Accepted Manuscript

ELSEVIER
SPEECH
COMMUNICATION
An international journal of the European Association for Signal Processing (EURASIP) and of the International Speech Communication Association (ISCA)

Greedy Double Sparse Dictionary Learning for Sparse Representation of Speech Signals

V. Abrol, P. Sharma, A.K. Sao

PII: S0167-6393(15)30031-5
DOI: 10.1016/j.specom.2016.09.004
Reference: SPECOM 2402

To appear in: *Speech Communication*

Greedy Double Sparse Dictionary Learning for Sparse Representation of Speech Signals

V. Abrol, P. Sharma, A. K. Sao
School of Computing and Electrical Engineering, Indian Institute of Technology, Mandi
e-mail: {vinayak_abrol, pulkit_s}@students.iitmandi.ac.in, anil@iitmandi.ac.in.

Abstract

This paper proposes a greedy double sparse (DS) dictionary learning algorithm for speech signals, where the dictionary is the product of a predefined base dictionary, and a sparse matrix. Exploiting the DS structure, we show that the dictionary can be learned efficiently in the coefficient domain rather than the signal domain. It is achieved by modifying the objective function such that all the matrices involved in the coefficient domain are either sparse or near-sparse, thus making the dictionary update stage fast. The dictionary is learned on frames extracted from a speech signal using a hierarchical subset selection approach. Here, each dictionary atom is a training speech frame, chosen in accordance to its energy contribution for representing all other training speech frames. In other words, dictionary atoms are encouraged to be close to the training signals that uses them in their decomposition. After each atom update the modified residual serves as the new training data, thus the information learned by the previous atoms guides the update of subsequent dictionary atoms. In addition, we have shown that for a suitable choice of the base dictionary, storage efficiency of the DS dictionary can be further improved. Finally, the efficiency of the proposed method is demonstrated for the problem of speech representation and speech denoising.

Keywords: Speech processing, dictionary learning, double sparsity, sparse representation.

1. Introduction

Sparse signal representations have recently drawn much interest in the field of speech processing for applications such as recognition [1–3], speech separation [4–7], synthesis [8, 9], denoising [10, 11] and voice activity detection (VAD) [12]. Sparse representation has been motivated by the desire to map the original signal/feature space in ways that are compact, efficient, robust to noise, and meaningful [13]. However, effectiveness of sparse representation is very much influenced by the choice of the dictionary, which requires to have minimum time and storage complexity, so that it can be used in various speech processing applications [14]. In literature, sparse representation of speech signals have been studied mostly using either complete or overcomplete dictionaries [15]. These are either signal independent (analytical) transforms e.g., discrete cosine transform (DCT) or signal dependent dictionaries e.g., principal component analysis (PCA) dictionary, learned using a dictionary learning (DL) algorithm [16]. It has been shown that, compared to an analytical dictionary, a learned dictionary provides sparser representation as they are relatively better for capturing the inherent structures and patterns in the speech signal [14, 16]. A detailed overview of some of the recent works exploring the application of sparse representation to transformation, analysis, and visualization of speech signals can be found in [14, 17–19] and references within.

For speech signals, the most popular approaches for DL are based on speech production mechanism (such as Linear Predictive (LP) model and Liljencrants-Fant (LF) model), where a speech signal is expressed as a sparse excitation signal in a dictionary build using the vocal tract impulse response (computed from LP coefficients) [20, 21]. These dictionaries have shown to provide good results for speech coding [20] and inference related applications e.g., VAD [12]. However, they require time-varying model parameters, which needs to be updated for each speech frame [20]. Alternatively, one can consider any DL method, where the