Unsupervised Spoken Word Retrieval using Gaussian-Bernoulli Restricted Boltzmann Machines

Pappagari Raghavendra Reddy, Shekhar Nayak and K Sri Rama Murty

Department of Electrical Engineering
Indian Institute of Technology Hyderabad, Hyderabad, India-502205
{ee12m1023,ee13p1008,ksrm}@iith.ac.in

10.21437/Interspeech.2014-397

Abstract

The objective of this work is to explore a novel unsupervised framework, using Restricted Boltzmann machines, for Spoken Word Retrieval (SWR). In the absence of labelled speech data, SWR is typically performed by matching sequence of feature vectors of query and test utterances using dynamic time warping (DTW). In such a scenario, performance of SWR system critically depends on representation of the speech signal. Typical features, like mel-frequency cepstral coefficients (MFCC), carry significant speaker-specific information, and hence may not be used directly in SWR system. To overcome this issue, we propose to capture the joint density of the acoustic space spanned by MFCCs using Gaussian-Bernoulli restricted Boltzmann machine (GBRBM). In this work, we have used hidden activations of the GBRBM as features for SWR system. Since the GBRBM is trained with speech data collected from large number of speakers, the hidden activations are more robust to the speaker-variability compared to MFCCs. The performance of the proposed features is evaluated on Telugu broadcast news data, and an absolute improvement of 12% was observed compared to MFCCs.

Index Terms: GBRBM, Representation learning, Joint density estimation, Subsequence matching, DTW.

1. Introduction

It is laborious job to manage and monitor increasing volumes of data on the internet. Resources can be efficiently managed by retrieving required information from this data. In case of speech data, we need to search and locate spoken query words in large volumes of continuous speech. This task is termed as Spoken Word Retrieval (SWR). Few applications of SWR include speech data indexing [1], data mining [2], voice dialling and telephone monitoring.

SWR task can be attempted through Automatic Speech Recogniser (ASR) by transcribing whole data, and using any text search engine for retrieval. But building a robust ASR is a difficult task, and it requires lot of manually labelled speech data. Moreover, an ASR system built using the data of one language cannot be used for other languages. To overcome these issues, attempts have been made to directly match the feature vectors extracted from test and query utterances. Although these approaches do not require any labelled data, their performance critically depends on the choice of features and the method employed to match them. A good feature should reflect the speech-specific characteristics and, at the same time, it should be robust to speaker and environmental variations. The template matching algorithm should be able to handle differences in the durations of the phonemes across the speakers. Dynamic time warping (DTW) or its variants are typically used for template matching. In this work, we focus on extracting good representative features suitable for developing an unsupervised SWR system.

Several attempts have been made to extract robust features for SWR system. However, most of these attempts use labelled speech data for feature extraction. Phoneme posteriors obtained from a pretrained ASR are used as features for SWR system [3][4]. While the performance of posterior features is satisfactory, they require a pretrained ASR built on a large amount of labelled speech data. A fully unsupervised feature extraction was attempted in [5][6][7], in which Gaussian Mixture Model (GMM) was used to generate posterior features. The posterior probabilities of individual Gaussian components was used as feature for SWR system. While the performance of these features is promising, the Gaussian assumption may not be valid in all the cases. Moreover, full covariance matrices need to be estimated for modelling correlated features using GMMs which require lot of data as well as computational resources.

In order to overcome these issues, we propose use Gaussian Bernoulli Restricted Boltzmann machines (GBRBM) to extract features for SWR system. A GBRBM is generative stochastic neural network model that can capture probability density of the feature vectors. In the last few years, GBRBM's [8] have been used for generative pretraining of deep neural networks which are used for phoneme classification. In this work GBRBM is used to estimate the probability density of mel-frequency cepstral coefficients (MFCCs), and hidden activations are used as features for SWR system. Subsequence matching has been done using DTW to match the proposed features, and locate the query word in the test utterance. It is observed the proposed features are more robust to speaker variability, and performed significantly better than raw MFCC features.

Rest of this paper is presented as follows. In section 2, unsupervised representation of spoken utterances using GBRBM is discussed. Subsequence matching is examined in section 3 to redeem the locations of spoken words. Experimental results are presented in section 4 and finally section 5 is furnished with conclusion and future work.

2. Unsupervised Feature Representation using GBRBM

2.1. Gaussian-Bernoulli Restricted Boltzmann Machines

A Restricted Boltzmann machine (RBM) is an undirected bipartite graphical model with visible units and hidden units. Intra-

layer connections between visible layers or hidden layers do not exist in RBM as opposed to Boltzmann machine. In RBM, visible and hidden units activations can assume only binary values. In general, speech data is represented in real valued features. In GBRBM visible units can have real values. The main difference between RBM and GBRBM is, in RBM visible units activations are sampled from Bernoulli distribution where as in GBRBM they are sampled from Gaussian distribution. Here we discuss only GBRBM as we are working with real valued MFCC features extracted from speech data.

GBRBM associates an energy with each configuration of visible units, hidden units and other parameters. The parameters learned in our work are weights, variances of visible units, hidden and visible biases. Parameters are updated such that the energy of GBRBM configuration reaches global minima on energy landscape. The energy function for GBRBM is defined as

$$E(\boldsymbol{v},\boldsymbol{h}) = \sum_{i=1}^{V} \frac{(v_i - b_i^v)^2}{2\sigma_i^2} - \sum_{j=1}^{H} b_j^h h_j - \sum_{i=1}^{V}\sum_{j=1}^{H} \frac{v_i}{\sigma_i} h_j w_{ij} \tag{1}$$

where $\boldsymbol{v}, \boldsymbol{h}$ are the real-valued vector of visible units and the binary vector of hidden units respectively, V and H are total number of visible and hidden units, v_i is the activation (real) of visible unit i, h_j is the binary state of hidden unit j, w_{ij} is the real-valued weight between visible unit i and hidden unit j, b_i^v is the real-valued bias into visible unit i, b_j^h is the real-valued bias into hidden unit j, σ_i is the real-valued parameter controlling width of parabola [9] formed by first quadratic term in the energy function.

The update equations for training GBRBM are

$$\Delta w_{ij} \propto \left(\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{recall}\right) \tag{2}$$

$$\Delta b_i^v \propto \left(\langle v_i \rangle_{data} - \langle v_i \rangle_{recall}\right) \tag{3}$$

$$\Delta b_j^h \propto \left(\langle h_j \rangle_{data} - \langle h_j \rangle_{recall}\right) \tag{4}$$

$$\Delta \sigma_i \propto \left(\langle \gamma \rangle_{data} - \langle \gamma \rangle_{recall}\right) \tag{5}$$

where

$$\gamma = \frac{(v_i - b_i^v)^2}{\sigma_i^3} - \sum_{j=1}^{H} \frac{h_j w_{ij} v_i}{\sigma_i^2}$$

and $\langle . \rangle_{data}$ denotes expectation over the input data and $\langle . \rangle_{recall}$ denotes expectation over recalled data.

Contrastive Divergence (CD) algorithm [10] is employed in training GBRBM . For each cycle of CD, the energy of the configuration decreases. After infinite cycles, the model is said to be reached thermal equilibrium where energy does not change. GBRBM is trained to estimate the underlying joint density of training data. GMM can also be used to estimate the density of data but limitations of GMM can be overcome by using GBRBM. Building GMM with full covariance matrix requires lots of data as it involves estimation of higher order correlations. So, in general, GMM is trained with diagonal covariance matrix. This limitation can be overcome through GBRBM. GBRBM has the ability to exploit the dependencies between each dimension even with small amounts of training data. High dimensional input data can be handled in GBRBM where as GMM can not. Generally in any mixture model for example GMM, the variances associated with each component allow for better modelling the given data. The importance of variance parameter in GBRBM is explained through experiments.

Intra layer connections between neurons are absent. So given one layer activations, the activations of other layer neurons are independent to each other. For example, given the hidden layer activations, all visible neurons are independent. So problem of modelling joint distribution of input data can be verified with modelled Probability Density Function's (PDF) of visible neurons (see section 4).

2.2. GBRBM Feature Extraction

The joint density modelled by applying GBRBM, can be interpreted as product of experts [11]. The probability that a given feature vector belongs to each expert k, where k can be any value from 1 to H, can be determined as the conditional probability that h_k is fired given visible neuron activations and it is calculated as

$$P(h_k = 1 \mid \boldsymbol{v}) = sigmoid(\sum_{i=1}^{V} \frac{v_i}{\sigma_i} w_{ik} + b_k^h) \tag{6}$$

Expert probability vector in other words, posterior vector for any input frame $\boldsymbol{v}$ is

$$EP = \{P(h_k = 1 \mid \boldsymbol{v})\} \tag{7}$$

for $k = 1, 2, ...H$. Expert posteriorgram is defined as a matrix of expert probability vectors arranged in a matrix for all input frames of test utterance. Now, expert posteriorgram represents the speech utterance in subsequence matching.

3. Subsequence Matching of GBRBM Features

Statistical behaviour of vocal tract system makes it impossible to generate two exactly similar speech segments in natural speech. So, no two speech utterances have equal durations of phonemes, they are distributed non-linearly. To locate a given spoken word in continuous speech, the template matching algorithm should be able to capture natural variations. DTW [12] was introduced for this task. Degree of similarity of two given utterances can be approximated through cost of optimal path of DTW. Limitation of DTW is that durations of two utterances used for matching should be comparable otherwise the computed similarity score denotes the closeness of long utterance with small utterance like spoken word which are not comparable. As name subsequence matching suggests, DTW is carried out on spoken word and subsequence i.e., chunk of test utterance. The subsequence is chosen sequentially starting from first frame of test utterance until last frame of test utterance is covered. Each comparison is represented with optimal score. The subsequence that best matches with given spoken word does have low dissimilarity score and is chosen as system proposed location for that test utterance.

We impose a constraint on DTW path. The path is constrained to move within a window. We have used window exploited by Sakoe et.al., in [13]. Window size can affect the SWR performance. Small window size constrain the warping path to diagonal. This does not work well because speech utterances are generally non-linearly warped in time. Large window size does not take into account the fact that each phoneme duration differ by only small amounts in normal speech. Optimal window size allows the natural variations in speech signal to produce best path.

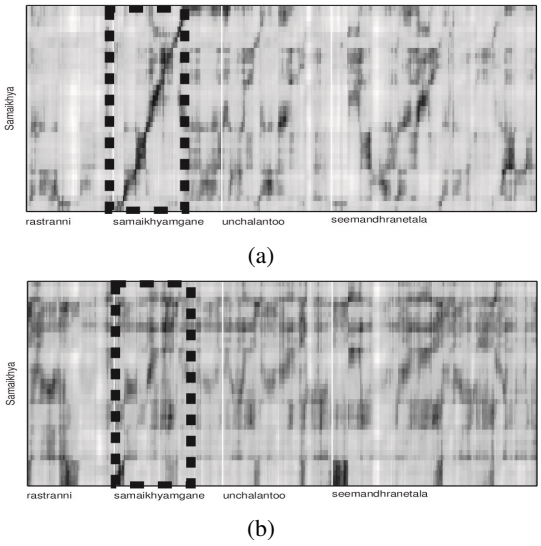

Figure 1: Similarity matrices for test utterance and spoken word spoken by (a) Same speaker (b) Different speaker which are represented by MFCCs

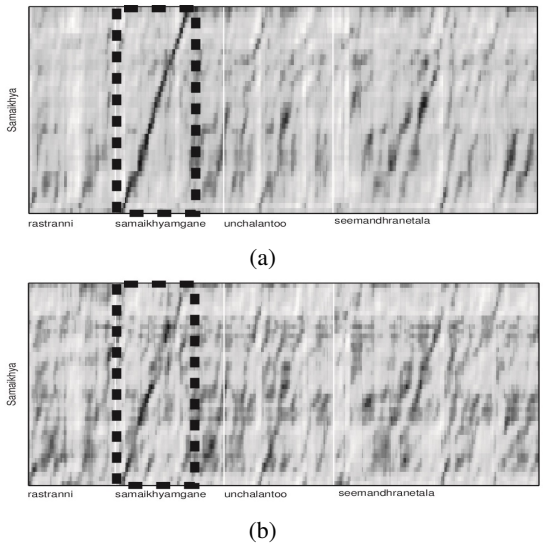

Figure 2: Similarity matrices for test utterance and spoken word spoken by (a) Same speaker (b) Different speaker which are represented by GBRBM posteriorgrams

4. Experimental Evaluation

4.1. Speech Corpus and System Setup

Developed SWR system is tested on six and half hours of Telugu data collected from news channels. Training dataset consists of five and half hours of data covering six female and six male speakers. Another one hour is used for testing consisting three male and three female speakers. Thirty words are chosen randomly from testing data so that minimum number of syllables in any word is 2 and maximum number of syllables is 8 and are listed in Table 4, scripted in IPA format. Each word is tested against 5-10 minutes news bulletin.

GBRBM is trained on 39 dimensional MFCCs extracted with 25ms window and 10ms shift. One hidden layer is used in the GBRBM architecture. Effect of number of hidden neurons on overall SWR system is analysed by experimenting with different number of hidden neurons. Visible layer contains 39 units as the input data is 39-dimensional MFCCs. As GBRBM is trained generatively, it estimates the joint density of the presented data. It is shown that introducing sparsity in the GBRBM architecture [14] improved the learning capabilities of GBRBM. We have used 0.3 as sparsity target. Only one CD cycle is used in training. From the joint density, expert posteriorgram using Eq. 6 and Eq. 7 is generated for test utterance and spoken word. DTW is applied on extracted GBRBM features with window size of seven frames. Window size is chosen empirically. Kullback-Leibler Divergence is used as local distance measure in calculating DTW similarity matrix.

4.2. Evaluation Measure and Results

The performance of SWR system is measured in terms of average precision (P@N), where N is number of occurrences of spoken word in test utterance. P@N is calculated as proportion of spoken words located correctly in top N hits from test utterance. System proposed location is chosen as hit if it matches more than 50% with reference location.

Similarity matrices for same and different speakers across test utterance and spoken word are shown in Figure 1a and 1b

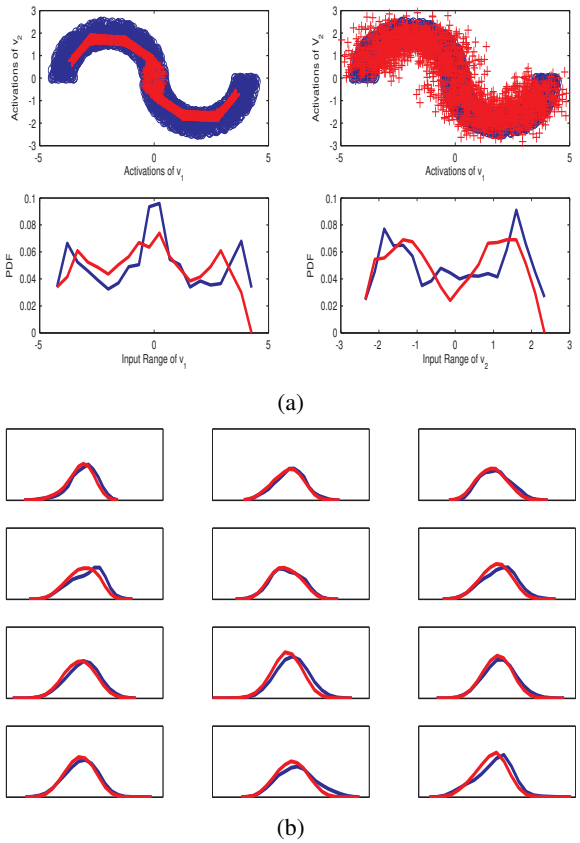

Figure 3: Illustration of distribution capturing capabilities of GBRBM

Table 1: Results of SWR system with various features used for subsequence matching

Metric	MFCC	GMM	GBRBM	GBRBM+GMM
P@N	45.68%	53.10%	57.64%	59.91%

Table 2: Effect of variance learning on SWR system performance

Metric	GBRBM-NV	GBRBM-NV+GMM	GBRBM+GMM
P@N	45.84%	53.86%	59.91%

respectively. Black color represents high similarity and white color represents low similarity. Similarity at marked area in Figure 1a and 1b is conspicuous. Similarity matrices with GBRBM features are plotted in Figure 2. The improved similarity at marked area in Figure 2b comparative to Figure 1b has shown the robustness of GBRBM extracted features.

Effectiveness of GBRBM model in capturing distribution of presented input data is illustrated in Figure 3. In Figure 3 blue color corresponds to input data and red color to estimated data. Synthetic data used as training data for GBRBM is plotted in left-top sub plot of Figure 3a. Figure 3a is the plot of visible units amplitude and mean of modelled density. As expected, the mean is around the middle of input data distribution. Comparison of input data and data sampled from GBRBM model is plotted in top-right sub plot of Figure 3a. The shape of the red color data is approximately following the input data. This means that the GBRBM model has modelled the input data efficiently. Bottom plots in Figure 3a shows the comparison of input data PDF and reconstructed PDF for two visible units. Figure 3a plots are generated using 2 visible neurons and 6 hidden neurons. For speech data modelled with 39 visible neurons and 50 hidden neurons, the modelled PDF's for first 12 MFCC coefficients are plotted in Figure 3b in comparison with input data PDF. We can observe that modelled PDF is closely following input PDF in every sub plot of Figure 3b indicating that GBRBM is successfully trained.

Comparison of SWR system performances for different parameters are tabulated in Table 1, 2, 3. Number of mixtures used in building GMM is chosen through experiments as 64. GMM is trained with the assumption that all dimensions are independent i.e., diagonal covariance is estimated for each mixture of GMM. GBRBM-NV represents GBRBM without variance learning. Also, GBRBM denotes the model with variance learning. The performance with GBRBM features is much superior than with MFCC features and GMM posteriorgrams [6] as shown in Table 1. Improvement in accuracy can be observed with concatenation of GMM posteriorgrams and GBRBM features. If variance is not learnt in GBRBM then the features extracted are no better than MFCC except for fractional improvement. Variance learning in GBRBM enabled us to improve the performance, this can be observed in Table 2. The features learned from GBRBM are much robust compared to MFCC and also GMM posteriorgrams. The number of hidden neurons in GBRBM also affects the SWR performance. Table 3 is featured with performances of SWR system for different hidden neurons count. The configuration with 39 visible neurons and 50 hidden neurons yielded best result.

Table 3: Results of SWR with various number of hidden neurons

Metric	GBRBM hidden neurons			
	30	50	70	90
P@N	50.83%	57.64%	56.58%	56.43%

Table 4: Descending order of P@N with associated spoken words

Spoken word	**P@N**(%)	**Spoken word**	**P@N**(%)
prənəbmuk^{h}ərji	100	saĭləɟa:na:t^{h}	60
teləngaːnə	90	digviɟəjsiŋg	60
səmaːveːsəm	86.36	ʋiʋəraːlu	59.91
kirənkumaːrreḍḍi	85.71	raːjiːnaːma	59.26
soːnijaːgaːnd^{h}i̊	83.33	prəb^{h}utʋəm	57.14
pənchaːjəti	82.76	səmaĭk^{h}jə	54.84
kaŋgres	78	vaːtaːvərənəm	50
neːpət^{h}jəm	76.92	adjəks^{h}uraːlu	50
paːrləmenţ	76.47	ḍi̊lli̊ː	41.67
bəngaːlaːk^{h}aːtəm	75	erpaːţu	40.91
poːliŋg	75	eṉṉikəlu	40.54
haĭdraːbaːd	71.43	ruːpaːji	40
alpəpiːḍənəm	69.23	kəmːţi	35.29
ad^{h}ikaːrulu	64.29	vib^{h}əɟənə	33.33
nirnəjəm	63.89	məntri	12.70

The accuracies of individual words are formulated in Table 4 in the descending order of their accuracy. As expected longer words are retrieved with high accuracy and words with less number of syllables are located comparatively with less accuracy.

5. Conclusion and Future Work

In this work, we presented the SWR system in unsupervised framework. Feature learning is done by applying GBRBM. GBRBM is trained with 39-dimensional MFCC vectors extracted from speech data. DTW is employed on features extracted from GBRBM. P@N is used as evaluation measure to determine the performance of SWR system. It is shown that GBRBM features are better suitable for SWR task compared to MFCC and GMM posteriorgrams. Importance of variance learning is explained through experiments. Accuracy is enhanced by 11.80% through variance estimation. Further, with the concatenation of GMM posteriorgrams and GBRBM features, 2.27% improvement is achieved. In total, 14.23% gain in accuracy is accomplished compared to the accuracy with MFCC features. We experimented with number of hidden neurons and and have chosen the best value. Our future work will focus to improve the performance by incorporating context dependence and by exploring different architectures of deep neural networks.

6. Acknowledgements

Authors would like to acknowledge Department of Electronics and Information Technology (DeitY), Ministry of Communications & Information Technology, Government of INDIA, for kindly sponsoring this work.

7. References

[1] J. Foote, "An overview of audio information retrieval," *Multimedia Systems*, vol. 7, no. 1, pp. 2–10, 1999.

[2] A. J. Thambiratnam, "Acoustic keyword spotting in speech with applications to data mining," 2005.

[3] G. Aradilla, J. Vepa, and H. Bourlard, "Using posterior-based features in template matching for speech recognition." in *INTERSPEECH*, 2006.

[4] T. J. Hazen, W. Shen, and C. White, "Query-by-example spoken term detection using phonetic posteriorgram templates," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 421–426.

[5] Y. Zhang and J. R. Glass, "Unsupervised spoken keyword spotting via segmental dtw on gaussian posteriorgrams," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 398–403.

[6] P. R. Reddy, K. Rout, and K. S. R. Murty, "Query word retrieval from continuous speech using GMM posteriorgrams," in *International Conference on Signal Processing and Communications - 2014 (SPCOM 2014)*, Indian Institute of Science, Bangalore, India, Jul. 2014.

[7] G. V. Mantena, S. Achanta, and K. Prahallad, "Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping," *IEEE/ACM Transactions on Audio, Speech & Language Processing*, vol. 22, no. 5, pp. 944–953, 2014.

[8] N. Le Roux, N. Heess, J. Shotton, and J. Winn, "Learning a generative model of images by factoring appearance and shape," *Neural Computation*, vol. 23, no. 3, pp. 593–650, 2011.

[9] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Computer Science Department, University of Toronto, Tech. Rep*, 2009.

[10] M. A. Carreira-Perpinan and G. E. Hinton, "On contrastive divergence learning," in *Proceedings of the tenth international workshop on artificial intelligence and statistics*. Citeseer, 2005, pp. 33–40.

[11] N. Wang, J. Melchior, and L. Wiskott, "An analysis of gaussian-binary restricted boltzmann machines for natural images," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2012, pp. 287–292.

[12] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, 1994, pp. 359–370.

[13] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 26, no. 1, pp. 43–49, 1978.

[14] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area v2," in *Advances in neural information processing systems*, 2008, pp. 873–880.