Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 70 (2015) 13–27

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Restoration scheme of instantaneous amplitude and phase using Kalman filter with efficient linear prediction for speech enhancement

Naushin Nower *, Yang Liu, Masashi Unoki

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

Received 12 August 2014; received in revised form 18 February 2015; accepted 23 February 2015
Available online 5 March 2015

Abstract

This paper proposes a restoration scheme for the instantaneous amplitudes and phases in sub-bands by using a Kalman filter with linear prediction (LP). A few important studies have already proved that the phase spectrum in the short-time Fourier transform plays an important role in speech enhancement. Thus, the proposed scheme concentrates on simultaneously restoring both instantaneous amplitudes and phases. The Kalman filter, which is an optimal estimator in this scheme, is used for both instantaneous amplitudes and phases in the sub-band representation to remove the effect of noise. We found that the effectiveness of the Kalman filter depended on accurate estimates of LP coefficients. We propose an effective LP training phase to derive gender and content independent LP coefficients as central processing for Kalman filtering. We carried out objective and subjective tests under various noisy conditions to evaluate the effectiveness of the proposed scheme and compared it with typical methods. The signal to error ratio (SER), perceptual evaluation of speech quality (PESQ), and SNR loss were used as objective measures in these simulations. The mean preference score was used in subjective evaluations. The results revealed that the proposed scheme could effectively improve these objective and subjective measures more than those with typical methods.

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

1. Introduction

The required speech signal in real world scenarios is frequently smeared by various kinds of noise. This noise not only degrades the perceptual aspects of speech quality and speech intelligibility but also reduces the performance of various automated speech systems, such as automatic speech recognition systems, speaker recognition systems, and hearing aids. Therefore, the quality and intelligibility of speech signals in noisy environments have to be enhanced.

Speech enhancement is concerned with improving the quality and intelligibility of corrupted speech in the presence of noise. Various methods of speech enhancement have already been proposed during the past two decades to remove the effects of noise from noisy speech to improve its quality. Of these, classical methods of speech enhancement, such as spectral subtraction (SS) (Boll, 1979), the Ephraim–Malah algorithm (MMSE-STSA estimator) (Ephraim and Malah, 1984), and the Scalart–Filho algorithm (Wiener filtering) (Scalart and Filho, 1996), have attracted a great deal of attention because of their simplicity and efficiency in estimating spectral magnitude. The SS method (Boll, 1979) subtracts the estimated noise magnitude spectrum from the noisy speech magnitude spectrum, where the noise spectrum can be estimated and updated during periods when speech is absent. The Wiener filter

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

http://dx.doi.org/10.1016/j.specom.2015.02.006
0167-6393/

(Scalart and Filho, 1996) algorithm filters noisy speech signals by using a filter derived based on the minimum mean-square error (MMSE) criterion. These methods employ the short-time Fourier analysis-modification-synthesis (AMS) framework for speech enhancement.

There are various statistical methods of model-based speech enhancement in the literature in addition to these. Modeling in the model-based approaches is done using the statistical properties of the speech signal over multiple frames. This modeling is performed using the hidden Markov model (HMM) (Ephraim et al., 1989; Ephraim, 1992; Zhao and Kleijn, 2007), the Gaussian mixture model (GMM), or codebook-based methods (Sriram et al., 2007). HMM-based speech enhancement is a renowned model-based technique and resolves common problems with classical methods of speech enhancement in dealing with rapid variations in noise characteristics (Veisi and Hossein, 2013). Nishikawa et al. (2003) combined an independent component analysis (ICA) based noise estimator with multi-channel-wise non-linear signal processing to reduce noise further in their method of noise reduction. However, all of their improvements were limited because they only considered the spectral domain.

Recent research has investigated the importance of speech enhancement in the modulation domain, such as the modulation-domain Kalman filter (MDKF) (So and Paliwal, 2011; Paliwal et al., 2012). So and Paliwal's method modeled temporal changes in the magnitude spectrum for both speech and noise without taking into consideration noise in the phase component (So and Paliwal, 2011). Consequently, the corpus based approach (Ji et al., 2011), model based speech enhancement with spectral estimation (Ruofei et al., 2012), and speech enhancement using nonnegative matrix factorization (NMF) (Mohammadiha et al., 2013; Sawada et al., 2013) are included in modern methods of speech enhancement. All the existing methods process corrupted speech signals by modifying or correcting speech in either temporal or spectral magnitude only and keeping the phase component unchanged. This is because the phase spectrum that is conventionally considered is unimportant and has been demonstrated not to contribute much toward speech enhancement. Wang and Lim emphasized this (Wang and Lim, 1982), which is perhaps the most cited work to justify the unimportance of phase in speech enhancement.

However, recent studies have reported that the use of the phase spectrum in the short-time Fourier transform (STFT) can significantly improve speech enhancement (Shannon and Paliwal, 2006; Paliwal and Alsteris, 2005; Roux et al., 2008). Shannon and Paliwal (2006) reported that they had carried out magnitude only and phase only experiments to investigate the effects of phase in speech enhancement. Clean magnitude was used in the magnitude-only experiment and the phase was set to a random value. In contrast, a clean phase was used in the phase-only experiment and the magnitude was set to one. Their results indicated that the phase spectrum also contained useful and important information. After that, Paliwal and Alsteris (2005) investigated whether the shape and length of the window function used in STFT for phase manipulation were important factors for speech enhancement.

Fardkhaleghi and Savoji (2010) investigated the role of the phase spectrum in speech enhancement using Wiener filtering and Martin minimum statistics. To do this, the noisy phase of each processed frame was corrected by applying three proposed optimization algorithms, which demonstrated that better results were achieved using phase correction. Zhang and Zhao (2013) also subtracted real and imaginary spectra separately in the modulation frequency domain to enhance magnitude as well as phase through spectral subtraction for better enhancement. Kleinschmidt et al. (2011) proposed a method of utilizing phase information by complementing it with conventional magnitude-only spectral subtraction speech enhancement through complex spectrum subtraction (CSS). They demonstrated that an average improvements of 20% in relative word accuracy were possible for automatic speech recognition systems when accurate estimates of the phase spectrum were available.

Paliwal et al. (2011) demonstrated that modifying the phase spectrum could greatly improve speech enhancement. By investigating various cases where different combinations of noisy, clean (noiseless), and compensated amplitude and phase spectra were considered. This suggested that significant speech enhancement could be possible if the clean phase was known or the compensated phase spectrum was available. They also studied the effect of mismatched or matched windows for both amplitude and phase spectra estimates during AMS in STFT. Their results indicated that a proper choice of an analysis window and AMS setting in the phase spectrum could significantly improve speech enhancement. All existing research on the phase spectrum has either emphasized the importance of phase in speech enhancement or investigated a suitable size and shape for the window to enable phase manipulation. Thus, we obviously need to consider both the amplitude and phase of the noisy signal to enable better speech enhancement.

It is well-known that all existing speech enhancement algorithms based on STFT-AMS can improve speech quality but not speech intelligibility (Loizou and Kim, 2011). The reasons for this are still unclear so that many researchers have investigated probable strategies for reducing distortions and enhancing features related to speech intelligibility. However, it has been found from psychoacoustical studies that the temporal envelope (TE) and temporal fine structure (TFS) are important cues for speech perception (Drullman, 1995; Moore, 2008). It has also been revealed that TE and TFS play an important role in improving the intelligibility of noise-degraded speech (Swaminathan, 2010; Swaminathan and Heinz, 2012). Therefore, AMS in the filterbank is a suitable framework for speech enhancement, rather than AMS in STFT. Hence, it is expected that temporal amplitude and phase