Correlating Speech Processing in Deep Learning and Computational Neuroscience

Shefali Garg, Department of Computer Science and Engineering
Smith Gupta, Department of Electrical Engineering

Abstract

Deep learning has being extensively used in various fields like computer vision, natural language processing, etc. but it has not been explored potentially in the field of auditory data classification. We are building up on the work of Lee[1] and presenting here an application of Convolutional Deep Belief Network (CDBN) on audio data for the task of digit classification.

Features are extracted from an unlabeled audio dataset, which is a version of TIDIGITS which contains speeches of 326 speakers (111 men, 114 women, 50 boys and 51 girls) each pronouncing 10 digit sequences.

We also present a comparison of our results with traditionally used approaches for audio classification i.e., MFCC and raw spectrogram methods.

Introduction

Processing of audio data in human brain has been of great interest since a long time. HMM and GMM were used in past for computation.

Applying deep learning methods to get more general model and to gather an insight as to how the neurons in the brain process speech.

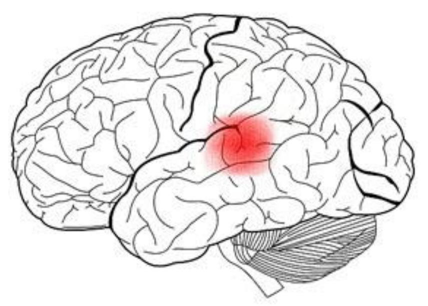
Learning sparse representation of auditory data results in filters which are similar to the neurons involved in audio processing in brains of mammals. For example, there was an amazing resemblance found in representations of auditory data and cochlear filters in the auditory cortex [2].

Speech Processing in human brain

Wernicke's Area

This is located in posterior section of the superior temporal gyrus (STG) in the left cerebral hemisphere and is involved in the understanding of written and spoken language.

source : Wikipedia



Feature Extraction :

According to a research[3] in 2014, human brain breaks up a speech signal into phonemes and extract features corresponding to these phonemes. Different neurons in the brain are responsive to different type of sounds, like plosives, fricatives, vowels and nasals.

So far, the task of digit classification is accomplished through use of baseline features (MFCC and spectrogram) and use of Deep Belief Network[4]. Here we apply CDBN on an unlabeled audio dataset and use the learned features on the task of digit classification. We compare our results with the traditional approaches.

Methodology

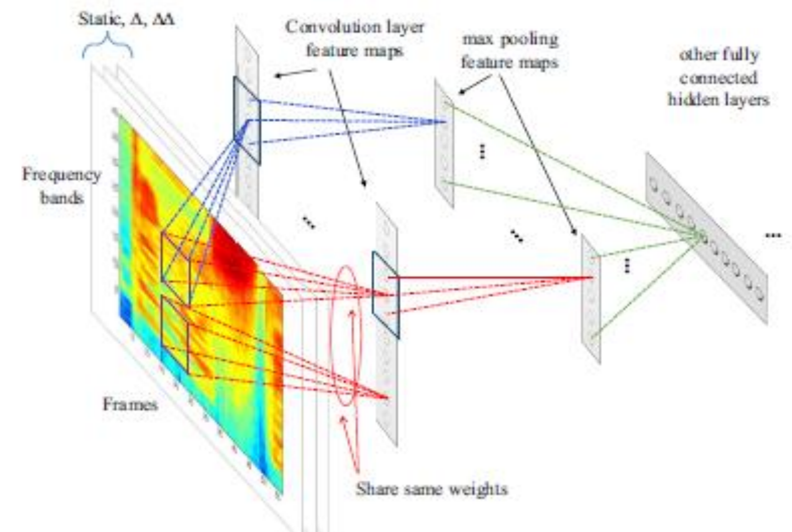
Audio Data
- Dataset used is a version of TIDIGITS

Unsupervised Feature Learning
- Feature extraction is done through MFCC and CDBN

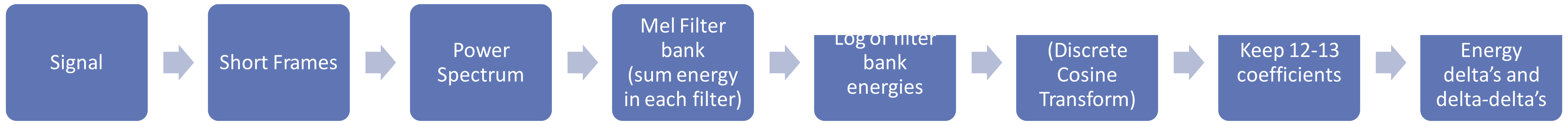
Classification
- SVM is used as a classifier

Convolutional Deep Belief Network (CDBN)

- PCA whitening (with 80 components)
- 300 first-layer bases (maps) , Filter size 6 , Max-pooling ratio of 3
- Local receptive field, weight sharing, pooling
- Invariance to small frequency shifts

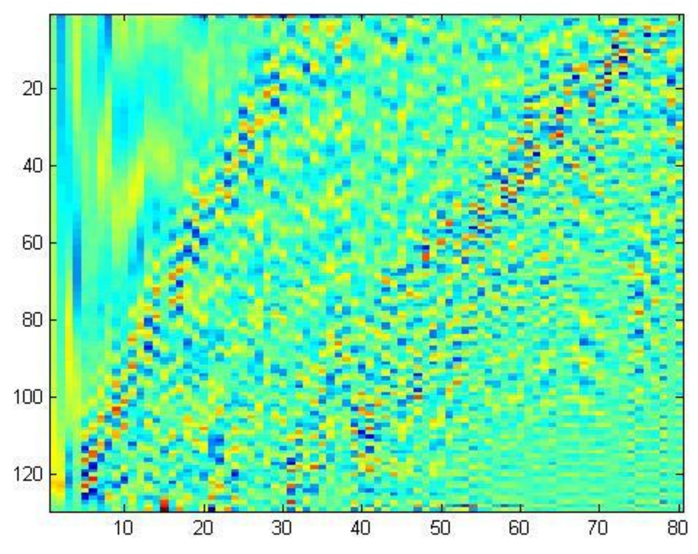


Mel-frequency Cepstral Coefficients(MFCCs)

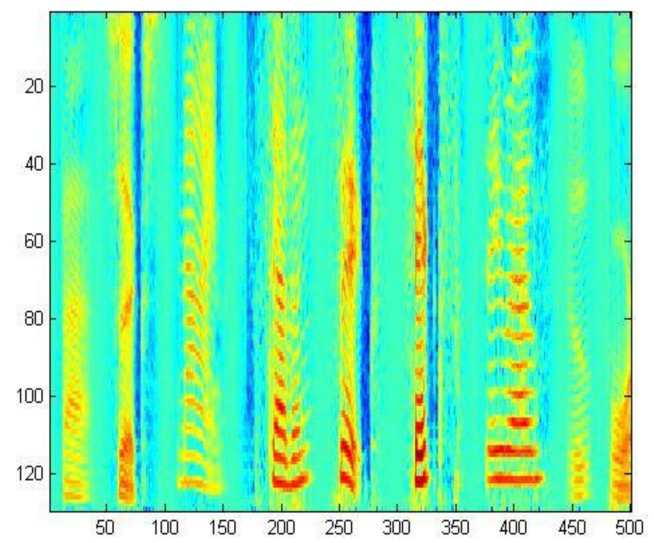


Results

	Spectrogram	MFCC	RBM	CDBN
Accuracy (%)	86.68%	92.76%	36.99%	53.42%



Feature reconstruction from PCA



Feature representation after PCA

Conclusions

CDBN is giving lower accuracy than MFCC (need more experiments and training)
MFCC features are specifically designed for audio data and particular task.
CDBN is general model, can be used for other speech processing tasks.
Maybe, because CDBN model is a bit closer to the human brain model.

In future,
Improve accuracy of CDBN by increasing number of training epochs and modifying parameters. Use GPU's to improve speed.
Combine other deep learning methods like Self Taught Learning, etc.
Try the model with other datasets and other speech classification tasks like speaker identification, phone classification etc.
Relate the activation of neurons in the Wernicke's area to the hidden layer activations in deep belief network.

References

1. H. Lee, P. Pham, Y. Largman, and A. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in Advances in Neural Information Processing Systems 2, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Cambridge, MA: MIT Press, 2009, pp. 1096–1104.
2. E. C. Smith and M. S. Lewicki. Efficient auditory coding. Nature, 439:978–982, 2006.
3. Nima Mesgarani, Connie Cheung, Keith Johnson, Edward F. Chang, Phonetic Feature Encoding in Human Superior Temporal Gyrus. (2014)
4. Audio Feature Extraction with Deep Belief Networks, API 2011
5. Abdel-Hamid, Ossama, Li Deng, and Dong Yu. "Exploring convolutional neural network structures and optimization techniques for speech recognition."INTERSPEECH. 2013.
6. Abdel-Hamid, Ossama, et al. "Convolutional Neural Networks for Speech Recognition." (2014).
7. G. Hinton, L. Deng, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath and B. Kingsbury, & ldquo, Deep Neural Networks for Acoustic Modeling in Speech Recognition,& rdquo, IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82-97, Nov. 2012