The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

Neural Networks-based Automatic Speech Recognition for Agricultural Commodity in Gujarati Language

Hardik B. Sailor and Hemant A. Patil

Speech Research Lab, Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT), Gandhinagar-382007, Gujarat, India

{sailor_hardik, hemant_patil}@daiict.ac.in

Abstract

In this paper, we present a development of Automatic Speech Recognition (ASR) system as a part of a speech-based access for an agricultural commodity in the Gujarati (a low resource) language. We proposed to use neural networks for language modeling, acoustic modeling, and feature learning from the raw speech signals. The speech database of agricultural commodities was collected from the farmers belonging to various villages of Gujarat state (India). The database has various dialectal variations and real noisy acoustic environments. Acoustic modeling is performed using Time Delay Neural Networks (TDNN). The auditory feature representation is learned using Convolutional Restricted Boltzmann Machine (ConvRBM) and Teager Energy Operator (TEO). The language model (LM) rescoring is performed using Recurrent Neural Networks (RNN). RNNLM rescoring provides an absolute reduction of 0.69-1.18 in % WER for all the feature sets compared to the bi-gram LM. The system combination of ConvRBM and Mel filterbank further improved the performance of ASR compared to the baseline TDNN with Mel filterbank features (5.4 % relative reduction in WER). The statistical significance of proposed approach is justified using a bootstrap-based % Probability of Improvement (POI) measure.

Index Terms: Agricultural commodities, Gujarati language, neural networks.

1. Introduction

Gujarat is one of the major agricultural crops provider states that increase the significant grossing in India's agricultural progress. Government of India is maintaining several websites to provide prices of agricultural commodities [1]. However, education-level and socio-economic backgrounds may create a hindrance to access the agricultural commodity prices, weather forecast, and various schemes for the farmers. Gujarati is one of the official Indian languages which is still in the low resource category. There are no standard available audio, transcription, and dictionary in Gujarati to build state-of-the-art ASR system compared to other high-resourced Indian language, such as Hindi.

To provide agricultural information to the farmers, we are developing the speech-based access system in Gujarati called as Mandi Information System (MIS). In this paper, mandi refers to the marketplace in India where farmers can sell their commodities. The project is based on using the telephone-based Interactive Voice Response System (IVRS) and Automatic Speech Recognition (ASR) systems. The farmers need to call a toll-free number to get information regarding the prices of commodities and the weather. This telephone-based system is helpful to all the farmers even with those who have variable educational-levels, since they can get information in their native regional language just by making a telephone call. In phase-I of the MeitY, Govt. of India funded consortium project, MIS was built for six Indian languages [2–7].

Recently, there is a huge surge of using neural networks for speech signal processing applications including the ASR task [8]. In particular, the earlier approaches use deep neural networks (DNN) for acoustic modeling and statistical N-gram models for language modeling (LM). Recently, many studies shows that the recurrent neural networks (RNN) perform significantly well when used along with n-gram LM in a form of rescoring [9], [10]. The detailed survey of RNN-based language models are presented in [11]. To model temporal dynamics in speech, Time Delay Neural Networks and Long Short-Term Memory (LSTM) are very effective in the ASR task [12], [13].

The objective of this paper is to build better acoustic and language models for ASR in agricultural-domain in the Gujarati language using neural networks. In our initial study, we used neural networks for only acoustic modeling and Convolutional Restricted Boltzmann Machine (ConvRBM) for filterbank learning on 400 farmers database [14]. In this paper, we used an entire agricultural database from 1005 farmers recorded from various villages belonging to Gujarat state. We proposed to use neural networks for language modeling, feature learning, and acoustic modeling. TDNN and LSTM are explored for the acoustic modeling. RNNLM are used along with bi-gram LM in a form of rescoring [9], [10]. The ConvRBM is used to learn filterbank directly from the raw speech signals [15, 16]. To the best of authors' knowledge, this is the first of its kind ASR system proposed for developing speech-based access system for an agricultural commodity in Gujarati.

2. Speech-Based Access for Agricultural Commodity

2.1. System Architecture

There are three major building blocks of the system, namely, information source, IVRS, and ASR system as shown in Figure 1. Everyday information regarding the agricultural commodities is fetched from the AGMARKNET webportal maintained by the Government of India [17]. The information regarding the weather forecast is fetched from the IMD website maintained by Ministry of Earth Science, Government of India [18]. Our local agricultural database is updated based on webcrawler program that automatically update the AGMARKNET and IMD data related to the Gujarat state. An IVRS is used to record the speech signals from the farmers via a telephone line (called as Primary Rate Interface (PRI) line). Based on the information from the database, the response is given to the farmer for a query recognized by the ASR system. The ASR system is one of the major components in a speech-based access system for an agricultural

10.21437/SLTU.2018-34

commodity. Since it is this component that identifies the required query, which is then passed to the IVRS. The success of the response to farmer's call is based on the accuracy of the ASR system.

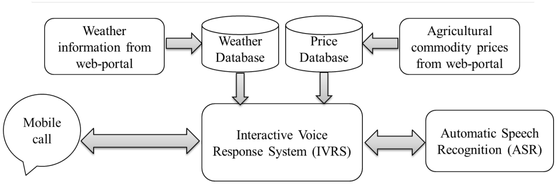
Weather information from web-portal
Weather Database
Price Database
Agricultural commodity prices from web-portal
Mobile call
Interactive Voice Response System (IVRS)
Automatic Speech Recognition (ASR)

Figure 1: *Functional block diagram of a speech-based access system for an agricultural commodity. After [4].*

2.2. Data Collection and Transcription

Speech data is collected with the help of mobile phone via a toll free number. The recording of speech signals is based on the Asterisk server configuration and prompts are stored in the server. The database includes the names of agricultural crops, mandi, weather information, and yes/no type of questions. The data collection includes the natural speaking styles, and dialects of the farmers with real environmental noises, such as vehicles, animals, babble, etc. The dataset also includes channel mismatch conditions since the recording was done with several mobiles of different companies. The data collection from 1005 farmers has been completed covering 21 districts of Gujarat state. There are a number of issues in data collection, such as (1) explaining the farmers about the task since they are hesitant to talk and response to the IVRS, (2) disfluencies in speech since many farmers are not much habituated to such mobile-based field recordings.

We prepared the dictionary (along with transcription) containing the names of crops, mandis and districts. Indian Language Speech sound Label set (ILSL) format has been used for transcription and dictionary preparation for Gujarati language [19]. The dictionary contains different varieties of commodities, mandi, names of villages, and districts. There are 25 districts, 328 markets, and 159 unique commodities (excluding variations) in the lexicon. The lexicon contains 5387 words including varieties in commodities, speaking market names and yes/no utterances spoken in various dialectal manner from the farmers. The example of dialectal variations in speaking commodity names are shown in Table 1. The speech signals were transcribed using semi-supervised transcription tool, namely, "Indic Language Transliteration Tool", provided by the MeitY ASR Phase-II consortium. The training and test set includes 37500 and 2989 utterances, respectively such that different speakers are used in both the sets.

Table 1: *An example of a word batxaakaa (potato in English) presented in the dictionary with different pronunciations*

Word	Phonetic Description (Spoken Form)
batxaakaa	b a t aa k aa
batxaakaa	b a tx aa k aa
batxaakaa	b a tx aa k u
batxaakaa	b a tx aa tx aa
batxaakaa	b a tx ae k aa

3. Neural Networks for Acoustic and Language Modeling

In this Section, we discuss for acoustic modeling using a hybrid DNN-HMM setup and RNN for language modeling.

3.1. Deep Neural Networks for Acoustic Modeling

In this paper, we consider to use DNNs that can model temporal context to model the temporal dynamics in the speech signal. Two such architectures are Long-Short Term Memory (LSTM)-based recurrent neural networks (RNN) [13] and Time-Delay Neural Networks (TDNN) [12]. The LSTM model is based on introducing self-loops to produce the paths, where the gradient can flow for longer durations. Using the gate controlled by the hidden unit, the time scale of integration can be changed dynamically [20]. Another DNN architecture which has been shown to be effective in modeling long range temporal dependencies is the time delay neural network (TDNN) recently proposed in [12]. In TDNN, initial layers learn representations using narrow context while higher layers learn representations using the wider context [12]. TDNN are one of the best performing system tested in the KALDI speech recognition toolkit for various ASR databases. We also tried recently proposed TDNN-LSTM system for acoustic modeling to get advantages of TDNN and LSTM models [21].

3.2. Recurrent Neural Networks for Language Modeling

Recurrent Neural Networks Language Models (RNNLM) represent the full history, $h_i = \{w_{i-1}, ..., w_1\}$ of a word w_i. The architecture of an RNNLM consists of three layers. The full history vector h_i is given to the input layer in the form of previous word w_{i-1} and vector v_{i-2} for remaining context. The hidden layer apply an activation function on the input. We used recently popular Gated Recurrent Unit (GRU) as activation function in the RNNLM. The output layer calculates the normalized RNNLM probabilities $P_{\text{RNNLM}}(w_i|w_{i-1}, v_{i-2})$ using a softmax layer. The information is also fed back into the input layer as the future remaining history to compute the probability of the next word, w_{i+1}. RNNLM can be trained using stochastic gradient-based back-propagation through time (BPTT) algorithm with cross-entropy (CE) objective function. In our study, the faster RNNLM training is employed using noise contrastive estimation (NCE) [22]. In state-of-the-art ASR systems, RNNLM are often interpolated with N-gram LMs to obtain better coverage of context and strong generalization. The LM probability using a linear interpolation is given by [22]:

$$P(w_i|h_i) = \lambda P_{\mathbf{nG}}(w_i|h_i) + (1-\lambda)P_{\text{RNNLM}}(w_i|h_i), \quad (1)$$

where λ is the weight given to the N-gram LM $P_{\mathbf{nG}}(\cdot)$.

4. Auditory Filterbank Learning

ConvRBM is an undirected probabilistic graphical model used for representation learning. It has two layers, namely, a visible layer and a hidden layer [15]. The input to a visible layer (denoted as $\mathbf{x}$) is a speech signal of length of *n*-samples. The hidden layer (denoted as $\mathbf{h}$) is divided into *K*-groups (i.e., number of subband filters). Each group weight has filter length of *m*-samples. The weights of ConvRBM are shared between the visible and hidden units in each group [15]. Denoting b_k as the hidden bias for the k^{th} group, the convolutional response for the k^{th} group is given as:

$$\mathbf{I}_k = (\mathbf{x} * \tilde{\mathbf{W}}^k) + b_k, \quad (2)$$

where $\mathbf{x} = [x_1, x_2, ..., x_n]$ are samples of the speech signal, $\mathbf{W}^k = [w_1^k, w_2^k, ..., w_m^k]$ is a k^{th} weight vector (i.e., subband filter) and $\tilde{\mathbf{W}}$ indicates a *flipped* array [15]. With a noisy leaky rectifier linear units (NLReLU), the sampling equations for hidden and visible units are given as:

$$\begin{gathered} \mathbf{h}^k \sim max(0, \mathbf{J}_k) + \alpha \cdot min(0, \mathbf{J}_k), \\ \mathbf{x}_{recon} \sim \mathcal{N}\left(\sum_{k=1}^{K}(\mathbf{h}^k * \mathbf{W}^k) + c, 1\right), \end{gathered} \tag{3}$$

where $\mathbf{J}_k = \mathbf{I}_k + N(0, \sigma(\mathbf{I}_k))$ with $N(0, \sigma(\mathbf{I}_k))$ is a Gaussian noise with mean zero and sigmoid of $\mathbf{I}_k$ as a variance. α is a leaky parameter which is generally set to 0.01 and c is a visible bias which is also shared among the visible units. In ConvRBM training, a dropout is applied before sampling the hidden units in both positive and negative phase of contrastive divergence (CD) learning [23]. Our earlier works showed that use of annealed dropout [24] when applied in ConvRBM resulted in an improved performance in speech recognition [16] and audio classification [25]. In an annealed dropout, the dropout probability of the units in the network is gradually decreased over the training period. The model parameters are updated using the Adam optimization method [26].

After ConvRBM is trained, the pooling (similar to windowing operation) is applied on the ConvRBM filter responses. During feature extraction stage, we have used absolute nonlinearity $|\mathbf{I}_k|$ as an activation function of the hidden units. The pooling operation reduces the temporal resolution from $K \times n$ samples to the $K \times F$ frames. Logarithmic nonlinearity compresses the dynamic range of features. In order to achieve robustness in the ASR task, we also used Teager Energy Operator (TEO) [27]. We have used TEO-based ConvRBM features proposed in [16]. The notations for ConvRBM and TEO-based ConvRBM filterbanks are CBANK and TEO-CBANK, respectively.

5. Experimental Setup

5.1. GMM-HMM System Building

The triphone GMM-HMM system was built from the 39-D Mel Frequency Cepstral Coefficients (MFCC) feature set by varying the number of Guassians and senones. MFCC features are obtained from 40 Mel subband filters. We have used the finite state transducer (FST)-based bi-gram LM trained from the agricultural commodity text data. Bi-gram is used for LM since the text contains maximum number of one-two words per line and few lines containing more than two words (e.g., names of commodities and answers of some basic questions while recording). The system is further improved using the linear discriminant analysis (LDA) and Maximum Likelihood Linear Transform (MLLT) on the triphone GMM-HMM system. The force-aligned labels are generated from the LDA+MLLT system to use as labels in the DNN training.

5.2. Training of ConvRBM and Feature Extraction

The ConvRBM is trained with an annealed dropout using dropout probability p=0.3 that decayed to zero (i.e., $p = 0$) during training. The learning rate schedule was chosen as suggested in [26] with initial value 0.001. The moment parameters in Adam optimization algorithm were chosen to be β_1=0.9 and β_2=0.999. The model is trained with 40 number of subband filters with a convolution window of length 8 ms. After the model was trained, the features were extracted from the speech signals. The Discrete Cosine Transform (DCT) was applied retaining only first 13-D coefficients and used the proposed features along with the MFCC feature set in the GMM framework. The delta and delta-delta features were also appended resulting in 39-D cepstral features (denoted as ConvRBM-CC). For DNN training, ConvRBM filterbank (CBANK) and TEO-based CBANK (TEO-CBANK) is directly used (without applying DCT) for DNN training. The system combination is performed using Minimum Bayes Risk (MBR) technique [28].

5.3. Training of RNNLM and DNN

The RNNLM is trained using the training corpus of the agricultural commodities in Gujarati. The number of hidden neurons and noise samples are selected based on performance on the ASR task. The RNNLM are trained using the faster-rnnlm toolkit [29]. The weights λ in the Eq. (1) are chosen to be 0.25, 0.5 and 0.75 for LM rescoring. We have explored recently proposed the Lattice-free Maximum Mutual Information (LF-MMI) in the HMM framework for sequence-discriminative learning [30]. The senones labels from the LDA-MLLT system is used to build the LF-MMI model that is later used in the hybrid DNN-HMM training in the KALDI toolkit [31]. The DNNs are trained using 120-D Mel filterbank features that include delta and delta-delta features (denoted as FBANK). CBANK and TEO-CBANK feature sets (120-D) are also used for DNN training. We trained TDNN, LSTM and TDNN-LSTM with different hidden units and layers. The learning rate for RNNLM was chosen to be 0.01 and for DNN models to be 0.001.

5.4. Performance Evaluation

The performance of ASR system is evaluated using % Word Error Rate (WER). The statistical significance of the one ASR system performing better than the other was assessed using the bootstrap technique [32]. Using bootstrap technique, we compute % Probability of Improvement (POI) for proposed approaches over the baseline system.

6. Experimental Results

6.1. Results using GMM-HMM System

Since there is no standard recipe for this ASR in Gujarati task, we varied the number of Gaussians and senones for GMM-HMM systems. The results are summarized in Table 2 for MFCC and ConvRBM-CC feature sets. The ConvRBM-CC performs better than the MFCC using the CD-GMM-HMM system with an absolute reduction of 1.09 % in WER. The ConvRBM-CC also significantly performs better than the MFCC using the LDA-MLLT system with an absolute reduction of 1.53 % in WER. Hence, ConvRBM-CC improved the ASR performance compared to the baseline MFCC feature set using the CD-GMM-HMM and LDA-MLLT systems. The alignments generated from the respective LDA-MLLT systems are used in the hybrid DNN-HMM systems.

6.2. Results of DNN Models and RNNLM

The experimental results of using LSTM and TDNN models are reported in Table 3. For initial experiments, we used the parameters of LSTM and TDNN models suggested in KALDI. The TDNN significantly performed better compared to the LSTM with 4.69 % gap in WER. We also tried using TDNN-LSTM architecture to see if it can improve upon TDNN. However, it did not perform well compared to the TDNN. Hence, we then

Table 2: *The summary of results using GMM-HMM systems in % WER for 1005 speakers*

Feature Set	Acoustic Model	Test
MFCC	Triphone (CD-GMM-HMM)	30.36
ConvRBM-CC	Triphone (CD-GMM-HMM)	29.27
MFCC	LDA-MLLT	26.98
ConvRBM-CC	LDA-MLLT	**25.45**

tried various parameters of a TDNN, such as number of hidden units and layers to further improve the performance. The lowest WER is achieved using TDNN with 800 hidden units and 6 hidden layers. We fix this architecture for further experiments using FBANK feature set.

Table 3: *The results of experiments on the test set in % WER performed using various DNN models with FBANK*

Model	**Specification**	**% WER**
LSTM	900×3	20.97
TDNN	900×6	16.28
TDNN-LSTM	800×(5 TDNN-3 LSTM)	19.27
TDNN	800×6	**15.82**
TDNN	700×6	15.98
TDNN	1000×6	16.56
TDNN	1500×6	17.33
TDNN	800×8	16.28
TDNN	800×4	16.35

The results of applying RNNLM rescoring are shown in Table 4 in terms of % WER. We have also shown the statistical significance of RNNLM compared to the 3-gram in terms of % POI. It can be seen that % WER is reduced compared to the bi-gram LM for various RNN configurations. RNN with 512 hidden units performed well with an absolute reduction of 1.18 % in WER and it has 99.98 % POI compared to the bi-gram LM. We also tried to use different noise samples in NCE training, however, 10 noise samples (k_{noise}=10) performed well. Hence, we used RNNLM with 512 hidden units and k_{noise}=10 for further experiments in LM rescoring.

Table 4: *The effect of parameters of RNNLM rescoring using TDNN trained on FBANK feature set. The results on the test set are presented in terms of % WER and % POI*

Language Modeling	**WER**	**POI**
bi-gram	15.82	-
RNNLM, units=128, k_{noise}=10	15.07	96.33
RNNLM, units=256, k_{noise}=10	14.75	99.87
RNNLM, units=512, k_{noise}=10	**14.64**	**99.98**
RNNLM, units=700, k_{noise}=10	14.71	99.93
RNNLM, units=800, k_{noise}=10	14.71	99.86
RNNLM, units=512, k_{noise}=5	14.69	99.92
RNNLM, units=512, k_{noise}=20	15.17	93.89

6.3. Results of Auditory Filterbank Learning

The experiments of using ConvRBM for filterbank learning and TEO-based features are shown in Table 5 for 3-gram LM. The CBANK and TEO-CBANK performed better with 1000 hidden units and 6 hidden layers in TDNN. Both the feature sets, CBANK and TEO-CBANK perform better than FBANK. In particular, CBANK gave the lowest % WER with 1.45 % relative reduction in WER and 71.98 % POI compared to the FBANK. The results of RNNLM rescoring using CBANK and TEO-CBANK feature sets are shown in Table 6. RNNLM rescoring significantly reduce % WER for FBANK compared to CBANK and TEO-CBANK. To get the possible complementary information from different feature sets, the system combination experiments are shown in Table 6. The combination of CBANK and TEO-CBANK with FBANK (S1⊕S2 and S1⊕S3, respectively) reduces the % WER compared to the FBANK alone. In particular, S1⊕S3 much reduce % WER compared to the S1⊕S2 with 98.19 % POI compared to the FBANK. The best results were achieved using system combination of all the three feature sets with 5.4 % relative reduction in WER. The proposed approach is statistically significant with 99.60 % POI compared to the FBANK alone.

Table 5: *The results of various feature sets using TDNN systems and 3-gram LM on the test set in terms of % WER and % POI*

Feature Set	**TDNN Spec.**	**% WER**	**% POI**
FBANK	800×6	15.82	-
CBANK	1000×6	15.59	71.98
TEO-CBANK	1000×6	15.76	59.69

Table 6: *The experimental results using various feature sets after RNNLM rescoring and their system combination in terms of % WER and % POI*

Feature Set	**% WER**	**% POI**
S1:FBANK	14.64	-
S2:ConvRBM-FBANK	14.90	25.91
S3:TEO-ConvRBM-FBANK	14.69	45.26
S1⊕S2	14.27	89.07
S1⊕S3	13.96	98.19
S1⊕S2⊕S3	**13.85**	**99.60**

7. Summary and Conclusions

In this study, the development of an ASR system in speech-based access for an agricultural commodity in low resource Gujarati language is presented using the neural networks. We used neural networks for three tasks: feature learning, language modeling, and acoustic modeling. The ConvRBM is used to learn an auditory-like filterbank from the speech signals. RNNLM rescoring significantly reduce % WER compared to bi-gram LM. TDNN performed better than LSTM and TDNN-LSTM for acoustic modeling. All the proposed approaches improved the performance of the ASR system. We also verified the statistical significance of the proposed approaches. Our future work include use of the neural networks in end-to-end deep learning framework [33].

8. Acknowledgments

Authors would like to thank the MeitY, Govt. of India for sponsored project ASR Phase-II. We also acknowledge NVIDIA for the hardware grant of TITAN X GPU to the Speech Research Lab, DA-IICT.

9. References

[1] AGRICOOP, "Ministry of Agriculture and Farmers Welfare, Government of India," URL: http://agricoop.nic.in, {Last Accessed on 19 June 2018}.

[2] A. Mohan, R. Rose, S. H. Ghalehjegh, and S. Umesh, "Acoustic modelling for speech recognition in Indian languages in an agricultural commodities task domain," *Speech Communication*, vol. 56, pp. 167–180, 2014.

[3] G. Mantena, S. Rajendran, B. Rambabu, S. Gangshetty, S. Yegnanarayana, and K. Prahallad, "A speech-based conversation system for accessing agricultural commodity prices in Indian languages," in *IEEE Joint Workshop on Hand-free Speech Communication and Microphone Arrays (HSCMA)*, India, 2011, pp. 153–154.

[4] J. Basu, S. Khan, R. Roy, and M. S. Bepari, "Commodity price retrieval system in Bangla: An IVR-based application," in *Proceedings of the Asia-Pacific Conference on Computer Human Interaction*, Bangalore, India, 2013, pp. 406–415.

[5] S. Shahnawazuddin, D. Thatoppa, B. D. Sarma, A. Deka, S. R. M. Prasanna, and R. Sinha, "Assamese spoken query system to access the price of agricultural commodities," in *National Conference on Communication (NCC)*, India, March 2013, pp. 1–5.

[6] T. Godambe and K. Samudravijaya, "Speech data acquisition for voice-based agricultural information retrieval," in *Proc. of 39th All India Dravidian Linguistics Association (DLA) Conference*, Patiala, India, 2011, pp. 1–5.

[7] T. G. Yadava and H. S. Jayanna, "A spoken query system for the agricultural commodity prices and weather information access in Kannada language," *International Journal of Speech Technology (IJST), Springer*, vol. 20, no. 3, pp. 635–644, Sept. 2017.

[8] G. Hinton, L. Deng *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[9] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," *Journal of Machine Learning Research (JMLR)*, vol. 3, pp. 1137–1155, March 2003.

[10] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *INTERSPEECH, Makuhari, Chiba, Japan*, 2010, pp. 1045–1048.

[11] W. D. Mulder, S. Bethard, and M.-F. Moens, "A survey on the application of recurrent neural networks to statistical language modeling," *Computer Speech & Language*, vol. 30, no. 1, pp. 61 – 98, 2015.

[12] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH, Dresden Germany*, 2015, pp. 2440–2444.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[14] H. B. Sailor, H. A. Patil, and A. Rajpal, "Unsupervised filterbank learning for speech-based access system for agricultural commodity," in *IEEE International Conference on Advances in Pattern Recognition (ICAPR), ISI Kolkata, India*, 2017, pp. 1–6.

[15] H. B. Sailor and H. A. Patil, "Novel unsupervised auditory filterbank learning using convolutional RBM for speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 12, pp. 2341–2353, Dec. 2016.

[16] H. B. Sailor and H. A. Patil, "Auditory feature representation using convolutional restricted Boltzmann machine and Teager energy operator for speech recognition," *Journal of Acoustical Society of America Express Letters (JASA-EL)*, vol. 141, no. 6, pp. EL500–EL506, June. 2017.

[17] AGMARKNET, "Ministry of agriculture and farmers welfare, government of India," URL: http://agmarknet.dac.gov.in/, {Last Accessed on 19 June 2018}.

[18] IMD, "India Meteorological Department (IMD), Ministry of Earth Sciences, Government of India," URL: http://www.imd.gov.in/pages/main.php, {Last Accessed: 19 June 2018}.

[19] K. Samudravijaya and H. A. Murthy, "Indian Language Speech sound Label set(ILSL12), version v2.1.3," *Indian Language TTS Consortium and ASR Consortium*, pp. 1–14, 2013.

[20] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. The MIT Press, First Edition, 2016.

[21] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 373–377, 2017.

[22] X. Chen, X. Liu, M. J. Gales, and P. C. Woodland, "Recurrent neural network language model training with noise contrastive estimation for speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Queensland, Australia*, 2015, pp. 5411–5415.

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

[24] S. J. Rennie, V. Goel, and S. Thomas, "Annealed dropout training of deep networks," in *IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, California and Nevada*, 2014, pp. 159–164.

[25] H. B. Sailor, D. M. Agrawal, and H. A. Patil, "Unsupervised filterbank learning using convolutional restricted Boltzmann machine for environmental sound classification," in *INTERSPEECH*, Stockholm, Sweden, 2017, pp. 3107–3111.

[26] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR), San Diego*, 2015, pp. 1–11.

[27] J. F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Albuquerque, New Mexico, USA, 1990, pp. 381–384.

[28] 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.

[29] Faster RNNLM, "Faster RNNLM (HS/NCE) toolkit," URL: https://github.com/yandex/faster-rnnlm, {Last Accessed: 18 March 2018}.

[30] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *INTERSPEECH 2016, San Francisco, CA, USA*, 8-12 September 2016, pp. 2751–2755.

[31] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Big Island, Hawaii, USA*, Dec. 2011, pp. 1–4.

[32] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Montreal, Que., Canada*, vol. 1, 17–21 May 2004, pp. I–409–12 vol.1.

[33] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *Proceedings of the International Conference on Machine Learning (ICML), Beijing, China*, 2014, pp. 1764–1772.