# CS229 Final Project

## Re-Alignment Improvements for Deep Neural Networks on Speech Recognition Systems

*Firas Abuzaid*

### *Abstract*

The task of automatic speech recognition has traditionally been accomplished by using Hidden Markov Models (HMM), which are effective in modeling time-varying spectral vector sequences [1]. Gaussian Mixture Models (GMM) have been used to determine how well each state of the HMM fits a frame or a short window of frames of coefficients that represents the acoustic input. However, recent research has shown that current speech recognition systems that use Deep Neural Networks (DNN) with many hidden layers can outperform GMMs on a variety of speech recognition benchmarks, sometimes by a large margin [2], [3]. We investigate further possible improvements of the DNN-HMM hybrid, by examining the role of forced alignments in the training algorithm.

### *Introduction*

In the context of automatic speech recognition, the GMM-HMM hybrid approach has a serious shortcoming – GMMs are statistically inefficient for modeling data that lie on or near a non-linear manifold in the data space. Because speech is typically produced by modulating a relatively small number of parameters, it has been hypothesized that the true underlying structure is much lower-dimensional than is immediately apparent in a window that contains lots of coefficients.

In contrast with GMMs, Deep Neural Networks have the potential to learn much better models of data that lie on or near a non-linear manifold. Many studies have confirmed this hypothesis, with DNN systems outperforming GMMs in ASR by approximately 6% for Individual Word Error Rates (IWER) [2]. These networks are trained to optimize a given training objective function using the standard error back-propagation procedure. In a DNN-HMM hybrid system, the DNN is trained to provide posterior probability estimates for the different HMM states. Typically, cross-entropy is used as the objective function, and the optimization is done through stochastic gradient descent (SGD) [4], [5]. For any given objective, the important quantity to calculate is its gradient with respect to the activations at the output layer. The gradients for all the parameters of the network can be derived from this one quantity based on the back-propagation procedure.

One significant hurdle in training speech recognition systems is determining the appropriate alignment between word sequences and acoustic observations. Typically, the acoustic data is divided into frames, with each frame approximately 15-25 ms in size. These frames must then be aligned with the sequence of words in the training set [1]. Usually, to determine these alignments, we start by obtaining an initial “forced” alignment, thus giving us a starting point to improve our posterior probabilities by training the DNN over multiple epochs. Specifically, we start by using an initial GMM-HMM baseline to assign feature vectors in our training set to different HMM states using the Viterbi algorithm. The Viterbi algorithm effectively chooses the most likely state sequences in our HMM model that corresponds to the correct observation sequences in our training data. Thus, the assignment is based on the most likely path through our initial composite HMM model.

Once we effectively “force” an alignment of the acoustics with the word transcripts in our training set, and we run the Viterbi algorithm to pass through those specific state sequences, these initial alignments can then be used by the DNN to train and improve our composite HMM model further (via back-propagation) and obtain more accurate log-likelihood probabilities for our HMM.

It has been shown that this forced alignment approach is the accepted method for HMM-based speech recognition systems. An interesting area of research, then, has focused on whether this same procedure can be used throughout multiple epochs of the DNN. Specifically, we’re investigating whether re-computing these forced alignments after every epoch of DNN training will lead to a further increase in

accuracy for our speech recognition system. Recent research – see Vesely et al.'s (2013) study [6] – demonstrates that computing these re-alignments could indeed be effective in reducing the Word Error Rate (WER) of these speech recognition systems.

### *Experiments*

We ran two sets of experiments to test our hypothesis. For our first set of experiments, we started with a smaller training set, simply to get an early indication of whether our hypothesis was correct. We trained the DNN on 60 hours of telephone conversation recordings from the Switchboard corpus, and we configured our neural network to have 5 hidden layers with 1200 neurons. We trained the DNN for 4 epochs.

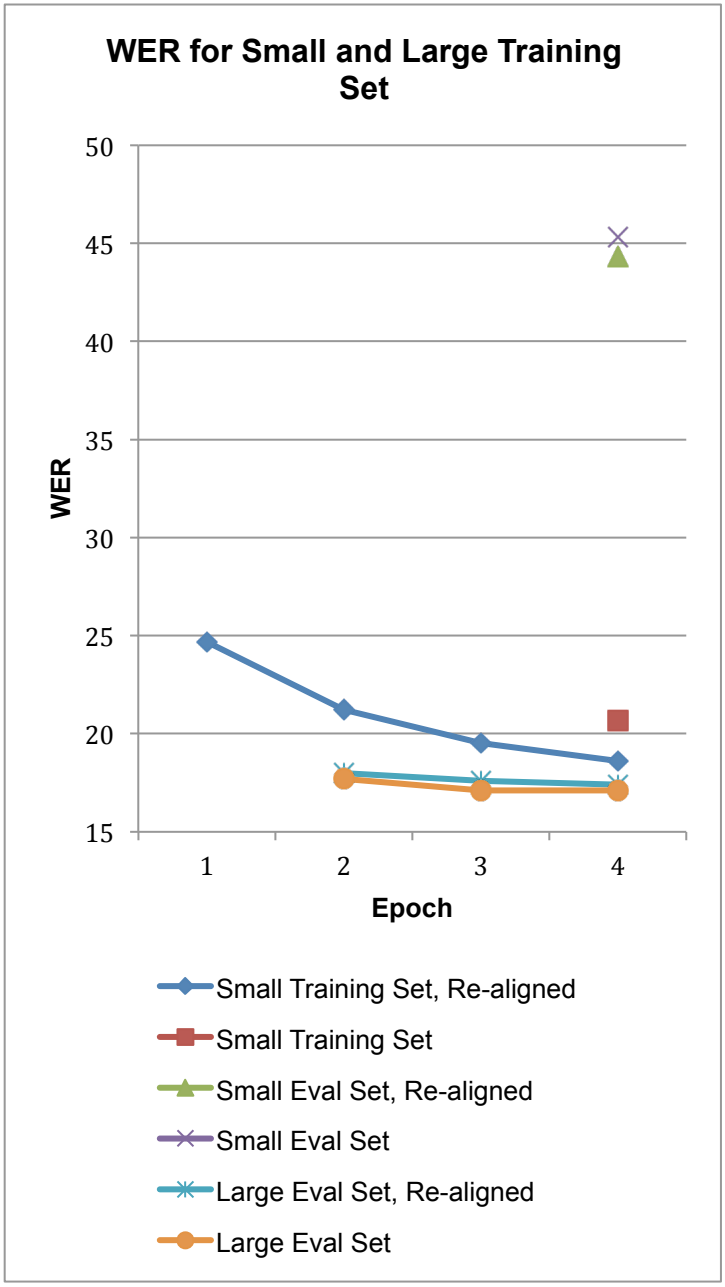
To test our hypothesis, we computed a re-alignment after each epoch, using the Viterbi algorithm that's used initially to compute the "forced" alignment before the first epoch. We simultaneously ran a separate DNN that did not compute these re-alignments after each epoch. We then compare the accuracy for both recognizers by evaluating them on the training set and the test set provided in the Kaldi open source toolkit [7].

Our second experiment was identical to the first, except we trained on a much larger corpus of training data – 300 hours from Switchboard. We also modified the neural network to contain 2048 neurons instead of 1200.

To implement our DNN speech recognizer, we used the Kaldi-Stanford codebase, a variant of the Kaldi open source toolkit, maintained by Professor Andrew Ng's Deep Learning research group.[1]

### *Results*

We measured the WER and IWER [8] for our speech recognizer, on both the small and large training set. Our recognizer was evaluated against its training set, as well against a corresponding evaluation set (either small or large) provided by the Kaldi toolkit. Our results for the two different experiments are as follows:




As the results above indicate, our hypothesis was correct for the smaller training set, but was incorrect for the larger training set, a somewhat surprising result.

### *Analysis*

Examining the results of our experiments, we see that there is a significant gap between the WER when evaluated on the small training set versus the small evaluation set; clearly, limited value can be gleaned from the experiments of the small training set. The results for the large training set, on the hand, do not

[1] Note: This project was done in conjunction with Professor Andrew Ng's research group, specifically with graduate students Andrew Maas and Christopher Lengerich.

give us a strong indication of whether computing re-alignments improves the accuracy of our speech recognizer. The WER delta between the re-aligned DNN and the non-re-aligned DNN is within the noise threshold and is thus statistically insignificant – we cannot say with any certainty that re-alignment improved our WER. More importantly, the rate of improvement over epochs between the experiments is nearly the same, which means that computing the re-alignments doesn't necessarily accelerate the training improvement of the DNN as we had hoped.

One potential explanation for this result could be that the hyper-parameters chosen for our larger neural network, particularly the number of neurons, need to be tuned. Modifying the network topology in this fashion could correct for any over-fitting in the Acoustic Model, which is more likely in the case of the larger training set with a larger network size and, thus, higher variance.

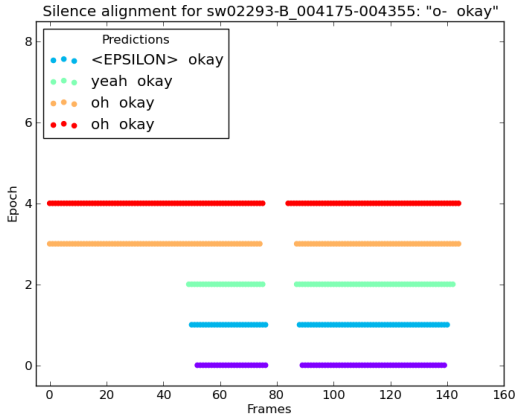
Another factor that needs to be examined is the number of epochs; it's unclear based on the data whether we've reached our optimization objective by the end of our training. The results for the small training set indicate that more epochs are needed, but the larger training set again is less clear in this regard.

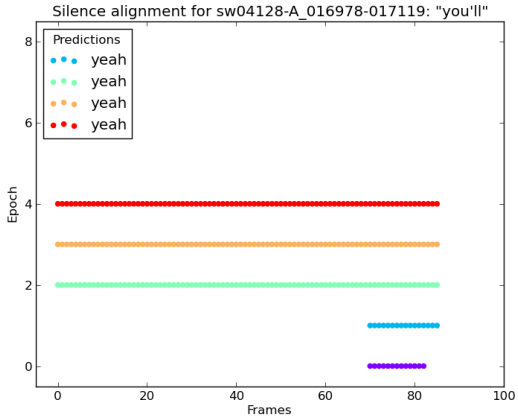
**Alignments**

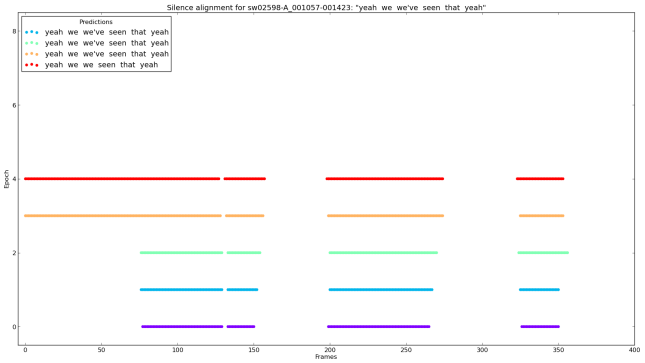
To better understand the performance of our DNN, we decided to closely examine the shift in alignments across epochs during the training stage. In the computed alignments, each frame gets assigned to a senone, which maps to a particular center phone. Analyzing the shift in phones between epochs is crucial in determining whether the re-alignments lead to a substantive reduction in WER.

In particular, the `silence` phone (denoted as `sil`) is quite significant, as these frames effectively become the demarcations between lexical tokens in our hypothesis. Our intuition is the following: if an alignment for a particular utterance undergoes a significant shift in its labeling of the `silence` phone, then this should lead to a more accurate prediction from the speech recognizer.

We generated alignment plots for the various utterances in our corpus to visually track how the alignment of the `silence` phone was shifting between epochs – the following are three example plots taken from the three utterances that exhibited the most variation in their alignments, along with the corresponding hypotheses from our speech recognizer for epochs 1 through 4 (note that this was only done for the small training set):



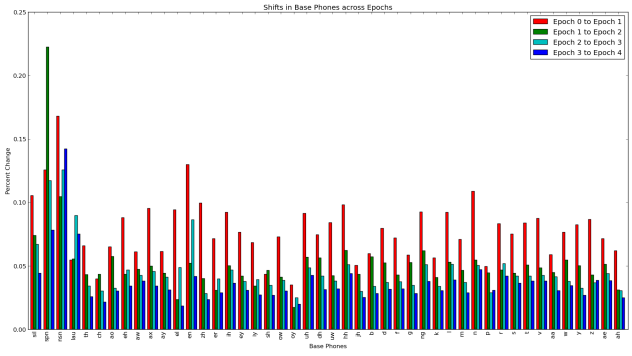




Based on this analysis, our theory is partially correct. The first example above certainly corroborates it – as the alignments change, we get a much more accurate prediction of the utterance. But the other two examples indicate this is not always the case; in the second example, the prediction doesn't change at all, despite the drastic shift in alignment, and in the last

example, the prediction becomes worse after a shift in the `silence` alignment.

To explain this behavior, there are a few possible theories to consider. One is that the length of the utterance must be factored into the analysis: utterances of shorter length (i.e. one or two words) tend be dominated by long stretches of silence followed by long stretches of non-silence, whereas longer utterances have smaller but more frequent moments of silence. We may see more dramatic shifts in silence for smaller utterances, but this doesn't necessarily lead to a difference in the Acoustic Model. On the other hand, a slight shift in silence could have a tremendous effect for larger utterances.

We also collected more general statistics about the shifts across epochs for all phones, not just the `silence` phone. The plot below shows the normalized frequency in base phone shifts for any given phone and epoch (note that this was only done for the small training set):




Several important things stand out from this data. Firstly, for the vast majority of the phones, the frequency of change decreases over the course of the epochs. Also of significance is that the `silence`, noise (both `spoken` and `non-spoken`), and `laughter` phones dominate these shifts, especially when compared to vowel and consonant phones. Particularly noteworthy is the abnormally high frequency in the first-second epoch transition (colored in green) for `spn`, which is spoken noise.

Amongst vowel and consonant phones, the `en` and `n` phones also shifted frequently, especially in the first epoch. In general, phones that had a combination of vowel and consonant sounds shifted more frequently, which could be attributed to the search for distinct syllables within words by our speech recognition system.

One possible experiment worth conducting in the future is to delay re-alignment computation until the second, third, or even fourth epoch. The magnitude of change in the earlier epochs indicates that perhaps these alignments are corrupted by the initial probabilities in the HMM model. Waiting until the second or third epoch, when the state probabilities in the HMM model are more robust due to the multiple epochs of training could lead to more accurate alignment shifts and, thus, better improvement in the WER. This experiment is especially worthwhile, since the re-alignment computation increases the runtime complexity of the training stage of our algorithm.

**Part-of-Speech Tagging**

As an additional vector of error analysis, we decided to measure the grammatical accuracy of our speech recognizers – how often were the erroneous hypotheses being made by our recognizer grammatically incorrect (e.g. confusing similar-sounding words such as 'money' and 'many' that have different grammatical meanings)?

Using the Stanford NLP group's Part-Of-Speech Tagger [9], we tagged each word in both the Reference utterance and the Hypothesis utterance, and then calculated the fraction of substitution errors that also differed in the corresponding part-of-speech tag. The results are as follows (note that this was only done for the small training set):

**Small Training Set – Re-alignment Experiments**

| Epoch | POS Error Rate |
|---|---|
| 1 | 84.04% |
| 2 | 83.26% |
| 3 | 82.46% |
| 4 | 81.40% |

The significance here is the magnitude of the error; the vast majority of the substitution errors made by our speech recognizer are grammar-related. (It is worth noting, of course, that human speech is often grammatically incorrect, which this analysis does not account for.) This analysis demonstrates that, although our experiments focus on addressing the Acoustic Model, perhaps our attention should shift to

improving the Language Model. In particular, this could easily address homonym confusion. This error, made by our recognizer in the small training set experiments, is particularly instructive:

| |
|---|
| ***Reference***: “we've been wanting to start camping again this year **too** uh my oldest” |
| ***Hypothesis***: “we've been wanting to start camping again this year **to** uh my oldest” |

In these situations, the Acoustic Model simply cannot distinguish between these two words – if we improve the Language Model, however, we could be able to address these errors.

### *Future Work*

In the future, we’d like to further examine the effect of further epochal training for our DNN, as well as fine-tuning the hyper-parameters of the neural network, particularly the network size. We’d also like to analyze the alignments in greater detail, and examine the relationship between various shifts in silence, noise and laughter alignments with utterance length. Also worth exploring are the effects of eschewing early stage re-alignment; perhaps we could see improvement in the WER by delaying re-alignment computation until the second, third, or even fourth epoch. Lastly, we’d like to explore alternative language models – the work of Mikolov et al. [10] demonstrates that the language model for DNNs can be improved through the use of Recurrent Neural Network (RNN) Language Models, so coupling those improvements with our re-alignments could yield a further increase in accuracy.

### *References*


[1] M.J.F. Gales and S.J. Young (2008). The Application of Hidden Markov
Models in Speech Recognition. Foundations and Trends in Signal
Processing.

[2] Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., et al. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition. IEEE Signal Processing Magazine, 29(November), 82–97.

[3] Dahl, G., Yu, D., Deng, L., & Acero, A. (2011). Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing, Special Issue on Deep Learning for Speech and Langauge Processing, 1–13.

[4] Povey, D., & Woodland, P. C. (2002). Minimum Phone Error and I-Smoothing for Improved Discriminative Training. ICASSP.

[5] Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., & Visweswariah, K. (2008). Boosted MMI for model and feature-space discriminative training. ICASSP (pp. 4057–4060). IEEE.

[6] “Sequence-discriminative training of deep neural networks”, K. Vesely, A. Ghoshal, L. Burget and D. Povey, to appear in: Interspeech 2013

[7] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Veselý, K., Goel, N., et al. (2011). The kaldi speech recognition toolkit. ASRU.

[8] Sharon Goldwater, Dan Jurafsky, Christopher D. Manning, Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates, Speech Communication, Volume 52, Issue 3, March 2010, pp. 181-200.

[9] Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003, pp. 252-259.

[10] Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky,Sanjeev Khudanpur: Recurrent neural network based language model, In: Proc. INTERSPEECH 2010