International Journal of Engineering & Technology, 7 (4.6) (2018) 133-137

SPC

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Speech Recognition Using Convolutional Neural Networks

D. Nagajyothi [1] , P. Siddaiah [2]

[1]Associate Professor of Electronics and Communication Engineering, Vardhaman College of Engineering, Shamshabad, Telangana, India.e-mail: nagajyothi1998@gmail.com
[2]Dean of Department of Electronics and Communication Engineering, University College of Engineering and Technology, Acharya Nagarjuna University, Guntur, India.
**Corresponding author E-mail: Siddaiah_p@yahoo.com*

Abstract

Automatic speech recognition (ASR) is the process of converting the vocal speech signals into text using transcripts. In the present era of computer revolution, the ASR plays a major role in enhancing the user experience, in a natural way, while communicating with the machines. It rules out the use of traditional devices like keyboard and mouse, and the user can perform an endless array of applications like controlling of devices and interaction with customer care. In this paper, an ASR based Airport enquiry system is presented. The system has been developed natively for telugu language. The database is created based on the most frequently asked questions in an airport enquiry. Because of its high performance, Convolutional Neural Network (CNN) has been used for training and testing of the database. The salient feature of weight connectivity, local connectivity and polling result is a through training of the system, thus resulting in a superior testing performance. Experiments performed on wideband speech signals results in significant improvement in the performance of the system in comparison to the traditional techniques.

***Keywords**: Neural Networks (NN),Convolutional Neural Networks(CNN).*

1. Introduction

Speech recognition is a system that translates spoken utterances into text. Text can be either in terms of words or word sequences or it could be in syllables, or it can be any sub-word units or phones, or even characters, but you're translating speech into its corresponding text form. Some of the well known examples are YouTube's closed captioning it has an ASR engine running producing the corresponding transcripts for the speech, the audio and the video clips. The voice mail transcription also has an ASR engine running. The older prototypes of ASR systems are the dictation systems. The dictation systems are the words are spoken out and then the corresponding transcripts are produced. Siri, Cortana, Google Voice, all of their front ends are ASR engines.ASR is strictly just translating the spoken utterances into text. The development of a good ASR system is very desirable as it fetches a lot of advantages [1]. A person can save his time which is mostly used in typing, so rather than typing he can speak to the devices. There is another kind of socially desirable aspect of building a good ASR system. As we know that the present technology has interfaces, if an ASR system is built into the interfacing device, then it can it can be used both by literate and the illiterate users. So even the users who cannot read or write in a particular language, can interact with the technology, if it is voice driven. Lots of languages are close to extinction, so if the technologies are built for such languages then it could be contributed towards the preservation of such languages. Building of an ASR system is quite difficult as it faces several sources of variability. One of them is the style of speech. Just the style of speech can have a lot to do with the performance of an ASR system. Rather than continuous speech individually, isolated words are much easier for the ASR systems. This is because in continuous speech words are flowing freely into one another. There is a phenomenon called coarticulation, where the preceding words affect the words that are coming and so on. This phenomenon becomes challenging for the ASR systems to handle itv[2].The other source of variability is environment. If a person is talking under noisy conditions or if the acoustics of the room such as the echo produced in the room becomes challenging. Background noise could be of two types, if the noise is of vehicles then it can be isolated but if the noise is made by people talking behind then that would be hard for the ASR system to pick the foreground voice. Various characteristics of speakers are also a challenging problem, as the main constraints are grammar, vocabulary and the language which might not have a written form, age also changes the characteristics of the speech.

2. Convolutional Neural Network

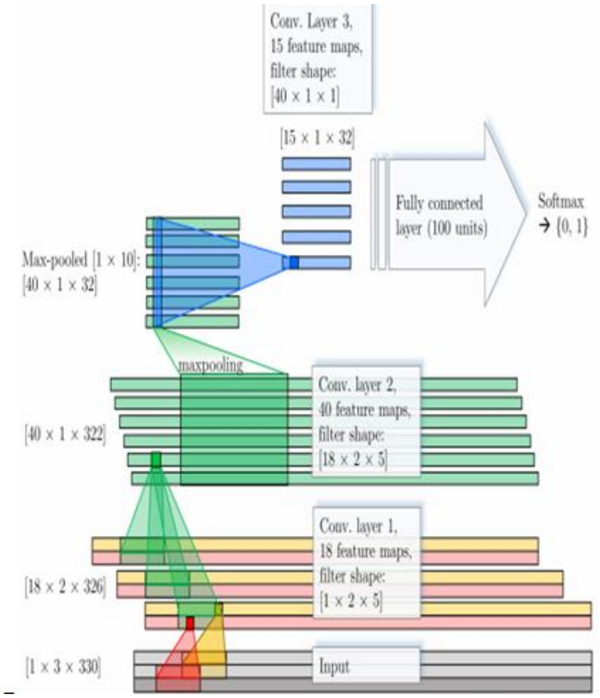


Fig.1: CNN structure

Convolutional Neural Network can do a lot of good things if they are fed with a bunch of signals for instance to learn some basic signals such as frequency changes, amplitude changes. Since, they are multi neural networks, the first layer is fed with this information. The second layer is fed with some recognizable features. To illustrate this, a signal of two-dimensional array of pixels is considered. It is a check board with each square on the board is either light or dark colour. By observing the pattern CNN decides whether it is a signal with frequency change or amplitude change.

The convolutional neural network match the parts of the signal instead of considering the whole signal of pixels as it becomes difficult for a computer to identify the signal when the whole set of pixels are considered [8][9]. The mathematics behind matching these is filtering. The way this is done is by considering the feature that is lined up with this patch signal and then one by one pixels are compared and multiplied by each other and then add it up and divide it with the total number of pixels. This step is repeated for all the pixels that is considered. The act of convolving signals with a bunch of filters, a bunch of features which creates a stack of filtered images is called as convolutional layer. It is a layer because it is operating based on stack that is in convolution one signal becomes a stack of filtered signals. We get a lot of filtered signals because of the presence of the filters. Convolution layer is one part.

The next big part is called as pooling that is how a signal stack can be compressed. This is done by considering a small window pixel which might be a 2 by 2 window pixel or 3 by 3. On considering a 2 by 2 window pixel and pass it in strides across the filtered signals, from each window the maximum value is considered. This passed through the whole signal. At the end it is found that by considering only the maximum values the size of the filtered signal is reduced [10]. The third part is normalization, in this if a pixel value is negative then the negative values are replaced with zeros. This is done to all the filtered signals. This becomes another type of layer which is known as a rectified linear unit, a stack of signals which becomes a stack of signals with no negative values. Now the three layers are stacked up so that one output will become the input for the next. The final layer is the fully connected layer.

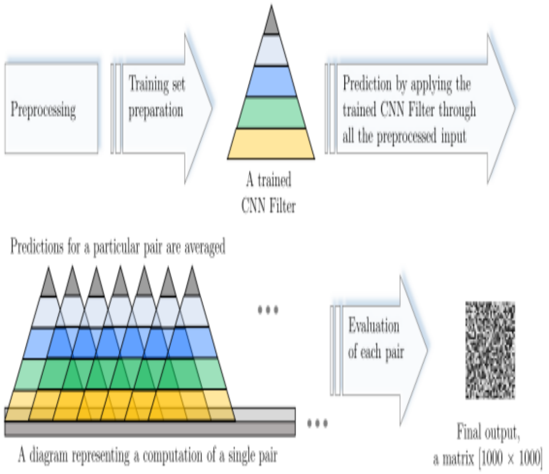

Preprocessing
Training set preparation
A trained CNN Filter
Prediction by applying the trained CNN Filter through all the preprocessed input
Predictions for a particular pair are averaged
Evaluation of each pair
A diagram representing a computation of a single pair
Final output,
a matrix [1000 × 1000]

Fig.2: CNN simplified workflow model.

The standard feed-forward fully connected Neural network (NN) is a computational model composed of several layers. An input to a particular unit is outputs of all the units in the previous layer (or input data for the first layer). The unit output is a single linear regression, to which output value a specific activation function is applied. Convolutional neural network (CNN) is a type of NN where the input variables are related spatially to each other [11]. To take into account very important spatial positions, CNNs were developed. Not only they are able to detect general spatial dependencies, but also are capable of specific patterns recognition. Shared weights, representing different patterns, improve the convergence by reducing significantly the number of parameters. CNN recognize small patterns at each layer, generalizing them (detecting higher order, more complex patterns) in subsequent layers. This allows detection of various patterns and keeps the number of weights to be learnt very low [12][13].

3. Two dimensional representation

The first step is to providing fragments in a two dimensional matrix. The filters of the first layer start doing recording for two dimensional values simultaneously [14]. The recorded values are then combined in the next, higher level layers. The CNN filter has different network behavior depending on its activity level. CNN was able to learn simple patterns and then predict more accurately by learning different combinations.

4. Selection of activation functions

One of the important decisions related to CNN was setting proper activation functions in the units. The most common activation function tanh is used in first Convolutional layer, while Rectified Linear Unit in the next two Convolutional layers. We do not want to allow an input from the previous layer [15].Because some patterns indicating a correlation.

5. Improvement of the CNN Filter

The first step is the input data was normalized to increase the learning speed. The second step is improve the network structure..The last step we used max pooling in the last Convolutional layer,. Since the errors are only propagated to the position of the maximally activated unit, it is highly probable that in a wider span there will exist a strong indication of a communication between cells [16].

6. Experimental Results

Input Data:
The input data for model is based on the airport enquiry system. The following questions are used as transcript.

Questions:

E : How do I book my flight?

T : Nenu vimana ticket ela book cheskogalanu / Nenu vimana ticket ela book cheskovali

నేను విమాన టికెట్ ఎలా బుక్చేసుకోగలను /నేను విమాన టికెట్ ఎలా బుక్చేసుకోవాలి

E : Can I make a group booking?

T : Nenu ekumandiki ela book cheskovali / Nenu ekumanidki ela book cheskovachu

నేను ఏకు మందికి ఎలా బుక్చేసుకోవాలి / నేను ఏకు మందికి ఎలా బుక్చేస్కోవచ్చు

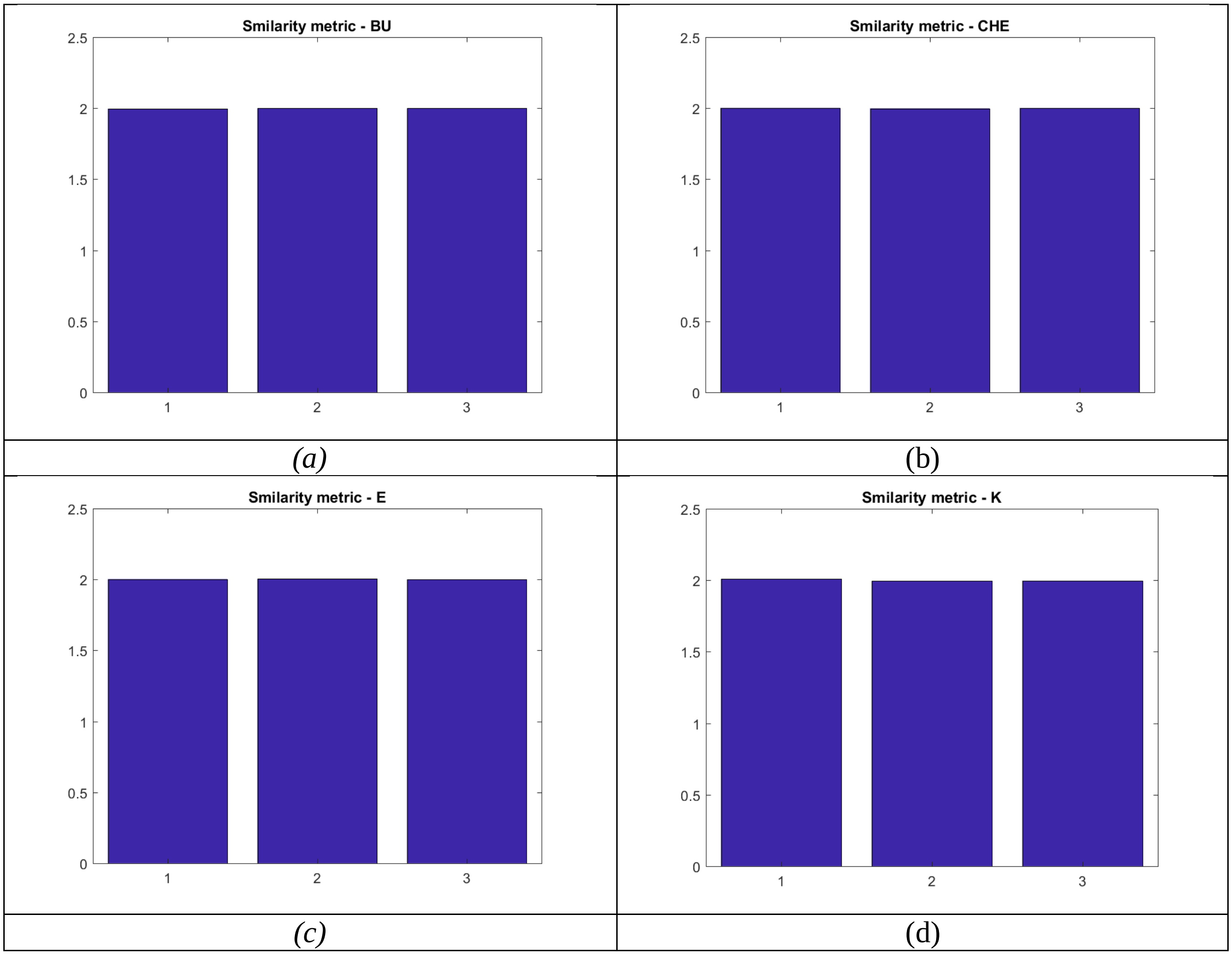
E : Can I book and hold a reservation and pay later?

T : Nenu ticket mundu book chesi dabbulu taravata katacha

నేను టికెట్ముందు బుక్చేసి డబ్బులు తరవాత కట్టచ్చా

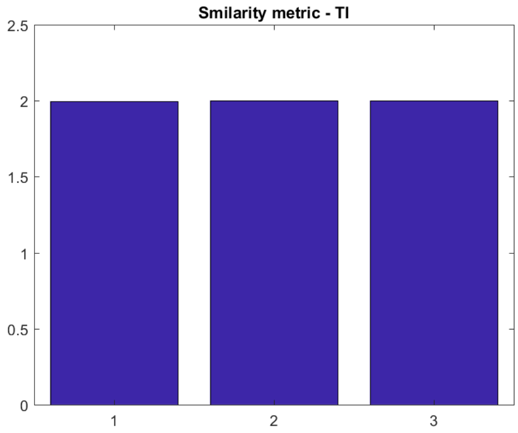
Etc

Total words trained for the process of recognition are 583. The total phone count resulted is 1765. The distinct phones are BU, CHE, E, GA, K, KE, KO, LA, LI, MA, NA, NE, NU, S, SKO, T, TI, VA, VI.

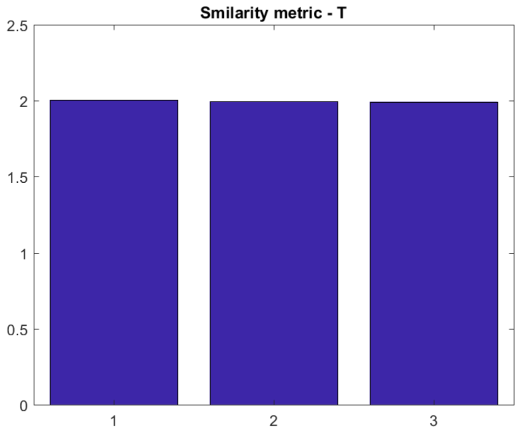


(a) (b)

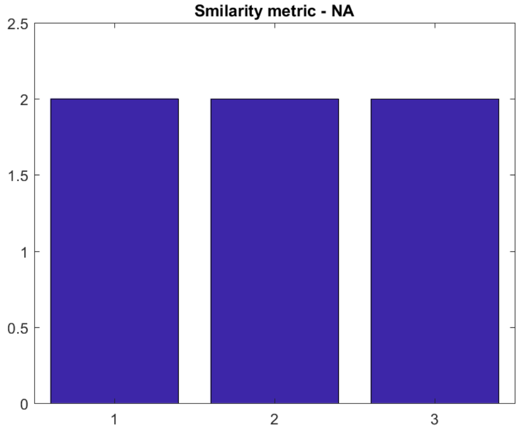
(c) (d)


Smilarity metric - TI
2.5
2
1.5
1
0.5
0
1
2
3

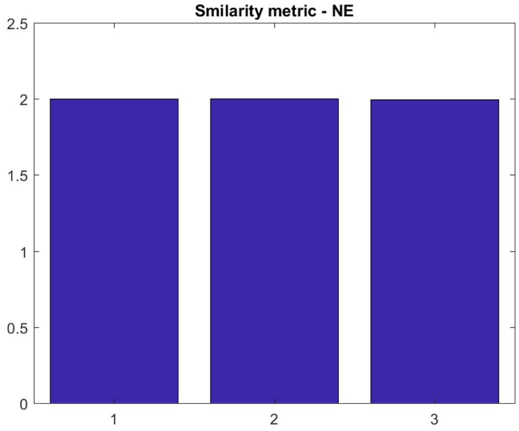
(e)


Smilarity metric - T
2.5
2
1.5
1
0.5
0
1
2
3

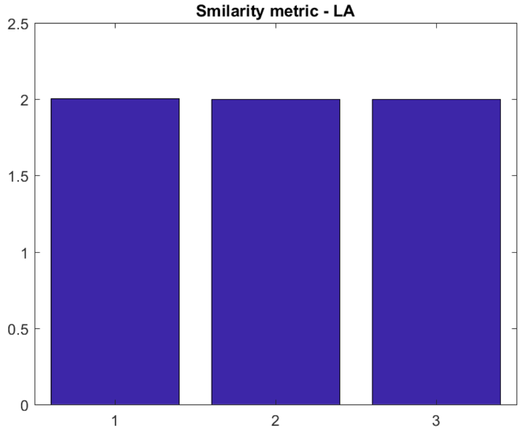
(f)


Smilarity metric - NA
2.5
2
1.5
1
0.5
0
1
2
3

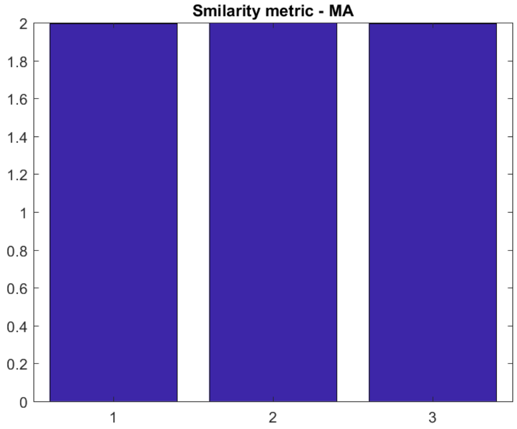
(h)


Smilarity metric - NE
2.5
2
1.5
1
0.5
0
1
2
3

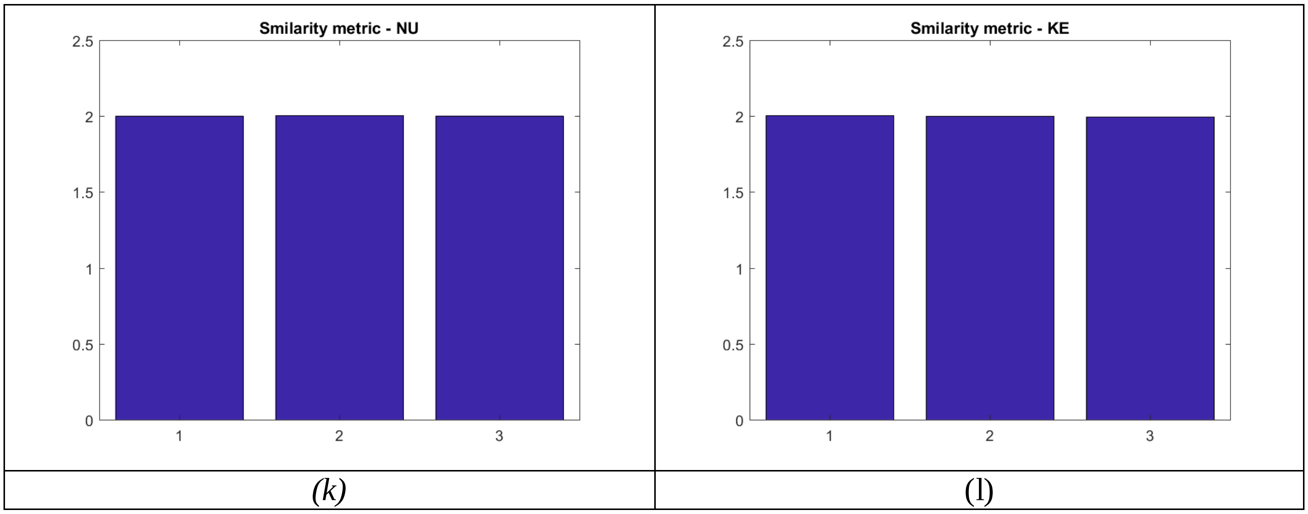
(g)


Smilarity metric - LA
2.5
2
1.5
1
0.5
0
1
2
3

(i)


Smilarity metric - MA
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
1
2
3

(j)



(k) (l)

Figure. (a) – (l) – Similarity matrices of phones 3 phones per sample.

7. Conclusion

In this paper, we investigated the Performance of an ASR based on CNNs, which takes raw speech signal, as input to large vocabulary task. Our analysis on wideband signals proved that the CNN based system is able to achieve good performance than the conventional Neural Network Techniques based system.

8. Acknowledgement

We sincerely thank Acharya Nagarjuna University (ANU) and Vardhaman College of Engineering (VCE) for allowing us to use resources to carry out this work.

9. References

[1] H. Jiang, “Discriminative training for automatic speech recognition: A survey,” *Comput. Speech, Lang.*, vol. 24, no. 4, pp. 589–608, 2010.

[2] X. He, L. Deng, and W. Chou, “Discriminative learning in sequential pattern recognition—A unifying review for optimization-oriented speech recognition,” *IEEE Signal Process. Mag.*, vol. 25, no. 5, pp. 14–36, Sep. 2008.

[3] L. Deng and X. Li, “Machine learning paradigms for speech recognition: An overview,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 5, pp. 1060–1089, May 2013.

[4] G. E. Dahl, M. Ranzato, A. Mohamed, and G. E. Hinton, “Phone recognitionwith the mean-covariance restricted Boltzmann machine,” *Adv.Neural Inf. Process. Syst.*, no. 23, 2010.

[5] A. Mohamed, T. Sainath, G. Dahl, B. Ramabhadran, G. Hinton, andM. Picheny, “Deep belief networks using discriminative features forphone recognition,” in *Proc. IEEE Int. Conf. Acoust., Speech, SignalProcess. (ICASSP)*, May 2011, pp. 5060–5063.

[6] D. Yu, L. Deng, and G. Dahl, “Roles of pre-training and fine-tuningin context-dependent DBN-HMMs for real-world speech recognition,”in *Proc. NIPS Workshop Deep Learn. Unsupervised Feature Learn.*,2010.

[7] G. Dahl, D. Yu, L. Deng, and A. Acero, “Large vocabulary continuousspeech recognition with context-dependent DBN-HMMs,” in*Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp.4688–4691.

[8] F. Seide, G. Li, X. Chen, and D. Yu, “Feature engineering in context-dependent deep neural networks for conversational speech transcription,”in *Proc. IEEE Workshop Autom. Speech Recognition Understand.(ASRU)*, 2011, pp. 24–29.

[9] N. Morgan, “Deep and wide: Multiple layers in automatic speechrecognition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no.1, pp. 7–13, Jan. 2012.

[10] A. Mohamed, G. Dahl, and G. Hinton, “Deep belief networks for phonerecognition,” in *Proc. NIPS Workshop Deep Learn. Speech Recognition Related Applicat.*, 2009.

[11] A. Mohamed, D. Yu, and L. Deng, “Investigation of full-sequencetraining of deep belief networks for speech recognition,” in *Proc.Interspeech*, 2010, pp. 2846–2849.

[12] L. Deng, D. Yu, and J. Platt, “Scalable stacking and learning forbuilding deep architectures,” in *Proc. IEEE Int. Conf. Acoustics,Speech, Signal Process.*, 2012, pp. 2133–2136.

[13] G. Dahl,D.Yu, L.Deng, and A. Acero, “Context-dependent pre-traineddeep neural networks for large-vocabulary speech recognition,” *IEEETrans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 30–42, Jan.2012.

[14] F. Seide, G. Li, and D. Yu, “Conversational speech transcription using context-dependent deep neural networks,” in *Proc. Interspeech*, 2011,pp. 437–440.

[15] T. N. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Novak,and A. Mohamed, “Making deep belief networks effective for largevocabulary continuous speech recognition,” in *IEEE Workshop Autom.Speech Recogn. Understand. (ASRU)*, 2011, pp. 30–35.

[16] J. Pan, C. Liu, Z. Wang, Y. Hu, and H. Jiang, “Investigation of deepneural networks (DNN) for large vocabulary continuous speech recognition:Why DNN surpasses GMMs in acoustic modeling,” in *Proc.ISCSLP*, 2012.