Modified Amplitude Spectral Estimator for Single-Channel Speech Enhancement

Zhenhui Zhai[1,b], Shifeng Ou[1,a], Ying Gao[1,c]

[1] School of Opto-electronic Information Science and Technology, Yantai University, Yantai, 264005, China

[a]email: ousfeng@126.com, [b]email:zhaizhenhui_2008@163.com, [c]email:claragaoying@126.com

Keywords: Speech Enhancement; Amplitude Spectral Estimation; Decision-Directed; Soft-Decision

Abstract. The widely used amplitude spectral subtraction for speech enhancement suffers from the unreasonable assumption that the phase spectra of clean speech and noise signal must being uniform. To avoid this problem, Zhu has presented an extensive analysis of amplitude spectral subtraction, and proposed a quantitative analysis based speech spectral recovery (QASSR) approach. However, the residual noise of QASSR approach is still highly perceptible due to the small suppression ratio in speech absent periods. In this paper, a modified algorithm for speech amplitude spectral estimation is proposed by incorporating the decision-directed (DD) method and soft decision technique. Firstly, the *a priori* SNR in each frame and frequency bin is estimated using the DD method, and then an over attenuation factor is derived from the soft decision technique to further suppress the residual noise in noise-only frames. The performance of our proposed algorithm is evaluated and compared with that of QASSR approach, and the simulation results showed the improved performance of the proposed algorithm in various noise conditions.

Introduction

The performance of most speech processing systems is degraded by various background noises. The easiest way to alleviate such interference of noise is to employ a speech enhancement technique, in which the noise is reduced as much as possible to recover the original clean speech. In the recent decades, a vast amount of speech enhancement algorithms have been proposed and developed. Such as singular value decomposition (SVD) or singular value decomposition (EVD)-based subspace method, Wiener filter method, minimum mean-square error (MMSE) spectral amplitude estimator approach and so on [1-6]. Among these speech enhancement methods, the most widely used algorithm is so-called amplitude spectral subtraction technique, which recovers the clean speech spectral by subtracting the noise spectral from noisy signal [7]. The traditional amplitude spectral subtraction works under the basic assumption that the phase spectra of noise signal and clean speech should being uniform. But many studies have shown that such assumption is commonly incorrect as the phase spectral of clean speech and noise are randomly distributed between $0 \sim 2\pi$. To avoid this unreasonable assumption, a QASSR method has been proposed to estimate the clean speech spectral based on a quantitative analysis for the effect of noise signal on speech spectra [8]. This method improves the speech enhancement performance effectively without any conditional assumption on the phase spectral. In the QASSR method, however, the minimum suppression ratio for the background noise is set to 0.25, which leads to much residual noise in the enhanced speech signals and thus degrades the performance of the speech enhancement system.

In this paper, we propose a modified algorithm to improve the performance of QASSR method by incorporating the DD and soft decision technique. Firstly, DD method is used to estimate the *a priori* SNR in each frame and frequency bin, and then an over attenuation factor is achieved by using the soft decision technique, which can further suppress the residual noise in speech absent periods. The results of experiments and simulations verify the superiority of our proposed approach compared to that QASSR method in different noise scenarios and SNR levels.

The Amplitude Spectral Subtraction and QASSR Approach

It is supposed that the noise signal is additive and let $y(t)$, $s(t)$ and $d(t)$ denote the noisy speech, clean speech and noise signals, respectively. Then, $y(t)$ can be written as

$$y(t)=s(t)+d(t) \tag{1}$$

Applying an N-point Discrete Fourier Transform (DFT) to the noisy speech signal, we can get

$$Y(k,m)=S(k,m)+D(k,m) \tag{2}$$

where m is the time frame index, and k is the frequency-bin index. The equation above can be expressed as the following polar form

$$\left|Y(k,m)\right|e^{j\theta_y(k,m)}=\left|S(k,m)\right|e^{j\theta_s(k,m)}+\left|D(k,m)\right|e^{j\theta_d(k,m)} \tag{3}$$

where, $|Y(k,m)|$, $|S(k,m)|$, and $|D(k,m)|$ denote the amplitude spectra of noisy speech, clean speech and noise signal, respectively. While $\theta_y(k,m)$, $\theta_s(k,m)$ and $\theta_d(k,m)$ are the corresponding phase spectra.

It is assumed that the phase spectra of clean speech and noise signal are uniform, i.e.

$$\theta_s(k,m)=\theta_d(k,m) \tag{4}$$

Then the speech amplitude spectral can be derived as follows

$$\left|\hat{S}(k,m)\right|=\left|Y(k,m)\right|-\left|D(k,m)\right| \tag{5}$$

After applying the phase spectral of noisy speech to the result above, the estimation of clean speech signals can be obtained in time domain using an N-point IDFT. This is the basic principle of amplitude spectral subtraction, which subtracts noise amplitude spectral from the noisy amplitude spectral based on the assumption that the phase spectra of clean speech and noise signal must being uniform. However, many researches have showed that both of the clean speech and noise phase spectral are randomly distributed between $0\sim 2\pi$, and the uniformity assumption is commonly incorrect. And thus, it will definitely bring lots of errors to the estimated speech amplitude spectral under thus unreasonable assumption. To solve this problem, a QASSR method was proposed in [8], which does not need any conditional assumption on the phase spectral. It improves the speech enhancement performance effectively by analyzing the effect of noise on the clean speech amplitude, and computes the amplitude spectral of clean speech by resolving a third-order polynomial equation. The QASSR method can be expressed as follows.

From equation (3), we can easily get

$$\left|Y(k,m)\right|=\left|S(k,m)\cdot e^{j\theta_s(k,m)}+D(k,m)\cdot e^{j\theta_d(k,m)}\right| \tag{6}$$

As the phase spectral of noise signal is randomly distributed between $0\sim 2\pi$, the expectation of noisy amplitude spectral can be written as (To keep it simple , the frame index and frequency-bin index m and k are omitted in the following equations)

$$\begin{aligned} E\{|Y|\} &= \frac{1}{2\pi}\int_{|D|} f(|D|)\cdot\left(\int_0^{2\pi}\left||S|\cdot e^{j\theta_s}+|D|\cdot e^{j\theta_d}\right|\right)d\theta_d\, d|D| \\ &= \frac{1}{2\pi}\int_{|D|} f(|D|)\cdot\int_0^{2\pi}\sqrt{|S|^2+|D|^2+2|S||D|\cdot\cos(\theta)}\,d\theta\cdot d|D| \end{aligned} \tag{7}$$

where $f(|D|)$ is the probability distribution function of $|D|$, and $\theta=\theta_s-\theta_d$ denotes the phase difference between the clean speech and noise. With the approximation that

$$E\left(F(|D|)\right)\approx F\left(E(|D|)\right)=F\left(|\bar{D}|\right) \tag{8}$$

where $F(|D|)$ being an integral function with respect to the phase difference, we can get

$$E\{|Y|\}\approx\frac{1}{2\pi}\int_0^{2\pi}\sqrt{|S|^2+|\bar{D}|^2+2|S||\bar{D}|\cdot\cos(\theta)}\,d\theta=\frac{|\bar{D}|}{2\pi}\int_0^{2\pi}\sqrt{1+r^2+2r\cdot\cos(\theta)}\,d\theta=|\bar{D}|\cdot Q(r) \tag{9}$$

where, $r=a/|\bar{D}|$, $Q(r)=\frac{1}{2\pi}\int_0^{2\pi}\sqrt{1+r^2+2r\cdot\cos(\theta)}\,d\theta$.

As the closed-form of $Q(r)$ can not be achieved in practice, it is proposed in [8] to use a third-order polynomial equation to approximate it

$$Q(r)\begin{cases}\approx T(r), \text{if}\,(0<r<2)\\ \approx r, \text{if}\,(r>2)\end{cases} \tag{10}$$

where

$$T(r)=0.9820-0.0075r+0.3761r^2-0.0461r^3 \tag{11}$$

Note that third-order polynomial equation above has three roots, but only one is in the range ($0<r<2$).

With the assumption that the noise amplitude spectral is known and approximating the expectation $E\{|Y|\}$ with the instant observation numerical, the clean speech amplitude $|S|$ can be achieved as follows:

$$T(r)\approx Q(r)=E\{|Y|\}/|\bar{D}|\approx |Y|/|\bar{D}|\Rightarrow r=T^{-1}\left(|Y|/|\bar{D}|\right),\ \text{and}\ |S|=|\bar{D}|\cdot r \tag{12}$$

The basic procedure of the QASSR approach in (12) is shown in Fig. 1. Comparing to the traditional amplitude spectral subtraction in (5), the QASSR approach sufficiently considers the phase difference of clean speech and noise signal and thus the improved accuracy of clean speech estimation can be obtained. From Fig. 1, however, we can see that the minimum suppression ratio for the background noise in QASSR approach is set to 0.25. It is to say that even during the speech absent periods, the residual noise in enhanced speech signals still makes up approximately a quarter of the background noise which definitely degrades the performance of the speech enhancement system.

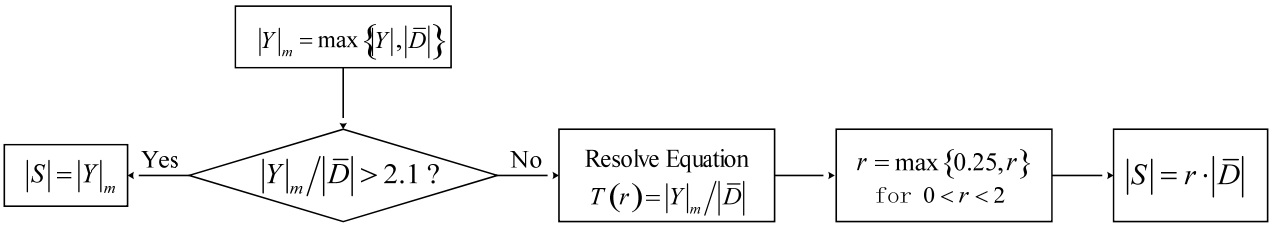

Fig.1. Flow diagram of QASSR approach

The Modified QASSR Based on SPP

In this section, to solve the residual noise problem in QASSR approach, a modified amplitude spectral estimator is proposed by incorporating the DD and soft decision technique.

Considering the fact that the clean speech signal is not always present at all times or in all frequency bins, we give two hypotheses H_i (i=0, 1) for the absence or presence of speech signal in DFT domain as

$$\begin{aligned}&H_0:\ \text{speech is absent, i.e. } Y=D;\\ &H_1:\ \text{speech is present, i.e. } Y=S+D.\end{aligned} \tag{13}$$

Then the estimator for the clean speech amplitude spectral can be modified as follows

$$\hat{S}=E\left(S|Y,H_1\right)P\left(H_1|Y\right)+E\left(S|Y,H_0\right)P\left(H_0|Y\right) \tag{14}$$

where $P(H_0|Y)$ and hypotheses $P(H_1|Y)$ denote the conditional probability of speech absence and speech presence given the noisy speech Y. it is obvious that when speech is absent, the estimation

$$E\left(S|Y,H_0\right)=0 \tag{15}$$

And so

$$\hat{S}=E\left(S|Y,H_1\right)P\left(H_1|Y\right) \tag{16}$$

According to Bayes rule, we have

$$P\left(H_1|Y\right)=\frac{p\left(Y|H_1\right)P\left(H_1\right)}{p\left(Y|H_1\right)P\left(H_1\right)+p\left(Y|H_0\right)P\left(H_0\right)}=\frac{q\cdot\Psi(Y)}{1+q\cdot\Psi(Y)} \tag{17}$$

where $\Psi(Y)$ is generalized likelihood ratio, which is defined as

$$\Psi(Y) = p(Y|H_1) / p(Y|H_0) \tag{18}$$

The coefficient q is defined as $q = P(H_1) / P(H_0)$.

It is assumed that spectra of clean speech and noise signal are all zero-mean and Gaussian distributed with variances λ_D and λ_S, respectively, we have

$$p(D) = \frac{1}{\pi\lambda_D} \exp\left(-\frac{D^2}{\lambda_D}\right) \tag{19}$$

$$p(S) = \frac{1}{\pi\lambda_S} \exp\left(-\frac{S^2}{\lambda_S}\right) \tag{20}$$

Then the probability distribution function of Y under hypotheses H_1 as well as the generalized likelihood ratio can be achieved as follows

$$p(Y|H_1) = \frac{1}{\pi[\lambda_D + \lambda_S]} \exp\left(-\frac{Y^2}{\lambda_D + \lambda_S}\right) \tag{21}$$

$$\Psi(Y) = \frac{1}{1+\xi} \exp\left\{\frac{\gamma\xi}{1+\xi}\right\} \tag{22}$$

where ξ and γ denote the *a priori* SNR and *a posteriori* SNR, respectively. As λ_D can be easily estimated using voice activity detection, the performance of speech enhancement system is mainly determined by the *a priori* SNR estimation. In this paper, the *a priori* SNR is computed by utilizing the famous DD method at each frame and frequency bin as follows

$$\hat{\xi}_k(m) = \alpha \frac{\hat{S}_k^2(m-1)}{\lambda_D(k, m-1)} + (1-\alpha)\max[\gamma_k(m) - 1, 0] \tag{23}$$

where α is the weighting factor. max[a] used here is to make the *a priori* SNR estimation being nonnegative. $\hat{S}_k^2(m-1)$ is the estimated speech spectral in the previous frame. Based on the result in [9], we set defined q as 1, which result in a smoothed speech absent probability for each frame. Then we will get

$$P(H_1|Y) = \frac{\Psi(Y)}{1+\Psi(Y)} \tag{24}$$

Here, we write the output speech amplitude spectral of original QASSR as S_p. As much residual noise is retained in S_p, we refine the estimation by multiplying an over attenuation factor $P(H_1|Y)$, and then the final estimation of the speech amplitude spectral can be obtained as

$$\hat{S} = S_p \cdot P(H_1|Y) \tag{25}$$

Performance Evaluation

In order to evaluate the speech enhancement performance of the proposed modified QASSR (MQASSR) algorithm, segmental SNR (SegSNR) measure, logarithm likelihood ratio (LLR) measure and perceptual evaluation of speech quality (PESQ) are used as objective quality evaluation metrics [10]. Ten utterances spoken by five female and five male speakers were used as clean speech data. Four types of noise signals from NOISEX-92: white, F-16, HFchannel and Destroyerops cockpit noises were added to the clean speech signals with the input SNRs of 0dB, 5dB and 10dB, respectively. The sampling frequency is 8 KHz, and a 256 samples Hamming window with 50% overlap was used. The weighting factor of the DD approach is set to 0.98 to reduce the musical noise effectively.

The SegSNR is measured by computing the SNR for each frame as well as frequency bin of speech and averaging these SNRs over all tested speech sequences. The SegSNR is defined as follows:

$$SNR_{seg} = \frac{10}{M}\sum_{m=1}^{M}\log_{10}\left(\frac{\sum_{n=(m-1)N+1}^{Nm} s^2(n)}{\sum_{n=(m-1)N+1}^{Nm}\left[s(n)-\hat{s}(n)\right]^2}\right) \tag{26}$$

where n is the index of signal samples, N is the frame length, m denotes the frame index, and M is the total number of frames. The logarithm likelihood ratio (LLR) reflects the mismatch degrees between the spectra of clean and enhanced speech, which can be calculated as:

$$LLR = \log_{10}\left(\frac{\alpha_{\hat{s}} R_s \alpha_{\hat{s}}^T}{\alpha_s R_s \alpha_s^T}\right) \tag{27}$$

where $\alpha_{\hat{s}}$ and α_s denote the linear prediction coefficient vector of enhanced and clean speech, and Rs is the linear prediction autocorrelation matrix of clean speech frame.

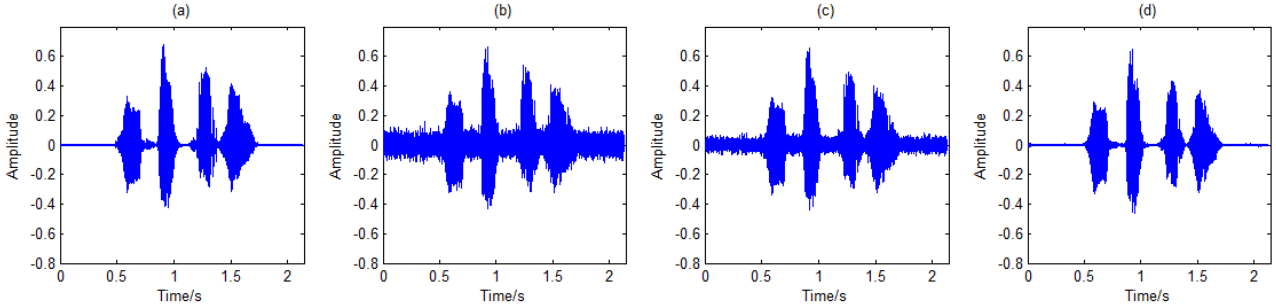

Fig.2. Waveforms of (a) clean speech, (b) noisy speech, and enhanced speech using the (c) QASSR approach, (d) the propose MQASSR approach

Table 1. Experimental results of output SegSNR, LLR and PESQ of each considered approach

Noise Type	Input SNR	SegSNR		LLR		PESQ	
		QASSR	MQASSR	QASSR	MQASSR	QASSR	MQASSR
White	0 dB	-4.0294	-1.3816	1.1151	0.8012	1.7961	1.9626
	5 dB	-1.7605	0.3259	0.7693	0.6862	2.0783	2.4085
	10 dB	0.6896	2.6386	0.5664	0.5252	2.4142	2.8841
F16	0 dB	-3.6949	-1.2658	1.0053	0.9082	2.0780	2.1946
	5 dB	-1.4719	0.4233	0.8453	0.7660	2.3843	2.5255
	10 dB	0.9970	2.8048	0.6969	0.6651	2.7188	3.1654
HFchannel	0 dB	-3.9291	-1.4816	1.0152	0.8113	1.8662	2.0616
	5 dB	-1.8601	0.2252	0.7783	0.6911	2.0493	2.3981
	10 dB	0.7191	2.5182	0.5461	0.5092	2.5002	2.8998
Destroyerengine	0 dB	-3.4871	-1.0167	0.8602	0.7900	2.0176	2.2191
	5 dB	-1.2243	0.6328	0.6981	0.6650	2.3085	2.6200
	10 dB	1.1877	2.8761	0.5542	0.5423	2.6048	3.0946

Firstly, we show the waveforms of clean speech, noisy speech, enhanced speeches using QASSR approach and our propose MQASSR approach in Fig. 2. The background noise is white noise and the input SNR is 5dB SNR. From the figure we can find that the enhanced speech using the QASSR approach still yields much residual noise, which leads to a degraded noise reduction performance. Due to an over attenuation factor being used in our proposed MQASSR approach, it is obvious that there is less background noise residual compare to the referred approach.

Secondly, we carry out the test based on the output SegSNR, WSS and PESQ measures. The average output results of each considered approach are listed in Table 1 for different noise conditions and input SNRs. According to the table, the proposed MQASSR method has an

improved performance as reflected by the higher SegSNR results in the tested environments. Meanwhile, the lower WSS are also obtained by the MQASSR method. Since lower WSS lead to less distortion, these results confirm that our approach is superior to the QASSR approach in reducing the speech distortion. It is well known that the PESQ measure is a perceptual evaluation of speech quality, the higher PESQ results in our approach corresponds improved speech quality when compared to the QASSR approach.

Conclusion

The QASSR approach can effectively avoid the unreasonable assumption in traditional amplitude spectral subtraction method, but suffers from much residual noise in enhanced speech signals. To overcome this problem, this paper presented a modified algorithm for further suppressing the residual noise in QASSR approach. The proposed method expressed the speech present probability as an over attenuation factor to refine the estimation of the QASSR approach. As the residual noise was further attenuated during noise-only frames, the output speech quality of the proposed method was markedly improved compared to original method. The results of simulation experiments have proven the modified performance of our approach in view of various evaluation measures.

Acknowledgement

This work is supported by the National Nature Science Foundation of China (Grant no. 61005021, 61201457) and the project of Shandong province higher educational domestic visiting scholar for young backbone teachers.

References

[1] Dendrinos M., Bakamidis S. Carayannis G.. Speech enhancement from noise: A regenerative approach. Speech Communication, 1991, 10(1):45-57.

[2] Jensen S. H., Hansen P. C. Hansen S. D., Sorensen J. A., Reduction of broad-band noise in speech by truncated qsvd. IEEE Transactions on Speech and Audio Processing, 1995, 3(6):439-448.

[3] Ephraim Y., H. L. Van-Trees. A signal subspace approach for speech enhancement. IEEE Transactions on Speech and Audio Processing, 1995, 3(4): 251-266.

[4] Lim J. and Oppenheim A.V. Enhancement and bandwidth compression of noisy speech. Proc. IEEE, 1979, 67(12): 1586-1604.

[5] Lim J. And Oppenheim A.V. All-pole modeling of degraded speech. IEEE Transactions on Acoustics, Speech, Signal Processing, 1978, ASSP-26(3): 197-210.

[6] Y. Ephraim, D. Malah. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Transaction on Acoustic Speech Signal Process, 1984, 32(6): 1109-1121.

[7] Boll S. Suppression of acoustic noise in speech using spectral subtraction. IEEE Transaction on Acoustic, Speech Signal Process, 1979, 27(6): 113-120.

[8] Q. Zhu and A. Alwan. The effect of additive noise on speech amplitude spectra: a quantitative analysis. IEEE Signal Processing Letters, 2002, 9(9): 275-277.

[9] Y. Park and J. Chang. A probabilistic combination method of minimum statistics and soft decision for robust noise power estimation in speech enhancement. IEEE Signal Process Letters, 2008, 15(1): 95-98.

[10] Quackenbush, S. R., Barnwell, T. P. Clements, M. A., Objective measures of speech quality. Englewood Cliffs, NJ: Prentice Hall, 1988.