Sparse coding based features for speech units classification

Pulkit Sharma, Vinayak Abrol, A.D. Dileep, Anil Kumar Sao
IIT Mandi, India
{pulkit_s,vinayak_abrol}@students.iitmandi.ac.in, {dileep,anil}@iitmandi.ac.in

Abstract
In this paper a sparse representation based feature is proposed for the tasks in speech recognition. Dictionary plays an important role in order to get a good sparse representation. Therefore instead of using a single over complete dictionary, multiple signal adaptive dictionaries are used. A novel principal component analysis (PCA) based method is proposed to learn multiple dictionaries for each speech unit. For a given speech frame, first minimum distance criterion is employed to select appropriate dictionary and then a sparse solver is used to compute sparse feature for acoustic modeling. Experiments are performed using different datasets, which shows the proposed feature outperforms the existing features in recognition of isolated utterances. Index Terms: Sparse representations, dictionary learning, speech recognition.

1. Introduction
Sparse representation (SR) based speech recognition systems proposed in recent works are shown to provide good recognition results [1–4]. In SR, any input signal is written as a linear combination of minimum number of basis functions from a resource called as dictionary. Existing work on the tasks in speech recognition using SR can be categorized into two classes: (i) exemplar based approaches, and (ii) feature based approaches. Exemplar based approaches model speech signal as a superposition of training speech exemplars [4]. Here an exemplar based dictionary for each class is stored and whenever a new frame comes, a sparse vector is obtained with respect to all the dictionaries. Minimum reconstruction error for all the classes is used as a metric to classify speech frames [4]. In feature based approaches, given a speech frame, a sparse vector is obtained using a sparse solver [1–3]. Now either this sparse vector [1] or the estimate of speech signal is used as a feature [2,3] for acoustic modeling. In exemplar based approaches sparse vector does the classification, on the contrary in feature based approaches, derived sparse vector is used as a feature for further classifier. In this work, we focus on sparse feature based approach for speech recognition.

In the literature, sparse coding based features are used for the task of speech recognition in [1–3], where a single overcomplete dictionary is learned [1] or multiple dictionaries are used [2]. A sparse solver is used to obtain the sparse vector [1] or the estimate of representation (of speech frame) [2], which is then used as a feature for the classification tasks. It is claimed in [1] that the information present in spectro-temporal patterns of speech signal is encoded into a sparse vector using a single overcomplete dictionary (learned using training data). On the other hand, different ways of learning dictionaries such as nearest neighbors, trigram language model etc. are explored in [2]. However, these approaches require entire training data to be stored which is used to seed dictionary atoms [2].

In this paper, we propose a novel SR based method to extract features from a speech signal for the tasks in speech recognition. Proposed feature extraction method consists of two stages: (i) dictionary learning (DL), and (ii) sparse coding. In the first stage signal adaptive dictionaries are learned using a DL method. In second stage a sparse solver is used to obtain the weights corresponding to the dictionary which are in turn used as a feature. Instead of learning a single overcomplete dictionary, we propose to use principal component analysis (PCA) based multiple dictionaries to capture the variations present in speech signal effectively. A minimum distance criterion is used to select a suitable dictionary for a given speech frame in the sparse coding stage. Proposed approach is similar to [1] where sparse vector is used as a feature, but multiple signal adaptive dictionaries are employed in this work as compared to a single overcomplete dictionary used in [1]. In addition, our method uses mel frequency cepstral coefficients (MFCC) to learn different signal adaptive dictionaries compared to spectro-temporal representations used in [1]. In this work only learned dictionaries corresponding to each speech unit are needed to be stored. We have also explored state-of-the-art greedy adaptive dictionary (GAD) [5] to obtain sparse features for speech signal.

Contributions of this work are: (a) a novel sparse representation based feature for speech signal, and (b) PCA and GAD based DL methods to learn multiple signal adaptive dictionaries for speech signal.

The organization of the paper is as follows: Section 2 describes basics of sparse coding for speech signals. Proposed technique of feature extraction is explained in section 3. Experimental observations are discussed in section 4, and finally the paper is concluded in section 5.

2. Sparse coding for speech signals
Speech signal captured using microphone has a lot of redundancy [6], which can be utilized for efficient and compact representation of the speech. Such representations can be achieved either using a compact code or a sparse distributed code [7]. For any signal $x \in \mathbb{R}^N$, a compact code ($x_c \in \mathbb{R}^N$) is a representation such that its dimension is less compared to the input signal i.e., $N_c < N$, and the loss in this description is minimum [7]. In case of sparse distributed code ($x_c \in \mathbb{R}^N$), the dimensionality of representations and input signal are equal, but the number of elements needed to describe the input are minimum. Sparse distributed codes require only $L (L \ll N)$ elements to represent the given input faithfully provided suitable basis functions are chosen [7]. These basis functions can be considered as generating causes for various natural signals (such as speech or images) and in general number of such causes contributing to a particular signal are always less [8]. Hence the information relevant to the underlying process of generating natural signal is generally sparse compared to the recorded observations [8].
Frames of Speech Signal

![Diagram](a): Dictionary Learning

![Diagram](b): Dictionary Selection

Figure 1: Graphical illustration of the proposed (a) dictionary learning approach, and (b) dictionary selection approach.

SR have recently drawn much interest in the field of speech processing [9–13]. SR based signal processing is backed by an assumption that signal is sparse in a given domain i.e., it can be represented with a few significant coefficients (in that domain). We assume that the speech signal $s \in \mathbb{R}^N$ can be sparsely expressed in $\Psi$ as $s = \Psi \alpha$ such that only $L$ ($L \ll N$) coefficients in $\alpha \in \mathbb{R}^N$ are significant, where $\Psi \in \mathbb{R}^{N \times N}$ is the dictionary. Given $s$ and $\Psi$ the corresponding sparse vector $\alpha$ can be obtained as

$$\hat{\alpha} = \arg\min_\alpha \|\alpha\|_0 \text{ such that } \|s - \Psi \alpha\|_2^2 < \epsilon,$$

(1)

where $\epsilon$ is a constant known as error tolerance. $\|\alpha\|_0$ denotes the $l_0$ norm problem, which is a complex combinatorial search and is an NP hard problem to solve. Therefore $l_1$ norm, which is convex relaxation to $l_0$ norm can be solved and hence equation (1) can be rewritten as

$$\hat{\alpha} = \arg\min_\alpha \|\alpha\|_1 \text{ such that } \|s - \Psi \alpha\|_2^2 < \epsilon.$$

(2)

The estimate of $\hat{\alpha}$ obtained using equation (2) can be used to generate estimate of speech signal as $\hat{s} = \Psi \hat{\alpha}$.

Some works exists in literature of speech recognition using SR based features solves for $s$ and uses its estimate ($\hat{s}$) as a feature for acoustic modeling [2, 3]. Method proposed in this paper is distinct from such methods in a way that instead of estimate ($\hat{s}$), sparse vector ($\hat{\alpha}$) is used as a feature. In the next section, we discuss the proposed sparse features for speech signal.

3. Proposed sparse feature for speech signals

The proposed approach uses sparse coding framework for deriving a feature from the speech signal. The weights correspond-

**Algorithm 1 Proposed algorithm for dictionary learning.**

**Inputs:** Matrix $S_i$ with speech frames as columns (for same speech signal) $S_i = [s_{i1}, s_{i2}, ..., s_{in}]$ and $\epsilon$.

**Outputs:** $K$ dictionary-centroid pairs $[\Psi_k, \mu_k]$ for each class.

1. Cluster $S_i$ into $K$ clusters using $K$-means clustering algorithm with $\mu_k$ being center of the $k^{th}$ cluster.
2. Learn a sub-dictionary $\Psi_k$ from cluster $S_k$ using the objective function:

$$\left(\Psi_k, \hat{\Lambda}_k\right) = \arg\min_{\Psi_k, \hat{\Lambda}_k} \|\alpha_k\|_1 \text{ such that } \|S_k - \Psi_k \alpha_k\|_2^2 < \epsilon,$$

where $\|\cdot\|_F$ is Frobenius norm and $\Lambda_k$ is the representation coefficient matrix of $S_k$ over $\Psi_k$.
3. 3.1. Adaptive dictionaries for speech signals

It has been observed that a single overcomplete dictionary is not able to capture the variations present in a signal [15]. Therefore in this paper multiple dictionaries are used to model the same speech signal (unit). Dictionary learning technique employed in this work is similar to method presented in [15] in context of images and is summarized in Algorithm 1. Multiple signal adaptive dictionaries for different frames (of same speech signal) helps in achieving better recognition results because such dictionaries can capture discriminative information better than a single overcomplete dictionary.

Consider $i^{th}$ speech class with $n_i$ training speech frames ($\{s_{ij}\}_{j=1}^{n_i}$) arranged in a matrix $S_i \in \mathbb{R}^{n_i \times N}$ as columns such that $S_i = [s_{i1}, s_{i2}, ..., s_{in}]$. Here a frame can also be a standard representation e.g., MFCC etc. In order to learn multiple dictionaries, we first cluster these frames into $K$ clusters denoted by
for testing. The CV segment recognition accuracy presented is the average classification accuracy along with 95% confidence interval obtained for 5-fold stratified cross-validation. Speech used for experimentation is sampled at a rate of 16kHz and features are extracted at a frame size of 25ms with 10ms overlap. For a speech frame corresponding 39-dimensional feature vector is extracted. Here first 12 features are MFCC and the 13th feature is log energy. The remaining 26 features are the delta and acceleration coefficients. For all the training examples corresponding to a speech signal, S, consists of this feature obtained from different frames as its columns, and multiple dictionaries are learned. During experiments we consider the number of clusters K as five and value of error tolerance (ε) is $10^{-3}$. The sparse vector is obtained using YALL1 $l_1$ solver [19]. Equation (5) is solved to obtain 39-dimensional weight vector, which is then used as a feature. Performance of CDHMM-based classifier using the proposed features labeled as $F_{GAD}$ and $F_{PCA}$ (corresponding to GAD and PCA dictionaries, respectively) is compared to CDHMM-based classifier using standard MFCC features and SVM-based classifier with HMM-based intermediate matching kernel (HMM-IMK) discussed in [17]. The comparison of results is shown in Table 1.

### 3.2. Selection of dictionary

After DL step, we have $KU$ pairs of $(\Psi_{ik}, \mu_{ik}), i = 1, ..., U; k = 1, ..., K$ associated with different feature vectors corresponding to different training signals of the same speech unit. In order to select the dictionary for a speech frame $s$, we use the minimum distance metric as follows:

$$k_{\alpha} = \arg\min_i \|s - \mu_{ik}\|_2.$$  

(4)

Corresponding to each speech frame $s$, the following minimization problem is solved (using $\Psi_{ik}$) to obtain estimate of sparse vector $\alpha$

$$\hat{\alpha} = \arg\min_{\alpha} \|\alpha\|_1$$  

such that $\|s - \Psi_{ik}\alpha\|_2 < \varepsilon$.  

(5)

Estimate of sparse vector obtained after solving equation (5) ($\hat{\alpha}$) for different speech frames is used as a feature for speech recognition. Thus we will have a sequence of sparse vectors for a given speech signal.

### 4. Experimental Observations

Continuous density hidden Markov model (CDHMM) based classifier is build to evaluate the performance of proposed features for different tasks in speech recognition. The tasks include recognition of isolated utterances of the E-set of English alphabet, recognition of consonant-vowel (CV) segments in Hindi and phoneme classification. The Oregon Graduate Institute (OGI) spoken letter database is used in the study on recognition of E-set [16]. The continuous speech corpus of broadcast news in Hindi [17] is used for the study on recognition of CV segments. TIMIT phonetic corpus [18] is used for phoneme classification. The E-set data [16] includes the 9 letters: B, C, D, E, G, P, T, V, and Z. The training and test sets include 240 and 60 utterances for each letter, respectively. Thus the E-set recognition accuracy is presented as the classification accuracy obtained for 540 test examples. For the recognition of CV segments in Hindi, we considered 103 CV classes that have at least 50 examples in the training data set [17]. The data set consists of 19,458 CV segments for training and 4,866 CV segments different tasks in speech recognition. The tasks include phoneme recognition in TIMIT dataset are shown in Table 2. The following observations can be made from Tables 1 and 2.

- **Performance of the proposed sparse features derived using both the dictionaries is better than MFCC feature.**
- **PCA based dictionary is performing better as compared to its counterpart GAD.** The reason can be understood from Figure 2, which shows atoms corresponding to both the dictionaries. It can be observed that the atoms corresponding to GAD dictionary are relatively sparse (see Figure 2 (a)) and thus the corresponding weights are less sparser as compared to obtained for PCA based dictionary. Hence discriminative information is better preserved in sparse vector derived using PCA based dictionary.
- **Sparse feature using PCA based dictionary is outperforming state-of-the-art features.** It shows sparse vector

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Feature</th>
<th>(N, Q)</th>
<th>Recognition Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDHMM</td>
<td>MFCC</td>
<td>(5, 3)</td>
<td>87.95 ± 0.77</td>
</tr>
<tr>
<td>SVM with HMM-IMK</td>
<td>MFCC</td>
<td>(5, 3)</td>
<td>95.93 ± 0.85</td>
</tr>
<tr>
<td>CDHMM</td>
<td>$F_{GAD}$</td>
<td>(5, 3)</td>
<td>94.07 ± 0.79</td>
</tr>
<tr>
<td>CDHMM</td>
<td>$F_{PCA}$</td>
<td>(5, 3)</td>
<td>97.38 ± 0.87</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature</th>
<th>Accuracy</th>
<th>MFCC</th>
<th>PLP [1]</th>
<th>$l_1$ [1]</th>
<th>$F_{GAD}$</th>
<th>$F_{PCA}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>64.5</td>
<td>66.9</td>
<td>61.7</td>
<td>66.31</td>
<td>69.83</td>
</tr>
</tbody>
</table>
preserve the discriminative information among different speech units.

5. Conclusions
In this work a novel feature extraction technique, based on principles of sparse representation, has been proposed. Proposed feature attempts to capture the discriminative information among different speech signals in terms of their sparse representations. For efficient computation of discriminative sparse representations, multiple signal adaptive dictionaries are used instead of single overcomplete dictionary. In addition, two dictionaries learning methods namely PCA and GAD are explored for the acoustic modeling of subword units of speech. It has been observed that PCA based dictionary gives sparser representation as compared to GAD and hence captures discriminative information among speech units. Experimental results using two databases support the claim that the sparse vector can be alternative to the state-of-the-art features.

6. References