A STATE-TYING APPROACH TO BUILDING SYLLABLE HMMS

Darryl Stewart, Ji Ming, Philip Hanna, F. Jack Smith

Queens University Belfast, N. Ireland

ABSTRACT

A severe sparse data problem is faced when building HMMs of syllable units due to the uneven distribution of syllables in natural speech. In this paper, we present a novel approach to building syllable HMMs which attempts to overcome this problem. The method involves tying the states in syllable models using a bottom-up clustering algorithm and a state similarity measure employing phonetic information. Experiments using the new approach on the TIMIT database show that it improves the recognition accuracy of syllable HMMs. We present encouraging results which show that the state-tying method is used in conjunction with a Multi-Model [11] approach to acoustic modeling a syllable identification accuracy of 53.4% can be achieved. This equates to a phoneme accuracy of 72.8% which is comparable with the best results achieved using triphone HMMs.

1. INTRODUCTION

Speech recognition systems have predominantly used phoneme-sized Hidden Markov Models (HMMs) to recognize the speech signal. Research using these models over the last 10-15 years has provided superb recognition accuracies on constrained tasks such as the Wall Street Journal database. However, the same performance has not been achievable on more difficult tasks like the Switchboard database. One of the reasons identified for the failure of these units is that they have extremely short durations and hence only contain a small amount of discriminative information. Furthermore, it is quite common for phonemes to be deleted in conversational speech, which makes mapping phoneme strings onto lexical pronunciations difficult.

It has been proposed in the literature since as early as 1975 that syllables would potentially be the most effective unit to use for automatic speech recognition compared to words or phonemes due to their strong links with human speech production and perception [1]. Nevertheless, phonemes offered a computationally less expensive alternative and provided reliable accuracies on the constrained recognition tasks attempted at that time. More recently, computing power has increased to the extent that the original incentive for using phonemes instead of syllables has been removed. This, coupled with poor recognition accuracies on the Switchboard database has led to a re-examination of syllables as the modeling unit within automatic speech recognition systems [2, 3, 4 & 5].

Researchers have shown that syllable HMMs can provide better recognition accuracies than monophone [3] and even triphone HMMs [4 & 6]. However, the same researchers and others [7] have also shown that syllables are unevenly distributed in speech data, with the result that some syllables appear frequently and some appear very infrequently. Indeed, it was shown that in the TIMIT database, some syllables that appear in the testing data are not seen at all in the training data. This raises the question of how to create accurate acoustic models for these infrequent or even unseen syllables. In the past, the problem has been avoided by only building models of a subset of all syllables, i.e. those which have high enough frequencies in the training data. However, it was recognized that this reduced the potential benefits that could be gained from any syllable-based speech recognizer. In this paper, we describe the development of a novel state-tying approach to building syllable HMMs, which attempts to overcome the sparse data problem, thereby allowing more accurate models of all syllables to be created.

The paper is organized as follows: the development of the new modeling approach is described in Section 2, followed by the details and results of an experimental investigation in Section 3. Finally a discussion of the results and the conclusions which can be drawn from them are provided in Sections 4 and 5 respectively.

2. SYLLABLE STATE-TYING

The problem of sparse data is not a new one in acoustic modeling. There is a similar problem faced when attempting to create HMMs of triphone units. The general solution to the problem has been to allow different models to share the same states through a process known as state-tying: thereby allowing the available training data to be shared more evenly amongst the models. In the case of modeling unseen triphone units, the states are effectively tied by comparing the phonetic context of the unseen states with the phonetic contexts of the other states in the system that have sufficient training data. The unseen states are then tied to those states which were found with the most similar context. Several measures have been employed to assess the similarity of phonetic contexts including mutual information [8], allophonic classes [9] and hierarchical phonetic classes [10]. The state-tying method developed is based upon a similar strategy to that used for triphones.

2.1 Model Topology

In this work, the number of states in each syllable model was related to the number of phonemes in the model. As in phoneme modeling approaches, each phoneme in the syllable models was allocated 3 states. For instance, the model for the syllable /r ah s v/ ("rest") was allocated 12 states. This state allocation
strategy was employed because it allows an assumption to be made about the phonetic context of each state in the model. For example, if the states are allocated according to this schema then the 4th state in the model of /r eh s t/ can be assumed to represent the 1st state of the /eh/ phoneme within the model and to have a left context of /r/ and a right context of /s t/, as illustrated in Figure 1.

The weights used in this work were calculated using an exponential decay function as shown in Equation 1.

\[ Weight = e^{-\mu(p-1)} \]  

Where:
- \( p \) is the priority given to a particular phoneme
- \( \mu \) controls the rate of decay
- \( e \) is the base of natural logarithms
- \( \mu \) is a parameter that can be adjusted to control the rate of decay.

2.3 Clustering algorithm

The algorithm used in this work to cluster the similar states together was based upon a nearest neighbor merge and sort process. The prerequisite for the clustering algorithm is a list of all the states in the system with their accompanying frequency of occurrence in the training data, sorted according to the frequencies. This list of frequencies can be obtained by saving the state/frame occupancies generated during the initial training of the models. Clustering is then performed using the steps shown below:

1. While frequency of cluster \( x \) is lower than \( T_{end} \)
   a. Compare \( x \) with all clusters with min frequency \( T_{start} \)
   b. Find cluster \( y \) such that \( Similarity(x,y) \) is maximum
   c. Merge cluster \( x \) into \( y \)
   d. Sort list of clusters

   Where:
   - Cluster \( x \) is the least frequent cluster.
   - \( T_{end} \) is the minimum frequency that all clusters must have at the end of clustering.
   - \( T_{start} \) is the minimum frequency that a cluster must have before other clusters can be merged with it.
   - \( Similarity(x,y) \) is the function returning the similarity between clusters \( x \) and \( y \).

The \( T_{end} \) parameter is data dependent and is used to permit a balance between the available training data and the modeling resolution of the system. The \( T_{start} \) parameter affords some level of control over the shape that clusters can take by preventing clusters that are entirely made up of very infrequent states.

2.4 Syllable State Seeding

It was found through experimentation that when the syllable models were trained, the states did not automatically represent the phonemes in the syllables in the way that was shown in Figure 1. This was because some phonemes in the syllables were greeder than others. However, the success of the new approach relies heavily on the validity of the state allocation assumption being made in Figure 1. Therefore, a state seeding step was introduced to the training procedure in order to strengthen the assumption. This involved seeding the states of the syllable models using the states of monophone models trained on the same data. This also ensures that the states represent the phonemes for which they were allocated.
The 4 stages involved in seeding, training and tying the states of syllable models are listed below:

1. An initial set of 3 state monophone models with single mixture PDFs are trained.
2. The state output distributions of these monophone models are used to initialize the states of the syllable models. For example, the 3 states for the monophone model /ah/ are used to initialize the 4th, 5th and 6th states of the syllable model /ah/. The syllable models are then trained using Baum-Welch re-estimation.
3. The syllable model states can then be successfully clustered and tied using accurate phonetic details for each state.
4. The number of mixture components in each state can then be incremented.

3. EXPERIMENTAL INVESTIGATION

The experiments described in this section were designed to investigate the effectiveness of the state tying approach for building syllable HMMs and in particular to test the underlying assumptions upon which the method was based and to collect information about how each parameter in the method influenced its performance.

The experiments described in the following sections made use of the syllabified hand-labelled transcripts of the TIMIT database which resulted from the work described in [7]. The total vocabulary of syllables which were used in the experiments consisted of 9414 distinct syllables. This vocabulary consisted of all the unique syllables found in the training and testing data of the database. The following was true for all experiments:

- The sentences used 10 MFCC features along with the energy feature, which were augmented with accelerated spectral coefficients resulting in feature vectors containing 33 elements.
- The phoneme models used to seed the states of the syllable models consisted of the standard 39 phonemes used in the TIMIT database. Each model used single mixture PDFs and was trained using the hand labeled transcripts.
- The syllable-based experiments were identification experiments rather than recognition experiments. This meant that the syllable HMMs were used to identify discrete syllables with known boundaries. This was because the syllable models will be combined in future work with an automatic syllable detection algorithm similar to those described in [5].

3.1 Building state-tied syllable models

The prerequisite list of all the states in the syllable models along with their accompanying frequency of occurrence in the training data was compiled from the state/frame occupancies which were saved during the training of a set of untied syllable HMMs. It was found from this list that 16.5% of states were unseen in the training data. Furthermore, 92% of the states had a frequency of 30 occurrences or less. The states were then clustered using the new similarity measure and clustering algorithm. The rate of decay μ weights used in Equation 1 was set at 1 and various values of the two threshold parameters $T_{\text{ned}}$ and $T_{\text{ned}}$ were examined. Syllable HMMs consisting of single mixture PDFs were trained and then used to identify the syllables in the test sentences. A selection of the syllable identification rates produced in each case are shown in Table 1 with the best accuracy of 38.9% being highlighted.

<table>
<thead>
<tr>
<th>$T_{\text{ned}}$</th>
<th>30</th>
<th>40</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30.5%</td>
<td>29.8%</td>
<td>25.4%</td>
</tr>
<tr>
<td>5</td>
<td>37.3%</td>
<td>38.9%</td>
<td>38.0%</td>
</tr>
<tr>
<td>10</td>
<td>36.6%</td>
<td>37.5%</td>
<td>36.2%</td>
</tr>
</tbody>
</table>

When the experiment which provided the best accuracy shown in Table 1 was repeated with μ taking the values 0.5 and 2, the syllable accuracies deteriorated to 31.7% and 35.1% respectively; thereby showing the importance of finding the correct set of weights for the contextual elements.

The experiment that provided the best accuracy in Table 1 was repeated again; this time with the state seeding phase omitted. The syllable accuracy was 31.3% which was a 7.6% absolute reduction in accuracy which highlights the importance of state seeding to the effectiveness of the proposed modeling approach.

Following this, the best single mixture models were increased to 16 mixture components and the experiments were repeated. This resulted in an improved accuracy of 49.1%.

3.2 Multiple Modeling approach

It was shown in [11] that an improvement in model accuracy for syllable-sized speech units can be gained by combining the IFD-HMM (Interframe Dependent HMM) approach along with the standard HMM approach with multiple mixture components. IFD-HMMs provide a method of capturing more temporal information in acoustic models than standard HMMs. Therefore a series of experiments were carried out using various IFDHMM time sequences along with the standard HMMs consisting of 16 mixture components. The optimum time lag sequence found in these experiments was {+6, +4, +2, +2, +4, +6}, which produced the best syllable accuracy obtained so far of 53.4%.

3.3 Syllable-to-phoneme accuracy conversion

If syllable accuracies are compared directly with phoneme accuracies then the phoneme accuracies are always higher. This is because syllable accuracies are penalized due to the increased amount of information which must be recognized in order to correctly identify them. For example, if the syllable /p ih t/ was expected and the syllable /h ih t/ was observed then the syllable recognition accuracy would be 0%. However, in a similar circumstance, if the three phonemes /p ih t/ were expected in a phoneme recognition experiment but the phonemes /h ih t/ were observed, the phoneme accuracy would be 66%. The accuracies of a phoneme recognition experiment and syllable recognition
experiment cannot therefore be compared directly and the syllable accuracy must be converted to phoneme accuracy. The conversion process is achieved by treating each expected and observed syllable as a string of individual phonemes. The phoneme recognition accuracy for each of these phoneme strings can then be calculated using the same Dynamic Programming method used to calculate phoneme accuracies in phoneme recognition experiments. Using this conversion process, the best syllable accuracy of 53.4% is equivalent to a phoneme accuracy of 72.8%.

4. DISCUSSION

To the best of the authors’ knowledge, there have been no other syllable identification results published for the TIMIT database. However, having converted the identification rates in these experiments to phoneme accuracies, it is possible to compare them to the large body of results, which have been published from research using phoneme models and triphone models.

The best published results for TIMIT were gained by Ming et al. [12] who reported a phoneme recognition score of 75.6% using Bayesian triphone HMMs. Robinson [13] achieved a phoneme recognition score of 75.0% using recurrent neural networks. Other results reported include, 74.5% using nearest-neighbor modeling by Chang et al. [14] and 70% using segmental models by Zue et al. [15]. It should be noted that in all of the above experiments context-dependent phoneme models (mainly triphones) were used along with bigram language models.

In light of these results, the phoneme recognition rate of 72.8% achieved using the syllable models compares quite favorably.

5. CONCLUSION

Syllables are distributed unevenly in speech databases and many are too infrequent to enable robust HMMs to be made. In this paper a novel state-tying approach to building syllable HMMs is proposed which makes use of a bottom-up clustering algorithm and a state similarity measure employing phonetic information. Experiments using the TIMIT database show that the method improves the accuracy of syllable HMMs. Using the state tying method along with a Multiple Model approach, gives the best syllable identification accuracy of 53.4%, which equates to a phoneme recognition accuracy of 72.8%. Although, this has only been a preliminary study of syllable modeling, it is already producing results that compare favorably with the best accuracies obtained using triphone units which have been highly optimized over many years of constant research. Therefore, it is felt that syllable HMMs shows great promise for future improvement.

7. REFERENCES