AUTOMATIC DETECTION OF NEW WORDS IN A LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION SYSTEM

Ayman Asadi
Northeastern University
Boston, MA 02115

Richard Schwartz and John Makhoul
BBN Systems and Technologies Corporation
Cambridge, MA 02138

ABSTRACT
In practical large vocabulary continuous speech recognition systems, it is nearly impossible for a speaker to remember which words are in the vocabulary. In this paper, we describe a preliminary investigation of techniques that automatically detect when the speaker has used a word that is not in the vocabulary. We developed a technique that uses a general model for the acoustics of any word to recognize the existence of new words. Using this general word model, we measure the correct detection of new words versus the false alarm rate. Experiments were run using the DARPA 1000-word Resource Management Database for continuous speech recognition. Preliminary results indicate a detection rate for new words of 74% with a false alarm rate of 3.4%.

1 THE NEW-WORD PROBLEM
Current continuous speech recognition systems are designed to recognize words within the vocabulary of the system. When a new word is spoken, the system recognizes other words that are in the vocabulary in place of the new word. When this happens, the user does not know that the problem is that one of the words spoken is not in the vocabulary. He assumes that the system simply misrecognized the word, and therefore he says the sentence again and again. Current systems do not tell the user what the problem is, which could be very frustrating.

Adding the ability to detect new words automatically is desirable and improves the performance of the system. Once a new word is detected, it is possible to add the word to the vocabulary with some extra information from the user, such as repeating the word within a carrier phrase and typing in the spelling of the word.

2 APPROACH
An obvious zero-order solution for detection of new words problem is to apply some rejection threshold on the word score. If the score reaches a level higher than the threshold then a new word is detected. However, when we examined the scores of words in a sentence, we found that the score of correct words varies widely, making it impossible to tell whether a word is correct or not. Therefore, this approach for detecting new words did not work well.

Our proposed solution is to develop an explicit model of new words that will be detected whenever a new word occurs. The word model should be general enough to represent any new word. It should score better than other words in the vocabulary where there is a new word; it must always score worse on words in the vocabulary than the model of that word. Given the above assumptions, we tried four acoustic models of new words which are described below.

2.1 Proposed Models For A New Word
All new-word models consist of sequences of phonemes. Each phoneme is represented by a 3-state Hidden Markov Model (HMM). The states are connected from left to right with "self-loops" on each state. There are transition probabilities on these connections. Also, associated with each state is a spectral distribution for the VQ clusters (256 clusters).

The first new-word model we tried consisted of four phonemes (see figure 1). It has 5 states and 4 identical phonemes with an adjustable flat spectral distribution. From now on we'll refer to this model as M1.

The second new-word model that we tried was a model that allows for any sequence of phonemes of at least two phonemes long (see figure 2a). The model has 3 states, all phonemes in parallel from the first state to the second state, all phonemes in parallel from the second state to the third state and all phonemes in parallel looping on the second state. The model has 3N phoneme arcs, where N is the number of phonemes used in the system (N=53 in our system). All phonemes are represented with context-independent phoneme models. Note that this is in contrast to the normal vocabulary of the system, which uses context-dependent phoneme models. The context-independent phoneme models in the new-word model are trained on the same data as the system vocabulary. We'll refer to this model as M2.

The third new-word model we used was similar to the second model (see figure 2b). It has 5 states with a minimum of 4 phonemes. The model has 5N phoneme arcs. All phoneme models are context independent. We'll refer to this model as M3.

The fourth word model was a diphone word model (see figure 3). It consisted of models of the phonemes in the context of the previous phoneme (diphones). It allows for any sequence of diphones with a minimum of two diphones. This model has
phoneme HMM with 3 flat spectral distributions

Figure 1: Flat New-Word Model (M1).

3-state new-word model
Requires 2 or more phonemes.

5-state new-word model
Requires 4 or more phonemes.

Figure 2: 3-State and 5-State New-Word Models (M2 and M3).

N+2 states, N is the number of phonemes used in the system and the 2 for the beginning and end word boundaries. It has N^2 + 2N left-context phoneme arcs. We’ll refer to this model as M4.

2.2 Grammar
We used a first order statistical class grammar. A statistical class grammar consists of class nodes, word arcs and arcs between classes. Each class node has a number of word arcs emerging from it. Word arcs in the same class have equal probabilities. There are transition probabilities between the classes that depend on the training given to the class grammar.

New words are more likely to appear in open classes than closed classes. Open classes are the classes that accept new words (e.g., ship names, port names) as opposed to closed classes that do not accept new words (e.g., months, week-days, digits). We created a separate new-word model for each open class to make the distinction whether the new word was a ship name or a port name, etc. Also, it is easy to add the open class words to statistical class grammars and to Natural Language syntax and semantics. The open classes for which we created new-word models are ship name, ship name possessive, port name, water name, capability, land name and track name.

3 EXPERIMENTS AND RESULTS
The experiments we describe here use the DARPA 1000-word Resource Management Database for continuous speech recognition and BYBLOS, the BBN continuous speech recognition system [1]. The database is limited to 1000 words and does not include data to test for new words. Therefore, we simulated new words in the system simply by removing words from the 1000 word vocabulary that occur in the test sentences.

In the following we give results for experiments that used the four word models for a new word. The experiments were run on 7 speakers from the speaker dependent portion of the database (BEF, CMR, DMS, DTB, DTD, JWS and PGH), 25 test sentences per speaker. We created a statistical class grammar from the remaining words in the vocabulary. We varied the perplexity of the statistical class grammars simply by changing the number of training sentences. A bias for new words was implemented in the case of the flat new-word model (M1); a bias against new words was implemented to reduce the false alarm rate in the 3-state, 5-state and diphone models (M2, M3 and M4). The bias is a scaling factor which is multiplied by the new word arc probability from an open class, reducing the probability of selecting the new word from that class with respect to the rest of the words in the class.
We described below, then followed by an example as an illustration:

- **classes**: the classes of new words allowed.
- **perp**: the perplexity of the grammar.
- **cor**: correct or exact detection rate as a percentage of number of new words. The removed word was exactly replaced by the new word model of the same class. In the following example, the lines marked **REF** is what was actually spoken; the lines marked **HYP** is what was recognized by the system. The sentence appears on more than one line for clarity. The words LAMPS, of class **capability**, and MOZAMBIQUE, of class **water name**, were removed from the vocabulary. The system recognized **NEW-CAPABILITY** and **NEW-WATER-NAME**, respectively, exactly in their places.

### SENTENCE (1237)

**REF**: how many LAMPS

**HYP**: how many **NEW-CAPABILITY**

**REF**: cruisers are in

**HYP**: cruisers are in

**REF**: MOZAMBIQUE channel

**HYP**: **NEW-WATER-NAME** channel

- **cls**: close call or close detection rate. That is, the new word was detected but there was an insertion or deletion in its vicinity. The new-word HAWKBILL, of class **ship name**, was recognized correctly as **NEW-SHIP-NAME**, but the word due DUE was deleted.

### SENTENCE (0464)

**REF**: when HAWKBILL DUE in

**HYP**: when **NEW-SHIP-NAME** *** in

**REF**: port

**HYP**: port

- **sw**: switch between classes, i.e., the new word was detected, but was assigned to the wrong class. In the following example the word PEORIA is a **ship name** and the system recognized it as a **ship name possessive**.

### SENTENCE (1006)

**REF**: when PEORIA last

**HYP**: when **NEW-SHIP-NAME** last

**REF**: in the atlantic ocean

**HYP**: in the atlantic ocean

- **det**: total detection rate, sum of cor, cls and sw. This is the rate of correctly detecting the existence of a new word in the vicinity of its occurrence. (While we would like the system to detect the exact location and the class of the new words, it is also useful to simply detect that a new word has occurred).
false alarm rate, percentage of number of false alarms to the total number of test sentences. A false alarm is a new word detected where there was no new words in that part of the test sentence. In the following example the word *AREA* was misrecognized as *NEW-TRACK-NAME*, which is a false alarm for a word from the *track name* class.

**Sentence (1025)**

**Ref:** WHEN DID Sherman last

**Hyp:** WHEN+S THE Sherman last

**Ref:** downgrade for asuw

**Hyp:** downgrade for asuw

**Ref:** mission AREA

**Hyp:** mission NEW-TRACK-NAME

| classes    | perp | cor | cls | sw  | det | fal |
|------------|------|-----|-----|-----|-----|-----|
| shiptname+ | 100  | 42  | 36  | 5   | 83  | 1.7 |
| shipname+  | 60   | 49  | 30  | 5   | 84  | 1.1 |
| portname   | 100  | 27  | 37  | -   | 64  | 0.6 |
| 7 classes  | 100  | 44  | 6   | 24  | 74  | 3.4 |

Table 1: Detection of new-words results using the model M2.

For the word model M2, the first experiment was detecting new words from the classes *ship name* and *ship name possessive*. The perplexity of the grammar was 100. We had a detection rate of 83% and a false alarm rate of 1.7%. In the second experiment we changed the perplexity of the grammar to 60 to measure the effect of the perplexity on the detection rate and the false alarm rate. There was no significant difference in the detection rate (84%) but the false alarm rate dropped to 1.1%, a reduction of 35% in the false alarm rate. Our third experiment was detecting new words from the class *port name* with grammar of perplexity 100. We had a detection rate of 64% and a false alarm rate of 0.6%. In the fourth experiment we tried to detect new words from 7 different classes, with a grammar of perplexity 100, the detection rate was 74% and the false alarm rate was 3.4%.

In Table 2, we compare the detection results using the new-word models M2, M3 and M4. The experiments were run on detecting the same set of removed words from 7 classes with a grammar of perplexity 100, and similar bias against the new words.

| model | cor | cls | sw  | det | fal |
|-------|-----|-----|-----|-----|-----|
| M2    | 44  | 6   | 24  | 74  | 5.4 |
| M3    | 37  | 8   | 26  | 71  | 4.0 |
| M4    | 21  | 14  | 41  | 76  | 8.6 |

Table 2: Detection of new-words results for M2, M3 and M4 with 7 new-word classes and grammar of perplexity 100.

The results for the model M2 are the same as the last row in Table 1. The results for the model M3 are 71% detection rate and 4% false alarm rate. As for the model M4, the results are 76% detection rate and 8.6% false alarm rate.

From Table 2 we can say that M2 is the best new-word model because it has the lowest false alarm rate and a high detection rate. M3 results are very close to M2 results, but M2 outperforms M3 in all categories. M4 has the highest detection rate but it has also the highest false alarm rate. The high false alarm rate is due to the fact that the diphone model M4 matches the existing words much better than the 3-state model M2. That is, even though the diphone model is somewhat better for new words, it is a much better model for existing words because it has been trained on the existing words. In fact, it matches existing words better than their own models because it does not have the sequential constraints that are in the models of the existing words. Therefore, it is hard to tune the bias such that the diphone model M4 detects only new words. In addition, the diphone model requires much more computation, since it consists of all possible diphones.

### 4 Conclusion

From the above results we have shown that the problem of detecting new words can be solved by selecting an explicit word model for new words. We tried 4 models for new words and compared their results. The 3-state model, consisting of two or more context-independent phonemes, has a high detection rate of 74% and the lowest false alarm rate of 3.4%. The 5-state model did not show any advantages for increasing the minimum length of new words to 4 phonemes. The 3-state model outperformed it in all aspects. The diphone model had a high false alarm rate because it models the existing words very well. Reducing the perplexity of the class grammar from 100 to 60 does not affect the detection rate significantly but reduces the false alarm rate.

Acknowledgment

The work reported here was supported by the Advanced Research Projects Agency and was monitored by the Office of Naval Research under Contract Nos. N00014-85-C-0279 and N00014-89-C-0008.

References

[1] Y.L. Chow, M.O. Dunham, O.A. Kimball, M.A. Krasner, G.F. Kubala, J. Makhoul, P.J. Price, S. Roucos, and R.M. Schwartz. "BYBLOS: The BBN Continuous Speech Recognition System." In *IEEE Int. Conf. Acoust.*, *Speech, Signal Processing*, pages 89–92. Dallas, TX, April 1987. Paper No. 3.7.