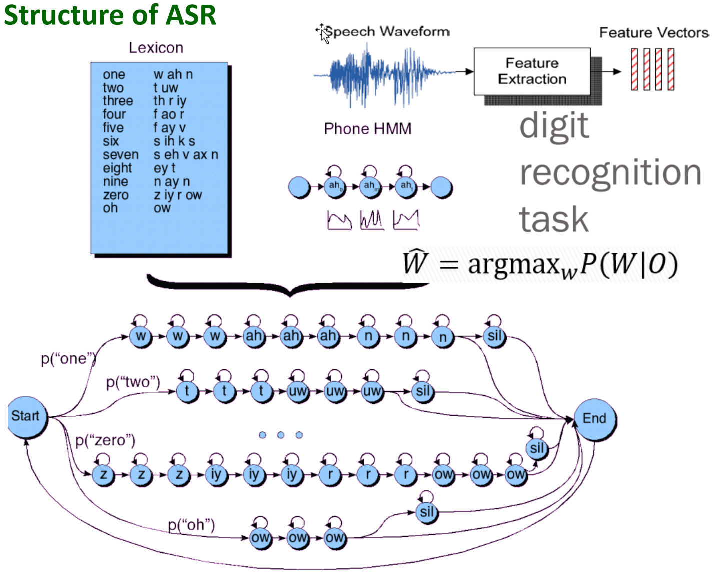
DEEP CONVOLUTIONAL AND LSTM NEURAL NETWORKS IN AUTOMATIC SPEECH RECOGNITION

Xiaoyu Liu
Pearson Education Inc., 4040 Campbell Ave, Suite 200, Menlo Park, CA, 94303, USA
xiaoyu.liu@pearson.com

Abstract

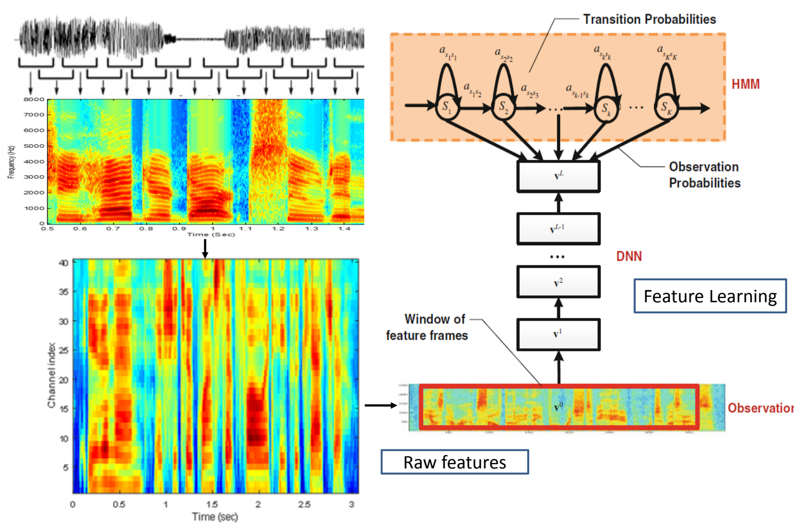
State-of-the-art Automatic Speech Recognition (ASR) systems have widely employed deep Convolutional Neural Networks (CNNs) as acoustic models. Also, deep Long-Short-Term-Memory (LSTM) recurrent neural networks are powerful sequence models for speech data. This work extensively investigates the effects of DNNs, deep CNNs, LSTMs and Bidirectional LSTMs (BLSTMs) as state-of-the-art acoustic models for various ASR tasks.

Structure of ASR



- P(W|O) = P(O|W)P(W): P(O|W) probability of a feature sequence given a word sequence, called acoustic model (AM), P(W) word language model (LM).
- Each word is decomposed into phonemes according to a lexicon, and each phone is modeled by a 3-state left-right Hidden Markov Model.
- Conventionally, the HMM state emission probability P(o|s) is modeled by Gaussian Mixture Models (GMMs). DNN, CNN, LSTMs have replaced GMMs.
- LM is usually a N-gram. Viterbi decoder puts together AM and LM at test time.

Raw and learned speech features by deep models



- Raw speech features (left column) are computed by Fourier Transform of each short frame (25 ms), and grouping to 40 perceptual frequency channels by a filterbank. A speech waveform is converted to a time-frequency (TF) plane.
- Learned features (right column) are computed by taking a context window of the raw speech, passing it through the deep model (DNN, CNN, LSTM) to maximize the correct HMM state probabilities P(s|o).
- The target HMM states are obtained by first training a HMM-GMM model, and searching for the most probable state sequence S given the feature sequence O as well as the word sequence transcription W.
- For each input frame, DNN reshapes the context window into a long vector, but CNN leaves the window as an image.
- LSTM can model the long term sequential dependency between frames, whereas DNN and CNN does frame-based training.

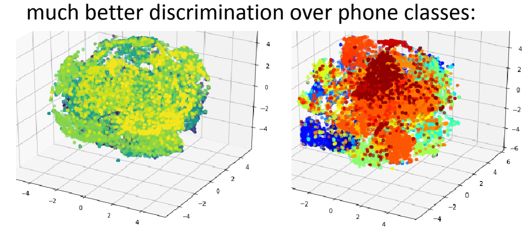
Continuous speech phoneme recognition and conclusions

(more details on parameter tuning and large vocabulary word recognition in report)

- Standard TIMIT database, 61 English phones (183 states), training set has 462 speakers (~5 hours), dev set has 50 speakers, and test set has 24 speakers, 8 sentences/speaker, all clean read data.
- Phoneme recognition accuracy (DNN and CNN context size = 31 frames; ResNet-17 and 33 refer to the depth; LSTM/BLSTM have 4 layers, 1024 memory cells per layer or per direction). CNN and LSTM greatly improve DNN. Deeper ResNet is better than shallower ResNet, and BLSTM improves single direction LSTM.

Model	Accuracy (%)
HMM-GMM	72.0
HMM-DNN	78.1
HMM-VGG	81.7
HMM-ResNet17	81.1
HMM-ResNet33	81.7
HMM-LSTM	80.5
HMM-BLSTM	81.6

- Feature space visualization by t-SNE: raw feature (left) has poor discrimination, and ResNet features (right) extracted from the last avg. pooling layer has much better discrimination over phone classes:



- Phoneme recognition accuracy for reverberated TIMIT data. Same conclusions as in clean data case.

Model	Accuracy(%)
HMM-GMM	57.2
HMM-DNN	71.9
HMM-VGG	75.6
HMM-ResNet17	74.4
HMM-ResNet33	75.2
HMM-LSTM	73.7
HMM-BLSTM	74.9