Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 60 (2014) 44–55

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A novel speech enhancement method based on constrained low-rank and sparse matrix decomposition

Chengli Sun [a,b,*], Qi Zhu [c], Minghua Wan [b]

[a] Science and Technology on Avionics Integration Laboratory, Shanghai 200233, China
[b] School of Information, Nanchang Hangkong University, Nanchang 330063, China
[c] Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

Received 7 September 2013; received in revised form 3 March 2014; accepted 4 March 2014
Available online 13 March 2014

Abstract

In this paper, we present a novel speech enhancement method based on the principle of constrained low-rank and sparse matrix decomposition (CLSMD). According to the proposed method, noise signal can be assumed as a low-rank component because noise spectra within different time frames are usually highly correlated with each other; while the speech signal is regarded as a sparse component since it is relatively sparse in time–frequency domain. Based on these assumptions, we develop an alternative projection algorithm to separate the speech and noise magnitude spectra by imposing rank and sparsity constraints, with which the enhanced time-domain speech can be constructed from sparse matrix by inverse discrete Fourier transform and overlap-add-synthesis. The proposed method is significantly different from existing speech enhancement methods. It can estimate enhanced speech in a straightforward manner, and does not need a voice activity detector to find noise-only excerpts for noise estimation. Moreover, it can obtain better performance in low SNR conditions, and does not need to know the exact distribution of noise signal. Experimental results show the new method can perform better than conventional methods in many types of strong noise conditions, in terms of yielding less residual noise and lower speech distortion.

Keywords: Speech enhancement; Matrix decomposition; Low-rank matrix approximation; Robust principal component analysis

1. Introduction

In real-world environments, speech signals are easy corrupted by adverse noise such as background noise, vehicle noise, channel noise, competing speakers, etc. The goal of speech enhancement is to improve the quality and intelligibility of speech by reducing noise (Loizou, 2007). This paper focuses on enhancement of speech corrupted by additive noise when using single microphone recordings. Over the last fifty decades, many algorithms have been proposed about this field. The typical algorithms including spectral subtraction (Boll, 1979), minimum mean square error (MMSE) estimation (Ephraim and Malah, 1985; Ephraim and Malah, 1984; Stark and Paliwal, 2011), Wiener filtering (WF) (Soon and Koh, 2000; Wiener, 1949; Plapous et al., 2006; Scalart and Vieira-Filho, 1996), and subspace methods (Moor, 1993; Ephraim and Van Trees, 1995; Doclo and Moonen, 2002; Hu and Loizou, 2003; Hermus et al., 2007).

Spectral subtraction and Wiener filtering were among the first introduced speech enhancement techniques. They remain popular today due to their reasonable performance and low computational complexity. Spectral subtraction performs subtraction of an estimated noise magnitude spectrum from a noisy speech magnitude spectrum, where the noise spectrum can be estimated and updated during

[*] Corresponding author at: School of Information, Nanchang Hangkong University, Nanchang 330063, China. Tel.: +86 079183863741.
E-mail address: sun_chengli@163.com (C. Sun).

http://dx.doi.org/10.1016/j.specom.2014.03.002
0167-6393/© 2014 Elsevier B.V. All rights reserved.

periods when the speech is absent (Boll, 1979). The major drawback of spectral subtraction is that it suffers from the problem of musical noise distortion because of inaccurate estimation of the noise spectrum (Lu and Loizou, 2008; Paliwal et al., 2010). The WF algorithm assumes that both the clean speech and additive noise have a Gaussian distribution, and performs filtering of a noisy speech signal by using a filter derived based on the minimum mean-square error (MSE) criterion (Soon and Koh, 2000). Recent research showed that the DFT coefficients of short time stationary clean speech signal might be better modeled by super-Gaussian distribution (Hai-yan et al., 2011), such as Laplace and Gamma distribution, and the corresponding minimum mean-square error based algorithm was presented for speech enhancement. An alternative popular approach is subspace method, which was originally proposed by Ephraim and Van Trees (1995). The main idea of subspace method is to consider the noisy signal as a vector in a p-dimensional vector space and to separate this space into two orthogonal subspaces: the signal-plus-noise subspace (with dimension smaller than p, corresponding to the clean signal), and the noise subspace, which is the orthogonal complement of the signal-plus-noise subspace (Hermus et al., 2007). Subspace decomposition is achieved by applying the Karhunen–Loeve transform (KLT) Ephraim and Van Trees, 1995 or singular value decomposition (SVD) to the noisy signal (Moor, 1993). The decomposition is under the assumption of a low-rank linear model for speech and an uncorrelated additive (white) noise interference. Speech enhancement is performed in the time-domain by removing the noise subspace and by estimating the clean speech signal from the remaining signal-plus-noise subspace.

Single-channel speech enhancement systems traditionally employ voice activity detection (VAD) to estimate and update the statistics of the noise signal during noise-only segments. However, current VAD approaches remain imperfect in low SNR conditions, which in turn causes speech enhancement methods to underperforms. Moreover, even if the VAD is reliable, changes in the noise spectrum occurring during active speech cannot influence the noise estimate in a timely manner, which would results in a poor estimate during long speech sentences with few noise-only excerpts available (Manohar and Rao, 2006).

In this paper we propose a novel speech enhancement method based on the principle of constrained low-rank and sparse matrix decomposition (CLSMD). The new method significantly differs from the previous methods in its motivation and methodology. The main idea behind our method is motivated by the recent development of robust principal component analysis (RPCA) theory (Wright et al., 2009; Candes et al., 2011). RPCA states that if an observation matrix is the superposition of a low-rank component and a sparse component, these two components can be perfectly recovered with an overwhelming probability under mild conditions. In time–frequency (T–F) domain, since noise signals within different time-frames are usually correlated with each other, the noise spectrogram can be assumed to be in a low-rank subspace. On the other hand, speech signals can be regarded as relatively sparse in the T–F domain (the same assumption of speech can be seen in Huang et al. (2012)). Thus RPCA can be exploited to reconstruct clean speech from noisy signal. However, RPCA may yield undesirable outcomes when prior knowledge of signal source is not applied. Therefore, we propose a CLSMD algorithm for speech and noise spectrogram separation by imposing the low-rank and sparsity constraints.

We believe the proposed CLSMD-based method will be a new promising direction for speech enhancement, as it has following excellent properties: (1) The proposed method is a non-parametric method. There were no specific assumptions made about the distribution of the spectral components of either speech or noise. It only requires that noise is low-rank and speech is sparse in the time–frequency space. (2) Due to the speech and noise spectra can be simultaneously recovered, the VAD process is needless in the CLSMD framework. This is superior to many traditional speech enhancement systems which rely on VAD for noise estimation. (3) The proposed method has the advantages of few tuning parameters and fast operation speed. Moreover, it can work well in strong noise conditions. Experimental results showed that the new method can obtain less residual noise and lower speech distortion than several most commonly used speech enhancement methods in low-SNR situations.

This paper is organized as follows. Section 2 briefly reviews the background information and related work. In Section 3, we introduce the principle of CLSMD and its optimization algorithm, and then present the CLSMD-based speech enhancement system. The experimental results are described and analyzed in Section 4. Finally, some conclusions are drawn in Section 5.

2. Robust PCA

Principal component analysis (PCA) (Jolliffe, 2002) is the most widely used statistical technique for dimensionality reduction. It assumes that the given high-dimensional data lie near a lower-dimensional linear subspace. The goal of PCA is to accurately estimate this low-dimensional subspace. Suppose the given data are arranged as the columns of a large matrix $M \in \mathbb{R}^{N \times K}$, PCA seeks the best rank-r matrix estimate of M by solving

$$\min_{L} ||M - L||_F, \quad \text{s.t. } M = L + E, \tag{1}$$

where $r \leqslant \min(N, K)$ is the target dimension of the subspace, $\|\cdot\|_F$ is the Frobenius norm and $||X||_F = \sqrt{X_{ij}^2}$, L is a low-rank matrix and E is a small perturbation matrix. This problem can be efficiently solved via the singular value decomposition (SVD) and may get the optimal estimate when the noise E is small, independent and identically distributed Gaussian.