Large-Vocabulary Continuous Speech Recognition with Linguistic Features for Deep Learning

CS 229/224N Joint Final Project
Peng Qi

ABSTRACT

Until this day, automated speech recognition (ASR) still remains one of the most challenging tasks in both machine learning and natural language processing. ASR research faces data with high variability, which requires highly expressive models be built. Recently, deep neural networks (DNN) have been successfully applied to various fields, including speech recognition. In this course project, We would like to investigate what are some possible linguistic features that would contribute to speech recognizers, as well as what machine learning techniques can be applied to such tasks.

1. INTRODUCTION

Deep neural networks have made great contribution to speech recognition research recently, pushing one step further the state of the art. A speech recognizer generally consists of two parts, an acoustic model that converts acoustic input into phonemes, and a language model that combines these phonetic information to form words and sentences. Deep neural network have been shown to improve the performance of both tasks, see, e.g. [2] and [4].

From preliminary analyses of our dataset (detials in Section 3), we observed that if the trigram language model is given perfect output from the acoustic model, the system word error rate (WER) is about 2%, while the state of the art on this dataset is about 20%. That comprises the basic motivation of focusing our project on improving the acoustic model. Specifically, we trained and analyzed a handful of different deep learning models on the dataset, and investigated how additional linguistic features such as conversation topic, speaker gender, as well as others would further affect the performance of acoustic modeling.

2. LITERATURE REVIEW

In [2], Hinton *et al.* introduced DNN-HMM hybrid systems for speech recognition, which achieved considerable improvements over the traditional GMM-HMM systems. In 2010, GoldWater et al. [1] conducted a thorough research on how various acoustic and linguistic properties might affect the performance of such systems. In that paper, the authors evaluated the effect of a myriad of properties including speaker gender, position near disfluency, pitch, etc. While in that work the authors evaluated feature effectiveness with independent word error rate (IWER), we will stick with senone[1] accuracy thoughout this project.

[1]Senones used in this project roughly correspond to triphone states of the successive HMM in the language model.

While reviewing related literature, we also found that a specific type of neuron activation function, namely *linear rectifiers*, are widely applied and achieved state-of-the-art performance in a number of recent publications. Hence in this project, we'll adopt a variant of linear rectifiers for our deep neural networks proposed in [3].

3. DATASET

In this project, the Switchboard speech recognition corpus[2] was chosen as our study dataset mainly because of two reasons. First, with about 2,400 telephone conversations from 543 speakers, this dataset contains a large amount of data that are highly diverse, which allows large deep neural networks trained supervisedly without the concern of heavy overfitting and poor generalization. The size of the corpus also relieves the burden to build a sophisticated language model. This allows us to focus on the acoustic model, and hopefully reducing the system WER by improving the senone (or frame) accuracy.

Another major reason for our choosing Switchboard (SWBD) over other datasets is that SWBD contains a number of well-documented linguistic features that were collected alongside the speech data, which would significantly help in verifying the idea that such features might help improve the performance of acoustic models. Below we will briefly describe the features used in our project, the rationale behind using them, and some basic statistics across the dataset. Before listing the linguistic features, it is worth noting that the input acoustic features should have been projected following a standard procedure to a subspace where speaker-dependent information are supposedly removed. However, due to the (conceptually) high nonlinearity of speech information with regard to its variability, we believe that some speaker-dependent information still exists within the acoustic features, and by introducing the corresponding linguistic features we can cancel out these "residuals" with highly non-linear deep neural networks to achieve better performance.

- **Speaker Gender.** Speakers of different sexes tend to present significant differences in pitch change, speaking speed (which affects the presense of senones related to repetition/deletion/insertion), as well as word choice (which affects the probability of presence of different senones).
- **Speaker Dialectic Region.** Speaker dialect tends to significantly affect the their pronunciation of phones.

[2]http://www.isip.piconepress.com/projects/switchboard/

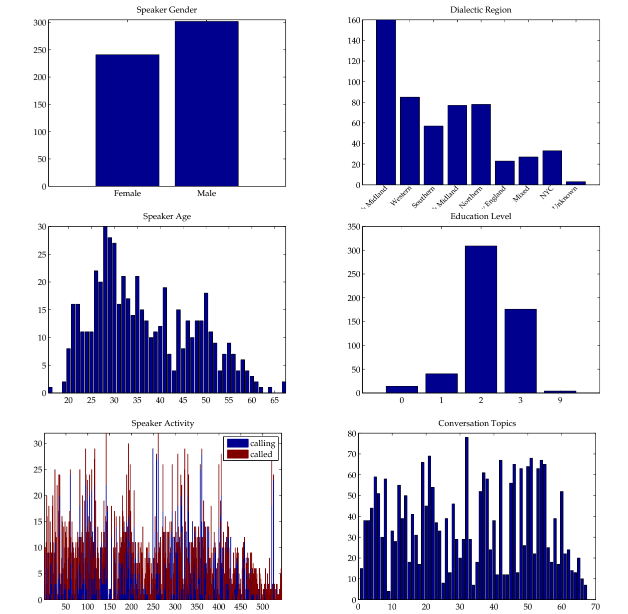

Figure 1: Linguistic features statistics of the Switchboard dataset

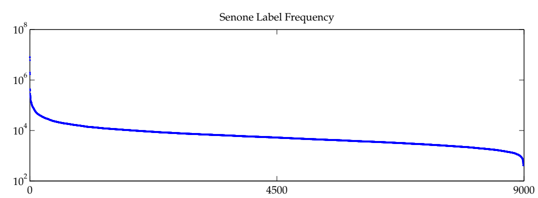

Figure 2: Senone label statistics of the Switchboard dataset (sorted by frequency)

- **Speaker Age & Education Level.** Both might contribute to word choice and/or pronunciation convention of the speaker.
- **Speaker Identity.** Apart from the information above, speaker identity might convey information about some speaker specific habits or personal marks of word choice, etc.
- **Conversation topic.** Apart from its evident effect on word choice, conversation topics might also affect speech speed, pitch change, etc.

In Fig. 1, we have drawn a number of statistics of the above-stated properties across the dataset. From the figure we can see that most linguistic features have a relatively even distribution, which is a good property for informative features as none of them will provide virtually "zero" information to the deep neural network.

4. BASELINING

Before introducing linguistic features, we briefly analysed the property of the dataset, and performed baseline training on several different deep neural networks that we will elaborate below. To balance between performance and training speed, the networks used in our project shared the same basic structure with 1,640 acoustic input units, three linear rectifier hidden layers of 2,048 units each, and a classification output layer with 8,986 senone classes where errors are back propagated from. The training set statistics of the senone labels is shown in Fig. 2 (log-scale).

From Fig. 2 it is evident that the senone labels follow a very skewed distribution in the training set, for which multiclass classifiers usually fail to achieve high accuracy. As a start, we trained standard softmax deep neural networks (DNNs) with cross-entropy cost function (CENet) on about 280 hours of speech data and tested on a separate 4.7 hours. With stochastic gradient descent, we trained the network on the whole training set with minibatches of 256 training examples. In the meantime, we considered it a good idea to attempt large-margin cost function (SVMNet), which conceptually should work better on multiclass classification tasks than CENet because it is purely discriminative rather than generative. Then, to account for the skewed distribution of the labels, we also tried to modify CENet with hierarchical classification. Specifically, after sorting the labels in decreasing order by their frequencies, we progressively classfied the top 2,000 (HCENet-2k) or 4,000 (HCENet-4k) senones against the rest until all labels are classified, and added the cost functions of these classifiers together to optimize with the DNN. Finally, we also attempted another scheme to address the skewness, reweighing cost functions. By reweighing the cost function softmax and large-margin networks with reciprocals of label frequencies, we obtained two final baseline networks RwCENet and RwSVMNet. The results of these baseline networks are shown in Table 1 after 5 epochs of training (usually took 3$\sim$7 days for each model with GNumPy[3]).

Surprisingly, CENet alone is capable of working well, while SVMNet, which ideally would have been better as a discriminative rather than generative model, turned out to be a lot worse. However, by looking at the reweighed models, we can see that RwSVMNet improves significantly based on SVMNet, which probably suggests that SVMNet's failure resulted from the imbalancement of training examples within each mini-batch of stochastic gradient descent[4], in which case the parameters for rare classes hardly got updated with enough positive examples, while reweighing the cost function alleviates this problem in gradient computation. On the other hand, reweighing didn't seem to help CENet, which is predictable as softmax classifiers are generative models, which works best if the prior knowledge of the data is correctly exploited. Also surprisingly, hierarchical classification scheme didn't work well on this dataset. This might suggest that the major challenge of the dataset is the distinction between some frequent class versus some infrequent ones, rather than among classes with similar frequency in the training set, in the sense that compared to CENet, the drop in performance resulted from the network's feature extraction capability misued on minor discriminations. These observations lead to potential future work directions on this dataset described in Section 6.

5. INCORPORATION OF LINGUISTIC FEATURES & ANALYSES

After baselining, we chose the standard softmax network, amongst others, as the baseline model for further analysis

[3]http://www.cs.toronto.edu/ tijmen/gnumpy.html
[4]For SVMNet we also attempted to use larger minibatches, but increasing minibatch didn't improve the performance of the network, either, before we ran out of GPU memory.

Table 1: Baseline model performances

Model	Train Accuracy* (%)	Test Accuracy (%)
CENet	**71.20**	**63.66**
RwCENet	52.63	48.56
HCENet-2k	38.57	36.09
HCENet-4k	47.58	43.91
SVMNet	7.90	8.16
RwSVMNet	52.79	48.73

* The training set accuracies are estimated on-the-fly during training, with $\alpha = 0.99\alpha + 0.01\alpha_{\text{minibatch}}$,where α is the overall accuracy estimation and $\alpha_{\text{minibatch}}$ the minibatch accuracy for the last-seen minibatch. The same technique is also applied to experiments in Section 5 to reduce computation time. Same procedure applied to models with linguistic features.

Table 2: Model performances with linguistic features

Model	Train Accuracy* (%)	Test Accuracy (%)
CENet	71.20	63.66
CENet-A	71.85	64.05
CENet-A2	72.03	64.18

with liguistic features. To assess the contribution of linguistic features that we introduced, we started with a basic augmented model, where the linguistic features are appended to the acoustic ones and fed together into the deep neural network (CENet-A). Specifically, the linguistic features were translated into binary or categorical features, and numerical feature like age is binarized by thresholding with its median value to ensure compatibility with the deep neural network model. To further ensure that the linguistic features take part in the training process of the DNN, we also developed a second network structure where the linguistic features were fed into each hidden layer as well as the output layer of the DNN, forcing each layer to accommodate the raw linguistic features when trying to minimize the model cost function (CENet-A2). The results from the models with linguistic feature incorporation are shown in Table 2, where the CENet results are also shown as a baseline.

Finally, we also attempted to train a DNN model that also predicts the linguistic feature themselves alongside the senone labels, which resembles an autoencoder in some ways, with the hope that this kind of structure can help us make sure that linguistic features are taking part in the representation of the DNN. Technically speaking, such models are called multitask learning systems (MTNet), which generally should reduce overfitting and improve model generalization ability. The potential logic behind such systems is that the local optima the softmax network alone achieves is possibly not as good as that for the multitasking network, or the dynamics of the latter could lead to a better local optima faster for the classification task with the help of extra information. This might not be generally true for most models, but for highly non-linear models such as DNNs where first-order gradient descent based methods are applied, it seems more reasonable to assume the existence of better local optima unreachable with simple optimization algorithms. However, by the time of the completion of this report, the complicated multi-task cost function significantly worsened the performance of the network on the original senone classification task. Though not much substantial improvements were achieved, this part of the project did suggest one of the future direction of our work.

From Table 2 it can be seen that the extra linguistic features did improve the classification accuracy of the senones, but it would be of interest to more closely examine how the features worked, and how much each individual type of extra information helped.

To perform error analysis on our models, the 8,986 senones are mapped back to their 46 different center phones, and the confusion matrix of these phones are shown in Fig. 3 top row (left). With this confusion matrix for the baseline CENet model, we can tell that the DNN is already performing impressively to correctly classify most of the phones, although some major anomalies do attract our attention. The most significant anomaly is that a major number of classification errors happened when spoken noise (spn), non-spoken noise (nsn), as well as in-word pause (lau) were misclassified as silence (sil), and in fact it is observed that a lot other phones are misclassified to silence as well. This is likely to result from the imbalanced distribution of the phones in speech, where silence appear in most utterances while specific phones appear much less. Some other observations include misclassfications of `en` as `n`, confusion among `k`, `g`, `p`, and `d`, between `eh` and `ae`, between `z` and `s`, as well as other common mispronunciations and mishearings that occur in speech. After the incorporation of linguistic features, the major results (confusion matrix) are similar, the change of the confusion matrix is analyzed instead. As it turned out, one of the improvements is that `ah`'s are significantly less recognized as `ae`. Other improvements include better differentiations between `s` and `z`, among `eh`, `aw`, `ay`, and `ae`, and among tailing consonants (`t`, `d`, `n`, `m`, etc). While intuitively the confusion of vowels might be majorly related to dialectic regions, the pronunciation habit of tailing consonants might also trace back to the speaker's age or educational level.

Next, we analyzed the feature effectiveness of the CENet-A model by plotting the average squared second norm of the weights for each class of linguistic features that were fed into the network. With the average value of all first-layer features plotted in dashed line and its one-standard-deviation range plotted in dotted line, it can be shown that age, dialectic region, and educational level are the most contributive linguistic features in this network, which endorses our reasoning in the analyses of confusion matrices. Identity and topical information helped less in this task, which probably results from their sparsity accross the dataset compared to the top three. To our suprise, gender information seems very unhelpful in this task, which suggests that the acoustic features that we use have successfully removed gender-related information in the transform, or that gender-related variabilities in the input is less of a problem given the representational power of deep neural networks.

6. CONCLUSION & FUTURE WORK

In this course project, we examined the effectiveness of various deep learning models with controlled experiments, and applied linguistic features to the softmax network, improving its performance in acoustic modeling, a crucial part and performance bottleneck of state-of-the-art speech recognition systems. We've demonstrated that with the incorporation of linguistic information when available, the performance of acoustic models can be improved, and analyzed the importance of each of the features.

One of the next steps of this project should intuitively be applying the linguistic feature-augmented deep neural networks to the full model of speech recognition, and examine whether word error rate could be lowered as a result.

Another potential future direction comes from our experience and observations during the project. While undertaking experiments for the project, the major bottlenecks for us were the efficiency for learning the deep neural networks, for which stochastic gradient descent is applied in line with the field of active research. However, our discoveries with large-margin cost functions as well as multi-task networks might suggest that we should research for more efficient and effective learning algorithms for deep learning models with a large number of parameters on such huge amount of data.

Finally, a potential future direction specific to speech recognition (and perhaps machine learning tasks with similar natures) is to apply structured classifiers to the DNN acoustic model. As was observed in our hierarchical classification task, if the representational capacity of the network is "wasted" on non-significant discrimination tasks, the model effectiveness would deteriorate. However, we would expect substantial improvement if such hierarchical information is used correctly. Specifically, we would like to apply the intuitive hierarchy that exists among the senone classes as a tree-structure classification target, where not only the task of the network is the discrimination of the senones themselves, but also their center-phones can be taken into account. We expect to see an improvement in senone classification accuracy as a result.

Acknowledgements

We would like to thank Prof. Ng for his in-class instruction, and the TAs for their feedback on this project. We would also like to thank Andrew Maas, Awni Hannun, and Chris Lengerich from the Stanford Deep Learning for Speech Recognition Group for providing source of data, for their insightful comments as well as helpful discussions.

7. REFERENCES

[1] Sharon Goldwater, Dan Jurafsky, and Christopher D Manning. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52(3):181–200, 2010.

[2] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97, 2012.

[3] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *ICML Workshop on Deep Learning for Audio, Speech, and Language Processing*, 2013.

[4] Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In *Proceedings of the international workshop on artificial intelligence and statistics*, pages 246–252, 2005.

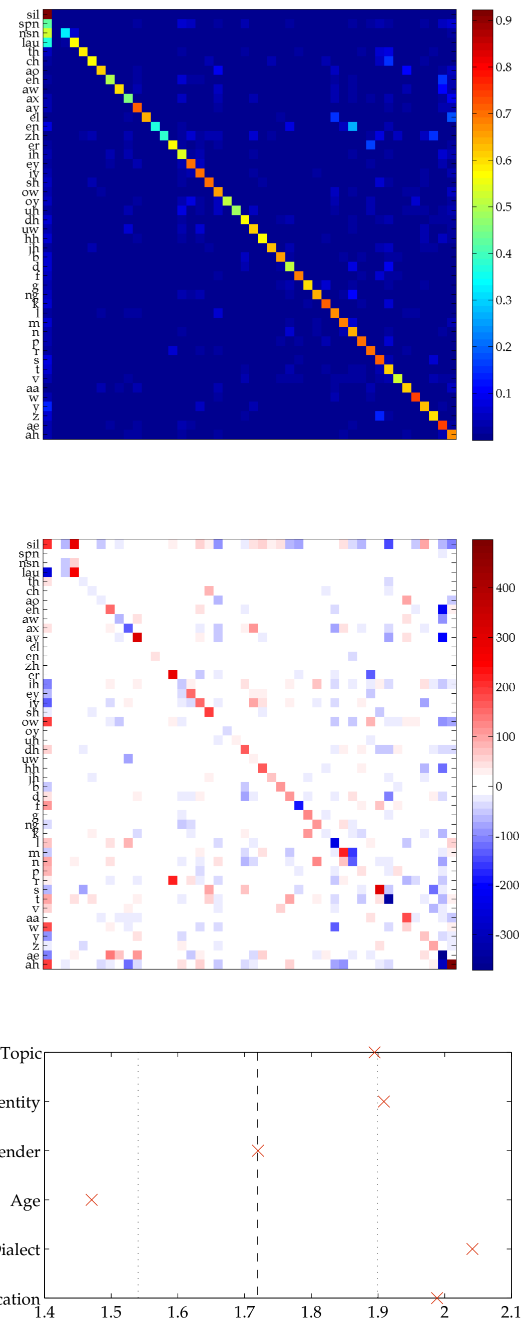

Figure 3: Analyses of the effect of introduced linguistic features