Speech Enhancement Using Magnitude and Phase Spectrum Compensation

Zhen Li, Wenjin Wu, Qin Zhang
Information Engineering School, Communication University of China, Beijing, China
lizhen@cuc.edu.cn

Hui Ren
Key Laboratory of Acoustic Visual Technology and Intelligent Control System, Ministry of Culture
Beijing,China

Shilei Bai
Beijing Key Laboratory of Modern Entertainment Technology
Beijing, China

***Abstract*—Background noise is a severe problem in speech related systems. In order to solve this problem, it is important to eliminate the noise from the noisy speech, which is called speech enhancement. Typical speech enhancement algorithms only operate on the short-time magnitude spectrum, while keeping the short-time phase spectrum unchanged for synthesis. Or only compensate the phase spectrum while keeping the magnitude spectrum unchanged. In this paper, we present a novel method by changing both magnitude and phase spectra to produce a modified complex spectrum. The test of an objective speech quality measure PESQ, and spectrogram analysis had showed that the proposed method can obtain better enhancement performance.**

Keywords—*speech enhancement; magnitude spectrum; phase spectrum compensation*

I. INTRODUCTION

Speech enhancement is a noise suppression technology It has important significance for solving the problem of noise disturbance. And it can improve the quality and intelligibility of voice communications. The purpose of speech enhancement is to restore the original signal from noisy observations corrupted by various noises [1].

Let us consider an additive noise mode

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

where $x(n)$, $s(n)$, and $d(n)$ denote discrete-time signals of noisy speech, clean speech, and noise, respectively. The discrete short-time Fourier transform (DSTFT) of the corrupted speech signal $x(n)$ is given by

$$X(n,k) = \sum_{m=-\infty}^{\infty} x(m)\omega(n-m)e^{-j2\pi km/N} \tag{2}$$

where k denotes the k th discrete-frequency of N uniformly spaced frequencies and $\omega(n)$ is an analysis window function of short duration. By using DSTFT we can obtain Eq.(1) as.

$$X(n,k) = S(n,k) + D(n,k) \tag{3}$$

where $X(n,k), S(n,k)$ and $D(n,k)$ are the DSTFTs of noisy speech, clean speech, and noise, respectively.Each of them can be described in terms of the DSTFT magnitude spectrum and the DSTFT phase spectrum. For example, $S(n,k)$ can be written in polar form as

$$S(n,k) = \left|S(n,k)\right|e^{j\angle S(n,k)} = A_k e^{j\alpha_k} \tag{4}$$

where $\left|S(n,k)\right|$ is the magnitude spectrum, and $\angle S(n,k)$ is the phase spectrum.

Most of the existing speech enhancement algorithms only change the magnitude spectrum of the noisy speech. The modified magnitude then recombined with the unchanged phase spectrum to produce a modified complex spectrum, which is the estimated clean speech spectrum. These algorithms are called magnitude spectrum based methods. Boll proposed the method of spectral subtraction(SSUB) in 1979. Its basic principle is to subtract the magnitude spectrum of the noise from the noisy speech magnitude spectrum, and obtain the estimate of the clean signal magnitude spectrum ,but the phase spectrum is unchanged[2].The MMSE estimator, which is presented by Ephraim and Malah in 1984. Its main idea is to minimize the mean-squared error (MSE) between the clean and estimated (magnitude or power) spectra[3]. Wiener filter[4]was proposed by Wiener. Hansen and Jensen first presented the Wiener method in the single-channel case enhancement[5]. Doclo and Moonen further extended the Wiener method in the mutil-channel case [6]. Ephraim and Van Trees proposed the linear predictive factors to estimate the pure speech signal[7].

The reason for ignoring the phase impact is that the phase spectrum has been found to have less perceptual effect at significantly higher signal to noise ratio (SNR) levels[8].But recently, it is found that the phase spectrum may be useful in speech processing applications[9]. Kamil Wójcicki et.al. proposed the speech enhancement method of phase spectrum compensation (PSC) in 2008[10][11].

This paper proposes a new method by changing both magnitude spectrum and phase spectrum to produce a modified complex spectrum. The proposed method obtains better performance in terms of an objective speech quality

- This work was financially supported by the national science and technology planning project "Study and application demonstration on the key technology of the stage effect"(Item Number: 2012BAH38F00) and Engineering Project of CUC (Foundation item: Research on Networked Control System of Stage Lighting, Project number: 3132015XNG1532).

ICIS 2016, June 26-29, 2016, Okayama, Japan

measure PESQ (Perceptual Estimation of Speech Quality) score compared with conventional methods.

The paper is organized as follows: Section II describes our proposed method. Section III describes the enhancement experiments. Section IV presents the results of the experiments and gives some analysis, and we conclude the paper in section V.

II. PROPOSED METHOD

Our method is based on modified magnitude and phase spectrum compensation. The block diagram of our method is shown in Fig.1

The magnitude estimation of the clean speech is[3]

$$\hat{A}_k = \frac{\sqrt{\pi}}{2}\frac{\sqrt{\nu_k}}{\gamma_k}\exp(-\frac{\nu_k}{2})\left[(1+\nu_k)\mathrm{I}_0(\frac{\nu_k}{2})+\nu_k\mathrm{I}_1(\frac{\nu_k}{2})\right]Y_k \tag{5}$$

where $I_0(\bullet)$ and $I_1(\bullet)$ denote the modified Bessel functions of zero and first order, respectively. ν_k is defined by

$$\nu_k = \frac{\xi_k}{1+\xi_k}\gamma_k \tag{6}$$

Where ξ_k and γ_k are defined by

$$\xi_k = \frac{\lambda_x(k)}{\lambda_d(k)} \quad \gamma_k = \frac{Y_k^2}{\lambda_d(k)} \tag{7}$$

We then obtain the phase spectrum compensation function from[10]

$$\Lambda(n,k) = \lambda\Psi(k)\left|\hat{D}(n,k)\right| \tag{8}$$

Where λ is a real-valued empirically determined constant, and in this paper $\lambda = 3.74$. $\Psi(k)$ is the antisymmetry function, and it is given by

$$\Psi(k) = \begin{cases} 1, & if\ 0 < k/N < 0.5 \\ -1, & if\ 0.5 < k/N < 1 \\ 0, & otherwise \end{cases} \tag{9}$$

where zero weighting is given to the values corresponding to non-conjugate vectors of DSTFT（i.e. the $k = 0$ value and possible singleton $k = N/2$ for N =even）. The next step is offsetting the complex spectrum of the noisy speech by the additive real-valued phase spectrum compensation function $\Lambda(K)$.

$$X_\Lambda(n,k) = X(n,k) + \Lambda(n,k) \tag{10}$$

The compensated phase spectrum is then obtained by

$$\angle X_\Lambda(n,k) = ARG[X_\Lambda(n,k)] \tag{11}$$

where ARG is the complex angle function.

From above we obtain the magnitude estimation and the compensated phase spectrum, and then we can get the modified complex spectrum by

$$\hat{S}(n,k) = \hat{A}_k e^{j\angle X_\Lambda(n,k)} \tag{12}$$

At last the IDSTFT is used to convert the frequency-domain $\hat{S}(n,k)$ to the time-domain representation. The resulting time-domain frames may be complex, in the PSC method the imaginary component is discarded. And then by employing the OLA（Overlap-Add） procedure, we obtain the enhanced time-domain signal, $\hat{s}(n)$.

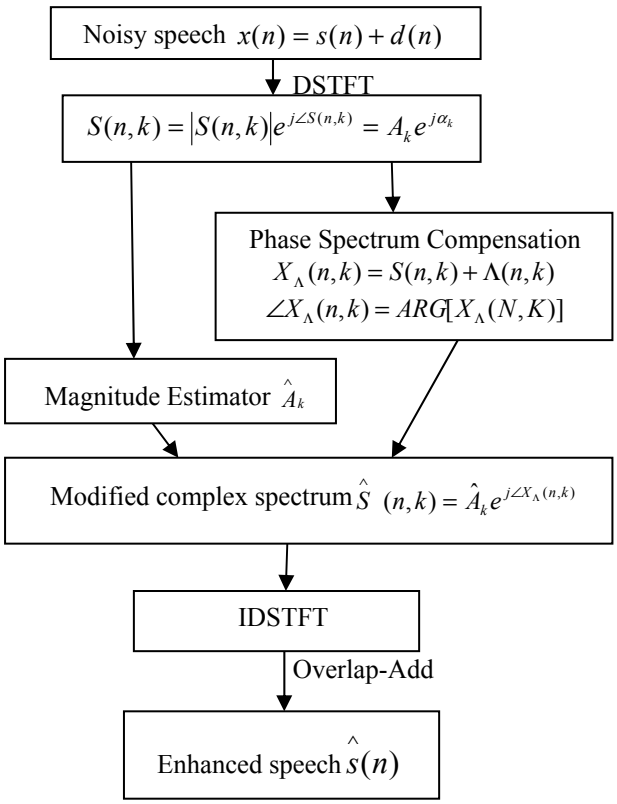


Fig. 1. Block Diagram of Proposed Speech Enhancement Method

III. ENHANCEMENT EXPERIMENTS

The enhancement experiments were carried out on the core test set of the NOIZEUS speech corpus which was composed of 30 phonetically-balanced sentences belonging to six speakers (three males and three females).There were 8 kinds of nonstationary noises at different SNRs except noisy speech with white noises. During the evaluation process, we generated a noisy speech set corrupted by additive white noise (taken from the NOISEX-92 noise database) at four SNR levels: 0 dB, 5 dB, 10 dB, and 15dB. Then the corrupted files were enhanced by the proposed method. Besides, we had

experiments with MMSE（minimum mean square error）method and PSC method separately. The other popular speech enhancement techniques, such as spectral subtraction method was used too.

For evaluation purpose, we employed an objective speech evaluation, that is the perceptual estimation of speech quality (PESQ).It is an enhanced perceptual quality measurement for voice quality in telecommunications. Its score is between 1.0 and 4.5, where 1.0 corresponds to *bad* and 4.5 corresponds to *distortionless. In our evaluations, we computed the mean PESQ scores for each method* and each noise case.

In order to show performance of these methods pictorially, the spectrogram analysis was used, too.

In our experiments, the samples of each of the sentence files have been normalized to be between -1.0 and 1.0. The Hamming window has been used as the analysis window .We set frame duration to 32 ms and the frame shift is 4 ms. We employ FFT length of 1024 samples. The anti-symmetric function given in (9) has also been used. The value $\lambda = 3.74$ in our evaluations was determined to maximize both PESQ and SNR scores.

IV. RESULTS AND DISCUSSION

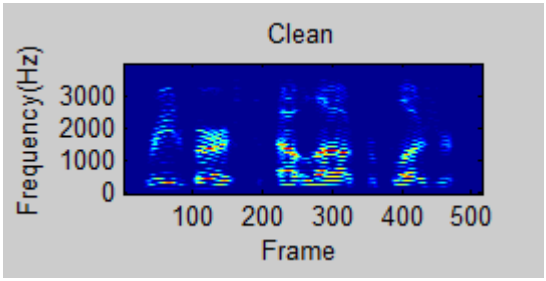
The results of the enhancement experiments, in terms of mean PESQ scores, are shown in Table 1. The results of spectrogram analysis are shown in Fig. 2. The Table 1 shows that compared to other methods, the proposed method performed well, it would obtain consistent improvements across all SNRs. The spectrograms in Fig.2 show a single utterance sp04.wav corrupted by white noise at 10dB SNR. As we can see, our proposed method had showed better enhancement performance, and it exhibited better noise cancellation than the other existing speech enhancement techniques.

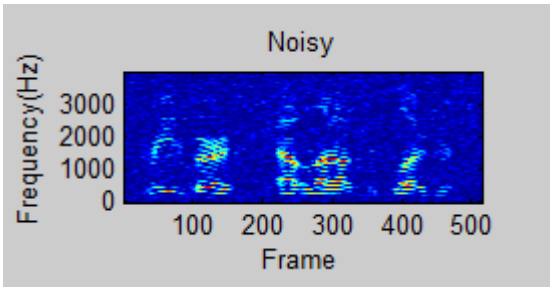
The performance of our proposed algorithm was compared with MMSE, SSUB and PSC methods. The results were analyzed mainly by mean PESQ scores and spectrograms to present the performance on objective quality and speech intelligibility, respectively. Through extensive experiments, we found that when averaged over all nine kinds of noises (TABLE I and fig. 2 only show the white noise type), our proposed algorithm achieved the best results in terms of mean PESQ scores at any input speech SNRs and can improve speech intelligibility for low SNR levels. This is because the phase spectrum compensation works at low SNR levels better, and other methods discarding phase factors only works well at high SNR levels. On the other hand, the spectrograms comparisons with other algorithms predicted that our proposed method was able to suppress noise and enhance the useful speech more effectively.

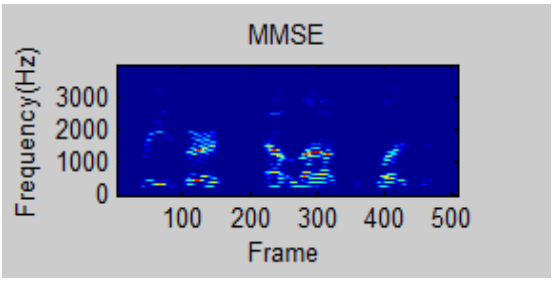
Further subjective tests are needed to verify the effectiveness of the proposed algorithm on improving both subjective quality and speech intelligibility. If it works better, it is worth mentioning that our proposed method can be used in the systems which may want speech as clean as possible even with some degree of speech distortion.

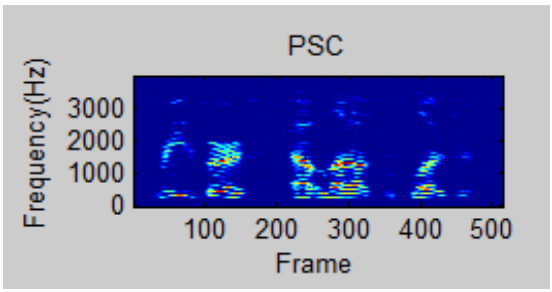
TABLE I. Mean PESQ scores of the white noise case for the MMSE,PSC, SSUB, and Proposed Method

INPUT SPEECH SNR(dB)	METHODS					
	CLEAN	*NOISY*	*SSUB*	*MMSE*	*PSC*	*Proposed Method*
0	4.50	1.59	1.76	1.96	1.93	2.08
5	4.50	1.83	2.22	2.30	2.30	2.43
10	4.50	2.14	2.64	2.62	2.66	2.78
15	4.50	2.47	3.06	2.92	3.02	3.13









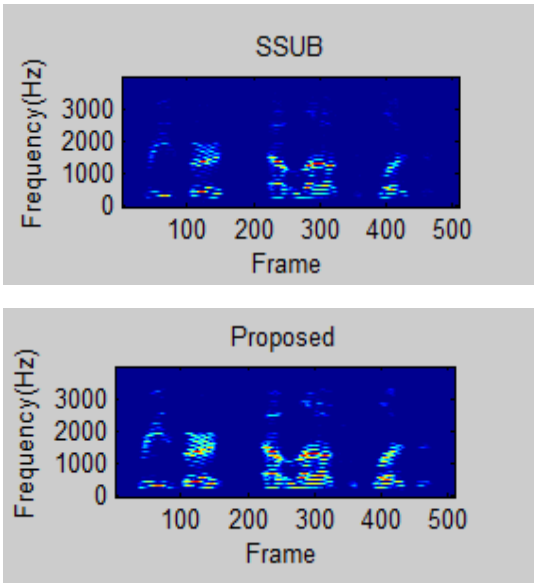


Fig. 2. Spectrograms of a sp04 NOIZUES utterance at 10dB SNR：“Read verse out loud for pleasure.”belonging to a male speaker.

V. CONCLUSION

In this paper we have proposed a new method of speech enhancement which is based on modified magnitude and compensated phase spectrum. Experiment results of the objective speech quality measure PESQ, and spectrogram analysis had showed that the proposed method achieve better speech quality than the conventional speech enhancement methods. The method can be used in the systems which need to cancel the background noises such as speech recognition, speech communication, etc, and it can further improve the speech quality and intelligibility.

REFERENCES

[1] P. Loizou, Speech Enhancement: Theory and Practice. Boca Raton,FL: CRC, 2007.

[2] BOLL S F. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Trans. Acoustics, Speech, Signal Processing, 1979, 27(2):113-120.

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

[4] N. Wiener, The Extrapolation, Interpolation, and Smoothing of Stationary Time Series With Engineering Applications. New York:Wiley, 1949.

[5] P. C. Hansen and S. H. Jensen, “FIR filter representations of

Reduced rank noise reduction,” IEEE Trans. Signal Process., vol. 46, no.6, pp.1737--1741, Jun. 1998.

[6] S. Doclo and M. Moonen, “On the output SNR of the speech-distortion weighted multichannel Wiener filter,” IEEE Signal Process. Lett., vol.12, no. 12, pp. 809--811, Dec. 2005.

[7] Y. Ephraim and H. V. Trees, “A signal subspace approach for speech enhancement,” IEEE Trans. Speech Audio Process., vol. 3, no. 4, pp.251–266, Jul. 1995.

[8] D.L. Wang and J.S. Lim, “The unimportance of phase in speech enhancements”, IEEE Trans. Acoust., Speech and Signal Process., Vol.30, pp. 679-681, Aug. 1982.

[9] K. Paliwal, L Alsteris, “Usefulness of phase in speech processing”,Proc. IPSJ Spoken Language Processing Workshop, Gifu, Japan, pp. 1-6, 2003.

[10] Kamil Wójcicki ,Mitar Milacic, Anthony Stark, James Lyons, Kuldip Paliwal. Exploiting Conjugate Symmetry of the Short-Time Fourier Spectrum for Speech Enhancement[A].IEEE Signal Process[C].Lett,2008,15:461-464.

[11] Stark, A., Wójcicki, K.K., Lyons, J.G. and K. Paliwal, “Noise driven short time phase spectrum compensation procedure for speech enhancement”, Proceedings of the 10th International Conference on Spoken Language Processing (INTERSPEECH-ICSLP), Brisbane.