Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 76 (2016) 218–229

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Model-based clustered sparse imputation for noise robust speech recognition

Mohammad Mohsen Goodarzi, Farshad Almasganj *

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

Received 8 January 2015; received in revised form 18 June 2015; accepted 19 June 2015
Available online 26 June 2015

Abstract

In the sparse imputation approach, missing spectral components of speech are estimated using the compressive sensing technique. For this purpose, a dictionary of clean speech components must be prepared. Noisy feature vectors are then reconstructed by the dictionary queries. In this approach, the dictionary elements should adequately cover all the possible varieties of the speech feature vectors; so, for a subject speech frame, there will be lots of irrelevant components inside the dictionary. These components make a huge size for the dictionary that in turn, slow down the estimation process and may produce artifacts in the final estimation. To face this problem, the current work proposes to cluster the dictionary queries in some smaller subspaces; the relevant subspace for a subject feature vector could be found through the posterior criterion. Moreover, this is shown that the likelihood of the Gaussian models developed for the subspaces could role as a regularization term and act as an extra prior knowledge in the estimation process and increases the final performance, significantly. To evaluate the benefits of the proposed methods, some well-designed ASR experiments are conducted on two different speech corpora, an English noisy connected digit database (Aurora 2) and a Persian continuous speech corpus (FARSDAT). The experiments show that the proposed methods not only increase the absolute word recognition accuracy but also make the entire process few times faster than the original sparse imputation approach.

Keywords: Compressive sensing; Dictionary; Robust automatic speech recognition; Sparse imputation; Subspace detection

1. Introduction

The performance degradation of Automatic Speech Recognition (ASR) system in background noise and unseen environment is a well-known story and lots of efforts have been performed to cope with it. These efforts range from acoustic and feature enhancement to design complex recognizers and take advantages of new emerging signal processing theories. One of these emerging theories is the Compressive Sensing (CS) which basically focuses on sampling signals with under Nyquist rate.

In the ASR field, there are studies which have used the CS concept in applications such as the speech recognition (Gemmeke and Virtanen, 2010), source separation (Raj et al., 2010; Schmidt and Olsson, 2007), feature enhancement (Gemmeke et al., 2011b) and missing feature estimation (BorgstrÖm and Alwan, 2009; Gemmeke et al., 2011b). Here, we focus on the later approach which tries to estimate unreliable missing spectral components of speech from available reliable components (Raj and Stern, 2005).

In the common CS approach, signal $\mathbf{x}$ is sampled using a random measurement matrix $\mathbf{\Phi}_{\text{measurment}}$; the goal of the CS approach is to recover $\mathbf{x}$ from the sampled signal $\mathbf{y} = \mathbf{\Phi}_{\text{measurment}}\mathbf{x}$ using the recovery algorithm of CS

* Corresponding author. Tel.: +98 6454 2372.
E-mail addresses: mm.goodarzi@aut.ac.ir (M.M. Goodarzi), falmas@aut.ac.ir (F. Almasganj).

http://dx.doi.org/10.1016/j.specom.2015.06.009
0167-6393/© 2015 Elsevier B.V. All rights reserved.

(Candes and Wakin, 2008). In this way, the optimum sparse vector is evaluated by

$$\tilde{\mathbf{s}} = \arg\min_{\mathbf{s}}\left\{\left\|\mathbf{\Phi}_{\text{measurment}}\mathbf{y} - \mathbf{\Phi}_{\text{measurment}}\mathbf{A}\mathbf{s}\right\|_2 + \lambda\left\|\mathbf{s}\right\|_1\right\}, \quad (1)$$

where $\mathbf{A}$ is a sparsifying basis for $\mathbf{x}$ and $\mathbf{s}$ is a sparse vector in $\mathbf{A}$ space.

In contrast, in the missing feature problem we have a noisy signal $\mathbf{y}$ that parts of it is destroyed by a random additive noise $\mathbf{n}$. By discarding the destroyed parts and modeling the effect of noise with a measurement matrix $\mathbf{\Phi}_{\text{noise}}$, we could recover clean speech $\mathbf{x}$ from $\mathbf{\Phi}_{\text{noise}}\mathbf{y}$ (reliable components of observed noisy signal), using an algorithm similar to the recovery algorithm of CS by replacing $\mathbf{\Phi}_{\text{measurment}}$ with $\mathbf{\Phi}_{\text{noise}}$. Here, $\mathbf{A}$ may be a sparsifying basis for $\mathbf{x}$ (BorgstrÖm and Alwan, 2009) or a dictionary consisted of exemplars of $\mathbf{x}$. A proper estimation of $\mathbf{x}$ could be evaluated as $\mathbf{A}\tilde{\mathbf{s}}$. Fig. 1 illustrates the block diagram of this approach.

This technique is dubbed as the Sparse Imputation (SI) method (Gemmeke and Cranen, 2008) and was developed in a series of work (Gemmeke and Cranen, 2009; Gemmeke and Virtanen, 2010; Gemmeke et al., 2011a,b); they created the $\mathbf{A}$ matrix from exemplars of clean speech frames to estimate missing speech spectral components.

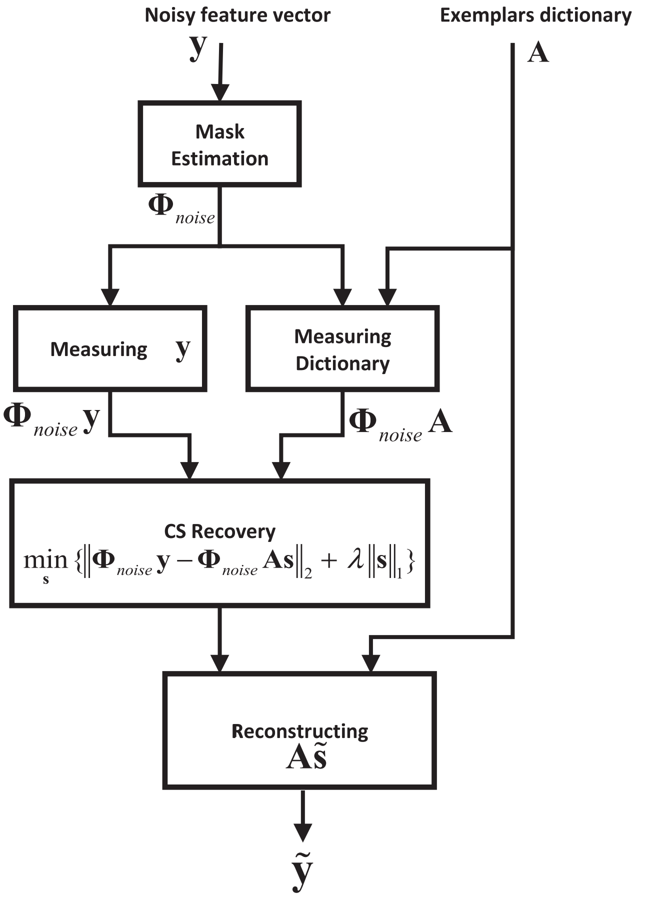

Fig. 1. Block diagram of the sparse imputation method.

Recently, a very comprehensive survey (Li et al., 2014) has been conducted on various robust ASR methods of past three decades. In this study, all methods have been categorized in five approaches. One of them, which includes the SI method, is the Feature-space approaches. This category comprises many common approaches such as spectral subtraction, wiener filtering, Cepstral mean and variance normalization and neural network-based methods such as TANDEM (Hermansky et al., 2000) and Bottle-Neck (BN) (Grezl et al., 2007) features. Also, the state-of-the-art acoustic models, referred to as the context-dependent deep neural network hidden Markov model (CD-DNN-HMM) is included in this category, because of its intrinsic layer by layer robust feature extraction structure (Li et al., 2014). Although the CD-DNN-HMM is not initially intended to deal with noise, a recent study (Seltzer et al., 2013) has shown that these DNN-based acoustic models can easily achieve the performance of the state-of-the-art methods in noisy conditions without any explicit noise compensation (in multi-condition training).

Generally, these methods perform well under some controlled experimental conditions. But, in real noisy conditions, they mostly fail because of their dependence on noise characteristics and the assumption of the stationarity of noise. On the other hand, in the reconstruction process of the missing feature approaches and specifically in the SI approach, there is no assumption about the noise characteristics. Another advantage of these approaches is that they could be used with a pre-trained acoustic model without the need of retraining acoustic model with the compensated clean data (sometimes referred to as the matched model). This positive characteristic makes it possible to use the SI method together with other robust noise compensation techniques such as DNN-based methods.

Although it has seen that the sparse imputation approach performs well in noisy condition speech recognition, but it requires solving (1) for all of the speech frames. Considering a dictionary with about 4000 exemplars, this would be a huge time consuming operation. It should be mentioned that this number of exemplars is enough for a connected digit recognition task as reported in Gemmeke et al. (2011b); but for an LVCSR task, as the number of context dependent phonemes grows dramatically, a very larger dictionary is needed. Reference Gemmeke et al. (2011b) uses a dictionary with 8000 exemplars for an average vocabulary size task. Various algorithms are proposed to solve (1); for instance, the least-angle regression (LARS) (Efron et al., 2004), as a fast one has a computational complexity of order $O(M^3 + NM^2)$, where M is the number of measurements (corresponding to the reliable components in the sparse imputation) and N is the size of the dictionary. So, we must keep in mind that in all of the CS-based approaches, larger dictionary means slower solving time.

Due to promising results of the sparse imputation, many studies focused on improving this method. For example, Tan et al. (2011) which uses an exemplar-based dictionary,