Automatic New Word Acquisition: Spelling from Acoustics

Fil Alleva and Kai-Fu Lee

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Abstract

The problem of extending the lexicon of words in an automatic speech recognition system is commonly referred to as the the new word problem. When encountered in the context of an embedded speech recognition system this problem can be be divided into the following sub-problems. First, identify the presence of a new word. Second, acquire a phonetic transcription of the new word. Third, acquire the orthographic transcription (spelling) of the new word. In this paper we present the results of a preliminary study that employs a novel approach to the problem of acquiring the orthographic transcription through the use of an n-gram language model of english spelling and a quad-letter labeling of acoustic models that when taken together potentially produce an acoustic to spelling transcription of any spoken input.

Introduction

This paper focuses on the problem of acquiring the orthographic transcription of new words and explicitly ignores the problems of identifying the presence of a new word and generating the phonetic base-form of the new word. The approach that we employ here is to map directly from the acoustic evidence to an orthographic transcription. In other words we model the acoustics of our training set based on the readily available orthographic transcription of the sentence instead of a phonetic transcription. The language model that we employ is the familiar n-gram model. Our model consists of a five gram with 27 tokens, *A* through *Z* plus *blank*. One may reasonably ask what led us to think that a reasonable level of performance would be possible. A question is the answer in this case. Ask yourself how many guesses you might require to get the fifth letter correct in a five letter sequence if you had been given the previous 4 letters? We guessed that a perplexity of english spelling might be somewhere between two and five for a five gram language model. A more detailed analysis of the perplexity of english spelling can be found in [Shannon 51]. Given such a low perplexity we believed it would be possible to overcome much of the inherent ambiguity in english spelling.

Acoustic Models

Signal Processing

The signal processing front-end used in this work is identical to the Sphinx front-end [Lee 89]. We computed power and 12 bilinear-transformed LPC cepstral coefficients, which are then quantized into three different codebooks: (1) 12 stationary coefficients, (2) 12 differential coefficients, and (3) power and differenced power. Each codebook has 256 entries, thereby reducing each centisecond of speech into three bytes.

HMM Inventory

In most speech recognition systems, subword units are based on phonemes, which we believe to be most appropriate for this task as well. However, deriving orthography from phonemic units requires a probabilistic phoneme-to-spelling component, as well as a complex search algorithm that satisfies orthographic as well as phonemic context constraints. This was considerably more effort than warranted for a preliminary study such as this one. Therefore, we compromised some accuracy by using letters of the alphabet and *blank* as our speech unit. In other words, a hidden Markov model represents each letter of the alphabet.

One serious problem with letter models is that letters are highly context-dependent -- much more so than phonemes. For example, the "h" in "sh", "ch", "th", "eh", are extremely different. In order to deal with this problem, we modeled the letters in a context-dependent fashion. Since there are only 28 units (26 letters, *blank*, and silence), we could afford to train very detailed units that model each letter in the context of its two left letters and one right letter. We shall refer to this model as the *quad-letter* model. For example, the letter "h" in the word "school" is aware that its two left neighbors are "s" and "c", and that its right neighbor is "o". From this information, the proper pronunciation of each letter can be inferred, and context can be modeled in the same spirit as triphones [Schwartz 85]. Since not all quad-letters occur frequently enough, we model only those that occur often enough. The less frequent ones would be merged into *tri-letter* (one left and one right context) or *bi-letter* (right context only) models. This resulted in a total of 1427 quad-letter models.

Another problem is that letters do not always have acoustic realizations. For example, the letters "g" and "h" are silent in *night*. In order to deal with this problem, we used a hidden Markov model that allowed the entire model to be skipped. Our model is shown in Figure 1. Since this skip probability is context-dependent, silent letters will have very high probabilities of being skipped.

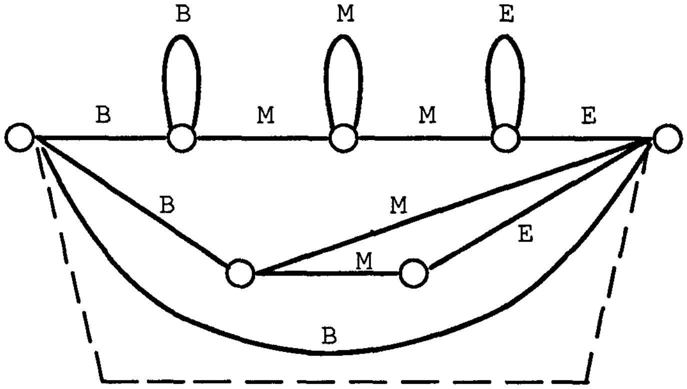

Figure 1: The Quad-Letter HMM

HMM Training

We used a total of 15,000 training sentences to train the quad-letter models. The training data included Resource Management, TIMIT, as well as locally-recorded Harvard sentences and General English database. These databases are described in [Hon 89].

First, an inventory of 1427 models was detemined by examining the frequencies of the quad-letters in the training set. These models are trained on sentences of connected speech, using an initialization from the context-independent letter models. During the quad-letter training, for each sentence, the corresponding quad-letters of each word are concatenated. Two words are connected by a *blank*, and silence models are used in the beginning and the end of the sentence. Thus, the *blank* model could correspond to silence, glottal stop, or nothing. Therefore, we model *blank* in a context-dependent fashion, in the same manner as quad-letters. A single silence model is used, but silence is used as a letter-context for quad-letters.

Two iterations of the forward-backward algorithm were run, and the resulting quad-letter models are smoothed by deleted interpolation [Jelinek 80] with context-independent letter models.

Language Models

The language model used here is a five gram model where we determine $P(l_4/l_1l_2l_3l_5)$ based on a training set. It is possible to train this five gram model because the size of the lexicon is only 27, ie. the letters *A* through *Z* plus *blank*. So there are potentially 27^5 (14,348,907) probabilities to determine. To train this model we chose a suitably large data base in the digital version of the Academic American Encyclopedia. The encyclopedia was preprocessed to remove tokens that were not part of the the 27 token lexicon as well as tokens that were likely to add noise to the language model. Examples of such 'noise' tokens would be those tokens that appeared between delimiters such as () and [] and those tokens that appeared in sentences that were fewer than ten tokens in length. After preprocessing 45 million tokens remained from which two language models were developed. The first language model included the *blank* token and modeled spelling across word boundaries. The second model did not contain the *blank* token and modeled the spelling of words in an unspecified context. Finally, in each of these language models, the five grams where interpolated with their four grams and the four grams with their three grams etc. when the number of observations was below a specified threshold. For the sake of simplicity in the recognizer five grams that were not observed are assumed to have a probability of zero. We note here that all of the five grams in our test happened to be modeled in the training set even though the two data sets were disjoint.

Recognition System

The recognition system used is a version of the one used in Sphinx [Lee 89] with the following additions. First we adapted it to compute the additional null transition in the acoustic models and second we adapted it to be able to manipulate our simplified five gram language model.

Results for continuous speech

Two experiments were performed. The first experiment was performed on a set of 25 general english sentences from 5 different speakers. For this experiment the language model that included *blank* was used. The results were a letter accuracy of 59.3% and an error rate of 54.3%. The spelling perplexity of this test set was 2.09. One problem we observed was that the system was unable to reliably find correct word boundaries and that this was a source of many errors. Below we present the two most accurately transcribed sentences from this experiment.

```
SENTENCE 1  (ge4214)
Correct =  77.1%, Errors =  43.8%

REF:he was also NOUr*IsH**inG A h*atRed** OF*** intellectu*als
HYP:he was also MARrY sTRAin* **hEat*ed A UNDED intellectuRals

SENTENCE 23  (ge4262)
Correct =  79.4%, Errors =  33.3%

REF:Hoffer EAr*nED his living asA dish*wash**e**R
HYP:Coffer *PrIn*G his living as**dish washIReD A

REF:Lumber**jackAND migrant
HYP:NumberG jack*IM*migrant
```

Results for end point detected embedded words

The second experiment perhaps is more indicative of the conditions an acoustic to spelling transcriber will be expected to operate in. In this experiment we identified 30 ship and place names from the 1987 Resource Management test set. Using the begin and end times identified by the Sphinx recognizer we analyzed this portion of the utterance with the second language model (the one that does not contain *blank*). The results were a letter accuracy of 72.7% and an error rate of 39.3% and a string accuracy of 21.1%. The spelling perplexity of the test set words was 4.04. It is not surprising that this perplexity is twice as large as the General English test set since the spelling for names is not as well constrained as the rest of english spelling. Despite this higher perplexity, the accuracy is much improved over the previous experiment though it remains to be seen if the end points of unknown words can be determined as accurately as the end points determined by Sphinx when it knew the word in question.

Summary

The performance is still too low to suggest that this approach be employed to address the new word problem. In future work we will address the use of phonetic units for acoustic modeling with an intermediate mapping from phonetic units to english spelling or perhaps to the syllable level and then to english spelling. Also the problems of identifying the presence of a new word and of creating the baseform for the new word must be addressed before we can fully integrate the new word into a speech recognizer.

References

[Hon 89] Hon, H.W., Lee, K.F., Weide, R.
Towards Speech Recognition Without Vocabulary-Specific Training.
Submitted to Eurospeech '89.
1989

[Jelinek 80] Jelinek, F., Mercer, R.L.
Interpolated Estimation of Markov Source Parameters from Sparse Data.
In E.S. Gelsema and L.N. Kanal (editor), *Pattern Recognition in Practice*, pages 381-397. North-Holland Publishing Company, Amsterdam, the Netherlands, 1980.

[Lee 89] Lee. K.F., Hon, H.W., Hwang, M.Y., Mahajan, S., Reddy, R.
The SPHINX Speech Recognition System.
In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. April, 1989.

[Schwartz 85] Schwartz, R., Chow, Y., Kimball, O., Roucos, S., Krasner, M., Makhoul, J.
Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech.
In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. April, 1985.

[Shannon 51] Shannon.
Prediction and Entropy of Printed English.
Bell Systems Technical Journal 30:50-64, 1951.