Information Sciences 310 (2015) 77–96

Contents lists available at ScienceDirect

Information Sciences

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

Dictionary evaluation and optimization for sparse coding based speech processing

Yongjun He [a,*], Deyun Chen [a], Guanglu Sun [a], Jiqing Han [b]

[a] Harbin University of Science and Technology, Harbin 150080, China
[b] Harbin Institute of Technology, Harbin 150001, China

ARTICLE INFO

Article history:
Received 21 May 2014
Received in revised form 13 February 2015
Accepted 8 March 2015
Available online 14 March 2015

Keywords:
Sparse coding
Speech denoising
Dictionary evaluation
Dictionary optimization
Speech recognition

ABSTRACT

Recently, sparse coding has attracted considerable attention in speech processing. As a promising technique, sparse coding can be widely used for analysis, representation, compression, denoising and separation of speech. To represent signals accurately and sparsely, a good dictionary which contains elemental signals is preferred and many methods have been proposed to learn such a dictionary. However, there is a lack of reasonable evaluation methods to judge whether a dictionary is good enough. To solve this problem, we define a group of measures for dictionary evaluation. These measures not only address sparseness and reconstruction error of signal representation, but also consider denoising and separating performance. We show how to evaluate dictionaries with these measures, and further propose two methods to optimize dictionaries by improving relative measures. The first method improves the efficiency of sparse coding by removing unimportant atoms; the second one improves denoising performance of dictionaries by removing harmful atoms. Experimental results show that the measures can provide reasonable evaluations and the proposed methods for optimization can further improve given dictionaries.

1. Introduction

One of the most important things in signal processing is to represent signals effectively. First, the resources for data storage and transmission are limited, which requires efficient representations of signals to save storage spaces. Second, signals are inevitably contaminated by noise, which needs the representations that are immune to noise. Third, for the task of signal analysis such as detection and classification, the representations are expected to capture more useful characteristics of signals. In brief, signal representation plays a fundamental role in signal processing.

Given a signal $\boldsymbol{x} \in \mathbb{R}^N$, the task of signal representation is to represent $\boldsymbol{x}$ by a linear combination of a set of elemental signals $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_K]$ with each $\boldsymbol{d}_i \in \mathbb{R}^N$, i.e., $\boldsymbol{x} = \sum_{i=1}^{K} w_i \boldsymbol{d}_i = \boldsymbol{D}\boldsymbol{w}$. The weight vector $\boldsymbol{w}$ is called a representation. The methods of signal representation in the early stage are linear resolution transforms, such as the fast Fourier transform (FFT), discrete cosine transform (DCT) and principal component analysis (PCA). In these methods, all elemental signals in $\boldsymbol{D}$ are used in the representation of each signal. Afterwards, the non-linear resolution transforms, including the short time Fourier transform (STFT), Gabor transform and wavelet analysis [30], etc., are developed to achieve better approximation by allowing signals to use different sets of elemental signals from $\boldsymbol{D}$. In the past few years, a new signal representation

* Corresponding author.
E-mail addresses: heyongjun@hit.edu.cn (Y. He), chendeyun@hrbust.edu.cn (D. Chen), sunguanglu@hrbust.edu.cn (G. Sun), jqhan@hit.edu.cn (J. Han).

http://dx.doi.org/10.1016/j.ins.2015.03.010
0020-0255/

method called sparse coding has been proposed and achieved great success [5,10,31]. This technique adopts an overcomplete dictionary which contains a large number of elemental signals (called atoms) to replace the transform. Each signal is represented by a linear combination of atoms from the dictionary. Most importantly, the amount of the used atoms for one representation is very small compared with the total number of the atoms in the dictionary, i.e., $\|\boldsymbol{w}\|_0 \ll K$.

Early contributions in sparse coding were made by Mallat and Zhang [31]. They introduced the concept of dictionaries to replace the more traditional and critically-sampled wavelet transform, proposed a greedy pursuit technique that approximates a sparse solution to an under-determined linear system of equations, provided characterization of dictionaries by their coherence measure, and more. Next, Chen, Donoho and Saunders introduced another pursuit technique that uses the l_1-norm for evaluating sparseness [10]. They showed that the quest for the sparsest solution could be tackled as a convex programming task, often leading to the proper solution. With these two contributions, the stage was set for a deeper analysis of these algorithms and their deployment in applications. After that, a crucial contribution was made by Donoho and Huo [14], who presented necessary theoretical backbone for the sparse model. With these progresses, compressed sensing [7,9,16] based on sparse coding emerged in 2006 by Candes, Romberg, Tao, Donoho and others that followed, sweeping many researchers and practitioners in excitement.

On the other hand, researches in cognitive science show that the neurosensory systems encode stimuli by activating only a small number of neurons out of a large population at the same time [3,35]. This indicates that the sensory perceptual system of human beings may process signals in a sparse manner. Further studies show that many natural signals, such as image and speech, are sparse or approximately sparse [12]. Sparseness seems to be an inherent property of signals and can be treated as priori knowledge. Sparse coding makes use of such knowledge and therefore can provide effective representations of signals. Through years of intensive research, sparse coding has been successfully applied to many tasks in signal processing, including signal representation, compression, separation, denoising and classification [18,27]. Recently, speech processing algorithms, which achieve improved performance by using sparse models [22,24,33,37,43], encourage researchers to make more and further explorations.

The question of the first importance in sparse coding is dictionary preparation. The dictionary can either be chosen as a predefined set of functions, such as the steerable wavelets [30], curvelets [6], contourlets [15] and bandelets [26], or be designed by adapting its content to fit a given set of signal observations [1,19,28,36,38,39,42,44,45,47]. Although a large number of dictionary learning methods have been proposed, there is a lack of appropriate methods to evaluate dictionaries. Two widely used measures in the literature are the reconstruction error and the sparseness degree that change with iteration number (e.g., [36,38]). On the one hand, current methods do not consider the relationship between reconstruction error and sparseness degree. On the other hand, these two measures only show two aspects of a dictionary. In different tasks, we need to evaluate the performance of a dictionary at different angles. At the angle of signal representation, we need to know whether a signal is sparse over a dictionary and whether the reconstruction error in sparse enhancement can meet the requirements. At the angle of denoising, we care about whether speech and noise are dense over the dictionary of each other. At the angle of speech separation, it is important to know whether two given dictionaries are suitable for the separation of two signals and whether two mixed signals can be separated easily. Unfortunately, to the best of the authors' knowledge, there is no evaluation strategy which covers the above-mentioned aspects.

In this paper, we define a group of measures to evaluate the representation ability of a dictionary. These measures address reconstruction error, sparseness degree (the number of non-zero elements in a sparse representation), as well as the number of atoms, etc. We also define another group of dictionary measures orientated towards speech denoising or speech separation. These measures consider the separability of two mixed signals and the distance between two dictionaries. Besides, we propose two dictionary optimization methods which can improve a dictionary by improving some of the measures. One is used to improve the efficiency of sparse decomposition and reconstruction; the other is used to improve denoising performance of two dictionaries.

Most current methods mainly focus on how to improve a dictionary learning method by solving some optimization problem. The inputs of these methods are training data, the outputs are dictionaries. The dictionary optimization in this paper refers to improving a given dictionary by improving one or more measures defined in this paper. Rather than repeating learning procedure again and again with the training data, we choose a subset of an overcomplete dictionary with part or all of training data. The data can be also any data from the application conditions. The input is a dictionary; the output is a better dictionary than the input. Therefore, the proposed optimization methods can be combined with any current dictionary learning method to further improve the learned dictionary.

Compared with the previous work, the contributions of this paper reflect on the following aspects:

(1) Define measures which consider the relationship between reconstruction error and sparseness degree. Considering that reconstruction error and sparseness degree vary with each other, we define one term on the premise of restricting the other term, leading to comparable measures with more detailed information.
(2) Define measures at different angles, allowing for evaluation in different tasks. Our measures cover the representation, reconstruction, denoising and separation of speech. To the best of the authors' knowledge, there is no method in the literature to measure the denoising performance of two dictionaries.