Improved Topic Classification and Keyword Discovery using an HMM-based Speech Recognizer Trained without Supervision

Man-Hung Siu, Herbert Gish, Arthur Chan, William Belfield

Raytheon BBN Technologies
Speech and Language Processing Department
Cambridge, MA 02421, USA

Abstract

In our previous publication [1], we presented a new approach to HMM training, viz., training without supervision. We used an HMM trained without supervision for transcribing audio into self-organized units (SOUs) for the purpose of topic classification. In this paper we report improvements made to the system, including the use of context dependent acoustic models and lattice based features that together reduce the topic verification equal error rate from 12% to 7%. In additional to discussing the effectiveness of the SOU approach, we describe how we analyzed some selected SOU n-grams and found that they were highly correlated with keywords, demonstrating the ability of the SOU technology to discover topic relevant keywords.

Index Terms: unsupervised learning, topic classification, speech recognition.

1. Introduction

Many speech applications require large amounts of transcribed audio for training speech recognizers. For some languages or domains, large amounts of transcribed audio can be difficult to come by. Different approaches for speech recognition training have recently been proposed for using various amounts of limited resources, such as converting models from related languages, or bootstrapping with a small amount of transcribed data [2], [3], [4]. In our previous work [1], we focused on the problem of automatic speech recognition without any resources such as transcription nor lexicons. We developed an iterative unsupervised HMM training strategy where the HMM was used to transcribe the audio into a sequence of self-organized speech units (SOUs) using only untranscribed speech for training, and the resulting unit sequences were used for topic identification (TID). One significant advantage of completely unsupervised training is that there will not be any mis-match between training and test, because the untranscribed test data can, if needed, be added for acoustic training.

In this paper, we report our recent improvements in both HMM training and classifier development. These include the addition of context dependent acoustic models and lattice rescoring in SOU recognition and lattice-based n-gram SVM features for TID. With these added improvements, our topic verification equal error rate (EER) is improved from 12% to 6.6%. To further understand the TID EER, we also performed an oracle experiment using the phoneme sequence from the true transcripts. We found that the top SOU n-grams features correspond to topic relevant keywords making the SOU TID system an unsupervised keyword discovery system which is related to [5].

This paper is organized as follows. In the next section, we briefly review the unsupervised HMM training. This is followed by a description of improved TID system in Section 3. We describe the experiments in Section 4 including a discussion of the relationship between SOUs and topic keywords. The paper is concluded in Section 5.

2. Unsupervised HMM Training

An HMM-based model can be specified by the acoustic model parameters, $\theta_{am}$, and the language model parameters, $\theta_{lm}$. We denote the HMM parameters by $\theta = [\theta_{am}, \theta_{lm}]$. Because the language and acoustic model parameters can be easily decomposed, the acoustic model likelihood, $p(X|W, \theta_{am})$, and language model likelihood $p(W|\theta_{lm})$ are maximized separately in typical supervised maximum likelihood (ML) HMM training. In this work, however, grouping them as a single parameter set $\theta$ is more convenient. With this notation, the ML parameter estimation finds the parameter set, $\hat{\theta}_{sup}$, that maximizes the joint likelihood of observation $X$ and the label sequence $W$, $p(X, W|\theta)$. That is,

$$\hat{\theta}_{sup} = \arg \max_{\theta} p(X, W|\theta).$$

(1)

In the case of unsupervised training in which the label sequence $W$ is not known, we maximize the joint likelihood by searching not only over the model parameters but also all possible label sequences. That is, $W$ becomes a variable to be optimized. The unsupervised ML parameter estimation becomes,

$$\hat{\theta}_{unsup} = \arg \max_{\theta} \max_{W} p(X, W|\theta),$$

(2)

$$= \arg \max_{\theta} \max_{W} p(X|W, \theta)p(W|\theta).$$

(3)

The maximization over both the label sequence and the acoustic model likelihood in Eqn 3 balances the acoustic likelihood and label sequence structure. At one possible extreme, one can use different symbols for each frame to maximize the acoustic likelihood at the expense of high entropy in the symbol sequence. At the other extreme, one can define only one symbol to maximize the language model likelihood but this will result in low acoustic likelihood. The balance between these two is

---

This work is partially supported by contract from Oasis and AFRL.
influenced by the choice of the initial label sequence and the complexity of the acoustic and language models.

2.1. Iterative Optimization

Eqn. 2 maximizes over two sets of variables, $\theta$ and $W$, which can be performed using iterative maximization. At each iteration, one set of variables is fixed while the other set is maximized. Then we alternate between them. So, at the $i$-th iteration, the two maximization steps are:

1. find the best parameter set $\theta_i$ on the previously found label sequence $W_{i-1}$.

$$\theta_i = \arg\max_{\theta} p(X, W_{i-1}|\theta). \tag{4}$$

2. find the best word sequence $W_i$ by using the previously estimated parameter set $\theta_i$.

$$W_i = \arg\max_W p(X, W|\theta_i). \tag{5}$$

Comparing Eqns 1 and 4, it is obvious that Step 1 (Eqn 4) is simply the regular supervised HMM training (both acoustic and language models) using the newly obtained transcription $W_{i-1}$ as reference. Finding the best word sequence in the second step would suggest a Viterbi recognition pass. Although recognition is usually viewed as finding the most likely label sequence over the posterior probability, $p(W|X, \theta)$, it is easy to show that the same sequence also maximizes the joint likelihood $p(X, W|\theta)$ as in Eqn. 5. So, Eqn. 5 expresses the recognition of a new transcription using the updated parameters $\theta_i$.

Note that because each step of the iterative process maximizes the likelihood $P(X, W)$ over a parameter space that covers the parameter set from the previous iterations, the process is guaranteed to converge.

3. Improved Topic Identification System

The training process of the TID system consists of three different stages. The first stage creates an initial sequence for HMM training. Multiple possible approaches are feasible including the use of a different recognizer trained from another language, or starting with a universal phone recognizer. These approaches can suffer from domain mis-match. We have taken a sequential learning approach to build a segmental tokenizer using polynomial segment models with audio from the domain of interest. The second stage is the unsupervised HMM training that iterates between optimizing the model parameters and the label sequences. The third stage tokenizes the training data into SOU sequences and builds the SVM topic classifiers.

3.1. Initial Labeling

Audio is first segmented based on its spectral discontinuities which are learned without supervision from the audio signal [6]. It is followed by fitting each audio segment with a polynomial (quadratic) trajectory in the cepstral space. The audio segments are then grouped into clusters of similar acoustics based on the distance between their polynomial trajectory parameters [7, 8]. The distance measure currently used on a pair of segments is the area between their polynomial trajectories. These segment clusters represent collections of sound units. Any individual cluster is a collection of variants of a particular sound and forms the basis for generating a segmental Gaussian mixture model (SGMM) with each mixture component representing a segment cluster. The SGMM is trained with the EM algorithm.

The SGMM becomes a speech tokenizer when, for an audio segment, it returns the mixture index by which the segment likelihood is maximized. After building the SGMM, it is used as a tokenizer for the training segments. These segment labels form the initial transcription for HMM training.

3.2. SOU HMM Training

We use the state-of-the-art BBN Byblos recognizer [9] for HMM training. Byblos includes advanced signal processing techniques, such as Vocal Tract Length Normalization (VTLN), Heteroscedastic Linear Discriminant Analysis (HLDA) feature transformation, context-dependent triphone and quiniphone models, multi-pass recognition, speaker adaptive training etc. Byblos uses “flat start” HMM training that does not require token time marks. Instead, iterative alignment and model estimation are carried out. While discriminative training is part of Byblos, our current experiments use only the maximum likelihood training. Details about the Byblos training can be found in [9]. In addition to acoustic models, initial bigram and trigram language models are constructed using the label sequences generated by the segmental tokenizer. Following Eqns 5 and 4, we iteratively maximize the model likelihood and find the best label sequence.

3.3. Tokenization and Topic Classification

With the trained acoustic models and language models, the tokenization of training (also test) audio into SOU sequences is no different from regular phoneme recognition. In our previous paper [1], we tokenized the audio using only non-crossword models with bigram language models. Because we used the phonemes (or SOUs) as words, using non-crossword models is equivalent to using context-independent (or mono-phone) models. To create context dependent models in Byblos, we need to create “phoneme” classes to drive the decision-tree based phoneme-state clustering. What “linguistic questions” can we use to drive the decision tree? In our current work, we cluster the 64 SGMM’s into 16 classes to act as “phoneme classes” for decision tree context clustering of quinphones. These quinphones are used to re-score SOU lattices generated by mono-phone models.

Next after tokenization is the SVM classifiers training. For topic classification, utterances are tagged as in-topic or out-of-topic. Note that data used for training the classifiers do not need to be the same data used for training the HMMs. We begin by tokenizing both the in-topic and out-of-topic audio into SOU sequences. We then extract from these SOU sequences the SVM classification features which are SOU n-gram statistics normalized by inverse-document-frequency (IDF). We then apply the more sophisticated SVM feature building techniques for generating useful long-span n-gram features [10].

In our previous work, the n-gram statistics were extracted from recognition 1-best hypotheses. In our current system, we extract n-gram statistics from the quinphone re-scored lattices. These n-gram statistics are soft counts based on the lattice posterior probabilities. The use of soft counts creates a problem for estimating the IDF weighting. We experimented with soft IDF weighting as proposed in [11] as compared to using a simple posterior threshold on the soft token counts. We found that the...
simple thresholding gave better EER.

The SOU TID evaluation process is quite similar to a typical TID system using words. The test audio cuts are first decoded by the HMM into SOU token sequences (or lattices). Classification features are extracted and passed to the SVM topic classifier to generate TID scores.

3.4. Keyword Discovery

The SVM feature building described above generates a set of long-span n-gram features by selectively expanding important lower order n-grams. The importance of any n-gram feature is measured by the absolute value of the corresponding SVM classifier weight as described in [10]. For each topic, this results in a ranked list of long-span n-grams, such as 5-gram and 6-gram. One question is whether these top SOU n-grams correspond to any topic specific keywords. By performing forced alignment on both the word transcription and SOUs, we can time align the SOU sequence with the word transcript. With this alignment, we can extract all English words that overlaps with any SOU n-gram of interest. In Section 4.3, we discuss how the top SOU n-grams relate to topic keywords.

4. Experiments

4.1. Experiment Setup

Our experiments were performed on the Switchboard-1 corpus, which consists of telephone conversations between strangers discussing one of 70 pre-assigned topics. Each conversation is approximately 5 minutes long. The most frequent ten topics were selected as target topics with the remaining 60 topics denoted as the non-target topic set. A set of 96 conversation sides (4 hours) was randomly selected from the non-target topic set for HMM training. The selected target topics, the number of training and test conversation sides (for TID) are listed in Table 1. TID decisions were made per conversation side. Target-specific impostor data were defined for each target topic. These included a set of 600 tests randomly selected from the combination of the non-target topic set which is not used in HMM training, as well as data from the other target topics.

The segmentation was based on spectral power from 14 frequency bands with a maximum segment duration of 50 frames. The details are described in [1]. Vocal tract length normalized (VTLN) cepstral features and their first derivatives were generated for each segment that was then modeled with a polynomial (quadratic) trajectory model. The clusters were created using the k-means algorithm with initial centroids obtained using binary centroid splitting. These k-means clusters were further refined with 3 iterations of EM training. This resulted in 64 mixture components.

The SGMMs were used to tokenize the acoustic training data as the initial label for HMM training. This was followed by multiple iterations of unsupervised HMM training. Byblos used 5-state left-to-right HMMs trained with the maximum likelihood criterion. Sixty dimensional acoustic features were generated by transforming the concatenation of 9 frames of 13-dimensional cepstrum and its energy via a HLDA transform. We used a 512 component state-tied-mixture model (STM) for the monophone model and a 64 Gaussian components for the cross-word state-cluster tied mixture (SCTM). Both speaker independent acoustic models and speaker adaptive training (SAT) models were created. During each iteration of unsupervised training, the acoustic training data were decoded using the newly learned parameters to generate new unit sequences. This decoding involved one pass of unadapted decoding followed by the HLDA adaptation, unsupervised MLLR adaptation and another pass of adapted decoding, all using only the STM's.

After the HMMs were trained, recognition was performed on all the classifier training and test data using both the STMs and the context dependent SCTMs as described in Section 3. We trained the SVM using the regression mode of the publicly available LIBSVM package [12].

4.2. Results

Various topic classification experiments were performed with different complexity of the unsupervised training as well as other benchmarks. The results are tabulated in Table 2. Rows 1 to 3 were results from using only the SGMM (Unsup. SGMM only), one iteration of Byblos training (Unsup. Byblos 1 iter) and five iterations of SOU training (Unsup. Byblos 5 iter). These are similar to the SOU results reported in [1].

Rows 4 to 7 show the effect of more complex modeling including context modeling as well as lattice-based features. By using a more detailed state-cluster tied-mixture acoustic model [9] and trigram language model, EER is reduced to 10.7% as shown in row 4. Adding context model in lattice rescoring further reduces the EER to 10.0%. The best performance of 6.6% EER comes from extracting n-gram features from the re-scored lattices.

To benchmark the above performance, we performed experiments using tokenizers trained with supervision as shown in rows 8 to 10, which include tokenizing with a BUT Hungarian phoneme recognizer (row 8), tokenizing with a one-hour supervised training Byblos English phoneme recognizer (row 9) and using the phoneme sequence derived from true test transcripts (row 10).

The gain from using more complex models is consistent with our insight that SOUs can leverage on progress from regular speech recognition. The particular gain in using lattice-based n-gram features is significant. It may be an indication that the recognized SOU sequences are not very sharp such that alternative hypotheses in the lattice contain many SOU sequences useful for TID.

The BUT Hungarian model has a matching channel (i.e., telephone channel training), trained with a good amount of data (10 hours) but was trained from a different language, while the one-hour English data matches the test. Even with an order of magnitude more data, the EER is 2.6% absolute higher than the one-hour Byblos training, probably due to the language mismatch. This is consistent with the results reported [13] in which the Hungarian recognizer almost doubles the classification EER compared to a similarly trained English phoneme recognizer.

Table 1: The 10 target topics and their num. of training and test conversation sides

<table>
<thead>
<tr>
<th>Topic</th>
<th># Train Conv</th>
<th># Test Conv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Car buying</td>
<td>85</td>
<td>20</td>
</tr>
<tr>
<td>Capital punishment</td>
<td>92</td>
<td>22</td>
</tr>
<tr>
<td>Recycling</td>
<td>102</td>
<td>25</td>
</tr>
<tr>
<td>Job benefits</td>
<td>90</td>
<td>21</td>
</tr>
<tr>
<td>News media</td>
<td>85</td>
<td>20</td>
</tr>
<tr>
<td>Public education</td>
<td>81</td>
<td>19</td>
</tr>
<tr>
<td>Drug testing</td>
<td>89</td>
<td>22</td>
</tr>
<tr>
<td>Exercise and fitness</td>
<td>81</td>
<td>19</td>
</tr>
<tr>
<td>Family finance</td>
<td>89</td>
<td>21</td>
</tr>
<tr>
<td>Family life</td>
<td>84</td>
<td>20</td>
</tr>
</tbody>
</table>
5. Conclusions

In this paper, we presented how we improved our previously proposed SOU-based TID system by using context dependent acoustic models and lattice-based n-gram features. We showed that both resulted in significant TID EER reduction. We further explored the relationship between some selected SOU sequences and their corresponding English words. We showed that the same SOU n-grams are consistently mapped to the same English keywords, and SOU n-grams selected as TID features were representing topic keywords. Thus, the TID system can also be used for unsupervised keyword discovery.

6. References


