Class Specific GMM Based Sparse Feature 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, {addileep,anil}@iitmandi.ac.in

Abstract—**In this paper, features based on the sparse representation (SR) are proposed for the classification of speech units. The proposed method employs multiple dictionaries to effectively model variations present in the speech signal. Here, a Gaussian mixture model (GMM) is built using spectral features corresponding to frames of all the examples of a speech class. Multiple dictionaries corresponding to different mixture are learned using the respective speech frames. Given a train/test speech frame, minimum spectral distance measure from the GMM means is employed to select an appropriate dictionary. The selected dictionary is used to obtain the sparse feature representation, which is used for the classification of speech units. The effectiveness of the proposed feature is demonstrated using continuous density hidden Markov model (CDHMM) based classifiers for (i) classification of isolated utterances of E-set of English alphabet, (ii) classification of consonant-vowel (CV) segments in Hindi language and (iii) classification of phoneme from TIMIT phonetic corpus. Experimental results reveal that the proposed features outperforms existing feature representations for various speech units classification tasks.**

Index Terms—**Sparse representation, speech recognition, dictionary learning.**

I. Introduction

In recent years, sparse representation (SR) based features are used for speech recognition where, given a segment of speech signal (frame) and a dictionary, a sparse feature vector is computed for the classification/recognition task [1], [2]. The SR based signal processing is supported by an observation that signal can be written as linear combination of minimum number of atoms of a dictionary [2]. In the literature, methods using SR for speech recognition can be broadly classified into two categories : (i) exemplar based approaches, and (ii) feature based approaches.

Exemplar based approaches directly uses the sparse vector for classification, while a feature based approach uses the derived sparse vector as a feature in a classifier. In particular, speech recognition in exemplar based approaches is performed either using the atom activations of the estimated sparse feature vector [3], [4], or using the minimum reconstruction error [5] between the test exemplar and its estimate. On the contrary, in feature based approaches, either the derived sparse vector [1] or the estimate of speech is used as a feature [6] for acoustic modeling. For computing the sparse feature vector, approaches in [3], [4] use a single overcomplete dictionary while [5] use multiple dictionaries corresponding to different speech units. A gradient descent approach is used to learn a single overcomplete dictionary using the spectro-temporal representation in [1], while mel frequency cepstral coefficients (MFCC) of training speech data (frames) are used to obtain dictionary atoms in [6] and [2]. However, for a given train/test frame, in [6] and [2] atoms for dictionary are seeded from the training data, which results in high computational complexity.

In this work, we propose a novel SR based method to derive features from a speech signal for the tasks in speech recognition. Proposed feature extraction method consists of two stages : (i) dictionary learning, and (ii) sparse coding. In the first stage, multiple dictionaries are learned for each speech unit and the second stage uses a sparse solver to obtain the sparse feature. The speech signal, being generated from a natural system has a lot of variations in it. In order to in effectively modeling the variations present in the speech signal the proposed method employs multiple dictionaries. This is achieved by clustering similar speech frames, and a single dictionary is learned for each cluster.

In contrast to existing approaches (such as in [5], [7]) which uses K-means clustering or K-nearest neighbors (KNN) respectively, we employed a Gaussian mixture model (GMM), which is a generative model, and is more efficient in modeling the variations among frames of same speech unit. GMM can be built using either raw speech or spectral feature e.g., MFCC as an initial representation. In this work, a GMM is used to model representations corresponding to all the frames in the training data of each class. A dictionary for each mixture in a GMM is learned from the respective frames, resulting in multiple dictionaries for each speech unit. For each train/test frame, a minimum spectral distance measure is employed to select an appropriate dictionary. The selected dictionary is then used to obtain the sparse vector, which is used as a feature representation for the classification task.

The proposed method is similar to [1] where sparse vector is used as a feature. However, we propose to use multiple dictionaries as compared to a single overcomplete dictionary used in [1]. In addition, our method uses both raw speech samples and MFCC to learn different signal adaptive dictionaries compared to spectro-temporal representation used in [1]. This work also employs and compares the performance of different dictionary learning algorithms namely greedy adaptive dictionary (GAD) [8], K-singular value decomposition (KSVD) [9], method of optimal directions (MOD) [10] and principal component analysis (PCA) [11] to learn dictionaries.

Contributions of this work are: (a) GMM to derive multiple dictionaries for each speech class, (b) comparison of

ISBN 978-0-9928626-7-1 © EURASIP 2017

dictionaries derived using PCA, GAD, KSVD and MOD based dictionary learning algorithms, and (c) deriving proposed sparse features using both MFCC and raw speech as initial representation.

The organization of the paper is as follows : Section II describes basics of sparse coding for speech signals. SR based proposed feature extraction method is explained in section III, and details about various experimental settings is provided in Section IV. A detailed discussion about various speech classification experiments performed and experimental observations is provided in Section V. Finally, the paper is summarized in section VI.

II. Sparse coding for speech signals

The speech signal corresponds to a high dimensional data captured using a microphone, however the total number of generating causes for signal are very less as compared to recorded observations [12]. Thus, the information relevant to the underlying process of generating speech signal is generally low dimensional as compared to the recorded observations [12]. This property can be exploited for estimating efficient representations for a speech signal and sparse coding is one of the methods to estimate such representations [13]. In recent years sparse coding based signal processing has been applied to various speech processing applications such as speech recognition [1], speech enhancement [14], speech coding [15] and voiced/nonvoiced detection [16].

The basic idea in SR based signal processing is supported by an assumption that the signal is sparse with respect to a suitable dictionary/basis. Assuming that a speech frame $\mathbf{s} \in \mathbb{R}^N$ is represented using a dictionary $\mathbf{\Psi} \in \mathbb{R}^{N \times N}$ as $\mathbf{s} = \mathbf{\Psi}\boldsymbol{\alpha}$, such that $\boldsymbol{\alpha} \in \mathbb{R}^N$ is H ($H \ll N$) sparse, i.e., $\boldsymbol{\alpha}$ has only H significant coefficients. Given $\mathbf{s}$ and $\mathbf{\Psi}$, the estimate of sparse vector $\boldsymbol{\alpha}$ can be obtained as

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\text{argmin}} \;\; f(\boldsymbol{\alpha}) \quad \text{s.t.} \quad \|\mathbf{s} - \mathbf{\Psi}\boldsymbol{\alpha}\|_2^2 \; < \; \epsilon, \tag{1}$$

where ϵ is the error tolerance constant, and function $f(.)$ is used to promote sparsity [17]. $f(.)$ can be solved using l_0 or l_1-norm, however, in this work l_0-norm is employed, and equation (1) is solved using orthogonal matching pursuit (OMP) [18], [19]. The obtained estimate of $\hat{\boldsymbol{\alpha}}$ can be used to estimate speech signal as $\hat{\mathbf{s}} = \mathbf{\Psi}\hat{\boldsymbol{\alpha}}$. Most of the existing SR based features in speech recognition use estimate of speech signal ($\hat{\mathbf{s}}$) as a feature for acoustic modeling [2], [6]. On the contrary, method proposed in this paper uses the obtained sparse vector ($\hat{\boldsymbol{\alpha}}$) as a feature.

III. Proposed sparse coding based features for speech signal

In this work, we propose a novel SR based feature for the tasks in speech recognition. The proposed method employs multiple dictionaries and thus in the dictionary learning step, a GMM is built for each speech class using MFCC. A dictionary corresponding to each mixture (obtained using GMM) is obtained using either MFCC or raw speech samples as a representation. Since multiple dictionaries are learned for each speech units an appropriate selection criterion must be used to select dictionary for each speech frame. Thus, the sparse feature extraction step employs MFCC from each train/test frame to select an appropriate dictionary using minimum spectral distance measure (between MFCC of each frame and means of Gaussian components). The selected dictionary is used to compute the SR, which is used as the feature representation for the speech units classification task. A pictorial representation of the proposed dictionary learning method is shown in Fig. 1. It shows a two dimensional representation corresponding to all the frames of training speech signals available for a single speech unit. These speech frames are modeled using a GMM, and this figure, symbolically shows four mixtures along with their corresponding means.

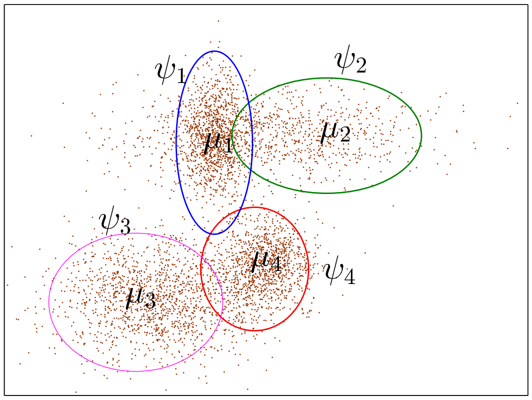

Fig. 1: Graphical representation of dictionary learning approach for a given speech unit class. Each point in two-dimension (2-D) represents a speech frame corresponding to a speech unit.

In GMM, each point has a soft assignment, i.e. it belongs to each cluster to a different degree. This degree is based on the probability of the point being generated from each cluster's (multivariate) normal distribution, with cluster center and cluster covariance as its mean and covariance. This soft assignment is more appropriate for clustering the data into different clusters as samples/points having similar probabilities for two clusters may be excluded from the data used to train dictionaries.

Consider M speech frames $\{\mathbf{s}_i\}_{i=1}^{M}$ of l^{th} speech class (obtained from the training examples of a class) arranged in a matrix $\mathbf{S}^l$ as columns such that $\mathbf{S}^l = [\mathbf{s}_l, \mathbf{s}_2, ..., \mathbf{s}_M]$. Here, $l = 1, 2, \ldots, L$ is the total number of classes. All the speech frames in $\mathbf{S}^l$ are modeled using a GMM $\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{S}_k|\mu_k, \Sigma_k)$ with $k = 1, 2, ..., K$ Gaussian mixtures, where $\boldsymbol{\mu}_k$ is the mean vector corresponding to k^{th} mixture. Let m_k denotes total number of training frames belonging to k^{th} Gaussian, arranged in a matrix $\mathbf{S}_k^l = [\mathbf{s}_l, \mathbf{s}_2, ..., \mathbf{s}_{m_k}]$, where $\sum_{k=1}^{K} m_k = M$. The following objective function can be used to learn a sub-dictionary $\mathbf{\Psi}_k^l$ for the data corresponding to each Gaussian mixture ($\mathbf{S}_k^l$):

$$\left(\hat{\mathbf{\Psi}}_k^l \;,\; \hat{\boldsymbol{\Lambda}}_k^l\right) = \underset{\mathbf{\Psi}_k^l \,,\, \boldsymbol{\Lambda}_k^l}{\text{argmin}} \;\; f(\boldsymbol{\Lambda}_k^l) \quad \text{s.t.} \quad \|\mathbf{S}_k^l - \mathbf{\Psi}_k^l \boldsymbol{\Lambda}_k^l\|_F^2 < \epsilon, \tag{2}$$

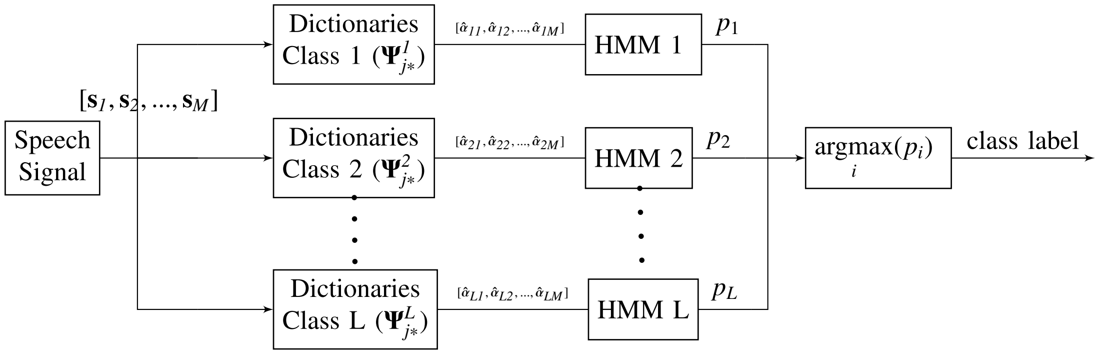
Speech Signal
$[\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_M]$
Dictionaries Class 1 ($\mathbf{\Psi}^1_{j*}$)
Dictionaries Class 2 ($\mathbf{\Psi}^2_{j*}$)
Dictionaries Class L ($\mathbf{\Psi}^L_{j*}$)
$[\hat{\alpha}_{11}, \hat{\alpha}_{12}, ..., \hat{\alpha}_{1M}]$
$[\hat{\alpha}_{21}, \hat{\alpha}_{22}, ..., \hat{\alpha}_{2M}]$
$[\hat{\alpha}_{L1}, \hat{\alpha}_{L2}, ..., \hat{\alpha}_{LM}]$
HMM 1
HMM 2
HMM L
p_1
p_2
p_L
$\underset{i}{\text{argmax}}(p_i)$
class label

Fig. 2: Block diagram representation of testing in the proposed approach. $[\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_M]$ represents M speech frames of a speech unit and $[\hat{\alpha}_{l1}, \hat{\alpha}_{l2}, ..., \hat{\alpha}_{lM}]$ represents feature representation derived using dictionaries of l^{th} class. p_l denotes the likelihood corresponding to l^{th} class HMM.

where ϵ is a constant, $||\ .\ ||_F$ is the Frobenius norm and $\boldsymbol{\Lambda}^l_k$ is the sparse weight matrix of $\mathbf{S}^l_k$ over sub-dictionary $\mathbf{\Psi}^l_k$. Equation (2) is a joint optimization problem of solving $\mathbf{\Psi}^l_k$ and $\boldsymbol{\Lambda}^l_k$, which can be solved by alternatively optimizing $\mathbf{\Psi}^l_k$ and $\boldsymbol{\Lambda}^l_k$. The mean vector corresponding to k^{th} mixture of l^{th} speech class is denoted as $\boldsymbol{\mu}^l_k$. Thus, K dictionary pairs $\left\{\mathbf{\Psi}^l_k\ ,\ \boldsymbol{\mu}^l_k\right\}$ corresponding to same speech unit are obtained. The same dictionary learning method is employed on the data corresponding to L different speech units resulting in KL ($KL = C$) total such pairs.

For a speech frame $\mathbf{s}_i$, the best fitted dictionary (corresponding to GMM of each class) is selected based on its minimum euclidean distance from the mixture means [1] i.e., $j^* = \underset{j}{\text{argmin}}\ ||\mathbf{s}_i - \boldsymbol{\mu}_j||_2 \quad , j = 1, 2, ..., k$. The corresponding sub-dictionary $\mathbf{\Psi}^l_{j^*}$ is used to obtain an estimate of the sparse vector $\boldsymbol{\alpha}_i$ corresponding to each speech frame $\mathbf{s}_i$ by solving:

$$\hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\text{argmin}}\ ||\mathbf{s}_i - \mathbf{\Psi}^l_{j^*}\boldsymbol{\alpha}_i||_2^2 \ \text{ s.t. } \ ||\boldsymbol{\alpha}_i||_0 \ < \ H. \qquad (3)$$

Sparse vector $\hat{\boldsymbol{\alpha}}_i \in \mathbb{R}^N$ obtained after solving equation (3) is used as a feature representation for various speech units classification tasks.

IV. Experimental Setup

The proposed features are used to build the continuous density hidden Markov model (CDHMM) based classifier for classification of (i) isolated utterances of E-set of English alphabet [20], (ii) consonant-vowel (CV) segments in Hindi language [20] and (iii) phoneme from TIMIT phonetic corpus [21]. All results reported in this paper are average results for 10 trials. The number of mixture in GMM during dictionary learning is five i.e., $K = 5$, and this number is obtained empirically. Orthogonal matching pursuit (OMP) [18] is used to solve equation (3) with a fixed value of sparsity (H) as $N/2$. A speech frame is excluded from the data used to train dictionary, if the difference in probabilities corresponding to two clusters (of the GMM) is less than 0.5 (this value is obtained empirically). Speech used for experiments is sampled at 16 kHz and is processed at a frame size of 25 ms with 10 ms shift. Both raw speech sample and MFCC are used as an initial representation to derive the proposed feature representation. The raw speech sample results in a 400-dimensional initial representation, while MFCC used is standard 39-dimensional. In MFCC, first 12 features are mel frequency cepstral coefficients and the 13^{th} coefficient is the log energy. The remaining 26 coefficients are the delta and acceleration coefficients.

In all the experiments, a left-to-right CDHMM is built for each class with varying number of states and the number of components for the state specific GMM. Here, we have considered diagonal covariance matrices for the state-specific GMM. During training L CDHMM models are built for L different speech classes. The testing strategy employed in this work is shown in Fig. 2. During the testing stage, for a given speech utterance, the sparse features are obtained corresponding to the dictionaries of all the classes. These features are then fed to the CDHMM corresponding to each class and the unit is classified to the class giving maximum posterior.

The performance of the proposed approach is evaluated using four types of complete dictionaries i.e., PCA based, KSVD, GAD and MOD. In order to have a fair comparison, the same approach is followed i.e., GMM is first used on the data of each class and these dictionaries are learned on data belonging to each mixture. For PCA based dictionary $\mathbf{\Psi}^l_k$, the equation 2 can be solved by conventional method of least squares while for other dictionaries standard dictionary learning algorithms are available. The results reported for E-set and TIMIT are the average classification accuracy, while for Hindi CV segments results are the average classification accuracy along with 95% confidence interval obtained for 5-fold stratified cross-validation.

V. Experimental Results

In this section, we evaluate the performance of proposed features for classification of individual speech units. The effect of initial representation (raw speech samples or MFCC) is studied at GMM modeling during dictionary learning and feature extraction step. The performance of the proposed feature is also compared with existing SR based features.

[1] In this work, these mixture means are also referred to as centroids of dictionaries.

TABLE I: Classification accuracy (in %) of CDHMM-based classifier with 5 states and 3 mixture in each state when different initial representations are used to derive the proposed feature representation. GMM_D and Feature Extraction represents the representation used for GMM modeling during dictionary learning and feature extraction respectively. Here, the dictionary used is PCA based.

Initial Representation		**Dataset**		
GMM_D	**Feature Extraction**	E-set	Hindi-CV	TIMIT
MFCC	MFCC	85.47	45.37 ± 0.73	61.27
Raw	Raw	75.19	37.51 ± 0.79	55.91
MFCC	Raw	98.13	65.07 ± 0.83	70.81

TABLE II: Classification accuracy (in %) of CDHMM-based classifier using the proposed features. N_s and Q indicates number of HMM states and number of GMM components in each state. $F_{R_{PCA}}$, $F_{R_{KSVD}}$, $F_{R_{MOD}}$ and $F_{R_{GAD}}$ are labels corresponding to the proposed feature deriveds using PCA, KSVD, MOD and GAD dictionaries , respectively.

			Dataset	
			E-set	**Hindi CV**
Classifier	**Feature**	(N_s , Q)	**Accuracy**	**Accuracy**
CDHMM	MFCC	(5 , 3)	87.95	48.87±0.77
SVM with HMM-IMK	MFCC	(5 , 3)	95.93	59.32±0.85
CDHMM	SR_{NN} [2]	(5 , 3)	97.58	64.03±0.73
CDHMM	SR_P [7]	(5 , 3)	97.38	61.08±0.87
CDHMM	$F_{R_{GAD}}$	(5 , 3)	86.14	52.03±0.79
	$F_{R_{KSVD}}$	(5 , 3)	96.53	63.45 ±0.76
	$F_{R_{PCA}}$	(5 , 3)	**98.13**	**65.07±0.83**
	$F_{R_{MOD}}$	(5 , 3)	97.47	64.79±0.71

TABLE III: Classification accuracy (in %) of CDHMM-based classifier with 5 states and 3 mixtures in each state using the proposed features for phoneme (TIMIT) classification. $F_{R_{PCA}}$, $F_{R_{KSVD}}$, $F_{R_{MOD}}$ and $F_{R_{GAD}}$ are labels corresponding to the proposed feature derived using PCA, KSVD, MOD and GAD dictionaries , respectively.

Feature	MFCC	PLP	SR_{NN} [2]	l_1 [1]	SR_P [7]	$F_{R_{GAD}}$	$F_{R_{KSVD}}$	$F_{R_{PCA}}$	$F_{R_{MOD}}$
Accuracy	64.5	68.47	70.74	69.38	69.83	62.31	68.15	**70.81**	70.13

A. Significance of initial representation

In this experiment, we analyze the effect of initial representation while deriving the proposed feature. Both raw speech samples and MFCC are used as initial representation for the GMM modeling during dictionary learning and feature extraction. Consider an example where MFCC is used for GMM modeling and raw speech samples are used for dictionary learning. Here, MFCC representation is modeled using GMM, so that a dictionary corresponding to data of each mixture can be learned. However, the dictionary learning and feature extraction is performed using the raw speech samples as an initial representation.

The results obtained with different initial representations for GMM modeling during dictionary learning and feature extraction are shown in Table I. These results are obtained using a PCA based dictionary. When raw speech samples are used to build GMM (during dictionary learning) there are more chances of choosing a sub-dictionary from GMM of different classes (for most of the frames), thus there is a reduction in classification accuracy. On the contrary, while building GMM from MFCC, the dictionary selection is more appropriate as now frames with similar spectral characteristics are clubbed into each Gaussian mixture. MFCC used in the feature extraction process model only spectral features based on auditory response. Another possible reason for better performance of raw speech samples while deriving sparse features (when MFCC are used to build GMM) is the possibility of capturing the inherent variations present in raw samples. Hence, for building the GMM (during dictionary learning) and selecting the dictionary, MFCC are used as feature. On the contrary, the proposed SR features are derived directly from raw speech samples.

B. Comparison with other features/approaches

The proposed features corresponding PCA, KSVD, MOD and GAD dictionaries are labeled as $F_{R_{PCA}}$, $F_{R_{KSVD}}$, $F_{R_{MOD}}$ and $F_{R_{GAD}}$, respectively. Performance of proposed feature is compared to CDHMM-based classifier using standard MFCC features and SVM-based classifier with HMM-based intermediate matching kernel (HMM-IMK) discussed in [20]. The comparison is also done with the SR based features proposed in [2], where N-nearest neighbors are used to seed dictionary atoms, with MFCC as initial representation (labeled as SR_{NN}). In addition, SR based features proposed in [7], where K-means clustering is used to learn multiple dictionaries (labeled as SR_P) is also used for comparison. The comparison of results for E-set and Hindi CV dataset is shown in Table II. The

TABLE IV: Classification accuracy (in %) of CDHMM-based classifier with 5 states and 3 mixtures in each state using the proposed features for speech corrupted by 0 dB babble noise. $F_{R_{PCA}}$ is label corresponding to the proposed feature derived using PCA dictionary.

	Classification Accuracy		
Feature	**E-Set**	**Hindi CV**	**TIMIT**
MFCC	27.37	16.12±0.78	19.94
SR_{NN} [2]	30.28	18.76±0.81	21.47
$F_{R_{PCA}}$	32.17	19.87±0.74	22.94

comparison of classification accuracy for TIMIT dataset is shown in Table III. For TIMIT dataset, MFCC, perceptual linear prediction (PLP) features and a SR based features proposed in [1] (labeled as l_1) are also used for comparison purpose.

These results reveal that the proposed feature representation derived using a PCA based dictionary outperforms existing features. The proposed method employs multiple dictionaries, where a single dictionary is learned for data belonging to a GMM mixture. This data in a single mixture is similar and can be modeled using a complete dictionary, making PCA an appropriate choice for dictionary. On the contrary, other dictionary learning methods e.g., KSVD perform better when the dictionary learned is overcomplete. This reason may be attributed for better performance of PCA based dictionary in our method.

The performance of the proposed features is also evaluated in noisy conditions. The speech segments of the datasets used are corrupted by additive babble noise taken from NOISEX-92 database at 0 dB signal to noise ratio (SNR) [22]. Comparison of the proposed features for classification of noisy speech units is given in Table IV. It can be observed that even in case of noisy speech, the proposed feature outperforms existing features.

VI. Summary

In this work, principles of sparse representation are used to obtain novel features for speech units classification. We propose to use multiple dictionaries for the computation of sparse vector which is used as a feature representation. Multiple dictionaries helps in effectively modeling the variations present in different speech signals corresponding to the same speech unit, will help in achieving better classification accuracy. In this work, four dictionary learning methods namely PCA, KSVD, MOD and GAD are employed to obtain dictionaries. The sparse feature vector corresponding to the PCA-based dictionary results in more discrimination as compared to other dictionaries. Classification results using three databases support the claim that the proposed sparse feature can be used as an alternative to existing features. In future, we would like to extend this work from classification of individual speech units to automatic speech recognition.

References

[1] G.S.V.S. Sivaram, S.K. Nemala, M. Elhilali, T.D. Tran, and H. Hermansky, "Sparse coding for speech recognition," in *IEEE Conference on Acoustics Speech and Signal Processing*, March 2010, pp. 4346–4349.

[2] T.N. Sainath, B. Ramabhadran, M. Picheny, D. Nahamoo, and D. Kanevsky, "Exemplar-Based Sparse Representation Features: From TIMIT to LVCSR," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2598–2613, Nov 2011.

[3] J.F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-Based Sparse Representations for Noise Robust Automatic Speech Recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, 2011.

[4] D. Baby, T. Virtanen, J.F. Gemmeke, and H. V. Hamme, "Coupled Dictionaries for Exemplar-Based Speech Enhancement and Automatic Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1788–1799, Nov 2015.

[5] E. Yilmaz, J.F. Gemmeke, and H. V. Hamme, "Noise Robust Exemplar Matching Using Sparse Representations of Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 8, pp. 1306–1319, Aug 2014.

[6] T.N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, and A. Sethy, "Sparse representation features for speech recognition," in *INTERSPEECH*. 2010, pp. 2254–2257, ISCA.

[7] P. Sharma, V. Abrol, A.D. Dileep, and A.K. Sao, "Sparse coding based features for speech units classification," in *INTERSPEECH*, 2015, pp. 712–715.

[8] M.G. Jafari and M.D. Plumbley, "Fast Dictionary Learning for Sparse Representations of Speech Signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 1025–1031, Sept 2011.

[9] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[10] K. Engan, S.O. Aase, and J.H. Husoy, "Method of optimal directions for frame design," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999, vol. 5, pp. 2443–2446.

[11] W. Dong, D. Zhang, G. Shi, and X. Wu, "Image Deblurring and Super-Resolution by Adaptive Sparse Domain Selection and Adaptive Regularization," *IEEE Transactions on Image Processing*, vol. 20, no. 7, pp. 1838–1857, July 2011.

[12] I. Tosic and P. Frossard, "Dictionary Learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, March 2011.

[13] M.V.S. Shashanka, B. Raj, and P. Smaragdis, "Sparse Overcomplete Decomposition for Single Channel Speaker Separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2007, vol. 2, pp. II–641–II–644.

[14] V. Abrol, P. Sharma, and A.K. Sao, "Speech Enhancement Using Compressed Sensing," in *INTERSPEECH*, August 2013, pp. 3274–3278.

[15] D. Giacobello, M.G. Christensen, M.N. Murthi, S.H. Jensen, and M. Moonen, "Speech coding based on sparse linear prediction," in *European Signal Processing Conference*, Aug 2009, pp. 2524–2528.

[16] V. Abrol, P. Sharma, and A.K. Sao, "Voiced/nonvoiced detection in compressively sensed speech signals ," *Speech Communication*, vol. 72, pp. 194 – 207, 2015.

[17] M. Elad, *Sparse and Redundant Representations - From Theory to Applications in Signal and Image Processing*, Springer, 2010.

[18] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient Implementation of the K-SVD Algorithm using Batch Orthogonal Matching Pursuit," CS Technion, 2008.

[19] J.A. Tropp and A.C. Gilbert, "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit," *IEEE Transactions on Information Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[20] A.D. Dileep and C.C. Sekhar, "HMM Based Intermediate Matching Kernel for Classification of Sequential Patterns of Speech Using Support Vector Machines," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2570–2582, Dec 2013.

[21] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus," 1993.

[22] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247 – 251, 1993.