Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 509–522

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

# Real and imaginary modulation spectral subtraction for speech enhancement

Yi Zhang, Yunxin Zhao *

*Department of Computer Science, University of Missouri-Columbia, Columbia, MO 65211, USA*

Received 11 October 2011; received in revised form 23 August 2012; accepted 5 September 2012
Available online 6 November 2012

**Abstract**

In this paper, we propose a novel spectral subtraction method for noisy speech enhancement. Instead of taking the conventional approach of carrying out subtraction on the magnitude spectrum in the acoustic frequency domain, we propose to perform subtraction on the real and imaginary spectra separately in the modulation frequency domain, where the method is referred to as MRISS. By doing so, we are able to enhance magnitude as well as phase through spectral subtraction. We conducted objective and subjective evaluation experiments to compare the performance of the proposed MRISS method with three existing methods, including modulation frequency domain magnitude spectral subtraction (MSS), nonlinear spectral subtraction (NSS), and minimum mean square error estimation (MMSE). The objective evaluation used the criteria of segmental signal-to-noise ratio (Segmental SNR), PESQ, and average Itakura–Saito spectral distance (ISD). The subjective evaluation used a mean preference score with 14 participants. Both objective and subjective evaluation results have demonstrated that the proposed method outperformed the three existing speech enhancement methods. A further analysis has shown that the winning performance of the proposed MRISS method comes from improvements in the recovery of both acoustic magnitude and phase spectrum.

*Keywords:* Spectral subtraction; Noise reduction; Speech phase; Modulation frequency

## 1. Introduction

The goal of speech enhancement is to improve speech quality in noisy environment, which requires finding a good tradeoff between noise reduction and speech distortion introduced during the enhancement process. Various speech enhancement techniques have been generated and applied to the real world noisy speech. In general, by using more hardware to acquire spatial information of a target speech source, multi-channel speech enhancement techniques (Farrel et al., 1992; Yellin and Weinstein, 1996) can provide enhancement performance superior to single channel enhancement methods. However, due to its convenient implementations, single channel speech enhancement has remained a hot spot in speech research. Some widely used single channel speech enhancement methods include spectral subtraction, Wiener filtering, and MMSE, etc.

Spectral subtraction is one of the most widely used speech enhancement techniques (Boll, 1979). Spectral subtraction methods typically focus on signal magnitude spectrum and use noisy phase spectrum in signal reconstruction, where the signal magnitude spectrum is estimated by subtracting an estimate of the noise magnitude spectrum from the noisy signal magnitude spectrum. A major drawback of the spectral subtraction approach is the introduced musical tone in the enhanced speech which is caused by the mismatch of the noise estimate and the true noise.

Wiener filtering (Chen et al., 2006) aims at reducing noise by minimizing the mean square error between the estimated and the clean speech signals. The major shortcoming of the Wiener filter approach is the requirement of a priori knowledge of the power spectrum of the clean speech.

* Corresponding author. Tel.: +1 573 882 3374; fax: +1 573 882 8318.
*E-mail addresses:* yzcb3@mail.missouri.edu (Y. Zhang), Zhaoy@missouri.edu (Y. Zhao).

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.09.005

MMSE (Ephraim and Malah, 1984) uses a Bayesian approach to determine the clean speech magnitude spectrum assuming Gaussian distributions for the speech and noise magnitudes. It is worth noting that under this assumption, noisy speech phase was proved to be the optimal phase for the enhanced speech, and hence only the magnitude MMSE has been used in speech enhancement applications.

Speech phase spectrum has been considered insignificant in perceptual speech quality (Wang and Lim, 1982), and so traditional enhancement methods focus on magnitude spectrum enhancement and use noisy phase spectrum in reconstructing speech. When SNR is high, noisy speech phase is indeed close to clean speech phase, and using noisy phase to replace clean phase would not introduce perceptual distortion. However, when SNR drops low, noisy phase plays a more apparent role in the enhanced speech. It has been indicated that when the spectral SNR is lower than approximately 8 dB for all frequencies, a mismatch in phase might be perceived as "roughness" in speech quality (Loizou, 2007), which means that under this condition, even if we had the exact clean speech magnitude spectrum, we would not be able to recover the clean speech signal with unperceivable distortion.

Recently, more interests in speech phase have been reported. Phase information was used to generate features for automatic speech recognition (Schluter and Ney, 2001; Zhu and Paliwal, 2004; Hegde and Murthy, 2007), and phase information was applied to improve perceptual quality of enhanced speech. Shannon and Paliwal (2006) investigated estimating the short time Fourier transform (STFT) phase spectrum independently from the STFT magnitude spectrum for speech enhancement applications and observed substantial improvements in noise reduction and speech quality. Wójcicki et al. (2008) proposed phase spectrum compensation to control the amount of reinforcement or cancellation that occurs during the synthesis of the enhanced signal by adding an anti-symmetry function to the noisy speech signal in the frequency domain. Aarabi and Shi (2004) proposed phase-error filtering based on the assumption that phase variations between multiple microphone channels after time delay compensation are due purely to the influence of the background noise, where the observed between-channel phase difference is used to filter noisy speech such that a larger phase difference results in a greater signal attenuation. Lu and Loizou (2008) proposed a geometric spectral subtraction approach that addressed the shortcomings of spectral subtraction concerning musical noise and speech-noise cross-term issues, where they used the phase differences between the noisy signal and the noise to estimate the cross-terms. Fardkhaleghi and Savoji (2010) investigated the role of phase spectrum in speech enhancement using Wiener filtering and minimum statistics and showed that better results are achieved using phase correction for different noise types. Kleinschmidt et al. (2011) proposed a novel method for acquiring phase information and used the phase information to complement the traditional magnitude-only spectral subtraction in speech enhancement, and they obtained good results in a 15–20 dB SNR environment.

In this current work, we propose a new approach to spectral subtraction for enhancing speech signal from noise, where the subtraction processing is performed on the real and imaginary spectra separately, and the separately enhanced spectra are used to recover the complex signal spectra. This approach is supported by our experimental observation that the real, imaginary, and magnitude spectra have similar time–frequency (T–F) characteristics. We carry out the subtraction processing in the modulation frequency domain for the purpose of reducing musical noise as proposed in (Paliwal et al., 2010). Differing from Paliwal et al. (2010) where the noisy speech acoustic magnitude spectra that contain the cross-terms of speech and noise were transformed to the modulation frequency domain for spectral subtraction, our separate transformation of the real and imaginary acoustic spectra to the modulation frequency domain does not carry the acoustic-domain speech-noise cross-terms. Furthermore, unlike many speech enhancement methods, our synthesis of speech signal from the modified acoustic spectra does not use the acoustic phase spectra of the noisy speech. We conducted comparative experiments using the criteria of segmental signal-to-noise ratio (SNR), PESQ, and ISD to evaluate the performance of the proposed MRISS method against three existing methods, including modulation-frequency domain spectral subtraction (MSS), nonlinear spectral subtraction (NSS), and minimum mean-square error (MMSE) estimator, in noisy conditions of five noise types (white, bable, pink, volvo, factory2 noises) and four SNR levels (−5, 0, 5, and 10 dB).

The organization of this paper is as follows. In Section 2, we discuss the background of conventional spectral subtraction algorithms; in Section 3 we introduce our proposed speech enhancement method; in Section 4 we present experimental results, and in Section 5 we give a conclusion.

## 2. Background of spectral subtraction

### 2.1. Acoustic domain spectral subtraction

A typical method of spectral subtraction performed in the acoustic frequency domain is the generalized frame-by-frame subtraction (Berouti et al., 1979; Boll, 1979) defined as:

$$|\widehat{S}(k,t)|^{\gamma} = \begin{cases} |X(k,t)|^{\gamma} - \alpha(k)|\widehat{N}(k,t)|^{\gamma} & if\,|X(k,t)|^{\gamma} > (\alpha(k)+\beta)|\widehat{N}(k,t)|^{\gamma} \\ \beta|\widehat{N}(k,t)|^{\gamma} & otherwise \end{cases} \quad (1)$$

where $|X(k, t)|$ is the noisy speech magnitude spectrum, $|\widehat{N}(k,t)|$ is the noise magnitude spectral estimate, $|\widehat{S}(k,t)|$ is the reconstructed speech magnitude spectrum, $k$ and $t$ are the frequency and the time indices, respectively; $\alpha(k)$ is an over-subtraction factor which is a function of segmental SNR (Kamath and Loizou, 2002), $\beta$ is a spectral flooring factor that controls the effect of over-subtraction and