Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language xxx (2017) xxx-xxx

www.elsevier.com/locate/csl

Speech enhancement using sparse dictionary learning in wavelet packet transform domain

Q2 Samira Mavaddaty*, Seyed Mohammad Ahadi, Sanaz Seyedin

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

Received 16 April 2016; received in revised form 5 January 2017; accepted 7 January 2017

Abstract

Sparse coding, as a successful representation method for many signals, has been recently employed in speech enhancement. This paper presents a new learning-based speech enhancement algorithm via sparse representation in the wavelet packet transform domain. We propose sparse dictionary learning procedures for training data of speech and noise signals based on a coherence criterion, for each subband of decomposition level. Using these learning algorithms, self-coherence between atoms of each dictionary and mutual coherence between speech and noise dictionary atoms are minimized along with the approximation error. The speech enhancement algorithm is introduced in two scenarios, supervised and semi-supervised. In each scenario, a voice activity detector scheme is employed based on the energy of sparse coefficient matrices when the observation data is coded over corresponding dictionaries. In the proposed supervised scenario, we take advantage of domain adaptation techniques to transform a learned noise dictionary to a dictionary adapted to noise conditions captured based on the test environment circumstances. Using this step, observation data is sparsely coded, based on the current situation of the noisy space, with low sparse approximation error. This technique has a prominent role in obtaining better enhancement results particularly when the noise is non-stationary. In the proposed semi-supervised scenario, adaptive thresholding of wavelet coefficients is carried out based on the variance of the estimated noise in each frame of different subbands. The proposed approaches lead to significantly better speech enhancement results in comparison with the earlier methods in this context and the traditional procedures, based on different objective and subjective measures as well as a statistical test.

Keywords: Speech enhancement; Dictionary learning; Sparse representation; Domain adaptation; Voice activity detector; Wavelet packet transform

1. Introduction

Environmental noise can significantly reduce the quality of the speech signal and results in an imperfect performance of hearing aid devices, automatic speech recognition systems, mobile phones etc. In this paper, we focus on single channel enhancement of speech corrupted by additive noise. The main problem in these kinds of speech enhancement methods arises from this fact that the capability of speech denoising, without any source distortion or source confusion, will be decreased when the speech signal is corrupted by nonstationary noises. Also, this difficulty

Q3 * Corresponding author.
E-mail address: 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.csl.2017.01.009
0885-2308/ 2017 Elsevier Ltd. All rights reserved.

will be prominent in case of speech-like noises where an essential overlap exists between the components of speech and noise signals in time-frequency domain.

1.1. Related sparse decomposition methods

Recently, there is considerable interest on the problem of dictionary-based speech enhancement algorithm based on non-negative matrix factorization (NMF). In this technique, a non-negative data matrix such as the spectrogram of the speech signal is factorized into two matrices that include a set of basis vectors and non-negative coefficients (Mohammadiha et al., 2011; Wilson et al., 2008). In Mohammadiha et al. (2011), the noise power spectral density (PSD) estimation algorithm using the constrained NMF is investigated and the speech and noise dictionaries are trained off-line. The proposed NMF procedure is employed based on the time correlation of the underlying noisy signal and provides smoother estimates of the nonnegative factors. Then, the estimated PSD in combination with a designed Wiener filter is utilized for denoising of the observed signal. A new version of NMF technique in combination with prior models of speech and noise signals for speech denoising is designed in Wilson et al. (2008). The statistical speech and noise models in this method provide the additional signal structure that results in the improvement of the denoising performance.

One of the categories of speech enhancement algorithms is based on robust principal component analysis (RPCA) (Sun et al., 2014a, 2014b; Chen and Ellis, 2013; Huang et al., 2014b). In Sun et al. (2014a), the constrained low-rank and sparse matrix decomposition (CLSMD) methods are proposed for speech enhancement. CLSMD uses an alternating projection algorithm to solve RPCA sub-problems in an iterative method and decomposes a noisy spectrogram by setting constraints on rank and sparsity of each input observation frame. In this method, the speech signal is considered as a sparse component and the noise signal is regarded as a low-rank component because of its correlated frames in the time-frequency domain. The RPCA-based approach proposed in Chen and Ellis (2013) uses the NMF method with a generalized KL-divergence algorithm to learn the speech dictionary without any noise model. The technique of alternating direction method of multipliers (ADMM) is employed in this algorithm to solve the optimization problem for spectrogram decomposition with adding different variables or parameters.

Other speech enhancement methods that have attracted much interest in recent years are exemplar-based procedures along with dictionary learning (Baby et al., 2015; Yılmaz et al., 2015a, 2015b, 2015c; Mohammadiha and Doclo, 2014). In Baby et al. (2015), features with lower dimension such as Mel-domain are regarded instead of the high dimension and full-resolution features in Fourier transform space to reduce computational time and problem complexity. Also, the noisy signal decomposition is carried out based on NMF procedure using the weighted summation of the speech and noise samples stored in learned dictionaries (Baby et al., 2015). The designed filters in a reduced resolution feature space require fewer parameters and have the ability to generalize better to unseen circumstances (Baby et al., 2015).

An adaptive noise dictionary learning approach to design a proper noise modeling for the noise robust exemplar matching procedure is proposed in Yılmaz et al. (2015a). The noisy signal in this method is approximated using sparse representation method with a linear combination based on the exemplars of multiple lengths. Using the tuned parameters to model the spectro-temporal content of the noise conditions, better recognition accuracy can be obtained (Yılmaz et al., 2015a).

1.2. DNN-based speech enhancement methods

Other types of speech enhancement approaches are based on deep neural network (DNN) methods that try to learn a non-linear mapping function between the pair of clean speech and noisy signals (Xu et al., 2014; Lu et al., 2013). These methods need a large size of training data to learn the generative structure of the mentioned mapping function based on the extracted features from speech and noise signals. By contrast, in our presented method, a small amount of training data is required. Also, a large number of hidden layers are utilized in DNN-based methods to get a generative mapping function to achieve a better capability in the mismatched test conditions. Moreover, a tradeoff between the size of training data and the size of the hidden layers exists, since if the size of training data is small, over-fitting may occur. This problem leads to an incorrect generalization because insufficient data exists for setting all parameters. It should be noted that the essence of dictionary-based and DNN-based methods is different. The input signal in