DETECTION AND TRANSCRIPTION OF NEW WORDS

B. Suhm, M. Woszczyna and A. Waibel

University of Karlsruhe
Carnegie Mellon University

ABSTRACT

This paper describes a model which enables a speech recognition system to automatically detect new words and to provide a rough phonetic transcription. In our approach to the new word problem the decision whether new words occurred in the speech input is not based exclusively on acoustic evidence but also on a language model designed to support the detection of new words. We describe preliminary experiments to create new word grammars on the Wall Street Journal task. Furthermore we present recognition results of our new word model using the recognition engine of the JANUS speech to speech translation system [1, 2], designed around the task of conference registration.

1. INTRODUCTION

The design of a speech recognition system for real applications requires addressing the new word problem since it is impossible to create vocabularies with 100% coverage of spontaneous input. The ability to detect and transcribe new words would enable a speech recognition system to handle their occurrence gracefully. Depending on the application it could replace the new word by its transcription, interact with the user, add frequently occurring new words to the vocabulary or transfer the new word directly to an output utterance.

In our approach to the new word problem new words are modeled as an arbitrary sequence of phonemes, similar to the system of Asadi [4]. However, the detection of new words based on acoustic evidence alone doesn't seem to be very promising [5]. We propose using a language model designed to capture possible locations of new words.

In the following section we examine some aspects of the new word problem which prove to be important to the design of a new word recognizer. In the third section we describe the concept of a new word grammar and introduce some perplexity measures, using text material from the Wall Street Journal as an example. Then we describe the implementation of a new word model in the recognition engine of the JANUS speech–to–speech translation system. Finally we present recognition results on recordings from the Conference Registration task.

2. ANALYSIS OF THE NEW WORD PROBLEM

We intentionally desist from the acoustics of a speech recognizer and consider the new word problem on a large text database, regarding all words outside a certain vocabulary as new words. We chose the Wall Street Journal (WSJ) database because its size makes it possible to examine new words in their natural frequency of occurrence.

For the experiments described below, we created vocabularies of different sizes by determining the unique words occurring in the training sentences. Additionly, we kept a separate test corpus consisting of 1500 sentences. Table 1 shows the number of sentences used to create the vocabularies, the vocabulary size, and the coverage on the test corpus.

name	sentences	vocabulary size	coverage
V1	250	1721	72.8%
V2	500	2460	77.6%
V3	1000	4143	84.7%
V4	3000	7968	91.7%
V5	5000	10549	93.9%
V6	9000	14072	95.8%

Table 1: Vocabularies

2.1. CLASSIFICATION OF NEW WORDS

It is obvious that even speech recognizers able to deal with new words should be designed such that their vocabulary obtains a high coverage of the expected speech input. To get an idea of what kind of words occur as new words we examined new words with respect to vocabulary V6.

We assigned each of the 1096 new words occurring in the test corpus to one of four classes: names, inflections, concatenations (of words from within the vocabulary) and other. The distribution is shown in Table 2.

As can be seen, names represent a considerable percentage in new words. Moreover, most new words are inflections from words in the vocabulary. To address the problem of inflections in a speech recognizer, one could store only the word roots and provide some al-

names	inflections	concatenations	other
27%	45%	6%	22%

Table 2: Classification of new words in *WSJ*.

gorithm to derive automatically the valid inflections, rather than storing all words of the vocabulary explicitly. Concatenations should be distinguishable from their separate components through the language model.

2.2. LENGTH OF NEW WORDS

In speech recognition, it is common to use detailed word models consisting of a sequence of phoneme models. The sequence is determined by the transcription of the word. These word models also determine the length of a word, used in the search for the sentence hypothesis to determine word boundaries. However, the length of new words is not known a priori.

We examined the length of new words with respect to different vocabularies in our test corpus. The length of a new word is the number of phonemes in its transcription which we get from a dictionary with full coverage of the test corpus. Figure 1 shows that the distribution of the length of new words is very similar over wide ranges in coverage, with a distinct maximum at 6 phonemes.

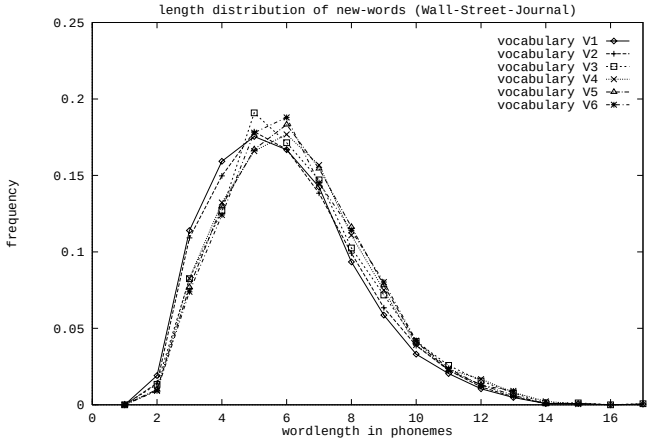

Figure 1: Length of new words in *WSJ*

Figure 2 compares the length distributions of new words, words within vocabulary V5, and all words occurring in the test corpus. Note that words are unique in a vocabulary, whereas they occur with rather different frequencies in the test corpus.

As shown, the length of new words is significantly longer than the average length of words in the text. This is because short function words which occur often are included in the vocabulary. Thus they do not occur as new words. On the other hand, the length of new words is very similar to the length of words within large

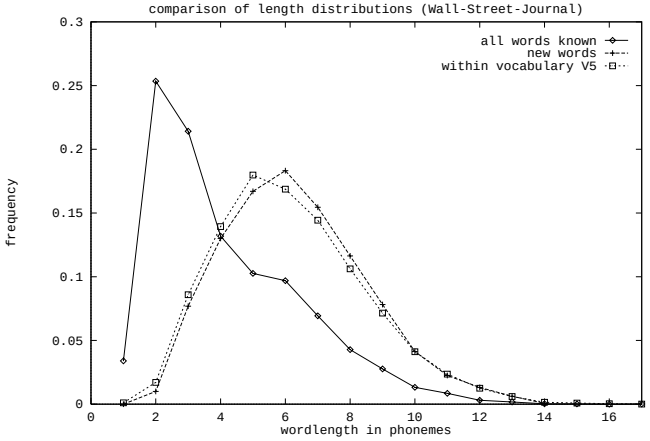

Figure 2: Comparison of word length distributions in *WSJ*

vocabularies, since there are only few unique short function words.

3. NEW WORD GRAMMARS

As mentioned above, we propose capturing possible locations of new words in a special language model, called a **new word grammar**. The new word grammar is basically a statistical grammar augmented with an out–of–vocabulary word class and is generated from large amounts of training text. This is done by mapping all words outside a given vocabulary onto the new word class.

The concept of perplexity [3] is commonly used as a measure for the degree of constraint in a language model. However, perplexity cannot capture the specific ability of a new word grammar to handle new words. Therefore, we introduce additional measures which we define below for the case of a trigram new word grammar. They can easily be extended to general N–gram new word grammars.

We assume that new word grammars are smoothed, i.e. no trigram is assigned a zero probability. The test corpus is represented by its sequence of words $w_1 \cdots w_M$.

We define **new word detection perplexity** $PP_{Detection}$ as a measure of the ability of a new word grammar to predict a new word based on the history of the two preceding words:

$$PP_{Detection} = 2^{LP} \quad \text{where}$$

$$LP = -\frac{1}{N} \sum \log P\left(\text{New Word} \mid w_{i-2} \quad w_{i-1}\right) \qquad (1)$$

The sum is taken over all trigrams $w_{i-2}w_{i-1}w_i$ where w_i is actually a new word; N denotes the number of new word occurrences in the test corpus.

To capture the tendency to predict new words where actually no new words occur, we define a **false alarm**

training sentences	4000	12000	36000
new words	2398	1802	1664
$PP_{Detection}$	18.0	15.7	10.9
$PP_{FalseAlarm}$	21.1	23.6	18.2
$PP_{PostNewWord}$	13.1	12.8	12.2
Trigram PP	258.6	238.6	197.6

Table 3: New word grammars on *WSJ*

perplexity $PP_{FalseAlarm}$:

$$PP_{FalseAlarm} = 2^{LP} \quad \text{where}$$
$$LP = -\frac{1}{M}\sum \log P(\text{New Word} \mid w_{i-2} \quad w_{i-1}) \quad (2)$$

In contrast to (1) the sum is taken over all trigrams $w_{i-2}w_{i-1}w_i$ where w_i is *not* a new word. M denotes the total number of trigrams occurring in the test corpus.

Having detected a new word, the correct *known* words following the new word have to be found. Therefore, we finally define a measure of the ability of the new word grammar to model the words following a new word, called **post new word perplexity**:

$$PP_{PostNewWord} = 2^{LP} \quad \text{where}$$
$$LP = -\frac{1}{2N}\sum \log P(w_i \mid w_{i-2} \quad \text{New Word}) + \log P(w_i \mid \text{New Word} \quad w_{i-1}) \quad (3)$$

We created new word grammars for the WSJ task on training sets with an increasing number of sentences. The definition for new words refers to vocabulary V5 with a high coverage on the test corpus.

Table 3 shows the number of training sentences, the number of new words occurring in the test corpus and the perplexity measures introduced in the previous section. Additionly, we computed the regular trigram perplexity, measured over all trigrams of the test corpus.

4. RECOGNITION OF NEW WORDS

The recognition engine of the most recent version of JANUS [6] uses context dependent LVQ triphoneme models. The search algorithm builds a sorted list of sentence hypotheses using a bigram language model. The system has been designed around the Conference Registration task (CR), a speech database of 12 read dialogs with a vocabulary of around 400 words in its English version.

For the design of a new word recognizer on CR we have restricted the definition of new words to some open word classes (names of persons, cities, streets, etc.). Here the vocabulary coverage in general is poor, as shown in Table 2. Because of the relatively small number of sentences available for CR, we trained a bigram new word grammar instead of a trigram new word grammar, using some word classes (digits, states, languages etc.).

Assuming that all short words are included in the system's vocabulary, our generic new word model allows for arbitrary sequences of at least 3 phonemes. A bias against entering the new word model reduces the number of false alarms.

In contrast to the context dependent modeling of words within the vocabulary, we use context independent LVQ phoneme models within new words for reasons of search efficiency. Additionly, we impose a triphone bias on transitions between phonemes within a new word. The triphone transition probabilites were trained on a 32000 word dictionary from Wall Street Journal.

To prevent the new word model from absorbing known words following new words we impose an additional length dependent penalty based on the distribution of the length of new words, as shown in Figure 1.

If a sentence hypothesis contains new words they are represented by the sequence of phonemes found in the search to match best the corresponding region in the speech input, providing a rough phonetic transcription.

5. RECOGNITION RESULTS

Preliminary tests have been performed on new recordings from the English Conference Registration (CR) database. The test set consists of 59 sentences (578 words), including 42 names. The set has been recorded by 4 speakers. All names have been removed from the system's vocabulary.

To measure the ability of the new word model to detect new words, one can define a detection rate (as a percentage of the number of new words occurring in the test sentences) and a false alarm rate (as a percentage of the total number of words in the test sentences). However, by lowering the bias against the new word model, one can trade a high detection rate against a high false alarm rate, and vice versa. Figure 3 shows the new word detection rate and the word accuracy as a function of the false alarm rate.

We suggest combining the new word detection rate and false alarm rate to find a Figure Of Merit (FOM). We define the FOM as the detection rate averaged over false alarm rates from 1% to 6%. The word accuracy is averaged accordingly.

The transcription accuracy is measured over the parts in the ouput hypothesis where the segmentation in known and new words was correctly determined. A simple DTW between actual transcription of the new word and the transcription provided by the new word recognizer is performed.

Table 4 summarizes the tests of our new word recognizer on the Conference Registration task. It compares the performance of the new word recognizer using

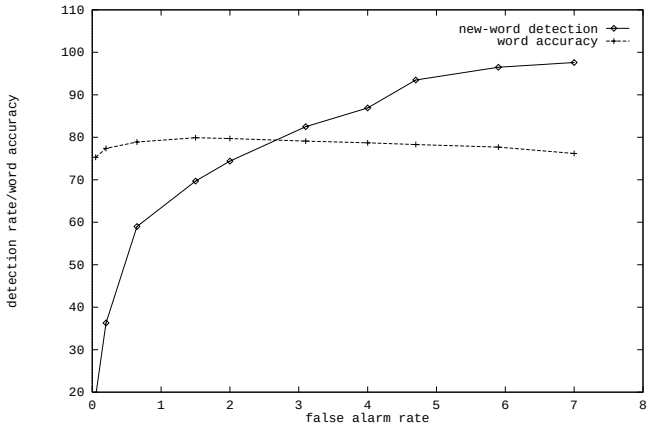

Figure 3: Operator response graph of the new word recognizer (using bigram new word grammar)

	bigram NWG	no NWG
FOM	82.9	55.6
transcription accuracy	37.4	39.7
word accuracy	78.6	74.6
all words known	88.5	88.5
no new word model	71.7	71.7

Table 4: Performance of the new word recognizer

a bigram new word grammar ("bigram NWG") to the performance without new word grammar ("no NWG"). Without new word grammar means that new words may occur after any word, and any word may follow a new word. Additionly, we tested the recognizer without new word model on the same test set and determined the word accuracy. In one experiment names were treated as new words ("no new word model"), in the other experiment names were treated as known words ("all words known"). For all experiments presented in Table 4, the underlying recognizer uses a bigram language model with perplexity 7.4 for transitions between known words.

These results indicate that a high FOM can only be achieved using a new word grammar. The new word model improves the word accuracy up to 6.8%. However, the word accuracy when no new word occur is still more than 10% higher. The transcription accuracy is low, indicating that providing a good transcription for new words is not trivial.

6. CONCLUSION

We have presented a new word model which detects new words both on acoustic and grammatical evidence. Thus, a substantial improvement in the word accurcy is achieved when the speech input contains out–of–vocabulary words.

The recognition results are based on the first–best hypothesis. Further improvements could be achieved using the N–best sentence search by reordering the list of sentence hypotheses using trigram new word grammars. Preliminary tests on CR yield promising results.

Research is in progress to redesign JANUS for the task of scheduling, where two partners try to arrange for a meeting in a spontaneous dialog. This task would provide the possibility to test the presented new word model on recordings where new words occur in their natural frequencies.

7. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the help of A. McNair, D. Roy, C. Wood and W. Ward. This project was carried out during a visiting appointment of the first author at Carnegie Mellon. The project is supported in part by grants from the NSF and ARPA.

8. REFERENCES

[1] Waibel, A., Jain, A., McNair, A., Saito, H., Hauptmann, A., Tebelskis, J.: *JANUS: A Speech–to–Speech Translation System Using Connectionist and Symbolic Processing Strategies,* ICASSP 91, Vol. 2, pp. 793–796

[2] Osterholtz, L., McNair, A., Rogina, I., Saito, H., Sloboda, T., Tebelskis, J. Waibel, A., Woszczyna, M.: *Testing Generality in JANUS: A Multi–Lingual Speech to Speech Translation System,* ICASSP 92, Vol. 1, pp. 209–212

[3] Jelinek, F.: *Self–Organized Language Modeling for Speech Recognition* In *Readings in Speech Recognition*, Morgan Kaufmann Publishers, 1990.

[4] Asadi, A., Schwartz, R., Makhoul, J.: *Automatic Modeling for Adding New Words to a Large–Vocabulary Continuous Speech Recognition System,* ICASSP 91, Vol. 1, pp. 305–308

[5] Young, S.R., Ward, W.: *Learning New Words from Spontaneous Speech,* ICASSP 93, Vol. 2, pp. 590–591

[6] Woszczyna, M., Coccaro, N., Eisele, A., Lavie, A., McNair, A., Polzin, T., Rogina, I., Rose, C.P., Sloboda, T., Tomita, M., Tsutsumi, J., Aoki–Waibel, N., Waibel, A., Ward, W.: *Recent Advances in JANUS: A Speech Translation System,* DARPA Speech and Natural Language Workshop 1993, session 6 - MT