Speech Communication 84 (2016) 1–14

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Speech enhancement of instantaneous amplitude and phase for applications in noisy reverberant environments

Yang Liu[*], Naushin Nower, Shota Morita, Masashi Unoki

School of Information Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan

ARTICLE INFO

Article history:
Received 30 October 2015
Revised 5 August 2016
Accepted 7 August 2016
Available online 16 August 2016

Keywords:
Speech enhancement
Instantaneous amplitude and phase
Kalman filter
Gammatone filterbank
Linear prediction

ABSTRACT

We previously proved that restoring the instantaneous amplitude as well as instantaneous phase on the output from Gammatone filterbank plays a significant role in speech enhancement. However, dereverberation is still a challenge since the previously proposed scheme can only work in noisy environments. In this paper, we extend our previously proposed scheme to general speech enhancement for removing both the effects of noise and reverberation by restoring instantaneous amplitude and phase simultaneously. Objective and subjective experiments were conducted under various noisy reverberant conditions to evaluate the effectiveness of the extension of the proposed scheme. The signal to error ratio (SER), correlation, PESQ, and SNR loss were used in objective evaluations. The normalized mean preference score and correctness in modified rhyme test (MRT) were used in subjective evaluations. We also tested how effective our proposed scheme is as a front-end for an automatic speech recognition (ASR) system in realistic noisy reverberant environments. The results of all evaluations revealed that the proposed scheme could effectively improve quality and intelligibility of speech signals under noisy reverberant conditions.

1. Introduction

In real environments, the quality and intelligibility of speech are always degraded due to background noise and reverberation. Specifically, the presence of background noise and reverberation might severely reduce the performance of applications such as automatic speech recognition (ASR) systems and speech coders. Therefore, these effects need to be simultaneously removed.

The short-time Fourier transform based analysis-modification-synthesis (STFT–AMS) framework is widely used in many speech enhancement methods in the acoustical spectral domain. It mainly consists of three stages: analysis, modification and synthesis. Among the many methods developed for noise reduction based on the STFT–AMS framework, the spectral subtraction (SS) has been shown to effectively suppress stationary noise (Boll, 1979). However, this method generates musical noise. This musical noise can be eliminated by recently proposed enhanced SS combined with a weighting function (Kaladharan, 2014). Ephraim and Malah (1984) derived the minimum mean-square error short-time spectral amplitude estimator (MMSE–STSA) that produces less musical noise than SS. Other methods, such as Wiener filtering (Scalart and Filho, 1984), have been considered and enhanced through the soft decision scheme. The denoising autoencoder (DAE) (Lu et al., 2013), which is based on deep learning, has recently been shown to effectively reduce noise. For dereverberation based on the STFT-AMS framework, cepstral mean normalization (CMN) (Wu and Wang, 2006), which could suppress the effect of reverberation by normalizing cepstral features, has been shown to be the simplest and most effective method. However, the dereverberation of CMN is not very effective when the late reverberation exists. To remove the late reverberation, a method based on SS was proposed in which late reverberation is removed as additive noise (Pacheco and Seara, 2006). Another method based on multiple-step linear prediction (MSLP) (Kinoshita et al., 2009) was proposed for microphones. First, this method estimates the late reverberation by long-term MSLP. Then these were suppressed by subsequent SS.

The analysis of speech signal in modulation domain has shown to be greatly significant by various psychoacoustical and physiological studies (Lu et al., 2010; 2011b; Zhang, 2000). Experiments showed that there are sub-bands in the auditory system that are tuned to detect modulation frequencies (Bacon and Grantham, 1989). Therefore, the low frequency modulation spectrum has been shown to accurately predict speech intelligibility (Drullman, 1995; Lu et al., 2011a), and many speech enhancement methods in the modulation domain have been proposed, such as band-pass filtering of the time trajectories of the cubic-root compressed short-time power spectrum (Lyons and Paliwal, 2008).

[*] Corresponding author.
E-mail addresses: yangliu@jaist.ac.jp (Y. Liu), naushin@jaist.ac.jp (N. Nower), s-morita@jaist.ac.jp (S. Morita), unoki@jaist.ac.jp (M. Unoki).

http://dx.doi.org/10.1016/j.specom.2016.08.002

Consequently, a corpus based approach (Ming et al., 2011) and speech enhancement using non-negative matrix factorization (NMF) (King and Atlas, 2010) have been proposed as modern speech enhancement methods. However, all existing methods process the smeared speech signals only by modifying the temporal or spectral magnitude.

Early studies have reported the unimportance of phase spectrum in perception. In recent studies, however, Paliwal et al. (2011) carried out modulation-phase-only experiments that proved modulation phase spectrum plays an important role in speech intelligibility. Moreover, a phase estimation method has been proposed by Mowlaee et al. (2012). It has been verified that replacing the phase of degraded speech by estimated phase could lead to improved speech quality. The importance of phase in speech enhancement has also been supported by many other positive results (Pobloth and Kleijn, 1999).

All the methods based on the STFT–AMS framework for speech enhancement improve quality rather than intelligibility of speech (Loizou and Kim, 2011). The reasons for this are still unclear, so many researchers have investigated strategies for reducing distortions and enhancing features related to speech intelligibility. On the other hand, psychoacoustical studies have found that temporal amplitude envelope (TAE) and temporal fine structure (TFS) are important cues for speech perception (Moore, 2008). They also confirmed that TAE and TFS are prominent in improving the intelligibility of speech with background noise (Swaminathan, 2010; Swaminathan and Heinz, 2012). Therefore, AMS in the filterbank is a more suitable framework for speech enhancement than AMS in the STFT. Hence, it is expected that TAE and TFS manipulations as the AMS in the filterbank can drastically improve quality as well as intelligibility of degraded speech.

Motivated by the existing literature on the effectiveness of phase manipulation, we previously proposed a speech enhancement scheme (Nower et al., 2015) that can significantly improve the quality and intelligibility of noisy speech. However, this scheme can only deal with additive noise without modeling the convolved noise such as late reverberant speech, due to the properties of convolved noise.

In this paper, we aim to propose a speech enhancement scheme based on the AMS framework in the filter to process the instantaneous amplitude and phase by Kalman filter to improve the quality and intelligibility of speech simultaneously in noisy reverberant environments. We concentrate on the derivation of the accurate transition matrices, which are quite important parameters in Kalman filtering, in the training phase. This is because the enhancement performance of the Kalman filter depends on the accuracy and reliability of transition matrices. These transition matrices of the state equation of both instantaneous amplitude and phase are unknown, so it is difficult to set these transition matrices in the Kalman filtering for suitable speech enhancement. We also considered the derivation of the observation noise since the convolved noise does not exist in the non-speech section. The main contribution of this paper is to extend the previous scheme to a general speech enhancement to remove the effects of noise and reverberation simultaneously with consideration of phase information. The novel points are that effects of noise corresponding to additive and convolved noises (late reverberant speech) on instantaneous amplitude and phase can be removed by Kalman filtering with efficient linear prediction (LP) and the early reflection effect can be removed by CMN. Our proposed scheme is expected to work well as front-end for ASR systems and hearing aids.

The rest of the paper is organized as follows. Section 2 describes the previous scheme for speech enhancement in noisy environments. Section 3 explains the details of the proposed scheme in noisy reverberant environments. Section 4 presents the subjective and objective evaluation results. Section 5 describes the experiment results in ASR system. Section 6 concludes with a summary and mentions future works.

2. Previous scheme

Our previous scheme (Nower et al., 2015) was designed to improve both instantaneous amplitudes and phases that are extracted from the output of the Gammatone filterbank (GTFB). In this scheme, the noisy speech $y_N(t)$, where $y_N(t) = x(t) + n(t)$, is only observed. Here, $x(t)$ is the clean speech and $n(t)$ is background noise. The output of the k-th sub-band, $Y_{N,\,k}(t)$, is represented as the analytical form by:

$$\begin{aligned} Y_{N,k}(t) &= Y_{N,1,k}(t) + Y_{N,2,k}(t), \\ &= A_{N,k}(t)\exp\left(j\omega_k t + j\phi_{N,k}(t)\right), \end{aligned} \tag{1}$$

where $Y_{N,\,1,\,k}(t)$ and $Y_{N,\,2,\,k}(t)$ are the sub-band components of $x(t)$ and $n(t)$, respectively. In addition, ω_k is the center frequency of the kth sub-band. $A_{N,\,k}(t)$ and $\phi_{N,\,k}(t)$ are the instantaneous amplitude and phase of the noisy speech, which are calculated as follows:

$$A_{N,k}(t) = |\tilde{f}(c,t)|, \tag{2}$$

$$\phi_{N,k}(t) = \int_0^t \left(\frac{d}{d\tau} \arg(\tilde{f}(c,\tau) - \omega_k) \right) d\tau, \tag{3}$$

where, $c = \alpha^{k-K/2}$, α is the scale of GTFB. $|\tilde{f}(c,t)|$ is the amplitude spectrum defined by the wavelet transform and arg $(\tilde{f}(c,t))$ is the unwrapped phase spectrum defined by the complex wavelet transform (Surhone et al., 2010; Unoki and Akagi, 1999). Then, the Kalman filter with trained LP is applied to remove the effects of noise on the instantaneous amplitude and phase because it is particularly effective in smooth prediction with the instantaneous amplitude and phase in sub-bands. Moreover, it can be regarded as the optimal estimator for both instantaneous amplitude and phase under non-stationary conditions. Finally, the restored signal, $\hat{x}(t)$ is resynthesized from the restored sub-bands components by inverse GTFB. However, this scheme cannot be applied in noisy reverberant environments because the reverberation is not considered in this model and the parameters in the Kalman filter need to be adapted for noisy reverberant environments.

3. Proposed scheme

The proposed scheme is an extension of the previous scheme, and the block diagram of the proposed scheme is shown in Fig. 1. The proposed scheme consists of three stages: analysis, modification, and resynthesis.

The noisy reverberant speech, $y_{NR}(t) = x(t) * h(t) + n(t)$, is observed. Here, $h(t)$ is the room impulse response (RIR). The RIR, $h(t)$, contains both effects of early reflection and late reverberation so that this can be represented as $h(t) = h_E(t) + h_L(t)$, where $h_E(t)$ is early reflection and $h_L(t)$ is late reverberation, as shown in Fig. 2. Then we have $y_{NR}(t) = x(t) * h_E(t) + x(t) * h_L(t) + n(t) = x_E(t) + x_L(t) + n(t)$, where $x_E(t)$ is early reverberant speech and $x_L(t)$ is late reverberant speech. Early reflection may not significantly degrade the quality or intelligibility of speech because humans cannot perceive these echoes as different sounds (Blauert, 1983; Plack, 2010), while late reverberation is detrimental to the quality and intelligibility.

The output of the kth sub-band, $Y_{NR,\,k}(t)$, is represented as the analytical form by:

$$\begin{aligned} Y_{NR,k}(t) &= Y_{NR,1,k}(t) + Y_{NR,2,k}(t), \\ &= A_{NR,k}(t)\exp\left(j\omega_k t + j\phi_{NR,k}(t)\right), \end{aligned} \tag{4}$$

where $Y_{NR,\,1,\,k}(t)$ and $Y_{NR,\,2,\,k}(t)$ are the components of $x(t)*h_E(t)$ and $x(t) * h_L(t) + n(t)$, respectively. $A_{NR,\,k}(t)$ and $\phi_{NR,\,k}(t)$ are the