# CS 224N/229: Joint Final Project: Large-Vocabulary Continuous Speech Recognition with Linguistic Features for Deep Learning

**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, and more importantly, how much they contribute to speech recognition, and how well these features generalize across different data instances.

## 1 Introduction

Deep neural networks have witnessed a resurgence over the past few years, and speech recognition is among the many fields where deep learning made great contribution to pushing one step further the state of the art. Generally speaking, a speech recognizing system consists of two parts, namely the acoustic model and the language model. The former converts acoustic input into a symbolic representation (syllabus), while the latter combines these symbols to form words and sentences. Deep neural network have been shown to work for both task, see, e.g. [2] and [4].

In this course project, I would like to focus on improving the acoustic model of speech recognizers. More specifically, I would like to investigate how additional linguistic features such as conversation topic, speaker gender, speaker education level, speaker age, speaker dialectic region, and speaker identity (which is related to personal habits in speech) would affect the performance of acoustic modeling, to what extent they contribute, as well as explore other possible ways of improving acoustic modeling with deep learning models in general.

## 2 Literature Review

In 2010, GoldWater et al. [1] conducted a thorough research on how various acoustic and linguistic properties might affect the performance of speech recognizing systems. In that paper, the authors evaluated the effect of a myriad of properties including speaker gender, position near disfluency, pitch, etc, covering a large set of linguistic and acoustic features that may affect speech recognition. While in that paper the authors benchmarked on a novel evaluation criterium called independent word error rate (IWER), in this course project I would like to stress more on the quality of the senones of the acoustic model, with reasons stated in Section 3.

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[1] 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. In fact, in this dataset, where the senones predicted perfectly from acoustic inputs, the HMM and trigram word/language model can achieve an word error rate (WER) of around 2%, significantly lower than the state-of-the-art performance of speech recognition systems on this dataset, which is around 20%. This allows us to focus on the acoustic model, and hopefully reducing the system WER by improving the senone[2] (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 of our project. Below we will briefly state 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 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 in the acoustic features, and by introducing the corresponding linguistic features we can cancel out these "residuals" with highly nonlinear deep neural networks.

- **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.
- **Speaker Age & Education Level.** Both might contribute to word choice and/or pronunciation convention of the speaker.
- **Speaker Identity.** Apart from the information above, 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 share the same basic structure with 1,640 acoustic input units, three linear rectifier hidden layer of 2,048 units, and a classification output layer with 8,986 senone classes. The training set statistics of the senone labels is shown in Fig. 2 (log-scale).

[1] http://www.isip.piconepress.com/projects/switchboard/
[2] senones used in this project roughly correspond to tri-phone states of the successive HMM in the language model.

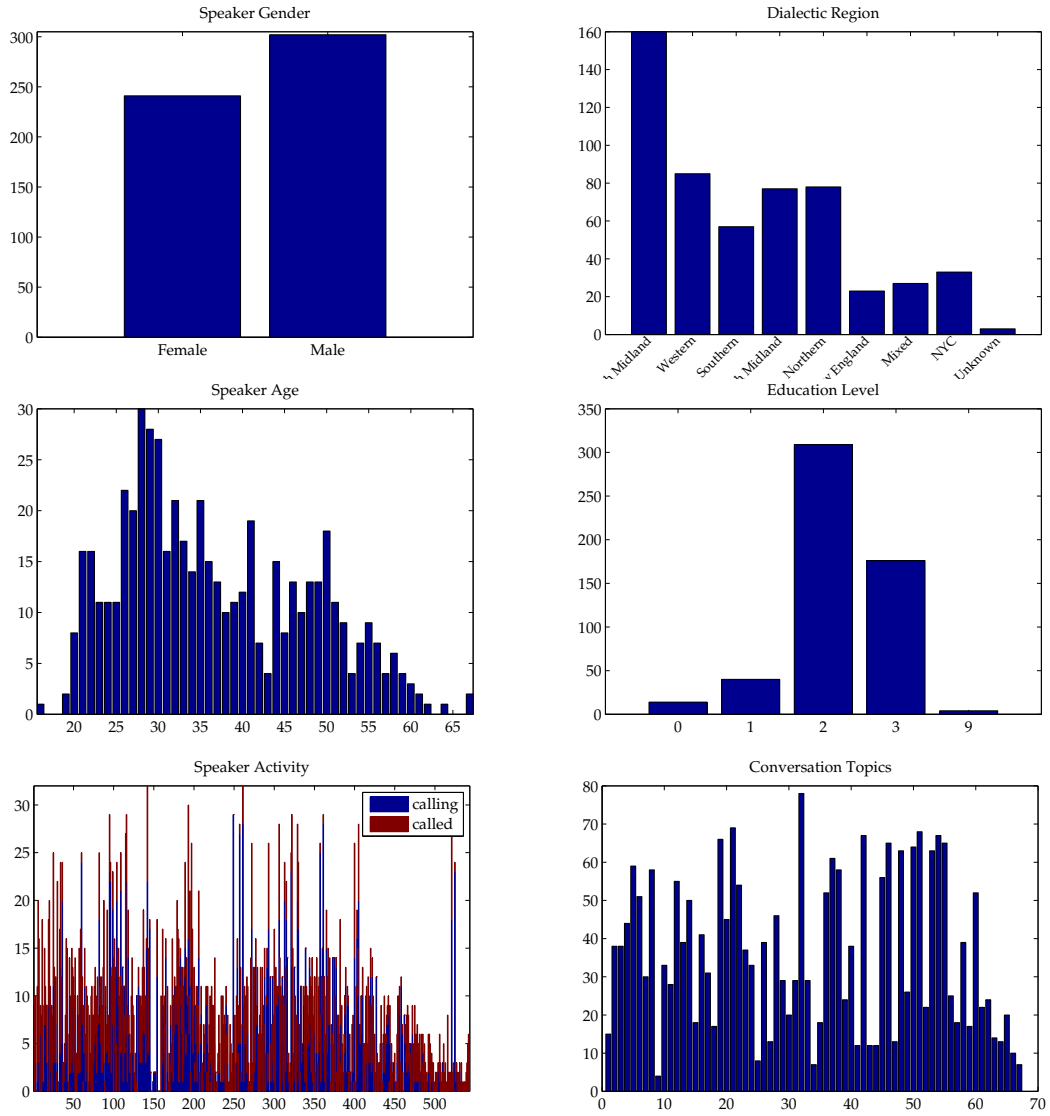



Figure 1: Linguistic features statistics of the Switchboard dataset

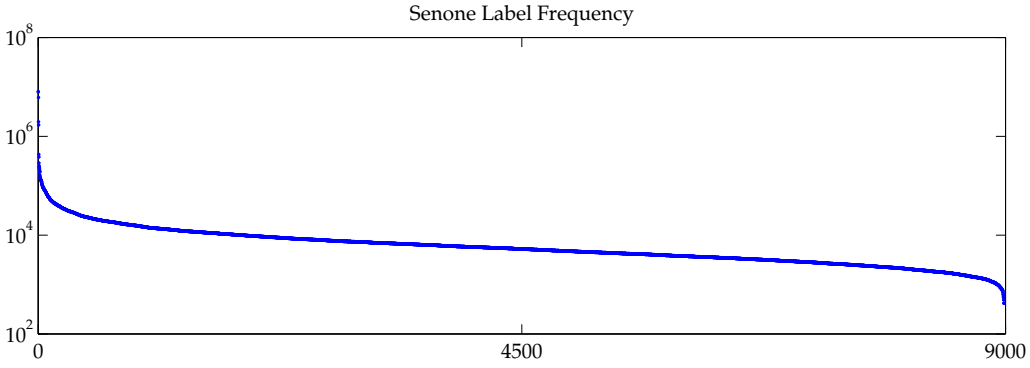



Figure 2: Senone label statistics of the Switchboard dataset (sorted by frequency)

Table 1: Baseline model performances

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

* The training set accuracies are estimated on-the-fly during training, with $\alpha = 0.99\alpha + 0.01\alpha_{\text{minibatch}}$,where $\alpha$ 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.

From Fig. 2 it is evident that the senone labels follow a very skewed distribution, for which multiclass classifier (layers) might struggle to achieve high accuracy. As a start, we trained standard softmax deep neural networks (DNNs) with cross-entropy cost function (codename: CENet) on about 280 hours of speech data and tested on a separate 4.7 hours. 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 5$\sim$10 days for each model with GNumPy).

Surprisingly, CENet alone is capable of working pretty well, while SVMNet, which theoretically would have been better, 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, in which case the parameters for rare classes hardly got updated with enough positive examples. 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 on this dataset. This might suggest that the major challenge of the dataset is the distinguish between some frequent class versus some infrequent ones, rather than among classes with similar frequency in the training set. 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 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). 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 and output layer of the DNN, forcing each layer to accommodate the raw linguistic feature 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.

To address our question in the problem proposal, we also attempted to train a DNN model that also predicts

Table 2: Baseline model performances

| Accuracy/% | CENet | CENet-A | CENet-A2 |
|---|---|---|---|
| Train | 71.20 | 71.85 | 72.03 |
| Test(dev) | 63.66 | 64.05 | 64.18 |

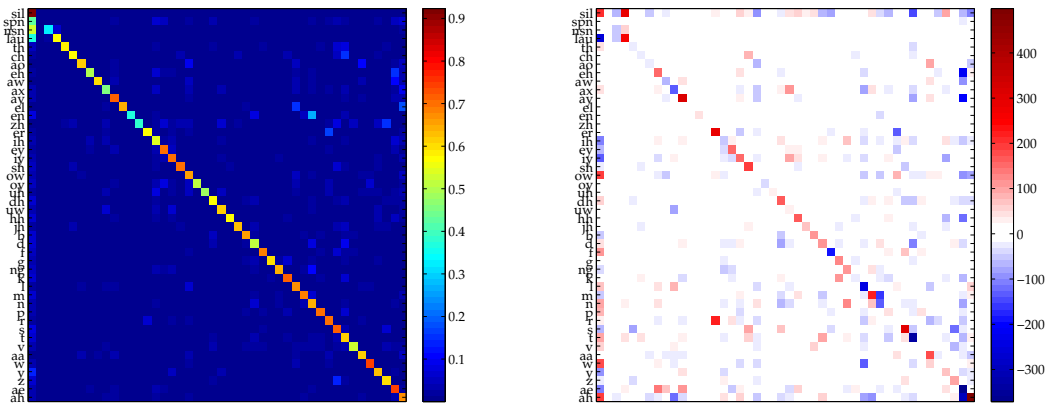
the linguistic feature themselves alongside the senone labels, which resembles an autoencoder in some ways, with the hope that this kind of structure can help us make sure that linguistic features are taking part in the representation of the DNN. Technically speaking, such models are called multitask learning systems (MTNet), which generally should reduce overfitting and improve model generalization ability[3]. However, as it turned out, the complicated multi-task cost function significantly affected the performance of the network, which hasn't yet been able to improve the results of senone classfication as this report is written. 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 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.

The 8,986 senones are mapped back to their 46 different center phones to perform error analysis, 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 points 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). Some other observations include misclassfications `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. After the incorporation of linguistic features, the major results (confusion matrix) are similar, thus we choose to analyze the change of the confusion matrix. 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 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 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 underpins our reasoning in the analyses of confusion matrices. Identity and topical information helped less in this task, which might result 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.

[3] In fact, this experiment roots more deeply in a sense of machine learning, under the assumption 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 generally true for most models, but for highly non-linear models such as DNNs where gradient descent based methods are applied, it seems reasonable to assume the existence of better local optima unreachable with simple optimization algorithms.



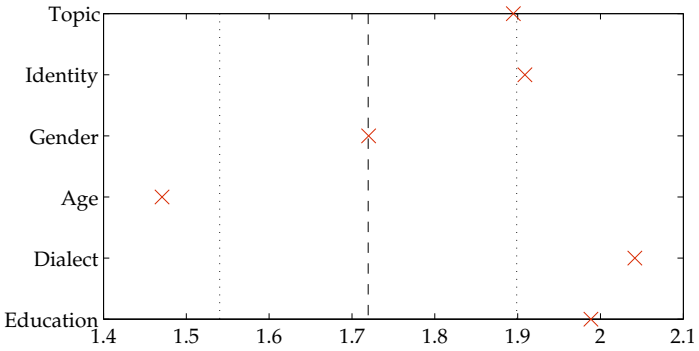




Figure 3: Analyses of the effect of introduced linguistic features

## 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.

## Acknowledgements

We would like to thank Prof. Manning 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.

## 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.