Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 76 (2016) 42–60

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

A novel speech enhancement method by learnable sparse and low-rank decomposition and domain adaptation

Samira Mavaddaty, Seyed Mohammad Ahadi*, Sanaz Seyedin

Electrical Engineering Department, Amirkabir University of Technology, 424 Hafez Ave, Tehran, Iran

Received 7 April 2015; received in revised form 2 November 2015; accepted 5 November 2015
Available online 1 December 2015

Abstract

This paper proposes a novel speech enhancement algorithm based on a low-rank sparse decomposition model. The sparse and low-rank components of a corrupted signal are considered as speech and noise in time-frequency domain, respectively. We use a new alternating optimization algorithm for accurate decomposition of the noisy observed data using sparse coding over the speech and noise dictionaries.

Adequate noise training frames with the same data size as speech data are provided by a noise estimation algorithm to learn the overcomplete noise dictionaries with low sparse approximation error. Since encountering non-stationary noises reduces the performance of speech enhancement methods, we take advantage of domain adaptation method which is a novel speech enhancement procedure to transform a learned noise dictionary to an adapted dictionary by data distribution captured in the enhancement process.

Using this step, an adapted separation scheme based on the current situation of noisy space is carried out, and the main drawbacks seen in the earlier dictionary-learning-based speech enhancement methods are solved. The proposed approach results in a significant reduction of noise, especially for non-stationary noises, in comparison with the earlier methods in this context and the traditional procedures, based on different objective and subjective measures.

Keywords: Speech enhancement; Dictionary learning; Sparse representation; Domain adaptation; Speech dictionary; Noise dictionary.

1. Introduction

The background noise can reduce both quality and intelligibility of the speech signals and results in a defective performance in many areas dealing with signal processing, such as mobile phones, hearing devices and voice recognition. This paper focuses on speech enhancement, when a single microphone records the noisy signals. The difficulties in these kinds of methods are more prominent when the speech signal is contaminated by non-stationary noise. In the case of speech-like noises, an indispensable overlap exists between speech and noise signals in time-frequency domain. The objective of speech enhancement is signal improvement to increase either intelligibility or signal quality. This process is supposed to be carried out by audio signal processing techniques via attenuating noise without causing any distortion in the speech signal. Various speech enhancement algorithms based on different basic approaches such as spectral subtraction (Kamath and Loizou, 2002), Wiener filtering (Scalart and Filho, 1996), statistical-based methods (Ephraim and Malah, 1985) and subspace techniques (Hu and Loizou, 2003) have been proposed over the years. The performance of the mentioned approaches is often dependent on the estimated noise, often captured during the speech absence. Spectral subtraction is a well-known method and among the very first approaches proposed for speech enhancement. However, an annoying artifact called musical noise is often observed in the processed signal (Kamath and Loizou, 2002). The filtering process using minimum mean square error (MMSE) in (Scalart and Filho, 1996) is performed with the assumption of a fixed Gaussian distribution for speech and noise signals, which is not always correct. The methods based on (Ephraim and Malah, 1985), similar to wiener filtering method, are inherently able to enhance the signals corrupted by stationary noises. Subspace algorithms are more promising than other categories to model non-stationary noises, but the assumption of orthogonality between the speech and noise subspaces leads to the degradation of the enhancement performance in the presence of different real world noises (Hu and Loizou, 2003). Thus, speech enhancement is still a challenging

* Corresponding author. Tel.: +98 21 6454 3336; fax: +98 21 6640 6469.
E-mail addresses: s.mavaddaty@aut.ac.ir (S. Mavaddaty), sma@aut.ac.ir (S.M. Ahadi), sseyedin@aut.ac.ir (S. Seyedin).

http://dx.doi.org/10.1016/j.specom.2015.11.003
0167-6393/© 2015 Elsevier B.V. All rights reserved.

open problem in realistic environments where speech information encounters non-stationary interferences.

In recent years, there has been an increasing interest in utilization of sparse representation techniques in speech enhancement (Sigg et al., 2010, 2012). In (Sigg et al., 2010), an ideal voice activity detector (VAD) has been applied to obtain training data for noise dictionary learning. This training data has been captured within the initial, middle or final sections of the noisy signal that is not usually enough for a desired convergence of the overcomplete dictionary. Later, generative dictionary learning method was proposed in (Sigg et al., 2012) with enough pure noise data according to the coherence measure. The main problem here is that the structure of noise in noisy frames does not necessarily contain an exact sparse representation in a dictionary trained on pure noise, and thus leads to speech distortion or too sparse coding. This malfunction appears due to coding some of the speech atoms using noise frames.

Moreover, the presence of multiple noise dictionaries in enhancement process causes a time consuming sparse coding, and may lead to either confusion or too dense coding because of inexact representation of the speech frames over different interferer atoms. On the other hand, the essential similarities between some atoms of different noises in concatenated dictionary can lead to sparse coding of noisy frames on undesired noise dictionary, and considerably increase approximation error of sparse coding. Recently, another speech enhancement technique based on robust principal component analysis (RPCA) is considered (Sun et al., 2014; Chen and Ellis, 2013; Huang et al., 2014). However, employing RPCA for speech enhancement in an unsupervised manner does not lead to the desired results, especially in the presence of non-stationary noises with low SNR levels. This problem may be alleviated by incorporating some knowledge about the statistics of the input sources in the component separation process. In (Sun et al., 2014), RPCA technique is used based on an alternating projection algorithm to decompose noisy spectrogram by setting constraints on rank and sparsity of each output component. The sparse and low-rank components in this method are obtained only from hard thresholding using shrinkage function and singular value decomposition (SVD) of the observed signal, respectively. As expected, better results are obtained in the presence of noises with high stationarity.

Another dictionary-based speech enhancement method is based on non-negative matrix factorization (NMF) and attempts to decompose a non-negative matrix, usually the spectrogram of the speech signal, into a set of basis vectors with non-negative coefficients (Mohammadiha et al., 2011; Wilson et al., 2008). In (Mohammadiha et al., 2011), NMF-based noise power spectral density (PSD) estimation is proposed. The speech and noise models are learned off-line and then a constrained NMF which is based on the time dependencies of the speech and noise signals is utilized to make a smooth estimate of the noise signal. The estimated noise PSD is used in combination with a Wiener filter to enhance the noisy speech.

In order to solve the mentioned problems in dictionary-based speech enhancement approaches, we propose a novel speech enhancement method that is based on learnable sparse and low-rank decomposition and domain adaptation (LSLDA). The pre-trained dictionaries for speech and different noises have been learned on clean speech and distorted speech training corpora, respectively. We use a noise estimation algorithm similar to (Rangachari and Loizou, 2006) to provide enough noise data for noise dictionary learning with lower reconstruction errors and better fitting. In the enhancement step, the learned noise dictionary, selected based on the type of the environment noise, is adapted to the new conditions according to the estimated data captured in the test domain. This step is carried out using domain adaptation method to yield better enhancement results, especially in the case of non-stationary noises in low SNRs values adapted from (Chen et al., 2012). Then, an alternating projection algorithm based on the idea of RPCA technique has been used to solve iteratively two sub-problems that are based on the sparsity of the speech signal and low-rank property of the noise signal in time-frequency domain.

It should be mentioned that the approach in this paper is proposed to address the shortcomings of the previous speech enhancement methods based on dictionary learning to achieve a robust algorithm. To this end, we utilized the ideas suggested in (Rangachari and Loizou, 2006) and (Chen et al., 2012), and modified and adapted them to our own method to obtain a more robust technique and better results. One of the fundamental difficulties in (Sigg et al., 2012) is that the composite dictionary in the enhancement step includes all noise dictionaries in which leads to increasing the approximation error for sparse representation of the noisy frames. The reason is that different noise types have similarities in their frequency contents in many cases that cause a noisy frame has been sparsely coded over several noise dictionaries. Also, a very time consuming enhancement procedure is yielded. Therefore, we propose considering only one noise dictionary for the proper representation of existing noisy conditions in the test step that results in an algorithm with much less computation time.

In addition, noise dictionary learning process needs enough noise data for better enhancement results in the presence of various non-stationary noises especially at low SNRs. Noise data provided from a VAD algorithm as proposed in (Sigg et al., 2010) is not enough for an exact sparse representation with low approximation error. Thus, we utilize the noise estimation algorithm in (Rangachari and Loizou, 2006) to obtain enough noise data with the same data size as speech signal resulting in a better fitting and lower reconstruction error in noise dictionary learning procedure. On the other hand, the time-frequency content of noise signal in the test step can be different from the estimated noise in the training step especially in the presence of noises with high non-stationarity characteristics. Hence, the selected noise dictionary may be inconsistent with the desired noise. In order to alleviate this mismatch, we adapt the learned noise dictionary to a new noise dictionary using the estimated noise in the test step inspired by the proposed algorithm in (Chen et al., 2012). Earlier, this technique was employed for image denoising and it has not been used in speech processing procedures such as speech enhancement so far. Thus, we adapted the algorithm in (Chen et al., 2012) appropriately for the case of speech enhancement. The sparse coefficients using this adapted dictionary are adjusted in a way to exactly represent noisy section of the