Single Channel Speech Enhancement Using Complex Kalman Filter in Noisy Reverberant Environments

Yang Liu[1], Yanmin Shi[2], Xinglu Ma[3]
{ yangliu_qust@foxmail.com[1]，shiyanmin312@163.com[2], Qdmxl@163.com[3]}
[1,2,3](College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061,China)

Abstract. The previous research has proved that restoring the instantaneous amplitude and phase individually by linear prediction with Kalman filter on Gammatone filterbank could improve quality and intelligibility of noisy reverberant speech significantly. However, this method still suffers from the double prediction errors caused by the separated estimation for instantaneous amplitude and phase. This paper aims to study feasibility for restoring both instantaneous amplitude and phase simultaneously by complex Kalman filter to improve previous method. The signal to error ratio (SER), perceptual evaluation of speech quality (PESQ) and SNR loss were used as objective evaluation metrics. Results showed that the proposed method could effectively improve these objective metrics more than the previous method.

Keywords: Speech enhancement, complex Kalman filter, Gammatone filterbank

1 Introduction

The quality and intelligibility of speech are very important for speech communication, which nevertheless are always degraded due to the interference such as background noise and reverberation. Particularly the performance of applications, which always contain only one microphone with the consideration of cost and size, such as speech coders, hearing aids and automatic speech recognition (ASR) systems, might be severely reduced. Therefore, single channel speech enhancement methods are of great necessity to be developed.

In the previous research, various single channel speech enhancement methods have already been considered to remove the effects of noise or reverberation to improve the quality or intelligibility of speech signals. Among the many proposed methods, short-time Fourier transform based analysis-modification-synthesis (STFT-AMS) was the most widely used framework for speech enhancement [1]. Based on the survey of noise reduction methods, the spectral subtraction (SS) method has been shown to effectively suppress stationary noise [2]. This method performs subtraction of an estimated noise magnitude spectrum from a noisy speech magnitude spectrum. The methods related to minimum mean-square error short-time spectral amplitude (MMSE-STSA) estimator, which was proposed by Ephraim and Malah [3] and the Wiener filtering algorithm proposed by Scalart and Filho [4] have also drawn a great deal of attention, because of simplicity and efficient spectral magnitude estimation.

In the reverberant environments, cepstral mean normalization (CMN) [5] based on STFT-AMS framework, which could suppress the effect of reverberation by normalizing cepstral features, has been shown to be the simplest and most effective method. However, this method

MOBIMEDIA 2018, June 21-22, Qingdao, People's Republic of China

DOI 10.4108/eai.21-6-2018.2276600

is not effective with the presence of the late reverberation. Another method based on multiple-step linear prediction (MSLP) was proposed by Kinoshita et al. [6]. This method estimates the late reverberation by long-term MSLP and then it could be suppressed by subsequent SS.

However, most existing methods cannot work well in noisy reverberant environments to completely remove the effects of noise and reverberation simultaneously. Recently the effectiveness of phase manipulation is of great interest for researchers in speech enhancement. Modulation-phase-only experiments have been carried out which proved the importance of modulation phase spectrum in speech intelligibility [7]. Liu et al. [8] tried to improve the quality as well as the intelligibility of speech signals in noisy reverberant environments by restoring instantaneous amplitude and phase individually. However, it is still challenging for the complex version of Kalman filter for restoring instantaneous amplitude and phase as a whole.

In this paper, complex Kalman filter is proposed to improve the quality and intelligibility of speech in noisy reverberant environments. The derivation of the accurate transition matrices was deeply concentrated because the speech enhancement performance of the Kalman filter depends on the accuracy and reliability of transition matrices. These matrices were obtained by autocorrelation approach [9] for the complex form of instantaneous amplitude and phase. In addition, the effects of noise corresponding to additive and convolved noise (late reverberant speech) and reverberation were removed simultaneously with consideration of complex Kalman filter with efficient LP and the early reflection effect can be removed by CMN process.

2 Proposed Method

Our proposed method for speech enhancement is the extension of previous method [8] based on ASM framework on a Gammatone filterbank. The model consists of three steps: (i) Analysis stage, the noisy reverberant speech is decomposed by the Gammatone filterbank [10]. (ii) Modification stage, where the speech components in each channel are enhanced by complex Kalman filter with linear prediction (LP). (iii) Re-synthesis state, the restored speech is synthesized by the inverse Gammatone filterbank [10].

2.1 Signal Definition

The noisy reverberant speech $y_{\mathrm{NR}}(t) = x(t) * h(t) + n(t)$ is observed. Here $h(t)$ is the room impulse response which consists of early reflection $h_E(t)$ and late reverberation $h_L(t)$, then we have $y_{NR} = x(t) * h_E(t) + x(t) * h_L(t) + n(t) = x_E(t) + x_L(t) + n(t)$, where $x_E(t)$ is early reverberant speech and $x_L(t)$ is late reverberant speech. The output of *k*-th channel from Gammatone filterbank is represented as:

$$Y_{NR,k}(t) = Y_{NR,1,k}(t) + Y_{NR,2,k}(t) = A_{NR,k}(t).\exp(j\omega_k t + j\varphi_{NR,k}(t)). \tag{1}$$

where $Y_{NR,1,k}(t)$ and $Y_{NR,2,k}(t)$ are components of $x_E(t)$ and $x_L(t) + n(t)$ respectively. ω_k is the center frequency in *k*-th channel. $A_{NR,k}(t)$ and $\varphi_{NR,k}(t)$ are the instantaneous amplitude and phase of noisy reverberant speech. $Y_{NR,k}(t)$ could be rewritten in complex form which is used as the input for complex Kalman filter as:

$$Y_{NR,k}(t) = A_{NR,k}(t).\cos(\omega_k t + \varphi_{_{NR,k}}(t)) + jA_{NR,k}(t).\sin(\omega_k t + \varphi_{_{NR,k}}(t)) . \tag{2}$$

In this paper, we focus on removing the late reverberation as well as the noise by complex Kalman filter and the early reflection could be suppressed by CMN process.

2.2 Complex Kalman Filter

Complex Kalman filter uses the state equation and observation equation to update the gain factors and predict the values. The state equation of *k*-th channel in complex Kalman filter is defined as:

$$S_k[m] = F_k.S[m-1] + W_k[m] . \tag{3}$$

where *m* is the sampling index. F_k is the transition matrix in *k*-th channel and $W_k[m]$ is driving noise which is assumed to be Gaussian white noise and the variance is Q_k . $S_k[m]$ is the state vector of the output of Gammatone filterbank at sampling point *m* and the size of vector is *p*, which is the LP order.

The observation equation in *k*-th channel is defined as:

$$O_k[m] = H_k.S_k[m] + V_k[m] . \tag{4}$$

where $O_k[m]$ is the observation value at sampling point *m* and H_k is the observation matrix. $V_k[m]$ is the observation noise whose variance is R_k . It should be mentioned that all the values in these equations are complex number which is different from our previous method. Moreover, the variance of driving noise and observation noise are separately estimated in each channel.

The process of complex Kalman filter is described as follows:

$$\hat{S}_k[m \mid m-1] = F_k.S_k[m-1 \mid m-1] , \tag{5}$$

$$P_k[m \mid m-1] = F_k.P_k[m-1 \mid m-1].F_k^{\mathrm{T}} + Q_k , \tag{6}$$

$$G_k[m] = \frac{P_k[m \mid m-1].H_k{}^{\mathrm{T}}}{H_k.P_k[m \mid m-1].H_k{}^{\mathrm{T}} + R_k} , \tag{7}$$

$$e = G_k[m](O_k[m] - H_k.\hat{S}_k[m \mid m-1]) , \tag{8}$$

$$P_k[m \mid m] = (I - G_k[m].H_k)P_k[m \mid m-1] , \tag{9}$$

$$\hat{S}_k[m \mid m] = \hat{S}_k[m \mid m-1] + e . \tag{10}$$

Table 1. Results of PESQ and SNR loss (averaged values)

SNR/T_R	**Methods**					
	Noisy Reverberant		Previous Method		Proposed Method	
	PESQ	SNR loss	PESQ	SNR loss	PESQ	SNR loss
10 dB/0.5s	1.81	0.89	2.16	0.69	2.77	0.66
10 dB/2s	1.33	0.92	2.02	0.71	2.35	0.67
0 dB/0.5s	1.39	0.93	1.97	0.75	2.31	0.71
0 dB/2s	1.02	0.94	1.91	0.76	2.12	0.74

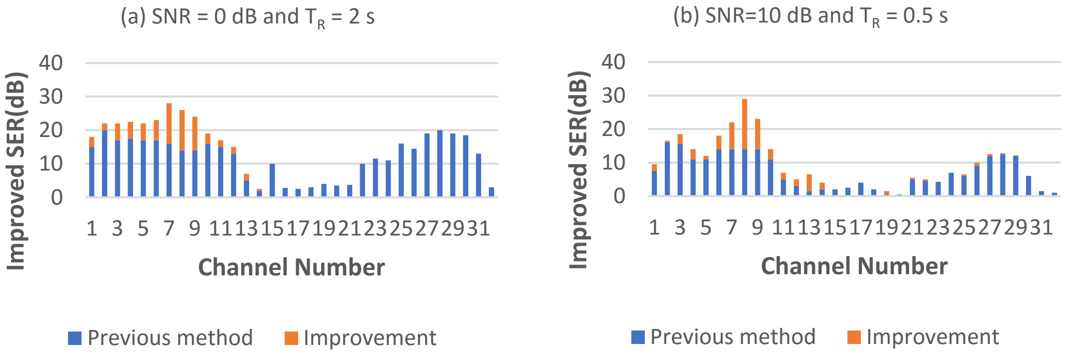

Fig. 1. Improved SER in different noisy reverberant conditions.

The initial state vector $S_k[1|1]$ could be set to random value because it will be close to the original state vector after a few iterations. $P_k[m|m]$ is the error covariance matrix which is calculated in equation (6) and updated in equation (9). $G_k[m]$ is the Kalman gain which is calculated in equation (7). *e* is the innovation and the final estimation of state vector could be obtained in equation (10). H_k is set to [0,0,0...,1] and F_k is estimated by LP method. We calculated the *p*-th order LP coefficient for each channel from a closed dataset and covert it to line spectral frequencies (LSF) for stability. After averaging the LSF coefficients, we convert it back to LP coefficients as trained coefficients which are used in the equations (5) and (6). The output of complex Kalman filter could be re-synthesized by inverse Gammatone filterbank and the late reverberation could removed by CMN process.

3 Experiments

To evaluate the effectiveness of the proposed methods, we carried out experiments using four Japanese sentences uttered by two male and two female speakers from ATR database [11] to train the LP coefficients and then chose ten different sentences uttered by five male and five female speakers for testing. The signal to noise ratios (SNRs) between *x*(*t*) and *n*(*t*) were fixed at 10 dB and 0 dB and the reverberation time T_R is set to 2 s and 0.5 s. We used a Gammatone filterbank to divide the signal into 32 channels (*K* = 32). We used the sampling frequency of 20 kHz and utilized a 25-ms-long rectangular window. The LP order, *p*, was set to 8.

We have evaluated the improvement of the restored speech by measuring the signal to error ratio (SER) which shows the level of the error that we can reduce. SER is defined as follows:

$$\mathrm{SER}(x_k, \hat{x}_k) = 10\log_{10}\frac{\int_0^T (x_k(t))^2\,\mathrm{d}t}{\int_0^T (x_k(t) - \hat{x}_k(t))^2\,\mathrm{d}t}. \tag{11}$$

where $x_k(t)$ is the clean speech of *k*-th channel and $\hat{x}_k(t)$ is the restored speech of *k*-th channel. Figure 2 shows the comparison of improved SER by previous method and proposed method under the best and worst conditions. The blue bars show the improved SER by previous method and the orange bars indicate the improvement between the previous and proposed methods. It could be observed that improvement in low frequency channels is higher than high frequency channels because high frequency components are only additive noise which has already been thoroughly removed by previous method.

Perceptual evaluation of sound quality (PESQ) [12] and SNR loss [13] were chosen for evaluating the quality and intelligibility of speech in the experiments. PESQ in the objective difference grades (ODGs) that covers from -0.5 (very annoying) to 4.5 (imperceptible) was used to evaluate subjective quality. SNR loss that ranges from 0.0 to 1.0 was used to evaluate intelligibility of speech. SNR losses (0 to 1.0) are corresponded to the percent correctness (100% to 0%). The results of objective measures are listed in Table 1, which indicate that the proposed method could improve more quality and intelligibility of speech than previous method.

4 Conclusion

We proposed a complex Kalman filter for speech enhancement on the Gammatone filterbank in noisy reverberant environments. The proposed method reduced the prediction error by dealing with instantaneous amplitudes and phases simultaneously. The results of objective evaluations revealed that the proposed method can greatly improve the previous method in terms of SER, PESQ and SNR loss, which are related with the quality and intelligibility of speech.

Acknowledgments. This work is supported by Shandong province key research and development project (2017GSF218088); Shandong key research and development fund project (2015GSF119016); Promotive research fund for young and middle-aged scientists of Shandong Province (BS2015ZZ006) and Natural Science Foundation of Tianjin (17JCQNJC00100).

References

[1] Parchami M, Zhu W P, Champagne B, et al. Recent Developments in Speech Enhancement in the Short-Time Fourier Transform Domain[J]. IEEE Circuits & Systems Magazine, 2016, 16(3):45-77.

[2] Boll S F. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Trans.acoust.speech & Signal Process, 1979, 27(2):113-120.

[3] Paliwal K, Schwerin B, Wjcicki K. Speech enhancement using a minimum mean-square error short-time spectral modulation magnitude estimator[J]. Speech Communication, 2012, 54(2):282-305.

[4] Scalart P, Filho J V. Speech enhancement based on a priori signal to noise estimation[C]// IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996. Icassp-96. Conference Proceedings. IEEE, 2002:629-632.

[5] Wu M, Wang D L. A two-stage algorithm for one-microphone reverberant speech enhancement[J]. IEEE Transactions on Audio Speech & Language Processing, 2006, 14(3):774-784.

[6] Kinoshita K, Delcroix M, Nakatani T, et al. Suppression of Late Reverberation Effect on Speech Signal Using Long-Term Multiple-step Linear Prediction[J]. IEEE Transactions on Audio Speech & Language Processing, 2009, 17(4):534-545.

[7] Paliwal K K, Alsteris L D. On the usefulness of STFT phase spectrum in human listening tests ☆ [J]. Speech Communication, 2005, 45(2):153-170.

[8] Liu Y, Nower N, Morita S, et al. Speech enhancement of instantaneous amplitude and phase for applications in noisy reverberant environments[J]. Speech Communication, 2016, 84(C):1-14.

[9] Goldfischer L I. Autocorrelation Function and Power Spectral Density of Laser-Produced Speckle Patterns[J]. Journal of the Optical Society of America, 1965, 55(3).

[10] Unoki M, Akagi M. A method of signal extraction from noisy signal based on auditory scene analysis[J]. Speech Communication, 1999, 27(3–4):261-279.

[11] Kurematsu A, Takeda K, Sagisaka Y, et al. ATR Japanese speech database as a tool of speech recognition and synthesis[J]. Speech Communication, 1990, 9(4):357-363.

[12] Cristobal E, Flavián C, Guinalíu M. Perceived e ‐ service quality (PeSQ)[J]. Journal of Service Theory & Practice, 2012, 17(3):317-340.

[13] Ma J, Loizou P C. SNR loss: A new objective measure for predicting the intelligibility of noise-suppressed speech[M]. Elsevier Science Publishers B. V. 2011.