Computers and Electrical Engineering 46 (2015) 56–64

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Joint Sparsity and marginal classification for improving Sparse Imputation performance in speech recognition ☆

Mohammad Mohsen Goodarzi, Farshad Almasganj *

Biomedical Engineering Department, Amirkabir University of Technology, 15857 Tehran, Iran

ARTICLE INFO

Article history:
Received 5 February 2015
Received in revised form 15 July 2015
Accepted 15 July 2015
Available online 25 Juy 2015

Keywords:
Sparse Imputation
Robust automatic speech recognition
Compressive sensing
Joint Sparsity
Self-similarity

ABSTRACT

Sparse Imputation (SI) is a relatively new method that reconstructs missing spectral components of noisy speech signal with the help of the sparse-based representation approaches. In this method, the redundancy of signal in the frequency domain helps to rebuild noisy spectral components from the remained reliable ones. On the other hand different parts of speech signal, despite time intervals between them, can be inherently similar to each other. In this paper, a major modification over the SI method is proposed that in addition to data redundancy property of speech signal in small regions, takes the advantages of its self-similarity nature over long intervals. By identifying mostly similar frames, using a method based on the marginal classification, the Joint Sparsity method is applied and a method dubbed as the Joint Sparse Imputation is presented. The experiments conducted on AURORA 2 data set show that the proposed method significantly improves the recognition results in different noisy conditions, compared to the original SI method.

1. Introduction

The Sparse Imputation (SI) is a method in the field of noise robust speech recognition which follows reconstructing missing spectral components of speech signal using the sparse representation-based approaches. In fact, this approach is a subset of the missing feature/data approaches [1,2] which concentrates on compensation of the additive noise effects. In such methods, with the help of the mask estimation methods [3–6], unreliable spectral components (on which the destructive effect of noise is significant) are first identified and tagged. These unreliable components are then reconstructed using the remained reliable ones on which the effect of noise is negligible.

The Sparse Imputation method was first introduced by Gemmeke [7]. In this method, the emerging theory of compressive sensing (CS) [8] was successfully utilized to reconstruct the unreliable components of noise corrupted signals. In the Sparse Imputation, the CS is utilized to compensate noisy feature vectors for the task of automatic speech recognition; but, there are other approaches that use the CS to enhance speech signal quality [9].

In the common CS approach, signal $\mathbf{y}$ is sampled using a random measurement matrix $\Phi_{measurment}$; the goal of the CS approach is to recover $\mathbf{y}$ from the sampled signal $\mathbf{b} = \Phi_{measurment}\mathbf{y}$ using the recovery algorithm of CS as given by

$$\hat{\mathbf{x}} = \arg\ \min_{\mathbf{x}}\{\|\Phi_{measurment}\mathbf{y} - \Phi_{measurment}\mathbf{A}\mathbf{x}\|_2 + \lambda\|\mathbf{x}\|_1\}, \quad (1)$$

☆ Reviews processed and approved for publication by the Editor-in-Chief.

* Corresponding author. Tel.: +98 6454 2372.
E-mail address: almas@aut.ac.ir (F. Almasganj).

http://dx.doi.org/10.1016/j.compeleceng.2015.07.013
0045-7906/

where $\boldsymbol{A}$ is a sparsifying basis for $\mathbf{y}$ and $\mathbf{x}$ is a sparse vector in the $\boldsymbol{A}$ space. The scalar λ is a constant weight used to balance between the fidelity (ℓ_2 norm) and the sparseness (ℓ_1 norm) terms. In contrast, in the missing feature [2] problem, we have a noisy signal $\tilde{\mathbf{y}} = \mathbf{y} + \mathbf{n}$ that parts of it is destroyed by a random additive noise ($\mathbf{n}$). Discarding the destroyed parts and modeling this effect of noise with a measurement matrix Φ_{noise} such as $\Phi_{measurment}$, we could recover y from $\mathbf{b} = \Phi_{noise}\tilde{\mathbf{y}}$ (reliable components of observed noisy signal), using an algorithm similar to the recovery algorithm of CS.

Here, $\boldsymbol{A}$ could be a sparsifying basis for $\mathbf{y}$ [10] or a dictionary consisted of exemplars of $\mathbf{y}$ [11]. The clean estimate of $\tilde{\mathbf{y}}$ would be $\boldsymbol{A}\hat{\mathbf{x}}$.

The main basis of the missing feature methods is established on the inherent information redundancy in speech signal. This means that the information in speech signal is not limited to a certain range of its time–frequency spectrum. For more explanation, it can be referred to the performed experiments in [12,13] showing that even with a small part of speech spectrum, still the remaining signal is understandable to the human listeners. On the other hand, in numerous references it has been pointed out that speech signal has considerable fractal properties in the time domain [14–16]. The fractal properties are closely related to self-similarity and in other words it means that different parts of speech, despite having time intervals from each other, could be potentially very similar and could have the same natures. This phenomenon motivated us to propose a method that leverages the self-similarity feature of speech signal to help better reconstruction of the missing spectral components of noisy speeches. The importance of this issue in the field of the missing feature is that if some of the spectral components of a frame are lost, one would hope that in another time position of the same signal, there are one or more less degraded frames with nearly similar components, which, along with reliable components of the given frame, can be used for better reconstruction of its unreliable components. In this manner, a higher efficiency of missing feature methods can be expected. In other words, the noisy frames that are similar in nature are reconstructed together, in a Joint manner. A similar approach, also known as the Multiple Measurement Vectors, is used in many applications such as the distributed compressive sensing [17,18] and the images noise removal [19].

In this paper, we aim to exploit the self-similarity property, available in speech signal using the Joint Sparsity approach to improve the performance of the Sparse Imputation method. For this purpose, we first offer a method based on the GMM clustering and marginalization process in order to be able to determine the similar speech frames, despite the destructive effect of noise on them. Then, the relations needed to use the Joint Sparsity approach inside the Sparse Imputation method are presented.

The paper is organized as follows, in Section 2.1 we introduce the Sparse Imputation method. In Section 2.2, details of the Joint Sparsity method and its mathematical relations are presented. In Section 2.3, the necessary relations for employing the Joint Sparsity are detailed. Moreover, the proposed method for determining the similar frames of speech is presented in this section. Section 3 specifies the exploited data set and the performed experiments. Implementation results and the discussion are presented in Section 4, and finally in Section 5 conclusions are presented.

2. Methods

2.1. Sparse Imputation

The Sparse Imputation method is mainly introduced to compensate the effect of noise over the speech features. For this purpose, Gemmeke [20] adapted the compressive sensing approach to estimate the unreliable features of the noisy speech. In this way, he created a dictionary $\boldsymbol{A}$ from clean speech exemplars, in which, each exemplar is the Logarithm of the (Mel) Filter Bank Energies (LFBE) of a windowed speech frame. Assume that the clean feature vector $\mathbf{y}$ is affected by the additive noise $\mathbf{n}$ and yielded the observed noisy feature vector $\tilde{\mathbf{y}} = \mathbf{y} + \mathbf{n}$; so, the effect of the noise may be presented by a binary vector, as given by

$$\mathbf{m}(i) = \begin{cases} 1 \overset{def}{=} \textit{reliable} & \textit{if } \mathbf{y}(i) - \mathbf{n}(i) > \theta \\ 0 \overset{def}{=} \textit{unreliable} & \textit{otherwise} \end{cases}. \tag{2}$$

where "1" means that the i-th frequency components of $\tilde{\mathbf{y}}$ is less affected by noise and is so reliable to represent a clean feature. On the other hand, "0" means that the information of corresponding frequency element is buried in the noise level and its clean counterpart should be estimated. The threshold value θ may be obtained empirically. This process is called "mask estimation"; there are many different approaches [3–6] to estimate $\mathbf{m}$, which are not in the scope of this paper.

The vector $\mathbf{m}$ is then converted to a measurement like matrix Φ_{noise} and is put in (1). To do this, diagonal matrix Φ_{noise} is constructed such as $Diag(\Phi_{noise}) = \mathbf{m}$. The estimated clean vector $\mathbf{y}$ is obtained by first solving

$$\hat{\mathbf{x}} = \arg \min_{\mathbf{x}} \{\|\Phi_{noise}\mathbf{y} - \Phi_{noise}\boldsymbol{A}\mathbf{x}\|_2 + \lambda\|\mathbf{x}\|_1\}, \tag{3}$$

and then, evaluating $\tilde{\mathbf{y}} = \boldsymbol{A}\hat{\mathbf{x}}$. Assuming that noise is additive, the estimated components are bounded by the observed noisy components, as an upper bound for them.

The matrix Φ_{noise} in (3) is responsible for picking the reliable elements of $\tilde{\mathbf{y}}$, and also keeping the corresponding row vectors of $\boldsymbol{A}$. In this method, dictionary $\boldsymbol{A}$ should contain enough number of exemplars of clean speech feature vectors. In [20], it