A Speech Denoising Algorithm Based on Harmonic Regeneration

To cite this article: Liu Zengyuan and Dong Anming 2019 *IOP Conf. Ser.: Earth Environ. Sci.* **332** 022042

A Speech Denoising Algorithm Based on Harmonic Regeneration

Liu Zengyuan[1] Dong Anming[2]

[1]Taizhou Institute of Science and Technology, Nanjing University of Science and Technology, Taizhou,Jiangsu, 225300 ,China

[2]School of Information Science and Engineering of Lanzhou University, Lanzhou, 730000，China

Corresponding author's e-mail: 840879885@qq.com

Abstract. Single channel speech enhancement in noisy environments is studied in this paper. Traditional speech enhancement methods that based on short time spectral analysis frequently introduce harmonic distortion due to imprecise noise spectral estimation. A kind of technology which is called harmonic regeneration is used to overcome the defect of traditional speech enhancement methods, and damaged harmonics are restored. Firstly, a non-linear operation is used on the harmonic distorted signal to generate an artificial signal which contains all of harmonics of pure speech signal. Secondly, the artificial signal is used to correct the traditional noise reduction gain function, and the purpose of harmonic reservation is reached. Theoretical analysis, objective and subjective results indicate that the harmonic regeneration algorithm can improve speech quality dramatically. The computational complexity of the new algorithm is rather low, and it can be easily implemented in real time applications.

1. Introduction

The traditional single-channel speech enhancement technology based on short-time spectral analysis relies on the estimation of the short-time noise suppression gain function, which is a function of the signal-to-noise ratio (SNR) estimation at each frequency point in the frequency domain, Therefore, the performance of the noise reduction technology depends on the quality of the SNR estimation. Voiced sound is characterized by its strong periodicity and its frequency domain is characterized by harmonic peaks [1]. One of the main drawbacks of traditional speech enhancement technology is that harmonic distortion will occur when the SNR is low [2-3]. Harmonic distortion mainly occurs in the frequency components with low SNR. When denoising according to the size of SNR, the high-frequency components with low SNR will be treated as noise to suppress, resulting in the loss of high-frequency components,so it will cause the quality of enhanced speech signal to decline.

In order to overcome the disadvantage of traditional speech enhancement method which causes distortion of harmonics, this paper introduces harmonic reconstruction technology to improve the traditional speech enhancement method. Firstly, the harmonic characteristic of speech signal is used to process the speech signal with harmonic distortion obtained by traditional enhancement method, and all the harmonic components of pure speech are obtained; Then, a new noise reduction gain which can retain the harmonic components is calculated by using the artificial reconstruction signal, and then the noise reduction gain is used to filter the noisy speech to obtain the noise reduction speech which retains the harmonic characteristics of the speech.

2. Traditional speech enhancement methods

In the classical additive noise model, the noisy speech signal is represented as:

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

In formula (1), $s(t)$ and $n(t)$ represent speech signal and noise signal respectively. Let $S(\lambda,k)$、$N(\lambda,k)$ and $X(\lambda,k)$ represent the k-th spectral component of the λ-th frame of speech signal $s(t)$, noise $n(t)$ and noisy speech $x(t)$ respectively. The process of noise reduction is to weigh the noise speech spectrum $X(\lambda,k)$ by using the spectral gain function $H(\lambda,k)$. The calculation of spectral gain requires estimating two parameters, a prior SNR and a posterior SNR. The priori SNR is defined as:

$$SNR_{prio} = \frac{E\left\{|S(\lambda,k)|^2\right\}}{E\left\{|N(\lambda,k)|^2\right\}} \tag{2}$$

The posterior SNR is defined as:

$$SNR_{post} = \frac{|X(\lambda,k)|^2}{E\left\{|N(\lambda,k)|^2\right\}} \tag{3}$$

The key of SNR estimation is noise spectrum estimation. There are many noise spectrum estimation methods, such as noise power spectrum estimation in silent area and minimum statistical noise power spectrum estimation,etc. Assuming that the estimated noise power spectral density is $|\hat{N}(\lambda,k)|^2$, the estimation of posterior SNR is given by the following formula:

$$\hat{SNR}_{post}(\lambda,k) = \frac{|X(\lambda,k)|^2}{|\hat{N}(\lambda,k)|^2} \tag{4}$$

A decision-making method can be used to estimate the prior SNR[4]:

$$\hat{SNR}_{prio}(\lambda,k) = \beta\frac{|H(\lambda-1,k)X(\lambda-1,k)|^2}{|\hat{N}(\lambda,k)|^2} + (1-\beta)P\left[\hat{SNR}_{post}(\lambda,k)-1\right] \tag{5}$$

In formula (5), $P[\cdot]$ is a half-wave rectification function, β is a constant factor, $0<\beta<1$.

The weighted gain is independent of the prior SNR and the posterior SNR.

$$H(\lambda,k) = h(\hat{SNR}_{prio}(\lambda,k), \hat{SNR}_{post}(\lambda,k)) \tag{6}$$

The gain function $h(\cdot)$ varies according to different criteria, such as spectral subtraction, Wiener filtering and MMSE,etc.The estimated speech signal spectrum is

$$\hat{S}(\lambda,k) = H(\lambda,k)X(\lambda,k) \tag{7}$$

From the above analysis, it can be concluded that we can construct a new noise suppression gain function $H_{harmo}(\lambda,k)$ to instead of $H(\lambda,k)$ in formula (7),which can retain the harmonic components, thus avoiding distortion.

3. Harmonic reconstruction

The harmonic component of the output signal $\hat{S}(\lambda,k)$ using the noise reduction method adopted above are distorted. In order to preserve the damaged distortion of harmonics, the principle of harmonic reconstruction is given below.

3.1. Harmonic Reconstruction Principle

Firstly, the speech signal obtained by traditional speech enhancement method (such as absolute value, maximum value, minimum value and threshold value) is processed in time domain by non-linear(*NL*) operation. The reconstructed signal $s_{harmo}(t)$ is given by the following formula.

$$s_{harmo}(t) = NL(\hat{s}(t)) \tag{8}$$

From the analysis later in this paper, we can see that the signal $s_{harmo}(t)$ has the same harmonic component as the pure speech. Using this signal to construct a new noise reduction gain function

$$H_{harmo}(\lambda,k) = h'(\hat{SNR}_{harmo}(\lambda,k), \hat{SNR}_{post}(\lambda,k)) \tag{9}$$

The function $h'(\cdot)$ may be any gain function proposed in the previous section. $\hat{SNR}_{harmo}(\lambda,k)$ in formula (9) is the estimation of prior SNR with harmonic information preserved.

$$\hat{SNR}_{harmo}(\lambda,k)=\frac{\rho\,|\hat{S}(\lambda,k)|^2+(1-\rho)\,|S_{harmo}(\lambda,k)|^2}{|\hat{N}(\lambda,k)|^2} \tag{10}$$

Parametric ρ （$0\le\rho\le1$） is used to control the degree of mixing of $|\hat{S}(\lambda,k)|^2$ and $|S_{harmo}(\lambda,k)|^2$ to offset the deviation caused by $|S_{harmo}(\lambda,k)|^2$ in magnitude. The final harmonic reconstruction speech enhancement signal is:

$$\hat{S}(\lambda,k)=H_{harmo}(\lambda,k)X(\lambda,k) \tag{11}$$

The noise suppression gain function $H_{harmo}(\lambda,k)$ can retain the harmonic components suppressed by other methods, thus avoiding distortion.

3.2. Harmonic Reconstruction Analysis

In order to analyze the steps of harmonic reconstruction, a special non-linear function $Max(a,0)$ is studied. If $a>0$, the function returns a, otherwise returns. 0 The function $NL(\cdot)$ in formula (8) is replaced by function $Max(\cdot)$:

$$s_{harmo}(t)=Max(\hat{s}(t),0)=\hat{s}(t)\,p(\hat{s}(t)) \tag{12}$$

The function $p(t)$ is defined as:

$$p(t)=\begin{cases}1, & t>0\\ 0, & t<0\end{cases} \tag{13}$$

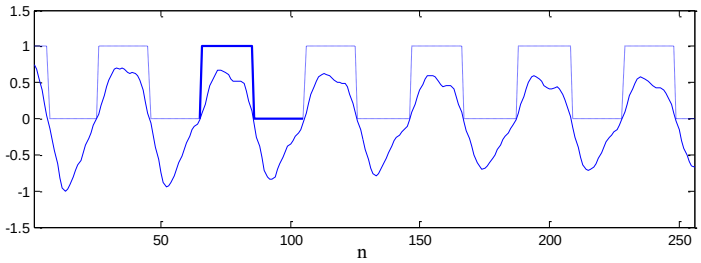

Figure 1. Harmonic reconstruction

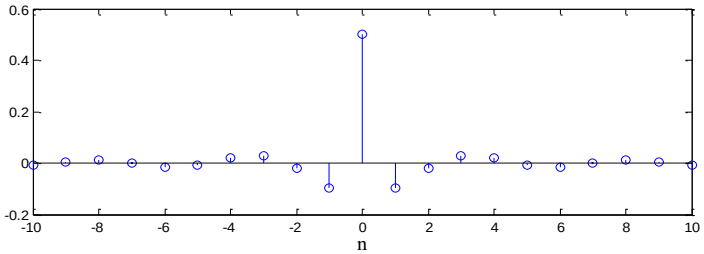

Figure 2. Spectrum of $p(\hat{s}(t))$

Figure 1 shows the waveforms of a frame of voiced enhancement signal $\hat{s}(t)$ (fine line) and corresponding $p(\hat{s}(t))$ (dotted line). In Figure 1, it can be found that the signal is composed of the basic unit waveform $p(\hat{s}(t))$ (heavy line) repeated appearances. The period of repetition of the basic unit is recorded as *T*, which is the same as the pitch period of the voiced sound. Assuming that speech is considered stationary in the duration of a frame, $p(\hat{s}(t))$ can be considered as a periodic signal whose Fourier transform is a discrete sampling pulse with $1/T$ interval.

$$\mathcal{F}\left[p(\hat{s}(t))\right]=\frac{2\pi}{T}\sum_{n=-\infty}^{+\infty}R(n)\delta(\omega-n\frac{2\pi}{T}) \tag{14}$$

Among them, the basic unit waveform Fourier transform is sampled periodically[5-7]. According to Formula (12), the Fourier transform is

$$\mathcal{F}\left[s_{harmo}(t)\right] = \mathcal{F}\left[\hat{s}(t)\right] * \frac{1}{T}\sum_{n=-\infty}^{+\infty} R(n)\delta(\omega - n\frac{2\pi}{T}) \tag{15}$$

Formula (15) shows that the spectrum $\hat{s}(t)$ of reconstructed signal $s_{harmo}(t)$ is convoluted with that of a harmonic comb filter. The comb filter has the same pitch rate as the processed speech signal, so it has the function of harmonic reconstruction. As shown in Figure 2, the spectrum amplitude of $p(\hat{s}(t))$ is symmetrical with respect to $n = 0$ and decreases rapidly with the increase of $|n|$. Therefore, only a few harmonic information adjacent to the lost harmonic can be used to reconstruct the lost harmonic signal.

The effect of the above algorithm on the unvoiced is discussed below. First, consider a mixed signal of unvoiced and voiced sound. It is assumed that the first half of the signal spectrum is the spectrum of voiced sound and the second half is the spectrum of unvoiced sound. The spectrum of $p(\hat{s}(t))$ is still comb structure, and its fundamental frequency is determined by the fundamental frequency of the voiced part. For the voiced part, the result is the same as formula (15). Because the envelope of comb harmonic filter decreases rapidly, the influence of harmonic components on the spectrum of the unvoiced part decreases rapidly. It can be found that the result of convolution of the spectrum of the unvoiced part and the spectrum of $p(\hat{s}(t))$ is still the unvoiced spectrum, and the harmonics of the voiced sound will not have a great influence on the reconstruction of the unvoiced part.

When the whole frame of voice is unvoiced, because the time domain waveform of unvoiced is similar to white noise, the shape of $p(\hat{s}(t))$ is irregular, and its spectrum does not have comb harmonic structure. As shown in Formula (15), it is concluded that the result of the convolution of the unvoiced spectrum and the spectrum of the uncertain shape is still the unvoiced spectrum.

In a word, the quality of the unvoiced part of speech signal is not degraded by the harmonic reconstruction process, so the harmonic reconstruction process can be applied to all voiced parts without distinguishing unvoiced parts.

3.3. Determination of parameter ρ

Parametric ρ is used to control the degree of mixing of $|\hat{S}(\lambda,k)|^2$ and $|S_{harmo}(\lambda,k)|^2$. When the estimated signal spectrum $|\hat{S}(\lambda,k)|^2$ of the traditional enhancement algorithm is reliable, the harmonic reconstruction process is not necessary. The value of parameter ρ can be approximately 1 in formula (10). When the estimated signal $|\hat{S}(\lambda,k)|^2$ of the traditional enhancement algorithm is unreliable, the speech signal spectrum must be modified by the harmonic enhancement process, and the parameter ρ should be close to 0.

According to the above principle, the value of parameter ρ can be selected as a fixed value, so that the fixed value can achieve a compromise between $|\hat{S}(\lambda,k)|^2$ and $|S_{harmo}(\lambda,k)|^2$. However, considering the non-stationarity of the actual speech signal, this paper uses a time-frequency dependent parameter $\rho(\lambda,k)$ instead of a fixed parameter ρ:

$$\rho(\lambda,k) = \frac{SNR_{prio}(\lambda,k)}{1 + SNR_{prio}(\lambda,k)} \tag{16}$$

The time-frequency parameter $\rho(\lambda,k)$ defined above satisfies the principle of harmonic mixing well. When the priori SNR is high, it shows that the signal component is strong and the possibility of harmonic distortion is small, $\rho(\lambda,k)$ is close to 1; when the priori SNR is small, the signal component may be suppressed, so the process of harmonic reconstruction is needed, $\rho(\lambda,k)$ is close to 0.

4. Experiments and results analysis

In the experiment, both traditional noise reduction gain function $H(\lambda,k)$ and harmonic reconstruction noise reduction gain function $H_{harmo}(\lambda,k)$ are processed by Wiener filter function [8]. The algorithm is validated by using MATLAB software simulation and DSP hardware test. The 32-bit floating-point arithmetic processor TMS320C6713 produced by TI company is selected for hardware test. The development environment is CCStudio 3.3 and the development language is C language. The sampling frequency of voice is 8kHz, the sampling points of each frame are 256, and the overlap between frames is 128, Hamming window is used. Software simulation is used to add white noise to the pure speech signal with a SNR of 5 dB. Hardware simulation uses computer sound card to output to the DSP processing system.

Figure 3 illustrates the waveforms in time and frequency domain before and after processing a frame of noisy speech signal using traditional speech enhancement method and harmonic reconstruction speech enhancement method respectively. It can be seen from the figure that the high frequency harmonic component of the signal is completely concealed by the noise spectrum, and the high frequency component of the speech spectrum is almost completely suppressed after the enhancement of the noisy speech using the traditional Wiener filtering method. When the harmonic reconstruction method is used, the lost harmonic frequency components are recovered.

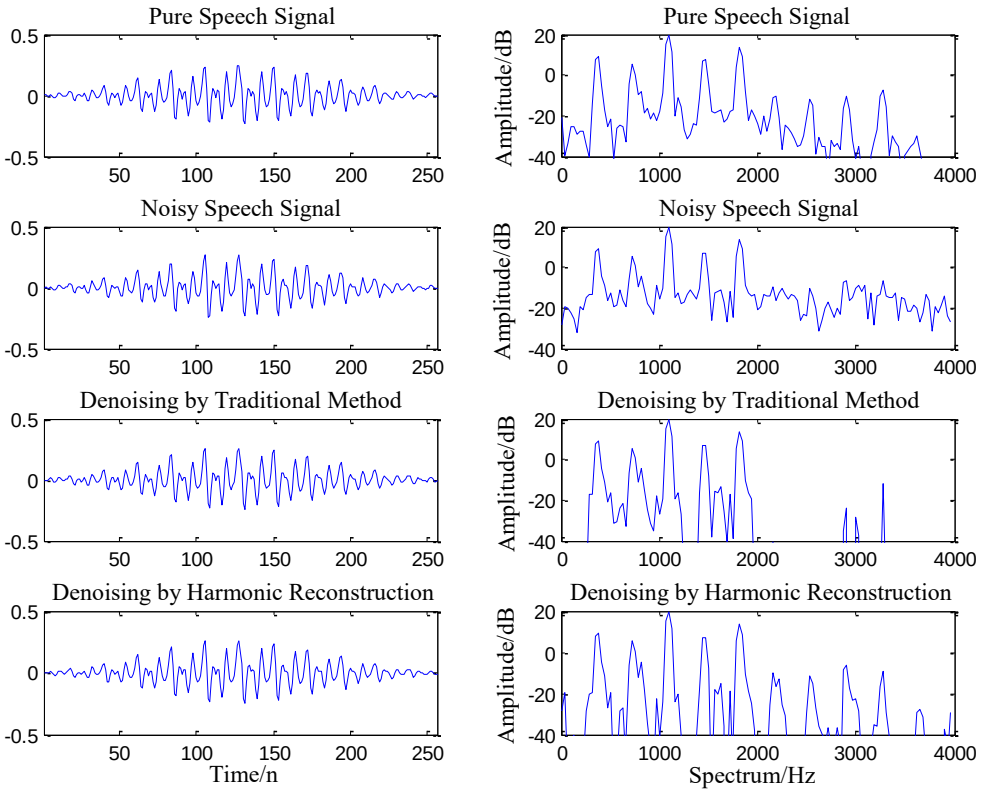

Figure 3. Time-frequency spectro graph after harmonic Reconstruction

Figure 4 shows the spectrum of traditional speech enhancement algorithm and harmonic reconstruction algorithm before and after enhancement. It can be seen from the spectrogram that the high-frequency harmonic component of the enhanced spectrogram is almost completely lost after the traditional method, while the harmonic component enhanced by the harmonic reconstruction method is well preserved.

In hardware implementation, the main frequency of DSP is set to 225MHz. The resource occupancy rate of this algorithm is 25%-30%, and the delay is 32 ms. For speech signals, the delay time of 32ms is within acceptable range, and the algorithm complexity is relatively low. Because C language is used for

programming and linear assembly language is not used, there is still much room to improve the performance of the algorithm in terms of the occupancy of DSP resources.

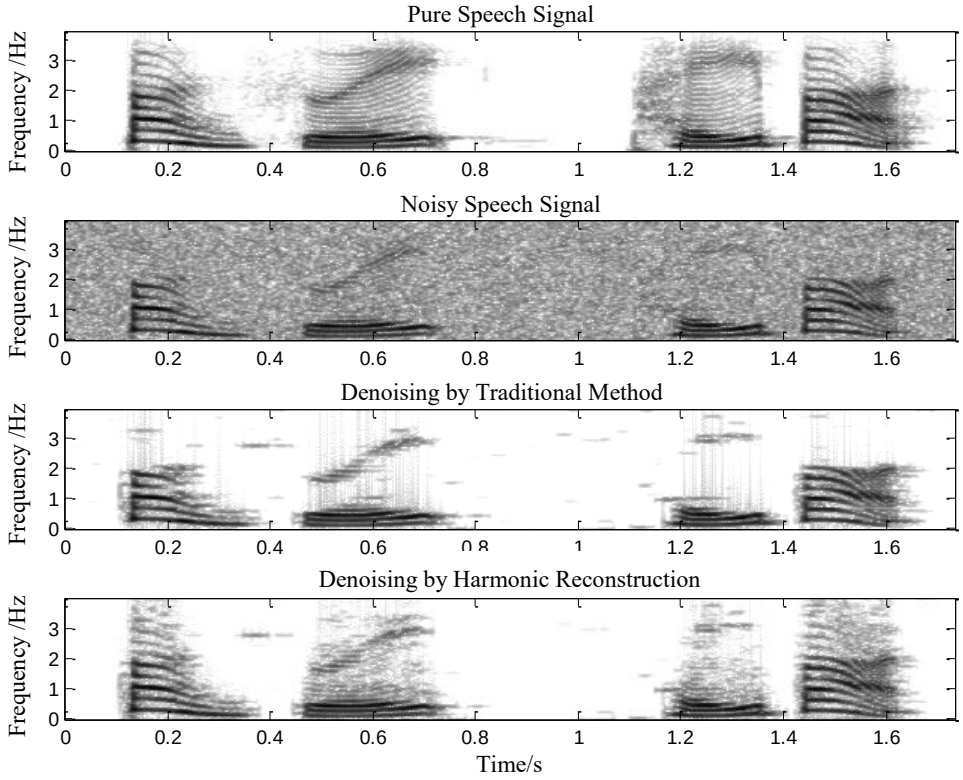

Figure 4. Spectrogram before and after harmonic reconstruction enhancement

5. Conclusion

In this paper, a harmonic reconstruction technique is used to improve the traditional speech enhancement algorithm. Traditional speech enhancement methods may cause great damage to the harmonics of speech signals. Harmonic reconstruction method can generate signals with all harmonic components by using non-linear operation, and use the generated signals to construct a noise reduction gain function that can retain the harmonic components of speech. Using this new noise reduction gain to enhance the noisy speech signal can effectively retain the harmonic components of the speech signal, thus improving the quality of the speech signal. At the same time, the algorithm of harmonic reconstruction has high computational efficiency and is easy to implement in real time.

References

[1] YagnXingjun. (1995) Digital Processing of Speech Signals. Publishing House of Electronics Industry,Peking
[2] Plapous Cyril; Marro Claude; Scalart Pascal. (2005) Speech enhancement using harmonic regeneration . In:IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '05. Philadelphia, United State , pp. I157-I160.
[3] Plapous Cyril; Marro Claude; Scalart Pascal. (2006) Improved Signal-to-Noise Ratio Estimation for Speech Enhancement. IEEE Transactions on Audio, Speech and Language Processing, Volume 14, Issue: 6,pp. 2098- 2108
[4] Y. Ephraim, and D. Malah, (1984) Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-32, No. 6, pp. 1109-1121.
[5] ZhengJunli.(2000) Signals and Systems. Higher Education Press, Peking.
[6] ChengPeiqing (1995) Digital Signal Processing Course.Tsinghua university press, Peking.

[7] HuGuangshu,(2003) Digital Signal Processing-Theory, Algorithms and Implementation. Tsinghua university press, Peking.
[8] J.S. Lim, A.V. Oppenheim. (1978) All-pole modelling of degraded speech. IEEE Transactions on Acoust, Speech and Signal Processing, Vol. 26, No.3: 197-210.