



Analysis of sparse representation based feature on speech mode classification

Kumud Tripathi, K. Sreenivasa Rao

Dept. of Computer Science and Engineering,
Indian Institute of Technology Kharagpur, India.

kumudtripathi.cs@gmail.com, ksrao@iitkgp.ac.in

Abstract

Traditional phone recognition systems are developed using read speech. But, in reality, the speech that needs to be processed by machine is not always in read mode. Therefore to handle the phone recognition in realistic scenarios, three broad modes of speech: read, conversation and extempore are considered in this study. The conversation mode includes informal communication in an unconstrained environment between two or more individuals. In the extempore mode, a person speaks with confidence without the help of notes. Read mode is a formal type of speech in a rigid environment. In this work, we have proposed a sparse based feature for speech mode classification. The effectiveness of sparse representation depends on the dictionary. Therefore, we have learned multiple overcomplete dictionaries by using parallel atom-update dictionary learning (PAU-DL) technique to capture the discrimination characteristics present in the considered speech modes. Further, sparse features correspond to the sequence of speech frames are derived using the learned dictionary by applying the orthogonal matching pursuit (OMP) algorithm. The proposed sparse features are evaluated on speech corpora consisting of six Indian languages by performing classification of speech modes. The results with the proposed sparse features outperform the standard spectral, excitation source and prosodic features.

Index Terms: Sparse representation, dictionary learning, speech mode classification

1. Introduction

Nowadays, humans are more dependent on machines. The conventional way of interaction between the human and the machine is possible using a keyboard, mouse, touch screen, etc. But, the natural way of communication for humans is speech. Therefore, it would be better to have a speech interface between human and machine for communication. This could be achieved by developing phone recognition system (PRS). The objective of PRS is to provide an efficient and accurate mechanism to transcribe human speech into the sequence of phones. In literature [1, 2], conventional phone recognition model is developed using read mode of speech corpus. However, in real scenarios speech may be one among the following given modes: conversation, extempore and read speeches. The performance of the phone recognition is optimized by developing PRS related to the speech modes. While analyzing the performance of the PRSs, the speech signal belongs to a particular mode need to be processed through the respective mode PRS. In view of this, we need to develop a robust model for speech mode classification (SMC). Further, the SMC will be located at the front-end of the PRSs.

In our earlier work [3], the speech mode classifier is developed for Bengali speech corpus using vocal tract, prosodic, and excitation source features. The performance of SMC was 98%

at very high computation cost. Therefore, in our current work, we have proposed a sparse based feature which has the better capability of discriminating between the modes at low computation cost in comparison to vocal tract, prosodic, and excitation source features discussed in [3].

In speech related task, features should focus on the discriminative details present in the sequence of frames. According to [4], the speech signal collected using microphones has very high dimensional data, but the responsible data dimension for representing the signal is very less. Thus, the low dimensional data can be obtained by applying a sparse code [5]. In recent works, sparse representation (SR) based feature outperform the conventional speech features for the task of speech recognition [6, 7]. In sparse representation of the speech, a frame is derived from a linear combination of atoms of a dictionary. A dictionary could be either single [6] or multiple [7] for deriving the sparse representation. Authors of [7] explored various algorithms for dictionary learning which uses complete training samples for initializing atoms of a dictionary. It is depicted in [6], a sparse vector encoded the information present in the spectral-temporal domain of speech frame. Therefore, representation of speech frames [7] or sparse vector [6] is derived using the sparse solver, which is further used as a feature for speech classification.

In this work, we illustrate the effectiveness of sparse representation (SR) as a feature for speech mode classification (SMC) using multilayer perceptron. Here, we considered six Indian languages: Bengali, Oriya, Manipuri, Gujarati, Marathi, and Telugu for training and testing the performance of the model. For each language, we developed an autonomous SMC. This study involves two stages for extracting proposed features: (i) learning the multiple overcomplete dictionaries from the mode-specific speech and (ii) obtaining the sparse feature from the learned dictionary using a sparse solver named as Orthogonal Matching Pursuit (OMP) [8]. SR transforms the speech signal into the useful, compact, robust and effective format. The efficiency of SR is affected by the dictionary. Therefore, to capture the mode specific information present in speech, we adopted parallel atom-updating dictionary learning (PAU-DL), and K-singular value decomposition (KSVD) based multiple overcomplete dictionaries instead of a single overcomplete dictionary used in [6]. In this study, raw speech frames and Mel frequency cepstral coefficients (MFCCs) are utilized to initialize the dictionaries as compared to spectral-temporal information used in [6].

The workflow of the paper is as follows. Section 2 describes the sparse representation for speech signals. The proposed feature extraction method is discussed in Section 3. Experimental results and performance analysis of the developed SMC models are discussed in Section 4. Conclusion and future work of this paper has been included in Section 5.

2. Sparse Representation of Speech Signals

In general sparse coding of speech signals X refers that signal has a sparse representation using an optimal dictionary θ . A sparse signal \hat{X} can be represented as a linear combination of least possible atoms of the dictionary. Assume a dictionary ($\theta = \{q_1, q_2, \dots, q_m\}$) and a sparse vectors ($B = \{b_1, b_2, \dots, b_n\}$) are given then the corresponding sparse signal \hat{X} can be estimated as follows:

$$\hat{X} = \sum_{i=1}^m b_i q_i = \theta B \quad (1)$$

In this work, dictionary ($\theta \in \mathbb{R}^{K \times m}$) represented as a matrix with each column as an atom $q_i \in \mathbb{R}^K$, a signal ($X \in \mathbb{R}^{K \times n}$) shown as a matrix where each column as a speech frame $x_i \in \mathbb{R}^K$ and a sparse coding ($B \in \mathbb{R}^{m \times n}$) represented as a matrix with each column as a sparse vector $b_i \in \mathbb{R}^m$ corresponding to each speech frame. With $m > K$, sparse representation trying to update an overcomplete dictionary. Therefore, the dimension of the sparse code B becomes larger than the dimension of the input signal X . But, the sparse code includes at most L (with $L \ll K$) non-zero elements for significantly representing a signal. Sparse coding estimates the representation (B) for a given signal (X) while learning the dictionary (θ) with m atoms. It can be mathematically derived as follows:

$$\hat{B} : \min_B \|B\|_0 \quad \text{Subject to} \quad \|X - \theta B\|_2 < \epsilon \quad (2)$$

where $\|\cdot\|_0$ stands for the l_0 pseudo-norm which counts the number of non-zero values in the sparse code. ϵ represents the constant value for error tolerance. Because of the combinatorial nature of l_0 norm that makes it hard to solve the problem. Several methods have been described for determining the sparsest possible solution for an input signal in a provided overcomplete dictionary namely, Orthogonal Matching Pursuit (OMP), Stage-wise Orthogonal Matching Pursuit (StOMP) [9], etc.

3. Sparse features for speech mode classification

In this section, we have discussed the approach based on sparse coding to derive a feature for speech mode classification. The sparse vector related to a frame of a speech derived using a learned dictionary is utilized as a feature in this study. The effectiveness of sparse coding is mainly affected by the decision of the dictionary (θ), which could be either analytical or learned. Analytical dictionaries such as, wavelets, contourlets, etc. have a quick mathematical execution property. The learned dictionaries such as, KSVD [10], and PAU-DL [11] are trained from the set of training examples and thus, adjust to the variations in training data successfully. Since, speech is produced because of variation in vocal tract system. Therefore, to capture the variation in speech, a learned dictionary is considered in this work.

3.1. Dictionary adaptation for speech signals

According to [12], the single overcomplete dictionary is not capable of obtaining the variation exists in a signal. Therefore, multiple overcomplete dictionaries are considered to discriminate among the speech modes in this work. The proposed dictionary learning approach is based on the concept of [11] for images and is elaborated in algorithm 1.

Consider n_c training speech frames from a speech mode c , $\{x_{ci}\}_{i=1}^{n_c}$, organized in a data matrix $X_c \in \mathbb{R}^{K \times n_c}$ as columns

Algorithm 1 Proposed dictionary learning algorithm

Input: Data matrix $X_c = \{x_{c1}, x_{c2}, \dots, x_{cn_c}\}$ of class c where columns are speech frames, target sparsity L and error tolerance ϵ .

Output: Learned dictionary $\theta_c = \{q_{c1}, q_{c2}, \dots, q_{cm}\}$ of class c .

- 1: Initialize dictionary θ_c with raw data or MFCCs.
 - 2: *Sparse coding:* Calculate sparse representation b_{ci} for each sample x_{ci} using OMP by solving:

$$b_{ci} : \min_{b_{ci}} \|x_{ci} - q_{ci} b_{ci}\|_2 \text{ s.t. } \|b_{ci}\|_0 \leq L$$
 where $i = \{1, 2, \dots, n_c\}$ represents number of speech samples of a class c .
 - 3: *Dictionary learning:* Dictionary θ_c can be updated by applying PAU-DL to speech frames using Eq. (3).
 - 4: Iterate above steps to construct dictionaries for each speech mode.
-

such that $X_c = \{x_{c1}, x_{c2}, \dots, x_{cn_c}\}$. For the sake of training multiple dictionaries, the frames are grouped according to the speech modes. A dictionary relating to a mode is learned to such an extent that the description over this dictionary is as sparse as could be expected under the circumstances. Dictionary θ_c can be updated by applying the given relation to speech samples x_{ci} :

$$(q_{ci}, \hat{b}_{ci}) : \arg \min_{q_{ci}, b_{ci}} \|b_{ci}\|_0 \quad \text{s.t.} \quad \|x_{ci} - q_{ci} b_{ci}\|_2 < \epsilon \quad (3)$$

where b_{ci} is the sparse vector of speech frames x_{ci} over learned dictionary θ_c . Multiple dictionaries for speech modes (c) are derived using Algorithm 1. Other popular dictionary KSVD is also explored for comparison with the proposed work. In the same manner, multiple dictionaries are trained using KSVD.

4. Experiments and results

In this study, the speech mode classifier is developed using multilayer perceptron (MLP) [13] to estimate the usefulness of the proposed feature. Here, we considered six Indian languages: Bengali, Oriya, Manipuri, Gujarati, Marathi, and Telugu for analyzing the performance of the model. We developed six monolingual speech mode classifiers independently, on data for each language. The speech dataset for the multiple Indian languages is collected through the project: prosodically guided phonetic engine for searching speech databases in Indian languages (PSI) [14]. The considered modes of speech in the study of classification are conversational, extempore and read. The Conversation speech is a type of intuitive, unconstrained communication between at least two individuals. The Extempore speech which is also known as lecture mode of speech is conveyed without the guide of notes. It is more vigorous, flexible and spontaneous. The Read speech includes news reading etc. where reader reads with the help of notes. It is more organized, arranged and planned well ahead of time. In this study, for each language the considered duration of the conversation, extempore and read speech is 1.16 hr. For the experiment, speech dataset is recorded from the 4-male and 4-female speakers. While training around 80% of the data and for testing remaining 20% of the data is used from different speakers.

The speech signal considered for experiments is sampled at a rate of 16 KHz with precision 16 bits per sample and parti-

Table 1: Classification accuracy (in %) of language-based speech mode classifiers in the presence of different features.

Language	Features (Performance in %)					
	MFCC	F_{ksvd}	F_{pau}	(MFCC+ F_{ksvd})	(MFCC+ F_{pau})	(MFCC+RMFCC+MPDSS+Prosody)
Bengali	90.15	89.43	91.13	96.54	98.24	98.04
Oriya	86.02	86.32	89.87	98.19	99.15	98.96
Manipuri	83.47	85.02	88.29	96.35	98.64	98.15
Gujarati	85.86	86.73	87.49	95.68	97.85	98.96
Marathi	92.88	90.23	92.91	98.81	99.19	99.23
Telugu	91.04	92.45	93.21	97.19	98.95	98.76

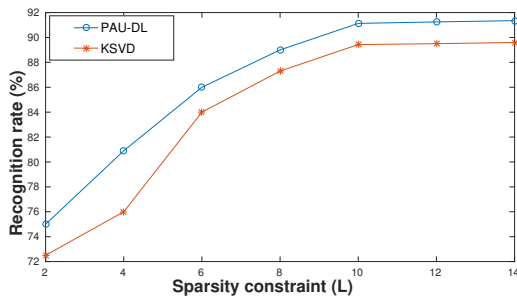
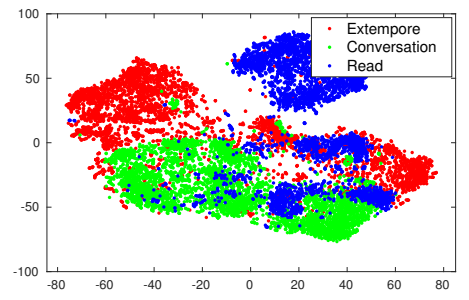


Figure 1: The recognition rate (%) of the SMC using different sparsity constraints (L) on Bengali speech corpus.

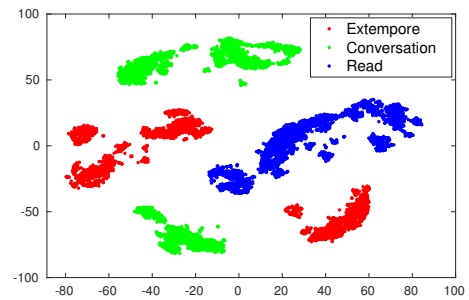
tioned into short frames of 25 ms with 10 ms overlap for feature extraction. In this work, both 39-dimensional MFCCs and 400-dimensional raw speech frames are used as the initial description of a dictionary. The sequence of speech frames without any pre-processing called as a raw speech signal. We observed that raw data are giving better performance as compared to MFCCs. In view of speech, MFCCs contains only spectral details while raw frames include all the inherent variation. This could be one of the reasons for better performance of raw samples. In this study, we have created the mode dependent dictionary for each considered language. Dictionary size is $K \times 3K$ where K is 39 and 400 when a dictionary is initialized using MFCC and raw speech frames, respectively. From the given input signal and dictionary, the $3K$ -dimensional sparse vector is estimated using OMP algorithm. Further, we analyzed the effect of sparsity (L) on the recognition of Bengali speech modes using PAU-DL and KSVD based dictionary in Figure 1. We can observe that the improvement in recognition rate is not significant beyond sparsity $K/4$. For all experiments, the value considered for error tolerance ϵ is 10^{-3} .

The proposed features corresponding to PAU-DL and KSVD based dictionaries are named as F_{pau} and F_{ksvd} , respectively. The performance of proposed features is compared with conventional MFCC and combination of spectral, prosodic and excitation source features discussed in [3]. The prosodic details are collected from pitch and energy contour of a speech signal whereas, MFCCs are representing the spectral information of speech. The excitation source feature is captured using residual Mel-frequency cepstral coefficients (RMFCCs) and Mel power differences of spectrum in sub-bands (MPDSS). The performance comparison of MLP based speech mode classifiers developed for each language independently using various features is listed in Table 1.

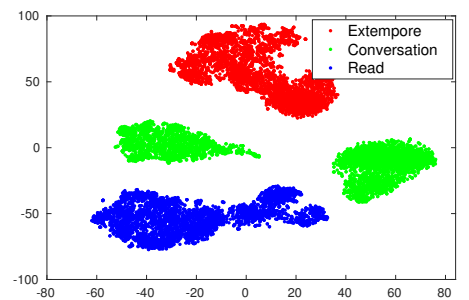
From Table 1, it is noted that the mode classification accuracy using the proposed sparse based features estimated using K-SVD and PAU-DL dictionaries is better than the mode clas-



(a)



(b)



(c)

Figure 2: 2-D t-SNE representation for (a) MFCCs (b) F_{pau} and, (c) (MFCCs+ F_{pau}) features of some random samples from each mode of Bengali speech.

sification accuracy using MFCCs. This represents the proposed feature contains better discriminative details about the speech modes. It is justified in Figure 2, by the 2-dimension (2-d) representation of MFCCs and proposed feature corresponding to PAU-DL using t-Distributed Stochastic Neighbor Embedding (t-SNE) visualizations [15]. T-SNE attempts to preserve the dis-

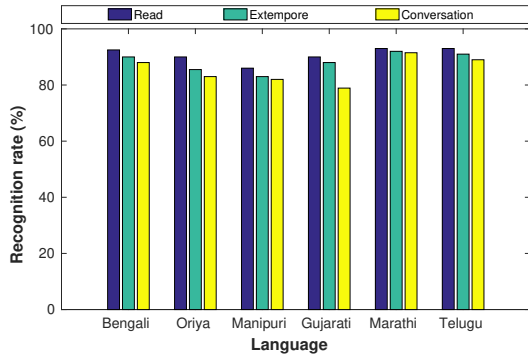


Figure 3: The recognition rate (%) of the three speech modes belongs to six different languages.

tances in the high-dimensional space, while dimension reduction for data visualization. The axes represent the x and y coordinates of the low-dimensional space. In Figure 2(a), MFCCs of data from three modes are overlapping while in Figure 2(b), sparse features of data from given modes are non-overlapping which makes it more suitable for SMC.

The performance of the PAU-DL based dictionary is better as compared to the KSVD based dictionary. This can be verified by the Figure 1. Hence, sparse vector estimated using PAU-DL based dictionary contains significant details about the class. Unlike KSVD, in PAU-DL each atom is partially updated before going to the next atom. By this way, updates in PAU-DL based dictionary is more reliable as compared to KSVD based dictionary [11]. This may lead PAU-DL to perform better as compared to KSVD. Further, the model developed with the concatenation of MFCCs and sparse features labeled as F_{ksvd} and F_{pau} , giving better performance compared to the model developed using the individual features. The concatenation of MFCCs and sparse feature derived from PAU-DL dictionary provides the best performance compared to all other features discussed in the Table 1. Figure 2(c), shows that combination of MFCCs and F_{pau} perfectly group the data belongs to a particular mode into a separate cluster. Here, each cluster preserves discriminative ability because of high inter-cluster distance and low intra-cluster distance. This demonstrates that MFCCs and F_{pau} preserve some complementary information which is responsible for upgrading the performance of the speech mode classifier. Experimental results show that SMC developed for six different languages using various features are independent of speakers.

In our earlier work [3], we had developed the speech mode classifier for Bengali speech corpus using spectral, excitation source and prosodic features. In our current work, we have explored five other languages for better comparison among the proposed sparse based features and previously explored features. From Table 1, it is clear that combination of MFCCs and proposed features are giving similar results as compared to the combination of spectral, excitation source and prosodic features. The major significance of our proposed features over the features described in [3] is (i) it reduced mathematical complexity by reducing the feature dimension and (ii) it gives better performance than spectral, excitation source and prosodic features.

In Figure 3, we studied the performance of three speech modes independently for each language using MFCCs. From the results, it can conclude that the model trained with differ-

ent languages giving better performance for read mode in comparison with conversation and extempore mode speech. In read mode, a person speaks formally in the constrained environment, but it is not necessary for extempore and conversation modes. It could be the explanation behind the better recognition of read mode when compared with other modes. It can be visualized in Figure 2(a), read mode data are classified better than other two modes, while, conversation and extempore modes get confused with remaining modes. In Figure 3, we observed that the difference in the recognition accuracy of speech modes belongs to Marathi and Telugu speech is less as compared to remaining languages. The reason is explored by listening to the speech from Marathi and Telugu languages, and they seem to be more distinguishable among modes than remaining languages. The classification accuracy of read mode in Oriya and Gujarati are quite similar whereas, Gujarati conversation speech has poor performance among all languages. This observation can be verified by listening to the speech from given languages.

5. Conclusion

In this work, we have proposed a feature extraction method based on the sparse representation for speech mode classification. We have considered six Indian languages: Bengali, Oriya, Manipuri, Gujarati, Marathi, and Telugu for better comparison of proposed features. The proposed feature utilizes the sparse representation to find the significant details among diverse speech modes: conversation, extempore and read. Here, two dictionary learning strategies such as KSVD and PAU-DL are investigated to get the features for speech mode classification. From the experimental analysis, it was observed that the proposed features performed better as compared to the spectral, prosodic and excitation source features. It was found that the PAU-DL based sparse representation contains more discriminative information as compared to the KSVD based sparse representation. The feature proposed in this study could be a substitute for the conventional speech features.

In this study, the monolingual speech mode classifier is developed using the sparse based feature, but in future, we intend to explore a universal multilingual speech mode classifier using the proposed feature. We will also study the effect of sparsity in the case of a multilingual model.

6. Acknowledgements

The authors would like to acknowledge the project entitled “prosodically guided phonetic engine for searching speech databases in Indian languages (PSI)” funded by the Department of Information Technology, the Government of India, for providing the database.

7. References

- [1] A.-r. Mohamed, G. Dahl, and G. Hinton, “Deep belief networks for phone recognition,” in *Nips workshop on deep learning for speech recognition and related applications*, vol. 1, no. 9, p. 39, 2009.
- [2] K. Manjunath, K. S. Rao, and D. Pati, “Development of phonetic engine for indian languages: Bengali and oriya,” in *International Conference on Oriental COCOSA held jointly with Conference on Asian Spoken Language Research and Evaluation (O-COCOSA/CASLRE)*. IEEE, 2013, pp. 1–6.
- [3] K. Tripathi and K. S. Rao, “Improvement of phone recognition accuracy using speech mode classification,” *International Journal of Speech Technology*, pp. 1–12, 2017.

- [4] I. Tasic and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, 2011.
- [5] M. V. Shashanka, B. Raj, and P. Smaragdis, "Sparse overcomplete decomposition for single channel speaker separation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2. IEEE, 2007, pp. II–641.
- [6] G. S. Sivaram, S. K. Nemala, M. Elhilali, T. D. Tran, and H. Hermansky, "Sparse coding for speech recognition," in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 4346–4349.
- [7] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, and A. Sethy, "Sparse representation features for speech recognition," in *INTERSPEECH*. ISCA, 2010, pp. 2254–2257.
- [8] 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.
- [9] D. L. Donoho, Y. Tsaig, I. Drori, and J.-L. Starck, "Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 1094–1121, 2012.
- [10] M. Aharon, M. Elad, and A. Bruckstein, "*rmk*-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311–4322, 2006.
- [11] M. Sadeghi, M. Babaie-Zadeh, and C. Jutten, "Learning overcomplete dictionaries based on atom-by-atom updating," *IEEE Transactions on Signal Processing*, vol. 62, no. 4, pp. 883–891, 2014.
- [12] W. Dong, L. 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, 2011.
- [13] S. Haykin, *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.
- [14] S. S. Kumar, K. S. Rao, and D. Pati, "Phonetic and prosodically rich transcribed speech corpus in indian languages: Bengali and odia," in *International Conference on Oriental COCOSDA held jointly with Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*. IEEE, 2013, pp. 1–5.
- [15] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.