Abstract

Our project looks into the ways that we can improve forced alignments. Taken as a whole, it introduces two approaches: one in which an expanded lexicon is used instead of a regular lexicon (in our case CMU Dict); the other in trying to see if acoustic model rigidity affects the accuracy in forced alignments.

This project uses a largely expanded lexicon that has been expanded via a variety of linguistic rules, and will be elaborated upon in the Lexicon section. It also uses three corpora: the Fisher English corpus, the Librispeech corpus, and the SCOTUS corpus. Our acoustic models were trained on these three corpora, and in testing our results, we use both the expanded and regular lexicon across all three acoustic models.

We discovered that having a large expanded lexicon decreased the accuracy of our results in label accuracy as well as boundary accuracy. We also discovered that the rigidity of the acoustic model does indeed have an effect on accuracy, as the more rigid corpora performed better than the more flexible corpora.

1 Introduction

The process of alignment aims to identify phones in sound files, recognizing both its label and its boundary. As hand alignment is extremely inefficient and time consuming, many are interested instead in the process of forced alignment. Through this process, one generates both a sequence of recognized phones and the times one calculates that these phones begin and end. Through a comparison with the original transcript - as well as with the known times that phones began and ended - one is able to judge the accuracy of their forced aligner.

The first step is to generate each pronunciation of each words in the transcript and to compile this information in a lexicon. Then, for each pronunciation, a Hidden Markov Model (HMM) is generated where each phone is decomposed into a series of hidden subphone states that are activated by acoustic features in short frames of speech following a pre-trained acoustic model. In decoding the speech signal, the Viterbi algorithm is used to find the maximum-likelihood path through the concatenated word-level HMMs, thus generating the alignment.

Our aligner relied on pronunciations generated both by the CMU Dict and an enriched lexicon, using different acoustic models that we trained that vary in flexibility. We hope to see the variability in results across acoustic models and the standard vs. enriched lexicon.

While the benefits of an enriched lexicon should help capture accents and conversational speech, much of which is not as particularly and eloquently pronounced as formal speech, its drawbacks include bloating the lexicon and accepting rarer phones.

2 Background/Related Work

Most of the inspiration for this project came from a previous 224s group project [1], which examined techniques to improve forced alignment in conversational speech. They explored different options in improving forced alignment, including enriching their lexicon and implementing boundary corrections for specific phones after alignment.

We found that several aspects of this paper were
of particular interest to us. The first was the enriched lexicon itself: we hypothesized that, by associating each of the words with more colloquial ways that they are pronounced in conversations, it would be easier to correctly identify these words, especially when they were not spoken in the exact same way as the expected CMU dict pronunciation. The second item of interest to us were the specific models that were used: we hypothesized that the accuracy of forced alignments would increase on stricter models. This became the foundation for our project. [1]

A second paper that was of interest [2] discussed the difficulty introduced by attempting to force align natural, unprompted speech. This was an intriguing concept, as our project also focused on aligning conversational speech and comparing accuracies between this type of speech and a more structured, clearly enunciated form of speech. It argued that data within the field of speech recognition, especially concerning phonetics, should shift to contain as realistic speech as one could get and should not be designed for specific investigation. This was an interesting idea, and became one of the reasons that we decided to train our models on corpora of varying rigidity. It also detailed the approaches that the authors took in order to try and improve their recognition models and how it ultimately decided that more work needed to be done in the area of natural realistic speech.

A last paper of interest [3] discussed different techniques for forced alignment and the way that each treats boundary alignment. More importantly, it analyzed techniques that it undertook in order to try and improve the boundary alignments in an HMM-based forced alignment model. For one, it introduced a more sophisticated boundary correction model and a wider array of features, including context and duration. It also took speaker-dependent features into account. This paper seemed very interesting, as it was similar to the premise for the work that we had proposed. However, while we were ultimately unable to incorporate this into our project, it did influence some of our ideas for future work that builds off of our results.

3 Approach

As demonstrated in the figure, forced alignment requires several inputs in order to run. For each sound clip that you want to force align, you need that wav file, that transcript, a lexicon mapping each word in the transcript to different pronunciations, in terms of phones, and an acoustic model. It outputs each phone that it recognizes, as well as the end time where it believes that that phone has stopped.

![Figure 1: Forced alignment flowchart](image)

In order to test our hypothesis that forced alignment would be more successful on more rigid models, we decided to train models on corpora of varying flexibility, using both the original CMU dictionary and our newly enriched lexicon.

Our approach itself was as follows: firstly, the expanded lexicon was created using the rules highlighted in the Lexicon section. Then, in order to test our hypothesis that forced alignment would be more successful on more rigid models, we decided to train models on corpora of varying flexibility, using both the original CMU dictionary and our newly enriched lexicon. We obtained the corpora and trained them without the expanded lexicon, as we decided to only use the expanded lexicon for testing and for making language models during testing. Though we trained neural networks for the acoustic models, we were unable to figure out how to use Kaldis forced aligner with a neural network model, so we used triphone models instead. However, because we are only using the models on a comparative basis, using triphone models across the corpora still works and illustrates our point.

4 Experiment

To begin, we first had to train models, define our lexicon, determine a set of correctly aligned
data on which to test our results, and determine a method through which we could test our results.

### 4.1 Models

In order to create models of varying flexibility, we decided to train them on corpora with differing levels of rigidity.

The first corpus that we chose was the Fisher English Corpus. This consists of 1000 hours of conversational speech and includes a variety of different speakers, accents, and topics. As this resulting model was the most flexible, we expected it to perform the worst.

The second corpus was the LibriSpeech Corpus, which consists of 1000 hours of read audiobooks from a variety of different genres. This model was more rigid, as the speakers were no longer conversational but instead enunciated clearly and formally, making their pronunciations more likely to line up with the standard pronunciations of the CMU dictionary. However, it still incorporates a very high number of speakers.

The last corpus that we used was the SCOTUS Corpus, which is made up of 38 years of recordings and transcripts of speeches given by the United States Supreme Court. We expected this one to be the most rigid model; not only is the language used in the recordings is formal and clearly enunciated, but there are only 9 speakers involved in the corpus.

### 4.2 Lexicon

We compared two lexicons: an original and an enriched. The original was based off of the CMU dict. The enriched lexicon was made according to these rules [1]:

- reduction of unstressed IY0 and UW0 to IH0 and UH0 respectively;
- centralization of any unstressed vowel to schwa, AX (AH0 in the default acoustic model);
- g-dropping, i.e. substitution of N for NG in the word-final sequence IH0 NG; simplification of t-initial alveolar obstruent clusters, i.e. deletion of T preceding S, CH or SH;
- deletion of a series of unstressed syllables starting at the left edge of a word;
- deletion of word-final T and D;
- deletion of word-initial HH;
- centralization of any vowel to schwa (in function words only);
- deletion of any vowel (in function words only).

Each of these rules defined a list of new pronunciations for each of the words in the original lexicon; our enriched lexicon compiled each of the new pronunciations generated.

While the benefits of an enriched lexicon should help capture accents and conversational speech, much of which is not as particularly and eloquently pronounced as formal speech, its drawbacks include bloating the lexicon and accepting rarer phones.

### 4.3 Test Set

For our test set we decided to use the ICSI Gold Standard, which comes from a small subset of the Switchboard corpus. Because this set of phones had been hand aligned, ensuring that each phone label and boundary were correct, compared the outputs of our models on this set to the hand aligned solutions to determine accuracy.

### 4.4 Evaluation and Testing Methodology

We tested several phenomena: label accuracy, boundary accuracy, and total accuracy.

**Label accuracy** was defined as the number of phones that our model had correctly recognized. This accuracy was measured by the Word Error Rate and tracked the number of insertions, deletions, and substitutions between our phones and the actual phones. We used the Levenshtein Distance to calculate WER, as this determines the similarity between two strings.

**Boundary accuracy** recognizes all generated boundaries that occur within x ms of any actual boundary, regardless of the phone label associated with it. For obvious reasons, if multiple boundaries occur within x ms of any particular boundary, only one of those is marked and counted as accurately defined.

**Total accuracy** was defined as the number of phones whose labels and boundaries were both marked as correct.
4.5 Results

After running our experiment, we produced these results:

<table>
<thead>
<tr>
<th>Model</th>
<th>Boundary Accuracy</th>
<th>Total Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Fisher</td>
<td>0.32</td>
<td>0.36</td>
</tr>
<tr>
<td>Fisher enhanced</td>
<td>0.23</td>
<td>0.41</td>
</tr>
<tr>
<td>Librispeech</td>
<td>0.32</td>
<td>0.52</td>
</tr>
<tr>
<td>Librispeech enhanced</td>
<td>0.24</td>
<td>0.41</td>
</tr>
<tr>
<td>SCOTUS</td>
<td>0.43</td>
<td>0.65</td>
</tr>
<tr>
<td>SCOTUS enhanced</td>
<td>0.43</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Figure 2: Model vs Boundary Accuracy (within 10, 20, 50, or 100 ms)

Interestingly, the results show that the number of insertion errors increased dramatically while the number of deletion errors decreased dramatically.

<table>
<thead>
<tr>
<th>Model</th>
<th>WER</th>
<th>Insertions</th>
<th>Deletions</th>
<th>Substitutions</th>
<th>Total Errors</th>
<th>Max Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fisher</td>
<td>0.46</td>
<td>0.54</td>
<td>0.61</td>
<td>0.03</td>
<td>0.56</td>
<td>2136</td>
</tr>
<tr>
<td>Fisher enhanced</td>
<td>0.44</td>
<td>0.44</td>
<td>0.54</td>
<td>0.01</td>
<td>0.51</td>
<td>1130</td>
</tr>
<tr>
<td>Librispeech</td>
<td>0.43</td>
<td>0.37</td>
<td>0.46</td>
<td>0.02</td>
<td>0.52</td>
<td>705</td>
</tr>
<tr>
<td>Librispeech enhanced</td>
<td>0.46</td>
<td>0.34</td>
<td>0.31</td>
<td>0.01</td>
<td>0.56</td>
<td>1857</td>
</tr>
<tr>
<td>SCOTUS</td>
<td>0.63</td>
<td>0.39</td>
<td>0.33</td>
<td>0.02</td>
<td>0.43</td>
<td>2044</td>
</tr>
<tr>
<td>SCOTUS enhanced</td>
<td>0.63</td>
<td>0.39</td>
<td>0.33</td>
<td>0.02</td>
<td>0.43</td>
<td>2044</td>
</tr>
</tbody>
</table>

Figure 3: Model vs WER (where the insertions, deletions, and substitutions are formatted as percentages of the total errors)

This means that the Viterbi algorithm chose a pronunciation that was much more likely to have phones added than it was to choose a pronunciation with phones deleted from it. Thus, Viterbi actually believed the words to be more complicated than they actually were. However, we do see that the difference between the insertion and deletion error rates between regular and enhanced lexicons decreases significantly when only looking at the SCOTUS model (which is the most rigid), so, as expected, it seems that the more flexible the corpus, the more noise and thus more inaccurate the results from Viterbi were.

In terms of boundary accuracies, we found that, in the cases of the Fisher and Librispeech model, the original lexicons performed better than the enhanced lexicons as well. This difference between the performance of the original and enhanced lexicon was absent within SCOTUS, and in the latter cases the enhanced lexicon even found more accuracies than the original lexicon. However, we must take into account that the 50 ms and 100 ms margins of error are very forgiving.

Figure 4: Model vs proportion of insertions, deletions, and substitutions

Figure 5: Model vs Boundary Accuracies (within specified margin of error in ms)

This phenomena where the enriched lexicon did not accurately identify as many boundaries as the original lexicon was not a surprise. Because the enriched lexicons was inundated with noise, the model could often choose the wrong pronunciation, which often consisted of a different number of phones than the original. For that reason, the boundaries within that word would not be able to line up and would often take over the boundaries that actually belonged to the phones of a different phone. As we increased the margin of error, the boundary accuracy did increase, as to be expected. However, the margin of error that we were most interested in was the 20 ms margin.

As demonstrated in the figure below about total accuracy, which requires both the boundary label and the phone label to be completely matching, SCOTUS, the most rigid corpus, outperformed the other two corpora. This is due to a combination of the above two factors; the more flexible mod-
els already had a smaller proportion of correctly identified labels boundaries. Within this subset there was a smaller proportion of correctly identified phones.

Figure 6: Model vs Total Accuracies (meaning that a phone has an accurate label and an accurate boundary)

However, the results also show that, in all cases, the original lexicons outperform the enriched lexicons. In the case of Fisher and Librispeech, it outperformed the enriched lexicon by a substantial amount. In the case of SCOTUS, the original also outperformed the enriched lexicon. However, we were genuinely surprised at how much better SCOTUS performed relative to the other corpora, and further exploration may be needed to look into the intrinsics of the results for total accuracy.

### 5 Conclusion

As we hypothesized, acoustic model rigidity did affect our results. Looking at the results, we immediately see that the original CMU dict lexicon performed better than the enriched lexicon across all corpora. The CMU standard pronunciation is good enough for most purposes, as evidenced by the fact that our Word Accuracy Rate for just the CMU dict produced completely accurate phones for between 60% to 66% of the data. The enriched lexicon offers multiple different pronunciations for each word, forcing the aligner to choose one of these many options, thereby lowering its word accuracy rate.

Although the Viterbi algorithm ensures that the most likely pronunciation for each word is given, this algorithm highly relies on the acoustic model to generate the likelihood of each pronunciation given the signal. However, the acoustic model can negatively affect the Viterbi algorithm if it is too flexible, and our results confirm this fact. The SCOTUS corpus is the most rigid of our corpora, as it uses only nine different speakers, whereas the Librispeech and Fisher English corpora use an increased number of speakers, thereby expanding the flexibility of the acoustic model. In addition, the Fisher English corpus and the Librispeech corpus talk about a much more expansive area of topics: the Librispeech corpus uses thousands of hours of audiobooks from many different genres, and the Fisher English corpus participants were specifically given randomly-assigned topics. In fact, we see that the Fisher English corpus is worse than the Librispeech corpus, precisely because Fisher English uses phone conversations, which are more colloquial and relaxed, rather than clearly and precisely enunciated. Many of these informal pronunciations are not accounted for in the majority of lexicons, while expanding the lexicons leads to the above problem.

This flexibility can be beneficial as it does allow for variability in signal. At the same time, its drawbacks include an inability to distinguish between differences in pronunciation variants, which is what occurred in our case. From our results, we see that the more flexible the model gets, the worse the results tend to be. We also see a growing gap in accuracy between the standard CMU dict and the enhanced lexicon as the flexibility of the corpus is increased.

We have also learned that blindly expanding the lexicon by a large array of linguistic rules is not enough to improve results in forced alignment, as can be seen by the decrease in results across all three of our corpora. Adding too many pronunciations creates too much noise, making it difficult for Viterbi to choose the correct pronunciation given an expansive list of different pronunciations. Essentially, the more pronunciations we added, the more chance that Viterbi that it chooses an incorrect variant of a word.

Ultimately, our results are useful because they suggest a correlation between the rigidity of an acoustic model and the quality of results for forced alignment, both in boundary accuracy and in label accuracy. In addition, we discovered that using a dramatically enhanced lexicon decreased the accuracy of our results for all three of the corpora on which we tested. Future research
can incorporate this knowledge to improve their forced alignment and generate better results, thus advancing acoustic speech recognition and the field of spoken language processing as a whole.

5.1 Areas for Further Study
We have seen that merely expanding the lexicon alone does not help with accuracy, and choosing the correct pronunciation of a word is difficult. We foresee a few ways that forced alignment can be improved for future results, even with an expanded lexicon.

Firstly, we can continue the process of finding a more rigid and less flexible acoustic model, as the results from our tests do demonstrate that the less flexible acoustic models tend to generate better results. The SCOTUS model, which is rigid in its formal speech and its, respectively, very limited, number of speakers, is much less flexible than phone conversations from a variety of different speakers. As proven, and as expected, the difference between the expanded and default results were much smaller.

Secondly, we can attribute weights in the form of different probabilities to the pronunciations in our lexicon, as each variant is not equally likely to occur. As we see from our results, the lexicon generated from the CMU Dict pronunciations alone is generally correct around 63% of the time, as measured by the Word Accuracy Rate. By having our lexicon greater emphasize the most likely pronunciation probabilities, it could greatly benefit the results by clearing away some of the noise.

Thirdly, we believe that adapting the enhanced lexicon by only using subsets of the rules could benefit the results as well. Adding each of the pronunciations that we generated by using all of the rules to the lexicon created too much noise; however, using only a subset of these pronunciations can be beneficial. This is because Viterbi can better figure out the correct pronunciation because there are not too many pronunciations. Also, the correct pronunciation is now more likely to be in the lexicon itself, which is not guaranteed when using only the CMU dict.

Finally, because we found that the expanded lexicon itself generally resulted in worse results, we can improve forced alignment by creating a lexicon-like CMU dict with one or two pronunciations maximum for each word and based on results from data that we know will be very similar to our testing data. This way, we can basically have a specific, CMU-like dict that is specifically tailored to our results. This may be expensive to create, but we think will result in greatly improved results, as it allows for Viterbi to choose between a much smaller set of pronunciations and increases the likelihood that the correct pronunciation is still included in the lexicon.

6 Acknowledgements
This project was supported by Simon Todd, whose previous research formed the foundation on which our project was based. We also thank him for providing us with the ICSI Gold Standard data, which became the test set for our current project, along with other tools that we incorporated into our modelling and testing.

References