On-line Learning of Acoustic and Lexical Units for Domain-Independent ASR

Giuseppe Riccardi

AT&T Labs-Research
dsp3@research.att.com

1. ABSTRACT

We are interested in on-line acquisition of acoustic, lexical and semantic units from spontaneous speech. Traditional ASR techniques require the domain-specific knowledge of acoustic, lexicon data and more importantly the word probability distributions. In this paper we propose an algorithm for unsupervised learning of acoustic and lexical units from out-of-domain speech data. The new lexical units are used for fast adaptation of language model probabilities to a new domain. We show that starting from the Switchboard corpus (lexicon and language model) we learn the most relevant language statistics of the "How May I Help?" task.

2. INTRODUCTION

We are interested in spoken language systems that adapts on-line to new domains. State-of-the-art research or deployed spoken dialog systems perform constrained tasks (e.g. travel information, stock quotes, etc.) and they achieve high task completion rates. These systems are programmed to answer users' queries as long as they belong to the apriori defined domain. However, the next generation spoken dialog systems should react to changes in the task (e.g. a new category for a call-routing task) and adapt to unseen speech and language events. The crucial features for an adaptive system are the acoustic and lexical units and the association between language and machine actions. In this paper we will cover the first two topics while the third is discussed in [3, 5].

The traditional large vocabulary speech recognition framework requires acoustic and language model to be trained on domain-specific data. This data is usually collected through Wizard-of-Oz protocol and speech utterances are transcribed for the purpose of acoustic and language model training. These models perform poorly in out-of-domain conditions and are not suitable for on-line learning of language. This is true despite the fact that large vocabulary lexicons can have millions of words and reduce the out-of-vocabulary rate to zero. However, the main reason for such poor performance is the mismatch of the language model probabilities.

An alternative approach to large vocabulary recognition is to model phone sequences. In the late eighties there has been active research on this topic and high phone recognition rate have been achieved for restricted domains and language [4, 7]. It is now widely accepted that phone recognition performance are inferior to word-based large vocabulary speech recognition tasks, where large databases are available to train language models. However, phone recognition allows in principle to perform task-independent speech recognition.

For these reasons, a task-independent speech recognizers should combine the accuracy of a word-based system and the acoustic event granularity of a phone-based system. In this paper, we propose a framework to learn on-line estimates of word probabilities for task-independent speech recognition. We decompose this problem into:

1. Search for the acoustic-lexical units of the new task (Phone recognition).
2. Map new acoustic-lexical units into words (Lexical access).
3. Estimate new-domain word-based language model from prior distributions and on-line estimates (Word Probability Adaptation).

In the first step we train a phone recognizer from an off-the-shelf database and acquire the variable length phone sequences (acoustic morphemes [3]) using information theoretic measures. In the second step we search for all possible words or phrases the acoustic morphemes are mapped into. In the last step we propose a new algorithm for on-line adaptation of word probability vectors and measure the perplexity improvements for the new model.

3. DATABASES

We describe the algorithms for on-line word probability estimation, in the context of adapting a spoken language system based on the Switchboard...
database transcriptions to the "How May I Help You?" (HMIHY) call routing task [2]. Both databases are collections of highly disfluent speech which has been collected with no constraints on the language. In particular, the "How May I help You?" database contains all the response utterances to the open-ended prompt "How May I Help You?" within a call-routing task [2].

The Switchboard database [8] has 3.1 million words (training set) and has estimated word probabilities $P_{\text{switchboard}}(w)$ for all the words $w$ in the dictionary $V$ with size 29.2K. The HMIHY database consists of 10K speech utterances which have been partitioned in training (8K), development (1K) and test set (1K). The dictionary size is 3.6K, the average length (words) of the speech utterances is 18 and the OOV rate is 1.6% and 30% at the token and utterance level respectively. The dictionary overlap between the two domains is 78.1% (of the HMIHY dictionary) and it covers 99.0% of the probability mass of the HMIHY dictionary.

4. PHONE RECOGNITION

The task-independent component of our language acquisition system is the phone recognizer which transcribes speech utterances from one domain (i.e. HMIHY) and is trained on another domain (i.e. Switchboard). The acoustic module of the phone recognizer is built upon an off-the-shelf context-dependent phone acoustic model trained on telephone speech different from the Switchboard and HMIHY database. The acoustic model performs well on the HMIHY task [2] in language-model matched conditions and very poorly with the Switchboard language model. The word accuracy on the HMIHY task for matched and unmatched language model conditions is respectively 52.7% and 13.4% (trigram language model) [2].

The second statistical model of the phone recognizer is the phonotactic model. We replaced each word in the Switchboard corpus with its most likely pronunciation and trained $n$-gram model based on the Variable Ngram Stochastic Automaton (VNSA) [6]. We varied $n$ between 1 and 6 and computed phone accuracy on the HMIHY test set. Phone accuracy saturates for $n \geq 6$ and this might be related to the fact that the average baseform length is 5.4 (phones), as computed on the HMIHY dictionary. In Fig. 3 we report the phone accuracies for different orders of the phonotactic models.

Noisy transcriptions (low phone accuracies) contain segments of speech which has been transcribed accurately or consistently (high correlation). These segments of speech might map into one or more words and correspond to variable phone length units (acoustic morphemes). Variable length unit models improve the predictive power of $n$-grams at no expense for the parameter space [2]. We acquire these phone sequences by computing information theoretic measures, such as the weighted mutual information $I_W(x, y) = P(x, y)MI(x, y) = P(x, y)logP(x, y)/P(x)P(y)$, where $x$ and $y$ are two phone sequences. Weighted mutual information compensates for probability estimate biases introduced by rare events in the mutual information, $MI(x, y)$. Alternative and computationally expensive methods for computing variable length units are entropy-minimizing units [2] and multigram approaches [1] [9].

We ran the 6-gram phone recognizer on the HMIHY training set and ranked all the substrings of the phone transcriptions by the $I_W$ weighted mutual information. An excerpts from the acquired variable length units are shown in Fig. 1. The first and second column contain the $I_W$ and $MI$ values, the third and fourth are the corresponding variable length phone sequence and its most likely word sequence. The top $N$ ($N = 500$) variable length units are incorporated in the stochastic phonotactic models which takes advantage of the long-spanning acoustic morphemes [2]. The acquisition of the units can be iterated, alternating phone recognition and morpheme acquisition as depicted in the block diagram in 2. The morpheme-based phonotactic model gives a relative improvement of 21% over the 6-gram model and this improvement carries over onto phone lattice accuracy (18.7%), as shown in Fig 3. Phone lattices provide a compact

---

**Table 1:** Excerpt from the set of acquired acoustic morphemes.

<table>
<thead>
<tr>
<th>$MI$</th>
<th>$WMI$</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.79</td>
<td>0.0611</td>
</tr>
<tr>
<td>16.98</td>
<td>0.0596</td>
</tr>
<tr>
<td>9.15</td>
<td>0.0421</td>
</tr>
<tr>
<td>14.57</td>
<td>0.0380</td>
</tr>
<tr>
<td>6.31</td>
<td>0.0374</td>
</tr>
<tr>
<td>16.66</td>
<td>0.0355</td>
</tr>
<tr>
<td>36.60</td>
<td>0.0277</td>
</tr>
<tr>
<td>3.63</td>
<td>0.0276</td>
</tr>
<tr>
<td>6.12</td>
<td>0.0227</td>
</tr>
</tbody>
</table>

---

Even though the word accuracy (52.7%) is small, the spoken language module of the HMIHY system achieves 80% correct call-type classification at 40% False Rejection Rate. [2]
and still accurate search space for the exact phone sequences. They can be used to calibrate the upper bound for the language acquisition problem (see section 6) or compute accurate posterior probabilities for semantic unit acquisition [5].

Variable length units can be mapped to zero, one or many word sequences, for a given baseform dictionary. In this experiment we took a conservative strategy to lexical access and considered only the exact match between baseforms in the dictionary and acoustic morphemes. The conservative strategy was dictated by the need to learn the most relevant lexical features of the domain with high precision and high rejection [3]. These lexical features will be used to adapt on-line an off-the-shelf (Switchboard) large vocabulary language model. The acquisition algorithm in section 4 has selected 500 units and we have matched them against their most likely word mapping drawn form the Switchboard dictionary. Hence, we have compiled a list of words and phrases (see Fig 1 with 45 items). While the list of lexical items is short, due to the conservative strategy, they cover 40% of the probability mass of the HMIHY corpus.

6. LANGUAGE MODEL ACQUISITION

The ultimate goal of our language acquisition model is to transform a prior probability distribution from one domain to another without speech transcriptions (on-line adaptation). This is the primary constraint of an adaptive spoken dialog system, which can react and recover to novel events in the syntactic and semantic input channel. Most of the literature in language model adaptation is for batch processing (with speech transcriptions) and assumes that the probability space is the same both for the source (i.e. Switchboard) and target (HMIHY) domain, namely the word sequence space.

Our model poses two challenges to traditional stochastic modeling. The first is to map phone sequence statistics into word statistics. If the most likely pronunciation of the word collect is K a e l K t, then P(collect) ≠ P(K a e l K t). In general terms:

\[ P_{\text{on-line}}(w) = \sum_{f_i \in B_w} P(f_i) \]  \hspace{1cm} (1)

where \( f_i \) is a baseform of the word \( w \) drawn from the set \( B_w \). In the rest of this paper we shall assume \( P_{\text{on-line}}(w) \approx P(f_l) \), where \( f_l \) is the most likely pronunciation. In Fig 4 we scatter plot \((\log P_{\text{HMIHY}}(w), \log P_{\text{on-line}}(w))\) (asterisk) and \((\log P_{\text{HMIHY}}(w), \log P_{\text{Switchboard}}(w))\) (circle), where both \( P_{\text{HMIHY}}(w) \) and \( P_{\text{Switchboard}}(w) \) have been estimated from the hand-labeled speech transcriptions. Most word probability estimates lie along the diagonal (dashed line) and closely approximate the true empirical distribution \( P_{\text{HMIHY}}(w) \). In order to get an upper bound on the on-line probability estimates we computed the phone sequence with lowest string edit distance from the reference phone transcription.
Thus, we have compiled the oracle phone transcription training set and estimated \( P_{\text{oracle}}(w) \). In Fig. 5 we scatter plot the log-probabilities for the on-line oracle estimates. From Fig. 5 we see that, the probabilities of the selected words and phrases have been estimated with high precision with respect to the best possible guess (oracle).

The second problem is to transform the word probability vector of the prior distribution using the word statistics learned on-line. Given the small lexical coverage and large probability mass of the acquired features we have merged the two distribution with following scheme:

\[
P_{\text{target}}(w) = \begin{cases} 
P_{\text{on-line}}(w) & w \in S \\ \frac{\alpha \beta}{P_{\text{Switchboard}}(w)} & w \notin S \end{cases}
\]

where \( \alpha = \sum_{w \in S} P_{\text{on-line}}(w) \), \( \beta = \sum_{w \notin S} V P(w) \) and \( S \) is the set of selected words. We have tested this model to measure the perplexity of the HMIHY test set perplexity for a unigram model. The unigram perplexity on the HMIHY test set, in matched language model condition \( (P_{\text{HMIHY}}(w)) \) is 128.3. The test perplexity using the Switchboard language model \( (P_{\text{Switchboard}}(w)) \) is 715.9. We computed the test set perplexity with the new word probability \( P_{\text{target}}(w) \) and achieved a 39.4% relative improvement (433.9) with respect to the Switchboard baseline.

![Figure 4: Log-Likelihood scatter plot for the on-line HMIHY and Switchboard probability estimates (X-axis) as compared to the HMIHY probability estimated from text transcriptions (X-axis).](image)

**Figure 4** Log-Likelihood scatter plot for the on-line HMIHY and Switchboard probability estimates (X-axis) as compared to the HMIHY probability estimated from text transcriptions (X-axis).

### 7. CONCLUSION

In this paper we have proposed an algorithm for on-line acquisition of language models for spoken language adaptive systems. The algorithm consists of three main steps. In the first step we train a phone recognizer from an off-the-shelf database and acquire variable length phone sequences (acoustic morphemes) using information theoretic measures. In the second step we search for all possible words or phrases the acoustic morphemes are mapped into. In the last step we estimate the target word probability distribution from the on-line phone sequence statistics. We have shown the new-domain word probabilities closely approximates the true distribution with respect to test set perplexity.

### REFERENCES


[8] Switchboard Database “http://morph.ldc.upenn.edu"