CORRELATING SPEECH PROCESSING IN DEEP LEARNING AND COMPUTATIONAL NEUROSCIENCE

Shefali Garg (11678)
Smith Gupta (11720)

MOTIVATION

- Speech Classification previously done through HMM and GMM[1]
- "Deep Learning" approaches not being extensively used for speech processing
- Task of Digit Classification done using DBN and MFCC features[2]
- Using proposed CDBN methods[3] for Digit Classification
- Relating extracted features and hidden units activation to neurons in brain

DATASET

- A version of TIDIGITS dataset will be used for implementation of digit classification
- Each speaker pronounces each digit twice

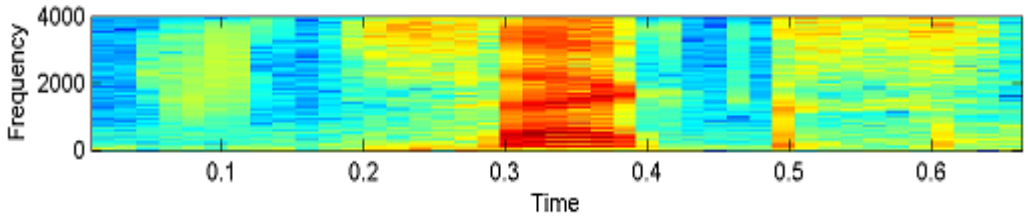
Methodology

- Audio Feature Extraction
  - Raw Features
  - MFCC
  - Deep Learnig
- Classification by SVM

Audio Feature Extraction

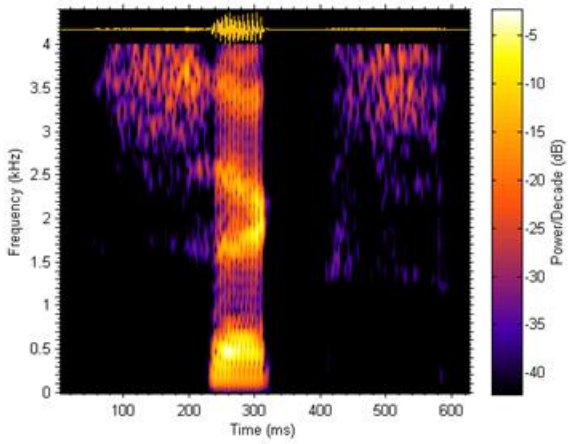
Raw Features

- wav file → spectrogram by FFT
- Spectrogram represents the power of different frequency bands over time
- Accuracy- 86.68% (Baseline)



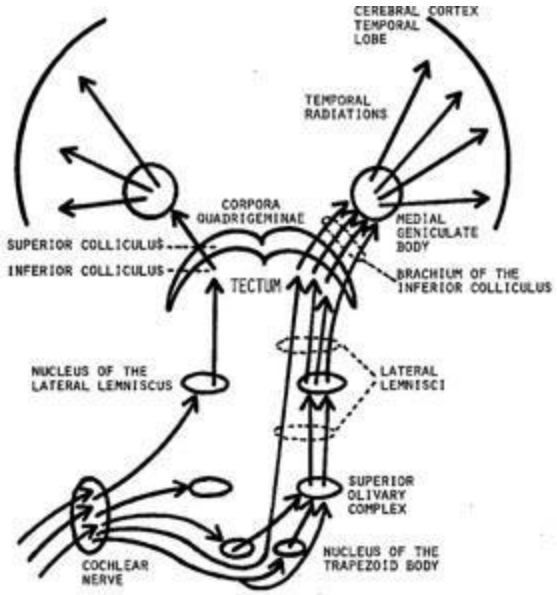
Mel-frequency Cepstral Coefficients(MFCCs)

- Take FFT of frame
- Map the powers of the spectrum obtained onto the mel scale
- Take DCT of the list of mel log powers
- 42-dim feature vector containing information of amplitude, frequency, temporal variance (delta's and delta-deltas) of spectrum
- Accuracy- 92.79%



Deep Learning

- when sparse coding models are applied to natural sounds (auditory signals), the learned representations (basis vectors) showed a striking resemblance to the cochlear filters in the auditory cortex



Deep Belief Networks

- Complete bipartite undirected probabilistic graphical model
- Network assigns a probability to every possible pair of a visible and a hidden vector via a energy function

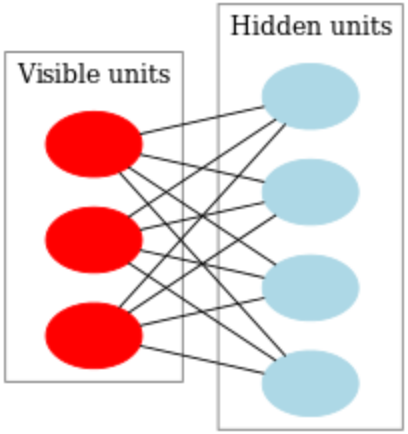


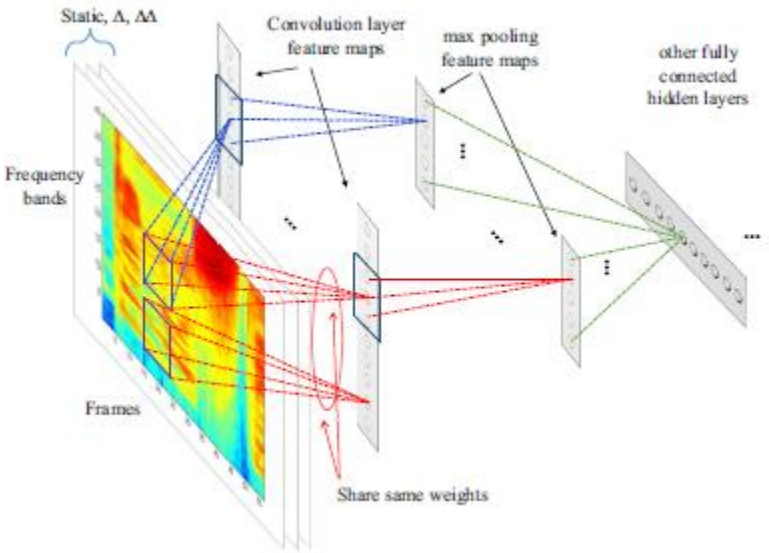
Image source : wikipedia

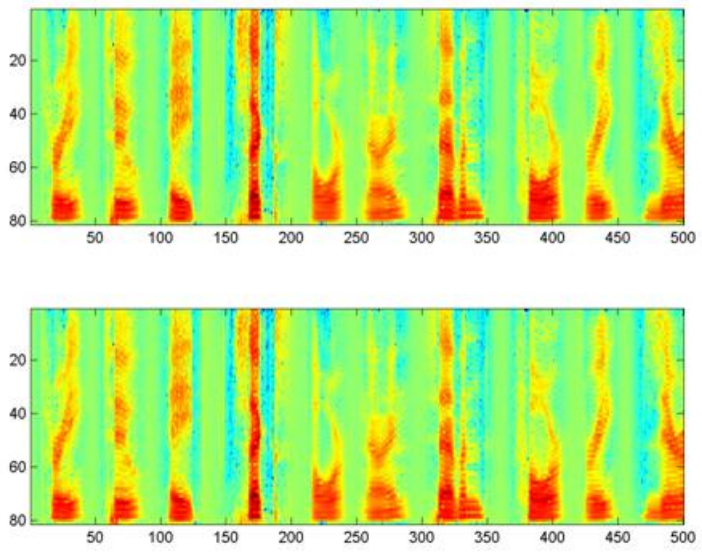
Convolutional Deep Belief Networks (CDBN)

- Each neuron receives input from local limited frequency range
- Hubel and Wiesel- cat's visual cortex cells are sensitive to small local receptive field
- Weight-sharing/Replicated Features- Neurons for same feature share weights
- Probabilistic max-pooling- maxima over small neighborhoods of hidden units computed in a probabilistically sound way.

- Invariance to small frequency shifts
- Sparsity, prevent overfitting (less number of parameters)
- Dimensionality reduction

- First layer bases of random file





FUTURE WORK

- Relating features extracted in neural nets to features extracted in human brain
- Broca‘s Area
- Wernicke’s Area

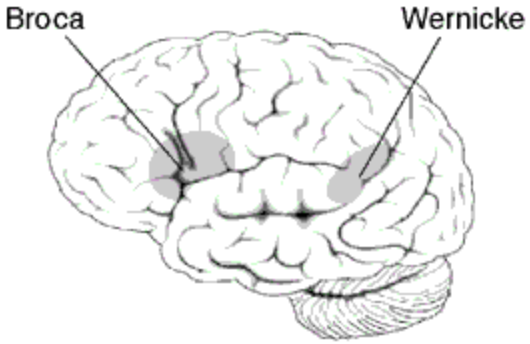


image source : wikipedia

- According to recent research[6] features are extracted based on
  - Plosives : p,t,k,b
  - Fricatives : s,z,v
  - Nasals : n,m

REFERENCES

- [1] 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.
- [2] Audio Feature Extraction with Deep Belief Networks Visit Page
- [3] 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 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, Eds. Cambridge, MA: MIT Press, 2009, pp. 1096–1104.
- [4] Abdel-Hamid, Ossama, Li Deng, and Dong Yu. "Exploring convolutional neural network structures and optimization techniques for speech recognition."INTERSPEECH. 2013.
- [5] Abdel-Hamid, Ossama, et al. "Convolutional Neural Networks for Speech Recognition." (2014).
- [6] Nima Mesgarani, Connie Cheung, Keith Johnson, Edward F. Chang, Phonetic Feature Encoding in Human Superior Temporal Gyrus. (2014)

THANK YOU!