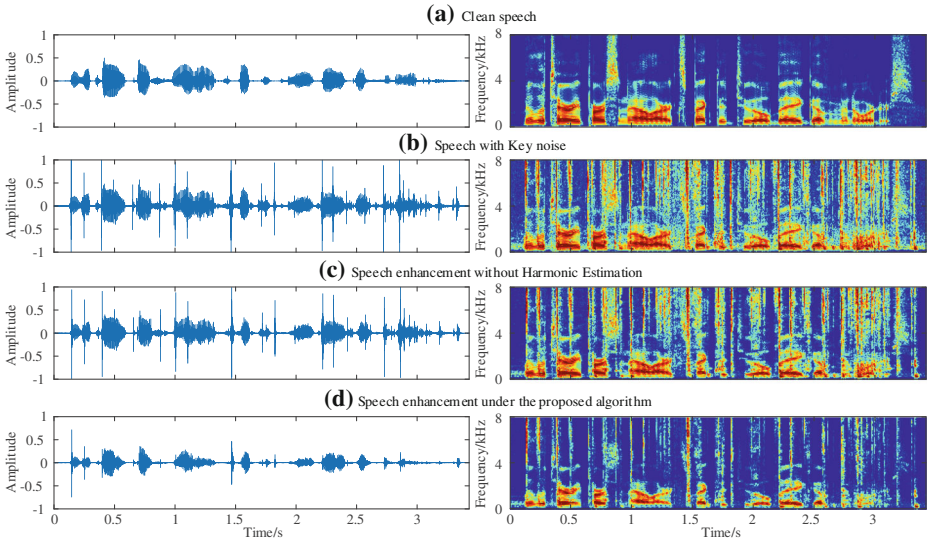


Fig. 3 The suppression of the transient noise

Table 1 Overall performance evaluation

algorithm		**Webrtc algorithm** [17]				**OM-LSA** [6]				**RNN-based algorithm** [37]				**Proposed algorithm**			
	SNR	**PESQ**	**segSNR**	**OQAI**	**STOI**	**PESQ**	**segSNR**	**OQAI**	**STOI**	**PESQ**	**segSNR**	**OQAI**	**STOI**	**PESQ**	**segSNR**	**OQAI**	**STOI**
Mean performance	−10	1.1	−1.49	1.33	0.51	1.02	0.03	1.34	0.48	1.19	**2.93**	**1.56**	**0.60**	**1.19**	−2.1	1.36	0.51
	−5	1.46	−0.02	1.59	0.6	1.5	2.27	1.69	0.59	1.50	**4.21**	**1.85**	**0.71**	**1.5**	−1.07	1.67	0.6
	0	1.88	2.27	2	0.67	1.97	5.05	**2.24**	0.69	**2.00**	**5.75**	2.22	**0.75**	1.82	0.38	2.21	0.69
	5	2.34	5.57	2.53	0.74	**2.38**	**7.85**	2.76	0.77	2.29	7.42	2.70	0.77	2.2	3.58	**2.76**	**0.77**
	10	2.75	8.93	3.02	0.81	**2.77**	**10.48**	3.22	0.83	2.65	9.16	3.11	0.84	2.61	7.06	**3.26**	**0.84**
Stationary noise	−10	1.09	−1.97	1.29	0.49	1	−0.36	1.29	0.47	**1.45**	**3.32**	**1.74**	**0.62**	1.18	−2.44	1.32	0.5
	−5	1.45	−0.58	1.54	0.58	1.49	1.94	1.65	0.58	**1.74**	**4.59**	**1.90**	**0.70**	1.49	−1.51	1.61	0.59
	0	1.87	1.68	1.96	0.66	1.97	4.8	2.21	0.68	**2.09**	**6.12**	**2.33**	**0.78**	1.81	−0.15	2.16	0.67
	5	2.33	5.02	2.48	0.73	2.39	7.67	2.73	0.76	**2.40**	**7.76**	2.73	**0.84**	2.19	3.1	**2.73**	0.76
	10	2.75	8.45	2.99	0.81	**2.78**	**10.37**	3.21	0.83	2.76	9.37	3.19	**0.88**	2.6	6.67	**3.24**	0.83
Transient noise	−10	1.34	2.59	1.92	0.7	**1.36**	**5.42**	**2.04**	**0.7**	1.13	2.54	1.55	0.59	1.35	5.2	1.99	0.69
	−5	1.64	5.07	2.28	0.76	1.62	6.95	2.38	0.76	1.47	3.84	1.80	0.70	**1.67**	**7.8**	**2.44**	**0.77**
	0	1.97	7.76	2.69	0.82	1.93	8.55	2.77	0.82	1.98	5.40	2.20	0.75	**2.04**	**10.5**	**2.89**	**0.84**
	5	2.34	10.3	3.14	0.87	2.31	10.33	3.11	0.86	2.23	7.10	2.68	0.77	**2.41**	**13.2**	**3.2**	**0.89**
	10	2.78	12.6	3.42	0.91	2.64	12.04	3.49	0.89	2.63	8.78	3.10	0.83	**2.76**	**15.7**	**3.6**	**0.93**
Runtime (1 s data)		20 ms				1400 ms				47 ms				45 ms			

six kinds of transient noise are not included in the training set of the RNN-based algorithm. Compared with the RNN-based algorithm in the transient noise suppression, the OQAI, the STOI and the PESQ increased by 24.6%, 13.2% and 8.4% respectively. Besides, compared with the Webrtc-based algorithm in transient noise suppression, these three indicators increased by 5%, 1.5% and 1.6%, respectively.

6.5 Real-time performance evaluation of the algorithm

To evaluate the performance of the two real-time algorithms, they were transplanted to the TI TMS320C6748 chip for evaluation. The main frequency of the chip is 456 MHz. The real-time algorithm takes 10 ms as one frame. The real-time test in the DSP chip showed the proposed algorithm and the Webrtc algorithm processed one frame in about 4.3 ms and 2.2 ms, respectively. So, the real-time requirements of these two algorithms have been met.

The actual waveform is shown in Fig. 4. The above is the signal of the proposed algorithm and below is that of the Webrtc algorithm. The whole signal is divided into three parts. The first part is the silent segment and includes mainly the background noise. The second part is the speech segment containing the keyboard sound and the third part is only the keyboard sound.

From the figure, the root mean square of the signal amplitude of the proposed algorithm decreases by 8.46 dB, 6.33 dB and 17.74 dB, respectively. In particular, the keyboard sound is eliminated.

7 Conclusions

To improve the ability to eliminate transient noise while ensuring eliminating stationary noise, an improved real-time single-channel speech enhancement algorithm is proposed. Based on the stationary noise estimation, the algorithm combines the transient noise detection algorithm to correct the noise spectrum. At the same time, to improve the accuracy of speech probability estimation, the algorithm proposes the method of combining speech features and harmonic analysis. Through weight adjustment, the algorithm not only improves the speech quality but also reduces the speech distortion caused by the misestimated speech presence probability.

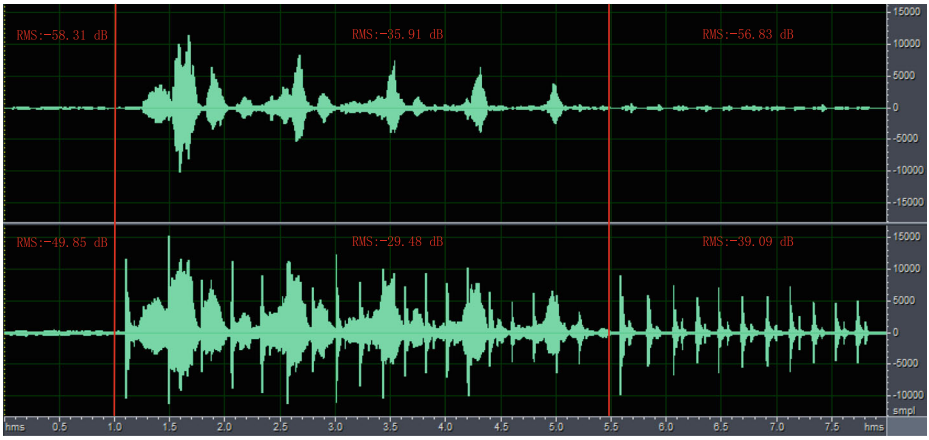


Fig. 4 Comparison of the real-time performance of the algorithm

Finally, based on the OM-LSA estimator, transient noise is effectively suppressed. The simulation experiment and hardware measurement results show that the improved algorithm has the best overall performance and can work effectively in real-time.

From the perspective of the application, the proposed algorithm can be applied in some real-time and high speech quality applications, such as audio and video conference system, hearing aid, hands-free telephone and other voice communication terminal equipment. Future research is planned to be carried out in two aspects: 1) further improving the transient noise suppression capability without introducing speech distortion, such as simulating the human ear characteristic to estimate noise and speech. 2) studying and combining current supervised learning algorithms. The joint can not only utilize the stability and generalization of noise suppression of the classical algorithm, but also the suppression performance of the non-stationary noise (such as babble speech) of the supervised learning algorithm.

Acknowledgments The authors would like to thank the reviewers for their valuable suggestions and comments. And they also thank Mr. ChaoHe for his excellent work in algorithm design and programming. The work was supported in part by the National Key Research and Development Program of China under Grant 2020YFC2004003 and Grant 2020YFC2004002, the National Natural Science Foundation of China under Grant No. 62001215.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

References

1. Boll SF (1979) A spectral subtraction algorithm for suppression of acoustic noise in speech. Acoustics Speech & Signal Processing IEEE Transactions on 27(2):113–120
2. Brockwell P, Dahlhaus R (2004) Generalized Levinson–Durbin and burg algorithms. J Econ 118(1–2):129–149
3. Cappé O (1994) Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE transactions on Speech Audio Processing 2(2):345–349
4. Chen J, Wang D (2017) Long short-term memory for speaker generalization in supervised speech separation. The Journal of the Acoustical Society of America 141(6):4705–4714
5. Cohen I (2003) Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Transactions on Speech & Audio Processing 11(5):466–475
6. Cohen I, Berdugo B (2001) Speech enhancement for non-stationary noise environments. Signal Process 81(11):2403–2418
7. Dahl GE, Yu D, Deng L, Acero A (2012) Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans Audio Speech Lang Process 20(1):30–42
8. Ephraim Y, Malah D (1985) Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics Speech & Signal Processing 33(2):443–445
9. Ephraim Y, Malah D (2003) Speech enhancement using a minimum Mean-Square error log-spectral amplitude estimator. IEEE Transactions on Acoustics Speech & Signal Processing 32(6):1109–1121
10. Fu S, Tsao Y, Lu X, Kawai H (2017) Raw waveform-based speech enhancement by fully convolutional networks, in 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). 006–012.
11. Gao T, Du J, Dai L-R, Lee C-H (2016) SNR-Based Progressive Learning of Deep Neural Network for Speech Enhancement, in INTERSPEECH. 3713–3717.
12. Griffin DW, Lim JS (1988) Multiband excitation vocoder. IEEE Transactions on acoustics, speech, signal processing 36(8):1223–1235
13. Hinton GE, Salakhutdinov RR (2006) Reducing the dimensionality of data with neural networks. Science 313(5786):504–507

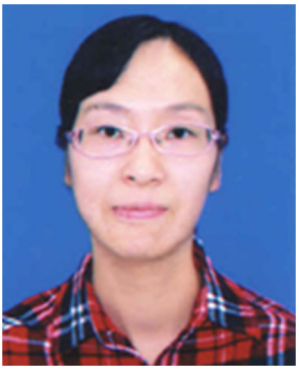
14. Hirszhorn A, Dov D, Talmon R, Cohen I (2012) Transient interference suppression in speech signals based on the OM-LSA algorithm, in IWAENC 2012; International Workshop on Acoustic Signal Enhancement. , VDE. 1–4.
15. Hu Y, Loizou PC (2008) Evaluation of objective quality measures for speech enhancement. IEEE Transactions on Audio, Speech & Language Processing 16(1):229–238
16. Hu G, Wang DL (2010) A tandem algorithm for pitch estimation and voiced speech segregation. IEEE Transactions on Audio Speech Language Processing 18(8):2067–2079
17. Inc. G. WebRTC. https://webrtc.org.cn/mirror/
18. Kennedy D, Corrsin S (1961) Spectral flatness factor and 'intermittency'in turbulence and in non-linear noise. J Fluid Mech 10(3):366–370
19. Kim M, Smaragdis P (2015) Adaptive denoising autoencoders: A fine-tuning scheme to learn from test mixtures, in International Conference on Latent Variable Analysis and Signal Separation. Springer. 100–107.
20. Kumar B (2018) Comparative performance evaluation of MMSE-based speech enhancement techniques through simulation and real-time implementation. International Journal of Speech Technology 21(4):1033–1044
21. Leng L, Zhang J, Xu J, Khan K, Alghathbar K (2010) Dynamic weighted discrimination power analysis: a novel approach for face and palmprint recognition in DCT domain. International Journal of Physical Sciences 5(17):467–471
22. Leng L, Li M, Kim C, Bi X (2017) Dual-source discrimination power analysis for multi-instance contactless palmprint recognition. Multimed Tools Appl 76(1):333–354
23. Li J, Wang S, Peng R, Zheng C, Li X (2014) Transient noise reduction based on speech reconstruction, in the 21st international congress on sound and vibration, Beijing/China. 1–8.
24. Lu X, Tsao Y, Matsuda S, Hori C (2013) Speech enhancement based on deep denoising autoencoder, in Interspeech. 436–440.
25. Manohar K, Rao P (2006) Speech enhancement in nonstationary noise environments using noise properties. Speech Comm 48(1):96–109
26. Michelsanti D, Tan Z-H (2017) Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification, in Proc. INTERSPEECH. 2008–2012.
27. Nongpiur RC (2008) Impulse noise removal in speech using wavelets, in 2008 IEEE international conference on acoustics, speech and signal processing. IEEE. 1593–1596.
28. Pandey A, Wang D (2019) TCNN: Temporal Convolutional Neural Network for Real-time Speech Enhancement in the Time Domain, in 44th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2019, May 12, 2019 - May 17, 2019. Institute of Electrical and Electronics Engineers Inc.: Brighton, United kingdom. 6875–6879.
29. Pandey A, Wang D (2020) Densely Connected Neural Network with Dilated Convolutions for Real-Time Speech Enhancement in The Time Domain, in ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 6629–6633.
30. Pascual S, Bonafonte A, Serra J (2017) SEGAN: Speech enhancement generative adversarial network, in Proc. INTERSPEECH. 3642–3646.
31. Rosen S (1992) Temporal information in speech: acoustic, auditory and linguistic aspects. Philosophical Transactions of The Royal Society B Biological Sciences 336(1278):367–373
32. Stahl V, Fischer A, Bippus R (2000) Quantile based noise estimation for spectral subtraction and Wiener filtering, in 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100). IEEE. 1875-1878.
33. Taal CH, Hendriks RC, Heusdens R, Jensen J (2011) An algorithm for intelligibility prediction of time–frequency weighted Noisy speech. IEEE Transactions on Audio Speech Language Processing 19(7):2125–2136
34. Takeuchi D, Yatabe K, Koizumi Y, Oikawa Y, Harada N (2020) Real-Time Speech Enhancement Using Equilibriated RNN, in ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 851–855.
35. Tan K, Wang D (2018) A convolutional recurrent neural network for real-time speech enhancement, in 19th Annual Conference of the International Speech Communication, INTERSPEECH 2018, September 2, 2018 - September 6, 2018. International speech communication association: Hyderabad, India. 3229-3233.
36. Tan K, Zhang X, Wang D (2019) Real-time Speech Enhancement Using an Efficient Convolutional Recurrent Network for Dual-microphone Mobile Phones in Close-talk Scenarios, in ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 5751–5755.
37. Valin J-M (2018) A hybrid DSP/deep learning approach to real-time full-band speech enhancement, in 20th IEEE international workshop on multimedia signal processing, MMSP 2018, august 29, 2018 - august 31, 2018. Institute of Electrical and Electronics Engineers Inc.: Vancouver, BC, Canada.

38. Varga A, Steeneken HJM (1993) Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Comm 12(3): 247–251
39. Wang DL (2017) Deep learning reinvents the hearing aid. IEEE Spectr 54(3):32–37
40. Wang D, Chen J (2018) Supervised speech separation based on deep learning: an overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing 26(10):1702–1726
41. Weninger F, Hershey J R, Le Roux J, Schuller B (2014) Discriminatively trained recurrent neural networks for single-channel speech separation, in proceedings 2nd IEEE global conference on signal and information processing, GlobalSIP, Machine Learning Applications in Speech Processing Symposium, Atlanta, GA, USA.
42. Weninger F, Erdogan H, Watanabe S, Vincent E, Le Roux J, Hershey J R, Schuller B (2015) Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR, in International Conference on Latent Variable Analysis and Signal Separation. Springer. 91–99.
43. Xishuang Y, Zhaoxiong L (2004) Implementation Summary of Mandarin Chinese Test.the commercial press
44. Xu Y (2015) Research on deep neural network based speech enhancement. University of Science and Technology of China
45. Xu Y, Du J, Dai L-R, Lee C-H (2014) An experimental study on speech enhancement based on deep neural networks. IEEE Signal Processing Letters 21(1):65–68
46. Zheng C, Chen X, Wang S, Peng R, Li X (2013) Delayless method to suppress transient noise using speech properties and spectral coherence, in audio engineering society convention 135. Audio Engineering Society
47. Zheng C, Yang H, Li X (2014) On generalized auto-spectral coherence function and its applications to signal detection. IEEE Signal Processing Letters 21(5):559–563

Ruiyu Liang received the Ph.D. degree from Southeast University (China) in 2012. He is currently an Associate Professor in Nanjing Institute of Technology. His major interests include speech enhancement and hearing aid signal processing.

Yue Xie is a doctoral student in the school of information and communication engineering from Southeast University, China. His research interests include speech signal processing and deep learning.

Jiaming Cheng is currently working towardthe Ph.D. degree from Southeast University, China. His research interests include speech enhancement and machine learning.



Guichen Tang received Mater degree in Communication and Information System from Hohai University (China) in 2004. Her research interest is speech signal processing.

Shinuo Sun is a graduate student in signal and information processing at Southeast University. His research interests include speech enhancement and speech signal processing.