SPARSE HIDDEN MARKOV MODELS FOR AUTOMATIC SPEECH RECOGNITION

Pranay Dighe        Afsaneh Asaei        Hervé Bourlard

Idiap-RR-28-2015

AUGUST 2015
A Dictionary Learning and Sparse Recovery Approach to Exemplar-based Speech Recognition

Pranay Dighe\textsuperscript{1,2}, Afsaneh Asaei\textsuperscript{1}, Hervé Bourlard\textsuperscript{1,2}

\textsuperscript{1}Idiap Research Institute, Martigny, Switzerland
\textsuperscript{2}Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

\{pranay.dighe,afsaneh.asaei,herve.bourlard\}@idiap.ch

Abstract

Stochastic speech recognition has been cast as a natural realization of the compressive sensing problem in this work. The compressed acoustic observations are subword posterior probabilities obtained from a deep neural network. Dictionary learning and sparse recovery are exploited for inference of the high-dimensional sparse word posterior probabilities. This formulation amounts to realization of a sparse hidden Markov model where each state is characterized by a dictionary learned from training exemplars and the emission probabilities are obtained from sparse representations of test exemplars. This new dictionary-based speech processing paradigm alleviates the need for a huge collection of exemplars as required in the conventional exemplar-based methods. We study the performance of the proposed approach for continuous speech recognition using Phonebook and Numbers’95 database.


1. Introduction

Exemplar-based (template-based) methods and stochastic approaches relying on hidden Markov model (HMM) are often considered as two distinct approaches to automatic speech recognition (ASR). While later enjoys many years of extensive development, exemplar-based methods have recently regained more serious attention \cite{1, 2, 3, 4}. In theory, having an “infinite” amount of exemplars, and the “right” distance measure, “optimal” recognizers could be sought \cite{5}. In practice however, development of an “optimal” exemplar-based system indicates massive memory and computational requirements.

A typical exemplar-based system uses spectral representation of speech signal and a collection of such training exemplars for ASR \cite{2, 3, 6, 7}. It has also been shown that constructing a dictionary of clean exemplars from the training data enables robust ASR if the noisy test exemplars are reconstructed as a sparse linear combination of the training exemplars. Enforcing the sparsity structure has shown to discriminate the subspaces of noise and clean speech leading to some de-noising or separation effect\cite{3, 8}. The exemplar-based sparse representation is also exploited to provide new features to a HMM-based ASR system \cite{2}. In this paper, we focus on exemplar-based sparse representation with two main distinctions: (1) Unlike the previous spectral exemplars, we use sub-word posterior probabilities estimated by a deep neural network (DNN) as exemplars. (2) Instead of taking the collection of training data for sparse representation, we exploit a principled way of dictionary learning for sparse representation.

The core assumption of this approach is that any possible realization of a linguistic unit (eg. a word) lies on a path (non-linear manifold) through low-dimensional union of subspaces spanned by exemplars already seen in the training set. Sparse recovery captures the low-dimensional representation of the concerned word at the frame or segment level. Integrating frame/segment wise sparse recovery within a HMM-like framework enables us to process the sequential evolution(path through the non-linear manifold) of the word. We show that this approach naturally develops a sparse HMM configuration (as in \cite{10}) for ASR where each hidden state models a point of the word manifold using a dictionary. The emission likelihoods of the data point are obtained from sparse representations with respect to the state dictionary. Using the posterior exemplars, we also demonstrate rigorously that the statistical speech recognition can be cast as a compressive sensing problem: The posterior exemplars are compressed observations and higher level word inference requires a high-dimensional sparse recovery where this linguistic compressive mechanism has to be learned from the training exemplars.

This paper is organized as follow: Section 2 explains the view on ASR using the CS principles relying on dictionary learning and sparse recovery. We explain how this approach amounts to realization of a sparse modeling HMM in Section 3. The experimental analysis is presented in Section 4 and the conclusions are drawn in Section 5.
2. Compressive Sensing Perspective

Each speech utterance is composed of a few words. If we consider the vector representation of an utterance in a linguistic space where each component corresponds to a unique word, this representation is high-dimensional whereas the informative components are highly sparse. The intuition behind this work is that the input features of the ASR system are compressed observations of this naturally high-dimensional representation problem. Hence, ASR can be cast as a sparse recovery problem while identifying the underlying compressive sensing mechanism.

In this paper, the compressed acoustic observations are sub-word conditional posterior probabilities obtained from a deep neural network. Let \( \{q_k\}_{k=1}^{K} \) denote the sub-word classes. Given an input feature vector \( x_t \) at time \( t \), the posterior probability vector \( z_t = [p(q_1|x_t)p(q_2|x_t) \ldots p(q_K|x_t)] \), is estimated using DNN. According to the marginalization rule of probabilities, the following relation holds

\[
\begin{bmatrix}
p(q_1|x_t) \\
p(q_2|x_t) \\
\vdots \\
p(q_K|x_t)
\end{bmatrix}
\begin{bmatrix}
p(q_1|w_1) & \cdots & p(q_1|w_L) \\
p(q_2|w_1) & \cdots & p(q_2|w_L) \\
\vdots & \cdots & \vdots \\
p(q_K|w_1) & \cdots & p(q_K|w_L)
\end{bmatrix}
= 
\begin{bmatrix}
p(w_1|x_t) \\
p(w_2|x_t) \\
\vdots \\
p(w_L|x_t)
\end{bmatrix}
\]

Dictionary: \( D = [d_1 \ldots d_L] \)

\( z_t \)

where \( \alpha_t \) is a binary vector with one component being 1 and the rest 0. Hence, \( z_t = d_t \). We model each column (atom) \( d_t \) of the dictionary as

\[
\begin{bmatrix}
p(q_1|w_1) \\
p(q_2|w_1) \\
\vdots \\
p(q_K|w_1)
\end{bmatrix}
\begin{bmatrix}
p(q_1|w^{sw_1}) & \cdots & p(q_1|w^{sw_N}) \\
p(q_2|w^{sw_1}) & \cdots & p(q_2|w^{sw_N}) \\
\vdots & \cdots & \vdots \\
p(q_K|w^{sw_1}) & \cdots & p(q_K|w^{sw_N})
\end{bmatrix}
= 
\begin{bmatrix}
p(w^{sw_1}|w_1) \\
p(w^{sw_2}|w_1) \\
\vdots \\
p(w^{sw_N}|w_1)
\end{bmatrix}
\]

Word manifold dictionary \( D_{sw} \)

\( d_t \)

where \( sw^{w_t} \) denotes the \( s^{th} \) state underlying the word \( w_t \), \( S_{w_t} \) represents the total number of (over-complete) “bases” to model the sub-space of word \( w_t \) and \( [\alpha_t^{w_s} p(w_s|x_t)]_{w_t = w_s} = \alpha_t \).

Equations (1) and (2) lead to a very intuitive and natural representation for continuous speech in terms of posterior features and word-to-subword hierarchical dictionaries obtained as

\[ D = [D_{w_1} \cdots D_{w_1} \cdots D_{w_L}] \]  

(3)

The dictionary \( D \) has an internal partitioning defined by the boundaries of individual sub-dictionaries \( D_{w_t} \). In addition, a sequence of posterior features \( Z = [z_1, \ldots z_T] \), extracted from an utterance of word \( w_t \), will have a hierarchical group structure underlying the individual sparse representation \( \alpha_t \) where all the coefficients tend to collaborate to activate a higher level group corresponding to \( w_t \). The sparse representation of \( Z \) yields a matrix \( A = [\alpha_1, \ldots, \alpha_T] \) where the support of the sparse coefficients hold a blocks structure as depicted in Figure 1.

![Figure 1: Given a sequence of acoustic features in Z, the sparse representation matrix A will have a block structure associated to the word-specific dictionaries where the inner block coefficients are sparse. This collaborative hierarchical sparsity structure is exploited in [9] to devise an efficient C-HiLasso algorithm for the sparse recovery objective.](image)

Based on the above formulation, ASR problem is an instance of a compressive sensing problem with the two key components:

- **Dictionary Learning**: Learning the dictionary for sparse representation of the word probabilities.
- **Sparse Recovery**: Structured sparse recovery for high-dimensional inference of word probabilities.

In the following section, we explain how this formulation meets the hidden Markov model framework.

3. Sparse Modeling HMM

In this section, we develop the above mentioned formulation as a sparse HMM configuration (shown as a graphical model in figure 2).

For brevity and simplicity, we assume that each hidden state models the manifold for exactly one word from the vocabulary. The goal is to infer the most probable sequence of hidden states for a given sequence of observation \( Z = [z_1 \ldots z_T] \). In equation (2), we characterize each word \( w_t \) with a dictionary where each atom defines an underlying (subspace) sub-word \( sw^{w_t}_s \). The hidden state corresponding to word \( w_t \) is thus modelled by the word manifold dictionary \( D_{w_t} \). The observation \( z_t \) associated to the state \( w_t \) is modeled as \( z_t = D_{w_t} \alpha_t^{w_t} \) where \( D_{w_t} \in \mathbb{R}^{K \times S_{w_t}} \) is an over-complete dictionary \((K < S_{w_t})\). The likelihood of observation \( z_t \) is given by the sparse sub-word posterior probability vector \( \alpha_t^{w_t} = [p(sw_t^{w_1}|w_t) \ldots p(sw_t^{w_N}|w_t)]^T \). The dictionary \( D_{w_t} \) characterizes the (non-linear) manifold associated to the state \( w_t \). Sparse recovery of \( \alpha_t^{w_t} \) using this dictionary indicates that the observation \( z_t \) lies on
minimization problem [10] which can be solved using Lasso solver. Summarizing the EM procedure as derived in [10] for learning the sparse modeling HMM parameters, we obtain the method of optimal directions (MOD) in sparse dictionary learning, which alternates between finding the sparse coefficients and Updating the dictionary [10]. This steps are summarized in Algorithm 1. In addition, we can leverage recent advances in dictionary learning for sparse representation. One such algorithm is KSVD [7] and sparse NMF [7]. In this paper, we study three approaches to dictionary learning for development of MOD-HMM, KSVD-HMM and NMF-HMM.

Algorithm 1 Online Dictionary Learning

<table>
<thead>
<tr>
<th>Require:</th>
<th>$Z$, $\lambda$ (regularization parameter), $D_0$ (initialization)</th>
</tr>
</thead>
<tbody>
<tr>
<td>for $t = 1$ to $T$ do</td>
<td></td>
</tr>
<tr>
<td>1: Sparse Coding of $z_t$ to determine $\alpha_t$: $\alpha_t = \arg\min_\alpha { \frac{1}{2} | z_t - D(t-1) \alpha |_2^2 + \lambda | \alpha |_1 }$</td>
<td></td>
</tr>
<tr>
<td>2: \text{Updating } D(t) \text{ with } D(t-1) \text{ as warm restart:} $D(t) = \arg\min_D \left{ \frac{1}{7} \sum_{t=1}^{T} \frac{1}{2} | z_t - D \alpha_t |_2^2 + \lambda | \alpha_t |_1 \right} $</td>
<td></td>
</tr>
<tr>
<td>end for</td>
<td></td>
</tr>
<tr>
<td>return $D(T)$</td>
<td></td>
</tr>
</tbody>
</table>

It may be noted that the sparse modeling HMM framework is fundamentally different than hybrid exemplar-based/HMM proposed in [3] in the following ways

1. Exemplars are posterior probabilities thus the probabilistic relation expressed in (1) amounts to direct estimation of word posterior probabilities.

2. EM derivation of the HMM state parameters requires dictionary learning. This key step has been ignored in all previous exemplar-based speech processing methods, to the extent of our knowledge.

4. Experimental Analysis

5. Conclusions

Posterior exemplars are central players in development of a novel statistical ASR framework that builds on dictionary learning to model a non-linear manifold of word posterior representations and sparse recovery for high-dimensional inference of the sparse word probabilities. We found this formulation analogous to realization of a hidden Markov model where the state distributions are characterized by a sparse coding dictionary. We also provided some examples that the exemplar-based sparse modeling of posterior features can be more accurate than the alternative probabilistic methods.
6. References


