EXPLOITING LARGE QUANTITIES OF SPONTANEOUS SPEECH FOR UNSUPERVISED TRAINING OF ACOUSTIC MODELS

Bhuvana Ramabhadran

Human Language Technologies
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598
bhuvana@us.ibm.com

Abstract

While large amounts of manually transcribed acoustic training data is available for well-known large vocabulary speech recognition tasks such as, the transcription of broadcast news and switchboard conversations, a significantly less amount is available for several large spoken collections such as the MALACH corpus (in multiple languages), meeting recordings, presentations at conferences, call center conversations, etc. However, these collections offer vast quantities of untranscribed spontaneous speech that can be used to improve recognition accuracies. Several narrow-band and broadband speech collections are currently available and carefully tuned speech recognition systems trained on several hundred hour of manually transcribed data are now able to achieve word error rates between 10% and 40%, depending on the difficulty of the collection. This paper studies the use of automatically recognized transcriptions at several levels of recognition accuracy to train acoustic models and the performance improvements obtained with such unsupervised training. This paper also proposes a recipe for selection of feature vectors at the utterance, word or fragment level for training acoustic models that provides the maximum gain in recognition accuracy. This paper demonstrates that a reduction in overall word error rate of up to 20% relative can be obtained with careful selection of acoustic training data.

1. Introduction

Current Automated Speech Recognition (ASR) systems require and consistently improve with large quantities of training data [5]. The manual transcription effort needed for carefully labeled data is tedious and prohibitively expensive. Several researchers [1, 2, 3] have explored the use of transcripts of untranscribed audio data generated by an ASR system trained with as little as one hour of data. In a recent review [4] presented a systematic study of the use of confidence measures and iterative unsupervised training with initial acoustic models bootstrapped from less than an hour of manually transcribed data on a broadcast news transcription task.

When building ASR systems for a new language or new domain, an existing recognizer trained on a different database is used to transcribe the new data automatically or a few hours of manually transcribed data is used to customize the models to the new domain. If the accuracy of the recognized transcriptions is reasonably high, it can be used to subsequently train acoustic models despite the transcription errors. In [2], lattice-based confidence measures were used to select feature vectors for training. A high threshold on the confidence measure selected words that were already modeled robustly and a low threshold on the confidence measure included more of the erroneously recognized words, thus providing a trade-off between robustness and increased modeling capacity and decreased recognition performance. It is not the aim of this paper to develop ASR systems with small amounts of labeled training data and iteratively train an improved recognition system. However, it is the goal of this work to determine the best method to rapidly fold in untranscribed data into training acoustic models and focus on the reduction in cost of rapidly building the overall system.

The MALACH corpus described in [7, 9, 10] naturally lends itself as an excellent testbed for LVCSR and NLP research. The initial acoustic models presented in this work were built off approximately 200 hours of manually transcribed data. This paper focuses on the unsupervised training of acoustic models when this data is increased by many folds. Therefore, it is impossible to compare the recognition performance against manual transcriptions as they simply do not exist for this corpus. Several researchers have reported a performance degradation of between 14% and 18% relative when using less than 100 hours of untranscribed versus manually transcribed data on the broadcast news transcription task [4]. The main focus of this work is to determine how best to incorporate the decoded transcriptions at different word error rates for training acoustic models while maximizing the performance of the ASR system. The rest of this paper is organized as follows.

Section 2 describes the MALACH database and the challenges for an ASR system on this data. Sections 3 and
3.2 describes several methods of data selection for unsupervised training of acoustic models. Section 4 presents the experimental setup and improvements in recognition accuracies obtained with unsupervised training. The paper concludes with a summary and potential applications for this work in other collections.

2. MALACH

The MALACH corpus consists of unconstrained, natural speech filled with disfluencies, heavy accents, age-related coarticulations, uncued speaker and language switching and emotional speech collected in the form of interviews from over 52000 speakers in 32 languages. Approximately 25000 of these testimonies are in English, spanning a wide range of accents, such as Hungarian, Polish, Yiddish, German, Italian, French, Croatian, Spanish, Ukrainian etc. A good number of words uttered in this corpus are foreign words or sequences of words spoken in a foreign language, unfamiliar names and places. The corpus consists of elderly speech, where the age of the interviewees range from 56 years to 90 years. In order to obtain training data for acoustic and language models, approximately 200 hours of the English portion of the MALACH corpus was manually transcribed and annotated with named entities. Transcription is challenging even for skilled annotators and they typically required 8 to 12 hours to transcribe a single hour of an English interview. The difficulties arise from unfamiliar names and places, multiple languages encountered during a single interview, coarticulations related to age, highly variable speaking rates, and heavily accented speech [7].

2.1. Infrequent Unlabeled Data

Given the large quantities of data available in this corpus, an important issue that arises with respect to rolling in large volumes of data is the word error rate usually incorporates recognition errors arising from OOVs. For illustration, we include here an example of the actual words spoken by different speakers during several segments, annotated with OOVs (shown in bold face).

because there was no normal teacher in Blashova so there was a teacher that ran more or less like a high school teacher ...

I was no longer able to stay in Flaten and neither was my f- the son of the Cookland who was the same age as me and we both all came and went back to live in Rectory Road in Hackney then the question came as to what I was going to do with my life ...

Moreover, these names occur in many variations (Hebrew names, Yiddish names, diminutives, first names only, nicknames, etc.).

While it is important for these words to be recognized it is reasonable to expect that they will not be hypothesized due to their absence in the lexicon or the fact that they are infrequently occurring words and phrases, and rare terms are inevitably modeled less well than more common ones. However, fragments of these words, such as a sequence of syllables or a combinations of phones and syllables will be decoded correctly. While this work focusses on the MALACH corpus, presence of these OOVs is universal to all spoken archives including call-center and meeting conversations. For example, in a 30K lexicon, the OOV rate can be as high as 10% thereby significantly contributing to the overall WER of nearly 40% [7]. Hence, it is important to analyze the level at which data units need to be selected to ensure maximum gains from unsupervised training.

3. Unsupervised Training Algorithms

In all the experiments presented in this paper, unsupervised training of the acoustic models constitutes retraining from scratch, i.e., all matrix transformations, decision trees and HMM parameters were recomputed with the new data. This section describes the different methods used to select data for unsupervised training.

3.1. Confidence Measures

The most commonly used method of data selection is based on confidence measures. In this work, the decoded transcripts are annotated with their word-based posterior probabilities. This confidence measure is computed using the forward-backward algorithm on the word-graphs. Based on two different thresholds on the confidence measures, the feature vectors used for training are either selected at the word, or utterance level. If the training happens at the utterance level, the silence frames obtained during dynamic alignment are also used, otherwise the word boundaries hypothesized by the decoder are used to select the feature vectors corresponding to the words and the silence frames are ignored. If the training happens at the fragment level silence frames are included depending on the use of within-word or cross-word context incorporated in the training procedure.

3.2. Selection of Training Units

Many infrequently occurring words in the manually transcribed training data will not be modeled well. For example, in the MALACH corpus, a person with the first name Alicia, has Alicia, Chana, Alice, Jadwiga, Alushia, and Alla as possible variations of the first name. The phonetic or syllabic representation of these words in the decoded transcripts are correctly recognized most of the time but the words are not hypothesized by the ASR system either because they do not exist in the lexicon or their infre-
quent occurrence causes them to be poorly modeled. Using fragments of these words will expand the cross word phonetic contexts seen by the decision trees and provide more robust acoustic models.

The selection of fragments was done in the following manner. A portion of the training corpus that was manually transcribed was used to build a phonetic language model and build a fragment-based decoder. The length of the fragment, i.e., syllables, phones or combinations thereof was determined by a threshold on the transducer generating the fragments and weighted by a language model on the phones. After experimenting with several fragment lengths and the resulting WER on a held out test set, the fragment length was restricted to less than 7 phones. While this choice may seem somewhat arbitrary, this fragment length is corpus dependent and relies on the spontaneous nature of speech seen in the data.

### 4. Experimental Setup

#### 4.1. Training and test corpora

The English training corpus was generated using 15-minute segments of an interview from 720 randomly selected speakers. Thus, a total of 180 hours of data was selected for manual transcription to serve as training material for ASR systems. Male and female speakers in this corpus were more or less equally distributed and a wide range of accents were covered (e.g., Hungarian, Italian, Yiddish, German, and Polish). The ASR test set consists of 30 minute segments taken from 15 randomly chosen speakers.

The manual transcriptions of 180 hours of training data was used to build language models using the modified Kneser-Ney algorithm [7]. The training data is relatively small (1.7M words), therefore, the language models built from Broadcast News (BN) and Switchboard (SWB) corpora (158M and 3.4M words, respectively) were interpolated with the LM built from this collection. The interpolated weights were optimized to achieve minimum perplexity on the held-out data from this collection. The perplexity of this task on the held-out test set is 72.3. The size of the lexicon is 30K.

For unsupervised training, upto 600 hours of data from overall 400 interviews was automatically transcribed. The audio signal was down-sampled to 16KHz from 44.1Khz and parameterized using 24-dimensional mel frequency cepstral coefficients (MFCC). Final acoustic features were derived using linear discriminant (LDA) and maximum-likelihood based linear transformations (MLLT). Speaker specific transformations (SAT and fMLLR) were used by the final system that gave the best WER.

#### 4.2. Decoding Strategy

The decoder used in our experiments is a Viterbi decoder operating on a fully flattened state-level HMM. A traditional decoding setup operates on a single HMM constructed from an overall language model. In contrast to this, in our experiments, the states are built for each speaker using a speaker-specific lexicon and language model. Detailed descriptions of the Viterbi decoder used is presented in [11].

### 5. Results

Table 1 shows the performance improvements obtained over the baseline system when trained with decoded transcriptions. All the transcripts including the incorrectly recognized words are used in the unsupervised training procedure. The WER on the decoded transcripts using the baseline system trained on 200 hours of manually transcribed data is 38.3%.

<table>
<thead>
<tr>
<th>Hours</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>38.0</td>
</tr>
<tr>
<td>400</td>
<td>37.5</td>
</tr>
<tr>
<td>600</td>
<td>37.1</td>
</tr>
</tbody>
</table>

Table 1: WER computed for ASR systems trained on varying amounts of decoded transcriptions

It can be seen that when the data is tripled, a gain of 1.2% absolute in WER can be obtained with the raw transcripts. The same experiment when repeated with transcripts at a higher WER of 46.2% (obtained at the speaker-independent level) showed less than 0.5% absolute gain.

The table 2 below illustrates the effect of confidence measure-based selection of acoustic data for different thresholds. The thresholds were selected in a straightforward manner from the histogram of word posterior probabilities averaged at the utterance level. Each utterance is obtained using a simple speech-silence segmentation scheme and was on an average 8-10 seconds long. As expected, significant gain in recognition accuracy can be seen when compared to the baseline where no selection of data was employed. Table 3 illustrates the effect of selecting units at the word level for training acoustic models thresholded by confidence measures as well as using all the decoded words. Two different thresholds are used for both cases. The best result was a reduction of 2% absolute in the WER using a word-based selection scheme. It can be seen that the effect of the threshold on the confidence measure disappears as more data is used in the training procedure.

Table 4 demonstrates the use of fragments in data selection for unsupervised training. A gain of 6.2% absolute is obtained; this is the best result seen so far.
Table 2: WER computed for ASR systems trained on varying amounts of decoded transcriptions selected at the utterance level using different thresholds on confidence measures.

<table>
<thead>
<tr>
<th>Hours</th>
<th>Utterances with conf. measure greater than 0.8 (%)</th>
<th>Utterances with conf. measure greater than 0.9 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>37.3</td>
<td>37.5</td>
</tr>
<tr>
<td>400</td>
<td>37.0</td>
<td>37.2</td>
</tr>
<tr>
<td>600</td>
<td>36.6</td>
<td>36.8</td>
</tr>
</tbody>
</table>

Table 3: WER computed for ASR systems trained on varying amounts of decoded transcriptions selected at the word level using different thresholds on confidence measures.

<table>
<thead>
<tr>
<th>Hours</th>
<th>Words with conf. measure greater than 0.8 (%)</th>
<th>Words with conf. measure greater than 0.9 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>37.3</td>
<td>37.5</td>
</tr>
<tr>
<td>400</td>
<td>36.9</td>
<td>36.8</td>
</tr>
<tr>
<td>600</td>
<td>36.4</td>
<td>36.4</td>
</tr>
</tbody>
</table>

Table 4: WER computed for ASR systems trained on varying amounts of decoded transcriptions selected at the fragment level.

<table>
<thead>
<tr>
<th>Hours</th>
<th>Fragment selection (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>35.3</td>
</tr>
<tr>
<td>400</td>
<td>33.4</td>
</tr>
<tr>
<td>600</td>
<td>32.1</td>
</tr>
</tbody>
</table>

6. Conclusion

This work has demonstrated that with additional training based on incorrect transcriptions at approximately 40% WER, performance improvements of up to 20% relative can be seen, provided the data is carefully selected for training of the acoustic models. This is particularly important as most ASR systems are trained on at least 200 hours of manually transcribed data and significant effects have been achieved only when the data is multiplied several fold with manual transcriptions [5]. Confidence measures play a crucial role selecting training units but better gains can be obtained when using a fragment-based selection procedure. The fragment-based selection scheme yields ASR systems with the best WER primarily because of better and increased use of untranscribed data as well as incorporating data from OOVs in an efficient fashion.

7. Acknowledgements

This project is part of an on-going effort funded by NSF under the Information Technology Research (ITR) program, NSF IIS Award No. 0122466. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

8. References


