Unsupervised Learning of Temporal Receptive Fields Using Convolutional RBM For ASR Task

Hardik B. Sailor and Hemant A. Patil

Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, Gujarat, India
sailor_hardik@daiict.ac.in and hemant_patil@daiict.ac.in

***Abstract*—There has been a significant research attention for unsupervised representation learning to learn the features for speech processing applications. In this paper, we investigate unsupervised representation learning using Convolutional Restricted Boltzmann Machine (ConvRBM) with rectified units for speech recognition task. Temporal modulation representation is learned using log Mel-spectrogram as an input to ConvRBM. ConvRBM as modulation features and filterbank as spectral features were separately trained on DNNs and then system combination is used. With our proposed setup, ConvRBM features were applied to speech recognition task on TIMIT and WSJ0 databases. On TIMIT database, we achieved relative improvement of *5.93%* in PER on test set compared to only filterbank features. For WSJ0 database, we achieved relative improvement of *3.63-4.3 %* in WER on test sets compared to filterbank features. Hence, DNN trained on ConvRBM with rectified units provide significant complementary information in terms of temporal modulation features.**

***Index Terms*—Convolutional RBM, temporal modulations, speech recognition, deep neural networks.**

I. Introduction

For any speech processing applications, representation of speech signal requires significant attention. As discussed in [1], a good representation of speech signal helps supervised pattern recognition tasks for several speech research problems. For speech recognition task, significant amount of studies have been done to design features of which widely used features include Mel Frequency Cepstral Coefficients (MFCC) [2] and Perceptual Linear Prediction (PLP) coefficients [3]. Many earlier attempts for feature vector design include handcrafted features which either perform better than MFCC (or PLP) or capture complementary information so as to improve the results of Automatic Speech Recognition (ASR) by appending it with MFCC (or PLP), i.e., feature-level fusion or using system-level combination. These features are mainly based on concepts of speech perception and/or speech production mechanism. However, they do not make use of the amount of information in speech data (to be used to train models) itself.

Representation learning for speech processing applications is an active area of research which makes effective use of speech data. Approaches for representation learning include supervised and unsupervised (where the labels of speech sound units are not available). Recent supervised feature learning techniques along with acoustic modelling include work reported in [4]–[7]. Unsupervised learning is most important form of deep learning as human learning is largely unsupervised [8]. For example, language acquisition by the infants during initial stages of their growth, is a type of unsupervised learning. There has been number of attempts for unsupervised feature learning either directly from raw speech signal [9]–[12], or from time-frequency representation of speech signal [13]–[17].

Recently, we have proposed unsupervised learning model based on convolutional RBM (denoted as ConvRBM) to learn filterbank directly from raw speech signals [12]. We have shown that ConvRBM is able to model human auditory system with subband filters resemble gammatone signals and shown to improve recognition performance in various ASR tasks [18]. In this paper, we have used ConvRBM to learn modulation representation from Mel spectrograms. Our model is different than the one used in [19] in terms of type of hidden units. We have proposed to use rectified linear units (ReLU) as hidden units activation function and for inference in the model. We have used a system combination framework for Mel filterbank features and modulation features learned by ConvRBM. Experiments on TIMIT and WSJ0 databases show significant reduction in error rates with this system combination framework over baseline DNN trained on filterbank features alone.

The rest of the paper is organized as follows: Section II presented the brief theory of ConvRBM. Analysis of the model and feature representation is presented in Section III. Details of system combination is given in Section III-C. Section IV and V present experimental setup and results. Finally, paper is summarized in Section VI.

II. Convolutional RBM

To improve the scalability of RBM [20], it was extended in convolutional framework following the idea of Convolutional Neural Networks (CNNs) [21]. In ConvRBM, weights between hidden units and visible units are shared among all the locations in hidden layer [14]. The input $\mathbf{v}$ to the ConvRBM is the time-frequency representation of speech (of entire utterance) with $s = 1, 2, .., S$ subbands and n_V frames. Hidden layer ($\mathbf{h}$) consists of *K* filters each of n_W-dimensional (i.e., $n_W - D$) filter weights $\mathbf{W}^K$ (also called as *bases* [19]). Hidden layer is divided into *K* groups of n_H-D (where $n_H = n_V - n_W + 1$ length of 'valid' convolution) array. All units in the k^{th} group share the weights $\mathbf{W}^K$. Biases are also shared in hidden layer

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

and visible layer denoted as b_k and c, respectively. The energy function for ConvRBM with visible units v_i and hidden units h_j^k is defined as [19],

$$E(\mathbf{v},\mathbf{h}) = \frac{1}{2}\sum_{i=1}^{n_V} v_i^2 - \sum_{k=1}^{K}\sum_{j=1}^{n_H}\sum_{r=1}^{n_W} h_j^k W_r^k v_{j+r-1} - \sum_{k=1}^{K} b_k \sum_{j=1}^{n_H} h_j^k - c\sum_{i=1}^{n_V} v_i. \quad (1)$$

The joint probability distribution in terms of this energy function is given by [14],

$$P(\mathbf{v},\mathbf{h}) = \frac{1}{Z}\exp(-E(\mathbf{v},\mathbf{h})). \quad (2)$$

The response of the convolution layer for the k-th group is given as [14]:

$$I_k = (\mathbf{v} * \tilde{\mathbf{W}}^k) + b_k, \quad (3)$$

where $*$ denotes 'valid' length convolution operation and $\tilde{\mathbf{W}}^k$ denotes filpped array (for convolution operation). The response of all K-groups $[I_1, I_2, ..., I_K]^T$ is $K \times n_V$-dimensional matrix. In case of sigmoid units, the conditional distribution of hidden layer is defined as follows [14]:

$$P(h_j^k = 1|\mathbf{v}) = sigmoid((\tilde{\mathbf{W}}^k * \mathbf{v})_j + b_k). \quad (4)$$

Recently, it is shown that rectified units perform better than sigmoid units in unsupervised as well as supervised networks [22], [23], [12]. Hence, we have used Rectified Linear Units (ReLUs) as activations in hidden layer instead of sigmoid units. For sampling from hidden units noisy ReLU is used as shown in [22]. Sampling equations for hidden and visible units are given as [12],

$$\mathbf{h}^k \sim max(0, I_k + N(0, \sigma(I_k))),$$
$$\mathbf{v} \sim \mathcal{N}\left(\sum_{k=1}^{K}(\mathbf{h}^k * W^k) + c, 1\right), \quad (5)$$

where $N(0, \sigma(I_k))$ is a Gaussian noise with mean zero and sigmoid of I_k as it's variance. During feature extraction stage (i.e., testing stage), we have used deterministic version of ReLU activation $max(0, I_k)$. The block diagram of our proposed model is shown in Fig.1. The input to ConvRBM is Principal Component Analysis (PCA) whitened log Mel-spectrogram extracted from the speech signal. Whitening the data using PCA gives approximation to sub-cortical processing which was observed in auditory cortex [24]. From Fig.1, we can see that ReLU nonlinearity force many hidden units to be zero and hence, increase *sparsity* in features [25]. Since the ConvRBM representation is *overcomplete* (more number of bases than dimension of input), sparsity penalty term is added [14]. Inference in ConvRBM is done using block Gibbs sampling. Gradient computation is performed using contrastive divergence which approximates the gradient term effectively [26]. Weights and biases are updated using gradient-descent algorithm as done in [12].

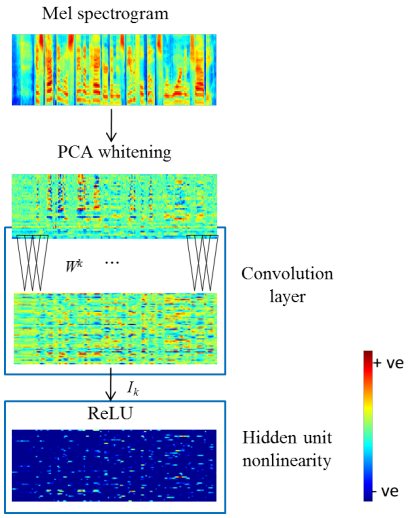

Fig. 1. Block diagram of input representation and proposed ConvRBM model.

III. Analysis and Feature Representation

A. Analysis of Learned Subband Filters

The filters learned in ConvRBM are visualized by applying inverse of PCA whitening on the ConvRBM weights. Since convolution is applied in temporal-domain (for each subbands), subband filters represent Temporal Receptive Fields (TRFs) [19]. Examples of TRFs learned on TIMIT database are shown in Fig. 2 where each block represents one TRF.

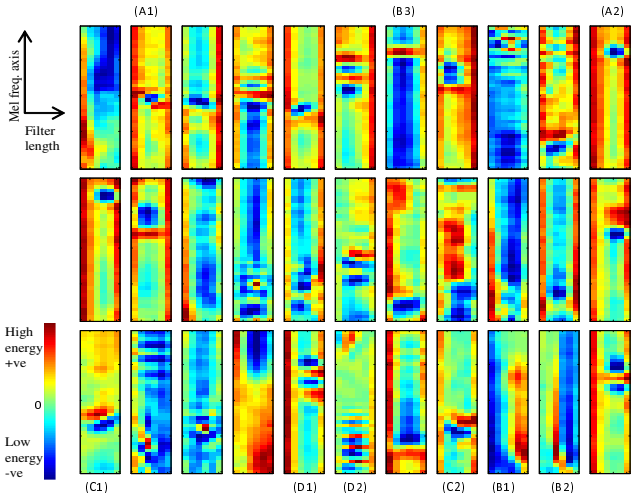

Fig. 2. Examples of ConvRBM filters.

Unlike cells in visual cortex, all Receptive Fields (RFs) in auditory cortex are not localized [27]. Receptive fields in A1 exhibit multiple characteristics as certain cells demonstrate responses from multiple frequencies [24]. Here, we observe similar behavior of TRFs. From Fig. 2, it can be seen that some of the filters are highly localized along Mel frequency-axis (e.g., *A1* and *A2*) while some filters are broadly distributed (e.g., *B1*, *B2* and *B3*). Some filters have strong localized exci-

tatory and inhibitory regions (e.g., *C1* and *C2*) while few have checkerboard-like pattern (e.g., *D1* and *D2*). Similar patterns of RFs were also found in Spectro-temporal Receptive Fields (STRFs) in auditory cortex [13]. Hence, ConvRBM subband filters capture temporal modulation information with different *subband modulation frequencies* from log Mel spectrograms [28]. As shown in [19] each subband filter may represent temporal variations in different phonetic units.

B. Rectified Linear Units (ReLU) in ConvRBM

We justify the use of ReLUs in ConvRBM by visualizing the reconstructions from the model using both non-linearity as shown in Fig. 3. It can be evident that reconstruction from sigmoid units is more noisy (shown in dotted circles) compared to original spectrogram and the one reconstructed using ReLU. This noise is due to saturation of neurons and vanishing gradient effect in case of sigmoid non-linearity which may affect ASR performance. In case of ReLUs, hidden units are not binary rather neurons can take any value from 0 to ∞ and hence, can better represent the input signal.

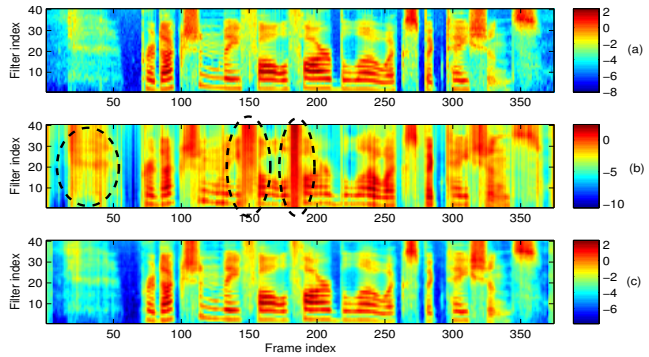

Fig. 3. Log Mel-spectrograms (a) original, (b) reconstructed using sigmoid units, (c) reconstructed using ReLUs. Dotted regions shows saturation effect in sigmoid units.

C. Feature Extraction and System Combination for ASR

Since ConvRBM filters capture temporal modulation information, we can use this along with standard spectral features, Mel frequency filterbanks (denoted as FBANK). We trained both features on separate DNN and use system combination technique. We have used Minimum Bayes Risk (MBR) technique for system combination which is very helpful when two different feature stream may represent complementary information [29]. Lattices generated by N different systems are combined to get optimal word sequence as follows [29]:

$$W^* = \underset{W}{arg\,min} \left\{ \sum_{i=1}^{N} \lambda_i \sum_{W'} P_i(W'|\mathbf{O}) L(W, W') \right\}, \quad (6)$$

where $L(W, W')$ is the Levenshtein edit distance between two word sequences, $P_i(W'|\mathbf{O})$ is the posterior probability of the word sequence W' given the acoustic observation sequence $\mathbf{O}$ and λ_i is the weight assigned to i^{th} system. Our feature extraction and system combination method is shown in Fig. 4. First pipeline is to extract FBANK features and their delta features to train DNN which we denote spectral feature trained DNN. Second pipeline is to extract temporal modulation features from trained ConvRBM and train DNN on these features which we call modulation feature trained DNN. Generated lattices from both DNN systems are combined (with $\lambda = 0.5$) using MBR decoding and then used for scoring.

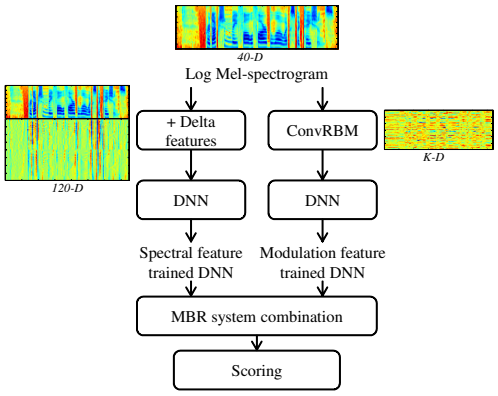

Fig. 4. Block diagram for feature extraction and system combination framework.

IV. Experimental Setup

A. Speech Databases

Two standard speech databases, TIMIT and subset of Wall Street Journal (i.e., WSJ0), were experimented in this paper. For phoneme recognition task, we have used phonetically balanced TIMIT database [30]. All SA category sentences (same sentences spoken by all speakers in database) were removed as they may bias the recognition performance. Training set includes *462* speakers, development set includes *50* speakers and core test set includes *24* speakers. WSJ0 SI-84 training set includes *7138* utterances spoken by *84* speakers [31]. Training data consists of *14* hours of speech data. Standard evaluation set contains *330* utterances from *8* speakers with *5K*-word WSJ vocabulary denoted as eval92_5K. We have also used *20K*-word WSJ vocabulary test set denoted as eval92_20K.

B. Training of ConvRBM

Log-Mel spectrogram was obtained from speech signal by framing it with a window length of *25* ms and shift of *10* ms using *40* Mel subband filters. PCA whitening was applied on all concatenated log-Mel spectrograms. Learning rate was chosen to be *0.01* and was decayed at each epoch. As suggested in [32], L^2-norm sparsity regularization on hidden units was applied with target sparsity *0.1* and regularization constant to be *1*. For first five training epochs, momentum was set to *0.5* and after that it was set to *0.9*. We trained model on different number of filters *60*, *80* and *120* with different filter lengths, namely, *6*, *8* and *10*. These parameters were optimized based on performance in speech recognition experiments. The learned weights and biases were used in inference to extract the features from ConvRBM. The notations for ConvRBM

using sigmoid and ReLU hidden units are ReLU-ConvRBM and Sigmoid-ConvRBM, respectively.

C. Hybrid DNN-HMM Systems

Monophone GMM-HMM systems were used to generate forced-aligned transcriptions for both databases. For these systems, MFCC features were extracted from speech signal using *25* ms window length and *10* ms window shift. *39*-D vector features were formed by *13*-D MFCC followed by delta and delta-delta features. Acoustic modeling was performed with hybrid DNN-HMM system in Kaldi [33]. DNN weights were randomly initialized. All experiments on TIMIT used bi-gram Language Model (LM) estimated from training set. For DNN training, *144* target labels (*48* phones with *3* states) were used. During final scoring, *48* phones were mapped to *39* phones as done in [34]. For WSJ0 database, language modelling is performed using tri-gram language model. For training of DNNs, *132* target labels (*44* phones with *3* states) were used with same learning rate and random weight initialization.

V. Experimental Results

A. ConvRBM Parameter Tuning

Parameters of ConvRBM were optimized using a single layer neural network trained on ReLU-ConvRBM features with *1500* hidden units and Context Window (CW) of *11* frames. The parameters of ConvRBM include number of filters and length of filter. Results of these experiments are reported in Table I on the TIMIT and WSJ0 databases in % Phone Error Rate (PER) and % Word Error Rate (WER), respectively. From Table I, for TIMIT database, *120* number of filters and filter length *6* gave lowest % PER. For WSJ0 database, *60* number of filters and filter length *6* gave lowest % WER. For large database, *60* and *120* filters yield almost similar WER and *60* filters are sufficient compared to double the number of filters for TIMIT database.

TABLE I
Results of ConvRBM parameter tuning experiments on TIMIT database

Number of filters	Filter length	Dev	Test	eval92_5K
120	6	**24.3**	**25.6**	7.14
80	6	24.5	25.6	7.57
60	6	24.6	25.8	**7.10**
120	8	24.5	25.7	7.15
120	10	24.9	25.8	7.20
60	8	24.8	25.9	7.15

B. Experiments on TIMIT Database

Two DNNs are trained with setup described in Section IV-C using FBANK (*120*-D) and ConvRBM (*120*-D) features and results are reported with *3* layers, *1500* hidden units and context-window of *11* frames. The performance of ConvRBM features alone and with our system combination setup is reported in Table II. ReLU-ConvRBM features which represent modulation information, perform similar as spectral features, FBANK. ConvRBM with sigmoid units (denoted as Sigmoid-ConvRBM) did not perform well compared to ReLU units. From eq. (6), the system combination (denoted using $\oplus$ symbol) of DNNs trained on ReLU-ConvRBM and FBANK features gave relative improvement of *7.73 %* and *5.93 %* on development and test set, respectively. System combination using Sigmoid-ConvRBM has very low improvement on development set (*4.1 %* relative improvement) compared to ReLU. This shows that DNNs trained on FBANK and ReLU-ConvRBM contain highly *complementary* information.

TABLE II
Results on TIMIT database in % PER. Numbers in bracket indicates relative improvement over FBANK-DNN

DNN System	Dev	Test
A: FBANK	22.0	23.6
B: ReLU-ConvRBM	22.3 (-1.36)	24.0 (-1.69)
C: Sigmoid-ConvRBM	23.4 (-6.36)	25.4 (-7.63)
A $\oplus$ B	**20.3 (7.73)**	**22.2 (5.93)**
A $\oplus$ C	21.1 (4.09)	22.6 (4.23)

C. Experiments on WSJ0 Database

Following the results of Table I, we have used parameters of ConvRBM for WSJ0 experiments. FBANK and ConvRBM features were trained using DNNs and results are reported in Table III in % WER with *3* layers, *1500* hidden units and context-window of *11* frames. From results, we can see that ReLU-ConvRBM features perform better than Sigmoid-ConvRBM. We can see that system combination of FBANK and ReLU-ConvRBM yield relative improvement of *4.3 %* for 5K word test set and *3.63 %* for 20K word test set over FBANK features. System combination of FBANK and Sigmoid-ConvRBM also improved over FBANK. However, improvement is less compared to ReLU-ConvRBM with relative improvement of *1.48 %* on 5K test set and *2.09 %* on 20K test set. Hence, both ASR experiments shows that ReLU-ConvRBM perform better than Sigmoid-ConvRBM and performance is improved using system combination with FBANK features.

TABLE III
Results in % WER and % Relative Improvements on WSJ0 database

DNN System	eval92_5K	eval92_20K
A: FBANK	6.07	14.32
B: ReLU-ConvRBM	6.52 (-7.4)	15.15 (-5.7)
C: Sigmoid-ConvRBM	7.44 (-22.57)	16.16 (-12.84)
A $\oplus$ B	**5.81 (4.3)**	**13.80 (3.63)**
A $\oplus$ C	5.98 (1.48)	14.02 (2.09)

VI. Summary and Conclusions

We have developed convolutional RBM with rectified units for unsupervised feature learning using log-Mel spectrograms. As convolution is applied in time-domain, it learns temporal receptive fields and hence, represents temporal modulations. Modulation features learned by ConvRBM is used with spectral features, i.e., FBANK in system combination framework

using DNN. Experiments on TIMIT and WSJ0 databases shows significant improvements when using ReLU-ConvRBM features in system combination with FBANK. Our future research work will be directed towards extending this for *2-D* ConvRBM incorporating convolution along frequency-axis. We also want to use stacks of ConvRBM where one ConvRBM is for learning filterbank from speech and another for learning modulation features giving the deep feature learning model. We would also like to apply this framework for low resource and noise robust ASR task.

ACKNOWLEDGMENT

The authors would like to thank Dept. of Electronics and Information Technology (DeitY), Govt. of India for sponsoring the consortium project, namely, Development of Text-to-Speech Synthesis Systems in Indian Languages. We also like to thank authorities of DA-IICT, Gandhinagar to carry out this research work.

REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug 2013.

[2] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on Acousti., Speech and Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[3] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acousti. Soc. of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[4] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, and O. Vinyals, "Learning the speech front-end with raw waveform CLDNNs," in *INTERSPEECH*, Dresden, Germany, 6–10 Sept 2015, pp. 1–5.

[5] P. Golik, Z. Tüske, R. Schlüter, and H. Ney, "Convolutional neural networks for acoustic modeling of raw time signal in LVCSR," in *INTERSPEECH*, Dresden, Germany, 6-10 Sep 2015, pp. 26–30.

[6] D. Palaz, M. Magimai-Doss, and R. Collobert, "Convolutional neural networks-based continuous speech recognition using raw speech signal," in *40^{th} International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD*, 2015, pp. 4295–4299.

[7] Z. Tüske, P. Golik, R. Schlüter, and H. Ney, "Acoustic modeling with deep neural networks using raw time signal for LVCSR," in *INTERSPEECH*, Singapore, Sep 2014, pp. 890–894.

[8] Y. B. Y. LeCun and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[9] J. Lee, H. Jung, T. Lee, and S. Lee, "Speech feature extraction using independent component analysis," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, 2000, pp. 1631–1634.

[10] M. S. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, pp. 356–363, 2002.

[11] N. Jaitly and G. Hinton, "Learning a better representation of speech soundwaves using restricted Boltzmann machines," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2011, pp. 5884–5887.

[12] H. B. Sailor and H. A. Patil, "Filterbank learning using convolutional restricted Boltzmann machine for speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016*, Shanghai, China, March 2016, pp. 5895–5899.

[13] D. J. Klein, P. König, and K. P. Körding, "Sparse spectrotemporal coding of sounds," *EURASIP J. Adv. Sig. Proc.*, vol. 2003, no. 7, pp. 659–667, 2003.

[14] H. Lee, R. B. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *26^{th} Annual International Conference on Machine Learning, ICML, Canada, June 14-18*, 2009, pp. 609–616.

[15] N. Jaitly and G. E. Hinton, "Using an autoencoder with deformable templates to discover features for automated speech recognition," in *INTERSPEECH 2013, 14^{th} Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29*, 2013, pp. 1737–1740.

[16] C. Zhang *et al.*, "Phone classification by a hierarchy of invariant representation layers," in *INTERSPEECH 2014, 15^{th} Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pp. 2346–2350.

[17] W. Wang *et al.*, "Unsupervised learning of acoustic features via deep canonical correlation analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, South Brisbane, Queensland, Australia, April 19-24*, 2015, pp. 4590–4594.

[18] H. B. Sailor and H. A. Patil, "Novel unsupervised auditory filterbank learning using convolutional RBM for speech recognition," submitted to IEEE Transcations on Audio, Speech and Signal Processing, 2016.

[19] H. Lee, P. T. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *23^{rd} Annual Conference on Neural Information Processing Systems, Canada, 7-10 December*, 2009, pp. 1096–1104.

[20] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1*, D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, Eds. Cambridge, MA, USA: MIT Press, 1986, pp. 194–281.

[21] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[22] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *27^{th} International Conference on Machine Learning (ICML)*, 2010, pp. 807–814.

[23] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and G. E. Hinton, "On rectified linear units for speech processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 3517–3521.

[24] A. M. Saxe *et al.*, "Unsupervised learning models of primary cortical receptive fields and receptive field plasticity," in *25^{th} Annual Conference on Neural Information Processing Systems, 12-14 December, Granada, Spain.*, 2011, pp. 1971–1979.

[25] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011, pp. 315–323.

[26] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[27] H. Terashima and M. Okada, "The topographic unsupervised learning of natural sounds in the auditory cortex," in *26^{th} Annual Conference on Neural Information Processing Systems 2012, December 3-6, 2012, Lake Tahoe, United States.*, 2012, pp. 2321–2329.

[28] C. Lee, F. K. Soong, and K. Paliwal, *Automatic speech and speaker recognition: advanced topics*. Springer Science & Business Media, 2012, vol. 355.

[29] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum Bayes risk decoding and system combination based on a recursion for edit distance," *Computer, Speech & Language*, vol. 25, no. 4, pp. 802 – 828, 2011.

[30] Garofolo *et al.*, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, p. 27403, 1993.

[31] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *The Workshop on Speech and Natural Language*, ser. HLT '91, Stroudsburg, PA, USA, 1992, pp. 357–362.

[32] H. Lee, C. Ekanadham, and A. Y. Ng, "Sparse deep belief net model for visual area V2," in *Proceedings of the 21st Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6*, 2007, pp. 873–880.

[33] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2011.

[34] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. on Acousti., Speech and Signal Process.*, vol. 37, no. 11, pp. 1641–1648, Nov 1989.