Speech Recognition with DNN-HMM framework

Group Members

1. Ankita Pasad (12D070021)
2. George Jose (153070011)
3. Nikunj Patel (143079002)
4. Hitesh Tulsiani (14307R032)

1. The main goals of your project

- Train a speech recognizer with DNN-HMM framework

In this report, we present results of an automatic speech recognizer (ASR) with GMM-HMM and DNN-HMM framework. We compare and contrast the results on TIMIT dataset [1]. For our work, we have used 'Kaldi' speech recognition toolkit[2]. Also, for this work we use simple phone bigram language model.

2. Related literature

The task of speech recognition is to convert an audio utterance to text. Any statistical ASR system is made up of two major components:

1. Acoustic Modeling
2. Language Modeling

Acoustic model capture the variability of smallest meaningful unit of sounds i.e. phones whereas language model controls the phone sequence. Acoustic model should capture both temporal and spectral variabilities inherent in speech. Traditionally, to capture spectral variabilities Gaussian Mixture Models (GMM) have been used and HMMs have been used to capture temporal variability. With the advent of Deep learning, nowadays DNNs have replaced GMM for modeling spectral variabilities.

Training for DNN usually comprises of 2 stages:

1. Pretraining: This is unsupervised training of DNN architecture considering it as Restricted Boltzman Machine (RBM). The main idea here is to allow the network to learn what speech is rather than actually discriminating between the phone.
2. Fine-tuning: This is supervised training of DNN using back propagation. The targets for training DNN are usually obtained with an initial GMM-HMM system

For testing, observation probabilities are obtained from DNN which are then used along with HMM and Language model to allow sequence of phones to appear in output text.

More details can be found in this reference[3]:
http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/38131.pdf

3. Database

For our work, we use TIMIT database [1]. The TIMIT corpus of read speech is designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. TIMIT contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. Each of 630 speakers utter 10 utterances. 2 utterances were common for all the speakers. The duration of data is roughly 5 hours.

Train set:
Training is done using data of 400 speakers. This amount to 4000 utterances.

Test set:
Testing is done on core 192 TIMIT set. There is no overlap between train and test set speakers and utterances.

4. Description of the set of approaches tried:

As stated earlier, we have tried both GMM-HMM and DNN-HMM frameworks for speech recognition.

For DNN-HMM system, we have done both pretraining and fine-tuning of the network. Fine-tuning was done using the alignments from baseline GMM-HMM system.

5. Experiments:

For our experiments, we have used Kaldi toolkit [2]. The toolkit is written in C language and has various binaries available for easy assess. The wrapper script to call the suitable binaries and for setting up the DNN network dimension were written by us in shell. Also, for language modeling we wrote the wrapper script in shell.

Link for wrapper codes:
Git Link: https://github.com/tulsianihitesh26/SR_DNN

For all our experiments, we have used I7 system with Nvidia GPU GeForce GTX 1080 (8 GB memory, 2600 cores). It took us around 2 hours for training each network configuration.

Results:

Our ASR engine is a DNN-HMM system implemented using Kaldi toolkit. We also have in place a parallel GMM-HMM system. Both the phone-based ASR systems are trained on the TIMIT dataset. The features used for GMM-HMM systems are obtained via LDA+MLLT transformations on 13 MFCC feature vectors spliced with 7 context frames whereas for DNN-HMM system we use filter bank energy features spliced with +/- 5 frames. We have context dependent, tied-triphone HMM models with number of senones limited to 1000. For the GMM-HMM system the number of Gaussian mixtures per senone are in the range 8-16 whereas

for DNN-HMM system we have 4 hidden layers each having 1024 neurons. The DNN-HMM system is pre trained and fine-tuned with same TIMIT dataset to minimize cross entropy criteria whereas GMM-HMM was trained using the MAP criterion.

Below are the results in terms of phone error rate that we get on core TIMIT test set (192 utterances):

	GMM-HMM	**DNN-HMM**		
Parameters	8-16 Gaussians per senone	2-layers; 1024 neurons each	3-layers; 1024 neurons each	4-layers; 1024 neurons each
Features	LDA + MLLT over MFCC (+/-3 frames)	MFBE (+/-5 frames)	MFBE (+/-5 frames)	MFBE (+/-5 frames)
Language model	Phone bigram	Phone bigram	Phone bigram	Phone bigram
Phone Error Rate (%)	29.65%	28.56%	23.42%	21.10%

We achieve 21.10% PER using 4 layer and 1024 neurons each. This is close to the state-of-the-art reported results of around 18.5% in literature.

6. Effort:

A large amount of time was spent in setting up the Kaldi toolkit to use available GPU. Also, a considerable amount of time was spent on providing the files in necessary format for Kaldi to work. After the set up, the toolkit was relatively easy to use. Obviously, writing the wrapper script for training DNN-HMM, GMM-HMM and language modeling required considerable effort from all of the team members.

The most challenging part of the work was to get the language model (LM) to work properly. LM is built using Finite State Transducers (FST). With no prior knowledge of FST, it was difficult to get it to work.

References

1. Garofolo, John, et al. **TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1**. Web Download. Philadelphia: Linguistic Data Consortium, 1993.
2. Povey, Daniel, Ghoshal, Arnab, Boulianne, Gilles, Burget, Lukas, Glembek, Ondrej, Goel, Nagendra, Hannemann, Mirko, Motlicek, Petr, Qian, Yanmin, Schwarz, Petr, Silovsky, Jan, Stemmer, Georg and Vesely, Karel, **The Kaldi Speech Recognition Toolkit**, Idiap-RR-04-2012

3. Hinton, Geoffrey, et al. **"Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups."** *IEEE Signal Processing Magazine* 29.6 (2012): 82-97.