Journal of Engineering and Applied Sciences 13 (16): 6680-6685, 2018
ISSN: 1816-949X
© Medwell Journals, 2018

Very Deep Convolutional Neural Network for Speech Recognition Based on Words

[1]Javier O. Pinzon, [1]Robinson Jimenez-Moreno, [1]Oscar Aviles, [2]Paola Nino and [3]Diana Ovalle
[1]Faculty of Engineering, Nueva Granada Military University, Bogota, Colombia
[2]Instituto Politecnico Nacional, Mexico DF, Mexico
[3]Farcisco Jose de Caldas District University, Bogota, Colombia

Abstract: This study presents the implementation of two very deep convolutional neural network architectures applied to speech recognition based on the usage of complete words for this case 12 specific words in order to evaluate their performance in two types of environments, one semicontrolled and another non-controlled. One of the architectures developed is based on the use of linear filters only in frequency while the other consists of linear filters in both frequency and time. It is proposed to use the power spectral density with its first and second derivatives as input of the network in order to strengthen the variety of feature maps that can be used in neural networks for speech recognition. Finally, in the tests performed in real time, the architecture with filters of frequency and time reaches an error rate of 16.67% in a semicontrolled environment while the other architecture obtained a 41.67%. This means that the architecture with the lowest error rate has better performance for word recognition, even with small databases and specialized in a particular group of people.

Key words: Deep convolutional neural network, speech recognition, power spectral density, proposed, CNN architecture

INTRODUCTION

In recent years, neural networks of deep learning have taken force in the field of artificial sensory development such as machine vision (LeCun *et al*., 1998). In the emulation of this sense, the developments evolve from the segmentation of objects into images (Ondruska *et al*., 2016; Girshick, 2015 object recognition (Schmidhuber, 2015) and localization of elements, e.g., localization of pedestrians (Orozco *et al*., 2016) where more recently, these techniques of usage of neural networks have been applied to voice recognition.

The initial developments that made use of neural networks in speech recognition, began to implement different types of basic networks such as time-delay neural networks (Weibel *et al*., 1989). Due to the low processing capacity that computers had in the 90's, the neural networks did not have great depth, however, thanks to the progress in processing speed, it was possible to start deploying ever deeper neural networks, making them increasingly efficient in pattern recognition (Deng *et al*., 2013) for this reason the interest has arisen in the application of these in tasks of speech recognition, even above other recognition techniques.

The introduction of Deep Neural Networks (DNN) for speech tasks began in the early 2010's (Seide *et al*., 2011; Hinton *et al*., 2012) developing DNN in combination with the Hidden Markov Model (HMM), since, it allows modeling the sequential structure of a speech signal (Deng and Li, 2013; Mohamed *et al*., 2012). Consequently, due to the high performance that Convolutional Neural Networks (CNN) have had in pattern recognition (Krizhevsky *et al*., 2012) it have been begun to apply the concepts of convolution in DNN to speech recognition (Abdel-Hamid *et al*., 2012, obtaining hybrids between convolution and fully-connected layers. However, at first they did not exceed 3 convolution layers combined with a higher number of fully-connected layers which limits the ability to acquire patterns and makes the performance not really high. Recently, architectures began to be developed deeper and in combination with other types of neural networks such as the recurrent neural networks (Hsu *et al*., 2016) which help in the temporal relationship that may exist between signal divisions, improving the reduction of error in phonetic recognition.

The architectures that have been developed, according to the state of the art have as input feature maps of phonemes which are used in large vocabulary continuous speech recognition (Sainath *et al*., 2015), however for more basic applications, the use of phonemes makes them more complicated for its implementation, so, an alternative to reduce the complexity is the recognition of certain number of words delimited by the application, making the implementation of a CNN more feasible and achieve a better performance by the amount of parameters that have to learn. For this reason in this work two CNN

Corresponding Author: Javier O. Pinzon, Faculty of Engineering, Nueva Granada Military University, Bogota, Colombia

architectures based on the recognition of complete words are built which have not been developed in the state of the art as a complement to this research area.

CNN architecture

Conventional convolutional neural network architectures: For applications of object recognition in images, different CNN architectures have been developed, ranging from the most basic that comprise only a convolution layer and a layer of pooling, followed by a fully-connected (Abdel-Hamid *et al*., 2012) to very deep architectures, consisting of up to 19 convolution layers (Simonyan and Zisserman, 2014) but these latter are mainly used for the recognition of images, depending on the robustness and the amount of features that are thought to have taking into account the complexity of the elements to recognize. However, for speech recognition applications it is not "So easy" to create an architecture compared to architectures created for images, since in the images it is possible to have an idea of the dimension of the object to recognize and based on this, start with filter kernels that allow to identify general characteristics of the objects in speech recognition the patterns that are to be recognized are not so obvious, since in this it depends on many factors, e.g., the types of feature maps that are used in the entry and their structure or size. Here, are some considerations for designing a CNN for speech recognition applications.

Configuration: Currently, some structures have been implemented for speech recognition. A typical convolutional neural network for speech tasks is based on the developments made by Abdel-Hamid *et al*. (2014) where it uses a convolution/pooling/fully-connected architecture achieving 6-10% error reduction, compared to DNNs. An improvement of this architecture is exposed by Sainath *et al*. (2015) replacing fully-connected layers by additional convolutions, obtaining deeper architectures for speech tasks, obtaining even an additional reduction of 2%. Additionally, the configurations of each of the layers tend to be 9×9 and 4×3 in the first and second convolutions and a pooling layer over time of size 1×3 (Sainath *et al*., 2015), since, this configuration has demonstrated a high performance. However, Qian and Woodland (2016), they implement very deep architectures with configurations of 3×3, 3×1 and 1×3 in their convolutions and pooling in both frequency and time, 2×1 and 2×2 in order to optimize speech recognition in a noisy environment.

Padding: In contrast to CNN configurations made for images in speech recognition it is not very common for layers to contain padding (Qian and Woodland, 2016). However because the input sizes are really small, each input volume of the coming layers will be very small, making the networks not very deep in terms of the number of convolutions to be made, basically making the network become a fully-connected. Therefore including padding in the layers, allows to maintain the size of the output volume to be able to add more convolution layers and to improve the network performance (Yoshioka *et al*., 2016), since, it allows to better process the information of the borders of the feature maps.

Input: As input to the network, multi-scale features are used, emulating the structure of an image in terms of the RGB channels (Sercu *et al*., 2016) for which the Mel-Frequency Ceptral Coefficients (MFCC) with its delta and delta-delta features are the most used.

MATERIALS AND METHODS

Input dataset: The main function of the network to be implemented is the recognition of 12 words recorded with a group of 4 persons (which are not native English speaking) for which unlike the developments that have been made where the phonemes are used as input, the input to the neural network is a 2 sec long audio containing the word to recognize this to test the performance of a CNN trained with full words. The words are divided into 12 different categories which are 'blue', 'car', 'cat', 'chicken', 'dog', 'duck', 'green', 'house', 'red', 'sun', 'table' and 'yellow', recorded at a sampling rate of 16 kHz.

For the extraction of features, the audio was processed with frames of 20 msec for a total of 100 vectors which are generated with 161 coefficients that represent how much energy there is in each frame at a sample rate of 320 Hz per frame. These coefficients are a representation of the Power Spectral Density (PSD) estimation by means of the Welch's method which allows to reduce the noise of the signal used (Gupta and Mehra, 2013) and two maps of additional features which are its first and second derivatives (delta/delta-delta). An example of the feature maps described above is shown in Fig. 1.

Architecture implemented: For this work, 2 architectures were implemented to test their performance in speech recognition. In general, both architectures use linear filters in time and frequency but only maxpooling in frequency. The configurations of the architectures are described.

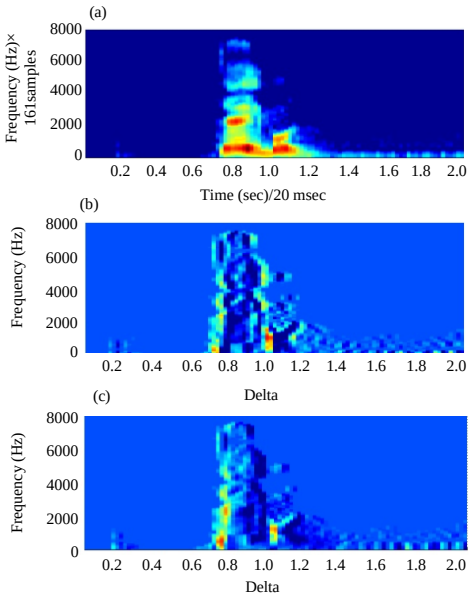

Fig. 1a-c): The PSD of the word 'Table' using the non-parametric Welch's method (top) with its respective first and second derivatives

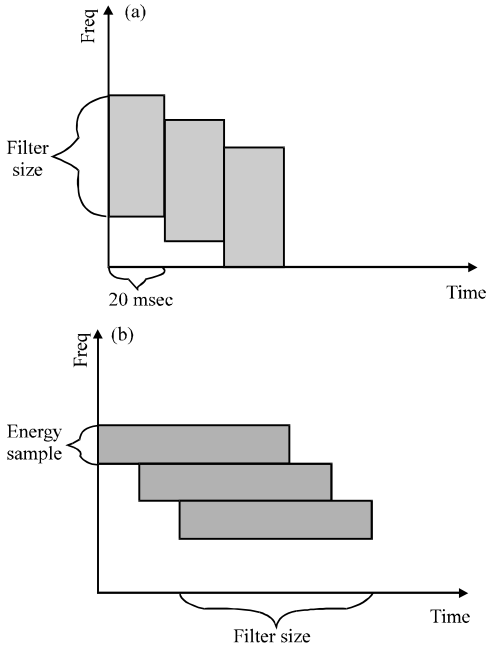

Fig. 2: a) Linear frequency filters and b) Linear time filter

Table 1: Architecture 1

Layer	Kernel		Filters
Input	----------161×100----------		-
Convolution	16×1	S = 1	32
Convolution	15×1	S = 1	32
Max pooling	2×1	S = 2×1	-
Convolution	10×1	S = 1	64
Convolution	10×1	S = 1	64
Max pooling	2×1	S = 2×1	-
Convolution	7×1	S = 1	256
Max pooling	2×1	S = 2×1	-
Fully-connected		1	256
Fully-connected		1	-
Softmax		12	-

Architecture 1: Its configuration consists of 2 convolution-convolution-maxpooling sets and a convolution-maxpooling set which use linear frequency filters (Fig. 2a) without padding with frequency downsampling and finally two fully-connected layers. The complete architecture is shown in Table 1 where S is the stride used.

Architecture 2: This is composed of 3 convolution-convolution-maxpooling sets where the first convolution uses linear frequency filters (Fig. 2a) and in the second, linear time filters (Fig. 2b). Each convolution has padding on the axis of time or frequency, depending on the filter to be used and only downsampling in frequency and finally, 2 fully-connected layers. Its configuration is shown in Table 2 where S is the Stride used and P the Padding applied.

Table 2: Architecture 2

Layer	Kernel		Filters
Input	----------161×100----------		-
Convolution	16×1	S = 1/P = 2×0	32
Convolution	1×9	S = 1/P = 0×2	32
Max pooling	2×1	S = 2×1	-
Convolution	8×1	S = 1/P = 1×0	64
Convolution	1×7	S = 1/P = 0×1	64
Max pooling	2×1	S = 2×1	-
Convolution	4×1	S = 1/P = 0	128
Convolution	1×7	S =1/P = 0	128
Max pooling	3×1	S = 3×1	-
Fully-connected		1	512
Fully-connected		1	-
Softmax		12	-

The idea of using linear filters is to obtain the characteristics in each axis (frequency and time) separately which allows a better characterization of the word. However, a problem that can be had when analyzing a complete segment of 2 sec lies in the computational cost, since, no downsampling is performed in the time axis its size varies from 86-100 parameters in the output volume of the last convolution as in the case of architecture 1 in which an input of approximately 230,000 parameters can be obtained for the fully-connected 1, causing the computational cost for the training to grow depending on the number of neurons that are to be used in said layer.

Architecture training: Each of the implemented architectures was trained with the input dataset that was set. To analyze the training behavior of each architecture

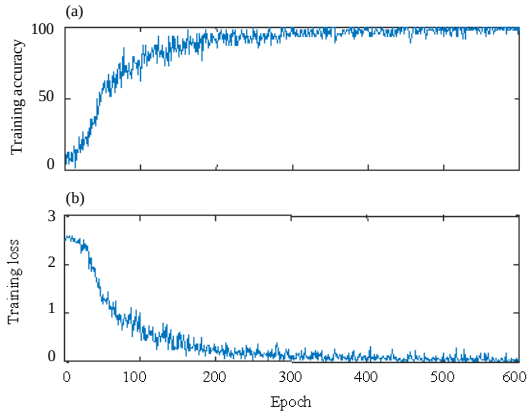

Fig. 3: Training results of the architecture 1

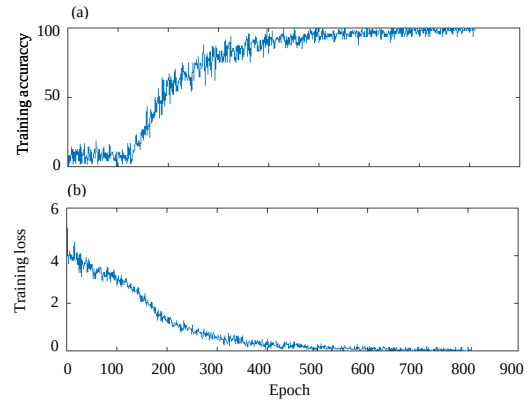

Fig. 4: Training results of the architecture 2

it is observed the learning speed and the reduction of cost per mistake or training loss. Figure 3 and 4 illustrate the resulting training behaviors of architecture 1 and 2, respectively where it can be seen that architecture 1 had a faster learning curve than architecture 2 due to the fact that as architecture 1 only visualizes the characteristics in terms of the frequency it did not require also to learn how the signal behaved in the time while architecture 2 had a very low learning for more than 100 epochs while recognizing patterns of feature maps, even this can be evidenced in its training loss that despite the accuracy did not rise, the cost per mistake did descend, i.e., the network was learning. However, the 2 architectures achieved a 100% training accuracy with very low training loss for which in practice is what is expected to ensure that the network will achieve an efficient performance in the task to which it will be destined.

RESULTS AND DISCUSSION

Experimental results: In order to determine the performance of each architecture there were evaluated

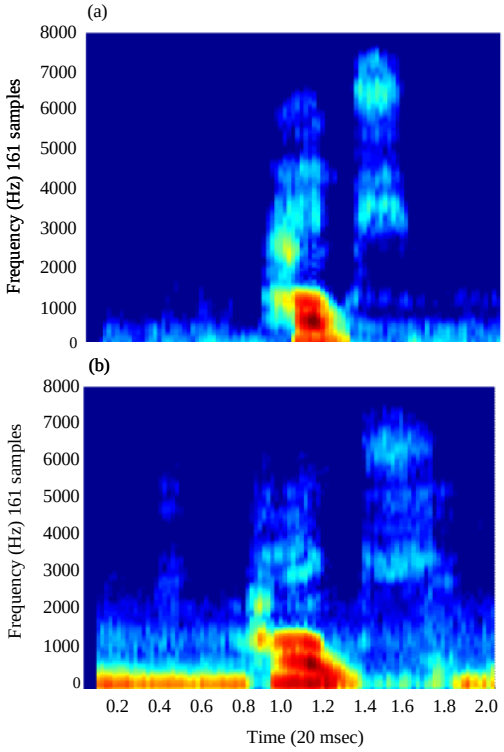

Fig. 5: PSD of the word "House" in a: a) Semi-controlled and b) Non-controlled environment

people from both belonging to the recording group and not belonging to it therefore, the validation of the network is performed with audios in real time, i.e., not pre-recorded.

For the tests, repetitions are made with the users, obtaining an error rate of the network applied in real time. Although, the training dataset was performed in a controlled environment (with very low external noise and some loud noises but of short duration such as a hit on a table), the audios of the tests are taken in two.

Scenarios: One having a semi-controlled environment (low external noise) shown in Fig. 5a and the other in an non-controlled environment (varied background noise such as people speaking, varied sounds, etc.) shown in Fig. 5b. This is done in order to observe which architecture behaves best even though it is not in the training environment.

Each test is performed with 5 repetitions per word per subject, resulting in the general Error Rate (ER) of the number of times there was a misrecognized word.

Semi-controlled environment: Within the tests in a semi-controlled environment whose results are shown in Table 3, architecture 1 which only has frequency filters,

Table 3: Controlled environment results

	Error rate (%)	
Subject group	Architecture 1	Architecture 2
Belonging to the dataset	41.67	16.67
Not belonging to the dataset	45.00	35.00

Table 4: Non-environment results

	Error rate (%)	
Subject group	Architecture 1	Architecture 2
Belonging to the dataset	58.33	40.00
Not belonging to the dataset	70.00	66.67

presented a recognition error of more than 40% for both the subjects belonging to the training group and the ones who was not part of it making this type of architecture not suitable for voice recognition because not taking into account characteristics of words over time is susceptible to confuse words with similar tones. On the other hand, architecture 2 presented an ER of 16.67%, mainly due to confusions between similar words such as "Car"-"Cat" and "Duck"-"Dog", since as subjects are not native English speakers their pronunciations tend to be similar between some words. In addition, in spite of the fact that the number of pronunciations was reduced to 4 persons in the training in terms of subjects not belonging to the dataset its error, although higher, recognized better than in architecture 1, even though presenting more serious problems with the two pairs of words mentioned above where it did present high confusions.

Non-controlled environment: In the tests in a non-controlled environment, ER increased in each architecture as shown in Table 4. Although, the recognition of words became more difficult for people talking in the background and a permanent noise, architecture 2 maintained a performance superior to that of architecture 1 within the persons belonging to the dataset but a very similar error within the non-belonging. It should be noted that although, the training audios were performed in an environment where noise was not significant, the recognition of words in architecture 2 can be promising, even increasing the database to have more different tones and varied pronunciations.

CONCLUSION

The two types of novel architectures were tested to shown their functionality in speech recognition applications thus, extending the variety of convolutional neural network configurations that can be used for these cases. In addition, the use of feature maps different from the Mel-Frequency Cepter Coefficients (MFCC) which are normally used allow a varied range of inputs that can be evaluated and compared in the architectures that are implemented for speech recognition in future works. For the case of this research, the power specter density by means of the non-parametric Welch's method was used as input to the network in a satisfactory way.

The use of convolution layers with frequency and time configuration presented better performance in speech recognition systems, obtaining an ER difference of 25% in a semicontrolled environment and 18.33% in a non-controlled one with respect to a convolutional architecture that does not use them which is a very wide gap. These results allow to determine that the analysis of only frequency does not acquire the evolution of the word in temporary terms, necessary for its recognition.

Because the audios obtained contain long dead times, i.e. where the word is not generated it is convenient to eliminate those times possibly by means of a main speech recognition algorithm in order to only analyze the said word and not the entire length of the obtained audio.

Given the tests performed, convolutional neural network architectures for whole-word recognition can be used in a variety of applications where determinate amounts of words are used, since, the implementation of a database of certain words is more feasible than creating one with phonemes for example in control of robotic agents for reaching objects where there are a number of specific words or in mobile agents to control their movements.

REFERENCES

Abdel-Hamid, O., A.R. Mohamed, H. Jiang and G. Penn, 2012. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 25-30, 2012, IEEE, Kyoto, Japan, ISBN:978-1-4673-0045-2, pp: 4277-4280.

Abdel-Hamid, O., A.R. Mohamed, H. Jiang, L. Deng and G. Penn *et al.*, 2014. Convolutional neural networks for speech recognition. IEEE. ACM. Trans. Audio Speech lang. Process., 22: 1533-1545.

Deng, L. and X. Li, 2013. Machine learning paradigms for speech recognition: An overview. IEEE. Trans. Audio, Speech Lang. Process., 21: 1060-1089.

Deng, L., G. Hinton and B. Kingsbury, 2013. New types of deep neural network learning for speech recognition and related applications: An overview. Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 26-31, 2013, IEEE, Vancouver, British Columbia, Canada, ISBN:978-1-4799-0356-6, pp: 8599-8603.

Girshick, R., J. Donahue, T. Darrell and J. Malik, 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, ACM, Washington, DC, USA., ISBN:978-1-4799-5118-5, pp: 580-587.

Gupta, H.R. and R. Mehra, 2013. Power spectrum estimation using Welch method for various window techniques. Intl. J. Sci. Res. Eng. Technol. IJSRET., 2: 389-392.

Hinton, G., L. Deng, D. Yu, G.E. Dahl and A.R. Mohamed et al., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE. Signal Process. Mag., 29: 82-97.

Hsu, W.N., Y. Zhang, A. Lee and J.R. Glass, 2016. Exploitingdepth and highway connections in convolutional recurrent deep neural networks for speech recognition. Proceedings of the International Conference on Interspeech, September 8-12, 2016, University of California, San Francisco, California, USA., pp: 395-399.

Krizhevsky, A., I. Sutskever and G.E. Hinton, 2012. Imagenet Classification with Deep Convolutional Neural Networks. In: Advances in Neural Information Processing Systems, Leen, T.K., G.D. Thomas and T. Volker (Eds.). MIT Press, Cambridge, Massachusetts, USA., ISBN:0-262-12241-3, pp: 1097-1105.

LeCun, Y., L. Bottou, Y. Bengio and P. Haffner, 1998. Gradient-based learning applied to document recognition. Proc. IEEE, 86: 2278-2324.

Mohamed, A.R., G.E. Dahl and G. Hinton, 2012. Acoustic modeling using deep belief networks. IEEE. Trans. Audio Speech Lang. Process., 20: 14-22.

Ondruska, P., J. Dequaire, D.Z. Wang and I. Posner, 2016. End-to-end tracking and semantic segmentation using recurrent neural networks. Master Thesis, Cornell University, Ithaca, New York, USA.

Orozco, I., M.E. Buemi and J.J. Berlles, 2016. A study on pedestrian detection using a deep convolutional neural network. Proceedings of the International Conference on Pattern Recognition Systems (ICPRS-16), April 20-22, 2016, IET, Talca, Chile, ISBN:978-1-78561-283-1, pp: 1-15.

Qian, Y. and P.C. Woodland, 2016. Very deep convolutional neural networks for robust speech recognition. Proceedings of the 2016 IEEE International Workshop on Spoken Language Technology (SLT), December 13-16, 2016, IEEE, San Diego, California, USA., ISBN:978-1-5090-4903-5, pp: 481-488.

Sainath, T.N., B. Kingsbury, G. Saon, H. Soltau and A.R. Mohamed *et al*., 2015. Deep convolutional neural networks for large-scale speech tasks. Neural Networks, 64: 39-48.

Schmidhuber, J., 2015. Deep learning in neural networks: An overview. Neural Networks, 61: 85-117.

Seide, F., G. Li and D. Yu, 2011. Conversational speech transcription using context-dependent deep neural networks. Proceedings of the 12th Annual International Conference on International Speech Communication Association, August 28-31, 2011, ISCA, Florence, Italy, pp: 437-440.

Sercu, T., C. Puhrsch, B. Kingsbury and Y. LeCun, 2016. Very deep multilingual convolutional neural networks for LVCSR. Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 20-25, 2016, IEEE, Shanghai, China, ISBN:978-1-4799-9988-0, pp: 4955-4959.

Simonyan, K. and A. Zisserman, 2014. Very deep convolutional networks for large-scale image recognition. Master Thesis, Cornell University, Ithaca, New York.

Weibel A., T. Hanazawa, G. Hinton and K. Shinkano, 1989. Phoneme recognition using time-delay neural networks. IEEE Trans. ASSP, 37: 328-339

Yoshioka, T., K. Ohnishi, F. Fang and T. Nakatani, 2016. Noise robust speech recognition using recent developments in neural networks for computer vision. Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), March 20-25, 2016, IEEE, Shanghai, China, ISBN:978-1-4799-9988-0, pp: 5730-5734.