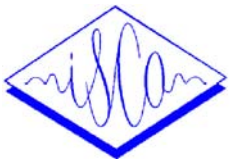
# OOV-DETECTION IN LARGE VOCABULARY SYSTEM USING AUTOMATICALLY DEFINED WORD-FRAGMENTS AS FILLERS



*Dietrich Klakow, Georg Rose, Xavier Aubert*

Philips Research Laboratories
Weisshausstrasse 2, D-52066 Aachen, Germany
{klakow,rose,aubert}@pfa.research.philips.com

6th European Conference on Speech Communication and Technology (EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999



## ABSTRACT

The problem of unknown words has been addressed using automatically generated filler fragments which augment the lexicon and are incorporated in the language model. These fragments are used to reduce the damage on in-vocabulary words, to detect OOV regions and to provide a phonetic transcription for these regions. The performance of this technique has been evaluated in terms of damage reduction error rate and OOV tagging rate. Significant improvements are reported on both measures. In particular, the influence of an appropriate tuning of the language model factor and word penalties is demonstrated as well as the usefulness of using cross-word triphones over fragments boundaries.

## 1. INTRODUCTION

No finite vocabulary can cover all words and hence out-of vocabulary word (OOV) handling is important. However, there is no well defined "OOV-task" and each individual setting of the problem requires completely different techniques. We want to distinguish between three main tasks: reduce the damage caused by an OOV on the words in the vocabulary (e.g. recognition errors adjacent to OOV words), detect or reject OOVs and finally identify semantic properties (e.g. class of OOV) of the OOV. To detect OOVs, confidence measures [1] and fillers for small [2] and large vocabulary [3, 4] may be used. In [5] and [6] the task is extended to classify the detected OOVs in a train-timetable information system. Classes considered here are *CITY* or *SURNAME*. Finally, [7] focuses on language model properties of the OOVs and improves the prediction of OOVs.

The key idea of this paper is to construct multi-phoneme word fragments to be used as fillers. Those fragments are automatically trained and are integrated in the lexicon and the language model the same way as in-vocabulary words. They are used to reduce the damage on in-vocabulary words in a continuous speech recognition task, to detect OOV regions and and to provide a phonetic transcription for these regions.

## 2. TRAINING WORD-FRAGMENTS

An easy way to get a full coverage of any utterance is to add all single phonemes to the lexicon. Those additional phonemes are the simplest examples of word-fragments. However, this simple approach amounts to a parallel phoneme recognizer for unknown words. As statistical properties of the phonemes typical for unknown words are not modeled, performance is poor. Hence, we want to construct better word-fragments and model their properties. To this end, we divide the vocabulary in three regions as illustrated in Figure 1.

The most frequent 5000 words are taken as in-vocabulary words and used as such. The next roughly 5000 words in the frequency sorted list are used as "typical" examples of OOV words. Those words are used to train and capture statistical properties of unknown words. We call those artificial OOVs. All words that are less frequent are true unknown words to which our approach should generalize.

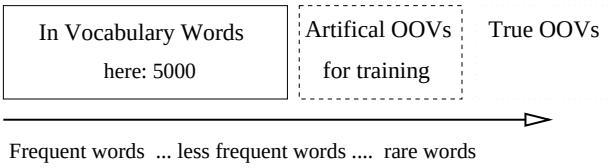



Figure 1: Definition of in-vocabulary and out-of-vocabulary words

In a first step, all artificial OOVs in the training corpus for the language model are replaced by their phoneme sequence taken from our background lexicon. Furthermore, the first phoneme of a word is marked by a special symbol, hence doubling the number of phonemes. This enables us to automatically detect word boundaries even in a sequence of unknown words. Now, the iterative scheme as described in [8] is used to form phoneme phrases which are interpreted as word fragments. To this end, a list of frequent phoneme pairs is created. There is a lower bound on the frequency which determines the number of fragments created. Those pairs that can be joined to one fragment without any conflict are kept in the list. The

appropriate replacement is performed on the corpus. This procedure is iterated until the list is empty.

## 3. ANALYSIS OF FRAGMENT PROPERTIES

A first measure of quality of the fragments is given by perplexity. In order to obtain a number that is relevant to the issue of unknown words, a modification is necessary. All scores that do not predict a fragment or do not contain a fragment in a fourgram history are removed from the perplexity calculation. Perplexity is normalized to all proper words covered by this procedure.

Figure 2 displays the results. The perplexity values appear very high. This is a consequence of the above definition of perplexity and illustrates the difficulty of predicting individual OOVs. A first observation is that there is an optimal number of fragments. Changing from bigrams to trigrams gives a very nice improvement. Fourgrams have also been considered. They gave no additional improvement. The reason for this will be given below.

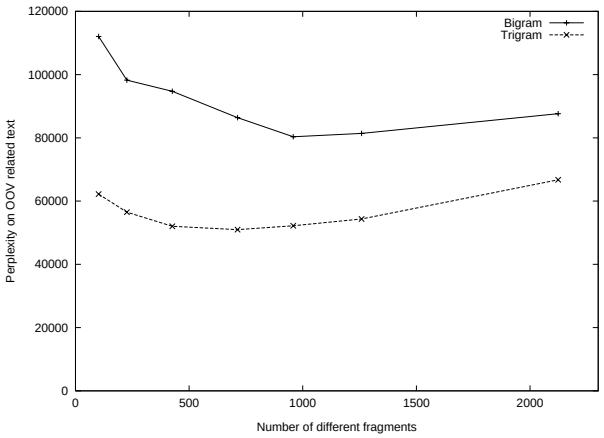



Figure 2: Perplexity for increasing number of fragments

Figure 3 shows the number of fragments that have to be concatenated on average to produce one unknown word as a function of the fragment inventory size. The very left data point indicates that the OOVs have an average length of six phonemes. Only half the number of fragments is needed to cover one OOV as we increase the number of fragments in the lexicon to $\approx 1500$. As trigrams are used in this work, it is obvious, that the language model context expressed in true words breaks down after an OOV. However, this is also true for a pure word LM, as the OOV is not seen in the history and backing-off has to be performed. To increase the language model span, distance modeling is needed which was not the focus of this paper.

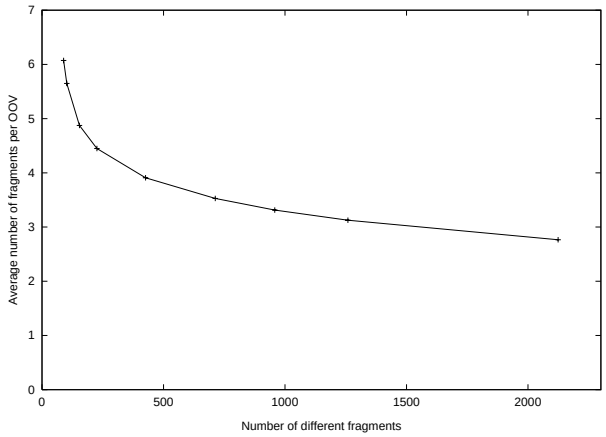



Figure 3: Number of fragments that make one OOV word on average.

## 4. RECOGNITION SET-UP

The recognition experiments were performed on the male part of the DARPA NAB'94 development set consisting of 155 sentences of 10 speakers. A reduced lexicon of only 5500 words plus all fragments was used resulting in an OOV rate of 10%. We used the Philips Large-Vocabulary-Continuous-Speech-Recognizer [9]. Mel-Cepstrum feature vectors spanning 25ms Haming windows at a frame shift of 10ms were calculated. We used speaker-independent acoustic models which were trained on the WSJ0+1 database (284 speakers) resulting in 183k densities with a globally pooled variance.

The very first experiments were performed with a within-word Viterbi decoder using triphone acoustic models. This led to very poor performance on the OOV regions. This is due to the mismatch of within-word triphones at the fragment boundaries. We next trained triphones and monophones in parallel giving in total 220k densities. For the fragment recognition we exclusively used monophones.

As a second approach we applied a cross-word decoder [10] using within-word acoustic references with the additional constraint that cross-word transitions are only applied over fragments. To be able to tune the total ratio of fragments and words we introduced three different word penalties: (a) a word penalty (WP) which is used between two words, (b) a garbage penalty (GP) for transitions from words to word-start fragments, (c) a OOV-length penalty (LP) for each fragment-fragment transition.

Using a bigram language model and the reduced lexicon without fragments we obtained a baseline a WER of 29.9% as opposed to WER of a 14.8% achieved with the full 64k lexicon, conforming to the rule of the thumb that each OOV results in 1.5 errors on average. The corresponding trigram results are 28.9% and 11.3%.

## 5. DAMAGE REDUCTION TASK

For many applications it is acceptable to produce deletions at OOV positions instead of substitutions and insertions. In order to investigate the performance of our fragment approach in terms of this garbage modeling we evaluated the damage reduction error rate (DRER) defined as the WER of the recognized text after removing all fragments.
Figure 4 shows the DRER as a function of the word penalty and OOV-length penalty both for bigram and trigram language models. For these investigations we only used the within-word decoder and the optimized garbage penalty and language model factor. Focusing on the word penalty an optimum of the DRER can be found at $5 \times 10^3$ and $10^4$ for bigrams and trigrams, respectively. For the length penalty we observe a flat minimum. Comparing with the baseline we achieved a optimal WER of 26.6% (24.4%) for the bigram (trigram) giving a relative improvement of 11% (16.3%).

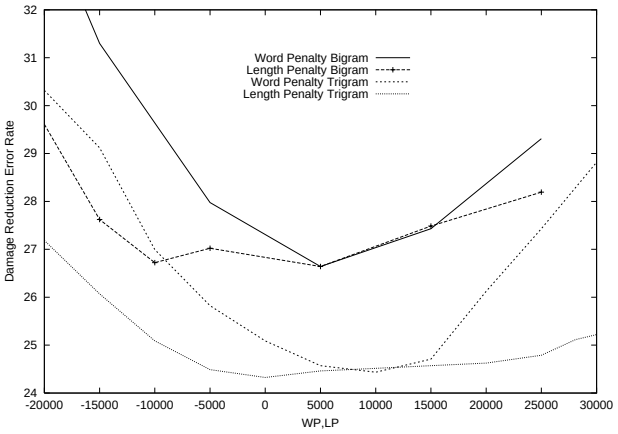



Figure 4: Damage reduction error rate on in-vocabulary words due to fragments for varying word insertion penalty and length penalty.

It is obvious that the garbage penalty strongly influences the frequency of fragment recognition. Using monophones for fragments and triphones for normal words the acoustic modeling of the fragments can be regarded as inaccurate. This explains the fact, that optimal garbage penalties have negative values: $-2.5 \times 10^3$ for bigrams and $-3.5 \times 10^3$ for trigrams (Figure 5).
In Figure 6 we finally investigated the influence of the language model factor on the DRER. At a value of about $10^4$ we obtain the best performance for bigram as well as for trigram.
In Table 1 we summarized the results for the DRER as well as for the OOV-tagging rate for different number of fragments in the lexicon. The other parameters where fixed using WP=0, GP=0 and LP=0. One can observe a small improvement for both tasks, increasing the fragment number from 713 to 1259 which does not seem to continue for an additional increase.

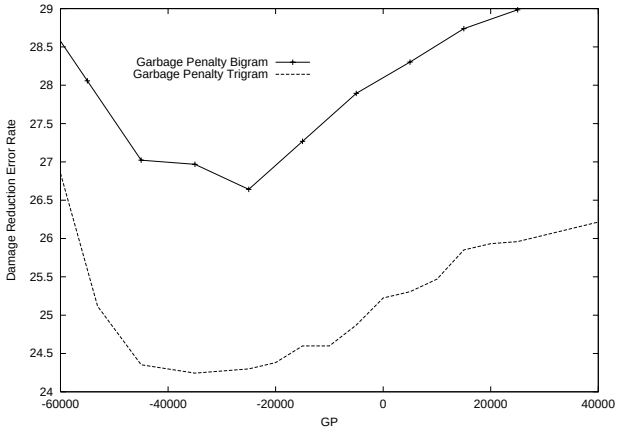



Figure 5: Damage reduction error rate on in-vocabulary words due to fragments for varying garbage penalty.

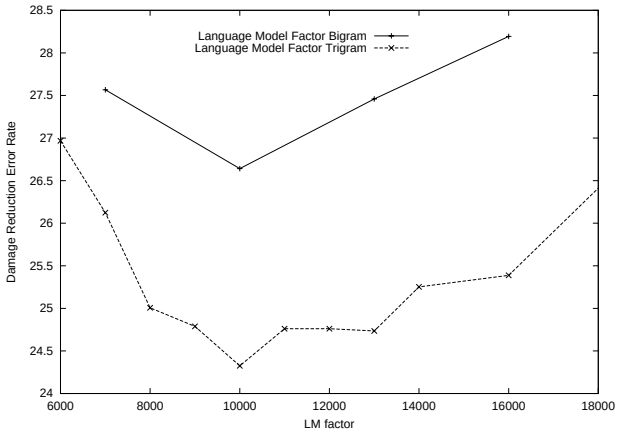



Figure 6: Damage reduction error rate on in-vocabulary words due to fragments for varying language model factor.

We compare the performance of the within-word and the cross-word decoder (WP=0 for bigram and WP=$10^4$ for trigram, GP=0, LP=0). As expected, due to a better acoustic match of the triphones at fragment boundaries the cross-word decoder significantly improves the performance of the fragment approach resulting in relative improvements of 16% for the damage reduction task with a bigram and of 13.5% with a trigram.

| # Frag. | Dam. Red WER | OTER |
|---|---|---|
| 713 | 29.2 | 14.1 |
| 1259 | 28.5 | 13.2 |
| 2124 | 28.6 | 13.2 |

Table 1: DRER and OTER for different number of fragments.

| Case | ww | cw |
|---|---|---|
| Dam. Red. BG | 28.5 | 26.0 |
| OTER BG | 13.2 | 11.3 |
| Dam. Red. TG | - | 25.0 |
| OTER TG | - | 9.0 |

Table 2: Influence of within-word and cross-word decoding for bigram and trigram language model.

## 6. OOV-DETECTION TASK

As a next step towards real OOV recognition, we investigated the OOV-tagging error rate (OTER) in terms of a WER of two class tags, i.e. we changed all in-vocabulary words to the string <word> and OOV-words to <OOV>, which can easily be done as we specially marked word-start-fragments.

We used the recognition results with a lexicon excluding fragments to obtain a baseline OTER of 15.6% and 15.9% for a bigram and trigram language model, respectively. As discussed above, we also analyzed the influence of the four parameters (WP,GP,LP and LM) on the OTER, resulting in an overall OTER of 10.5% for the bigram and 9.4% for trigram language model after tuning. As the number of fragments is increased we observe a saturation in OTER (Table 1) despite the fact that there is a clear minimum in perplexity (Figure 2) due to over-training.

The best tagging rate was achieved using a cross-word decoder and trigram language model giving 9% OTER. This is a relative improvement of 43.4%. However, it was not optimized with respect to the penalty parameters and language model factor and we expect an additional improvement after tuning.

## 7. CONCLUSION

The main advantage of the proposed word-fragments for OOVs is the reduction of errors on words adjacent to OOVs. Here, a reduction of 24% relative has been achieved.

Concerning the OOV-detection task using a trigram language model and cross-word decoding, we can correctly detect 47% of the OOVs, having only 10 false alarms. This amounts to a relative improvement of OTER by 43.4%.

Future work will focus on reducing the search effort for the current set-up due to the presence of numerous short fragments (in particular single phonemes).

Beyond this, we plan to tackle the problem of converting phoneme sequences to proper words.

## 8. ACKNOWLEDGMENT

We would like to thank P. Beyerlein for helpful discussions.

## 9. REFERENCES

[1] A. Bayya, "Rejection in Speech Recognition Systems with Limited Training", *Proc. ICSLP*, (1998) 305.

[2] H. Bourlard, B. D'hoore, J.-M. Boite, "Optimizing Recognition and Rejection Performance in Wordspotting Systems", *Proc. ICASSP*, (1994) I-373.

[3] R. A. Sukkar, "Subword-Based Minimum Verification Error (SB-MVE) Training for Task Independent Utterance Verification", *Proc. ICASSP*, (1998) I-229.

[4] R. El Mliani and D. O'Shaughnessy, "Accurate keyword spotting using strictly lexical fillers", *Proc. ICASSP*, (1997) 907.

[5] E. Nöth F. Gallwitz and H. Niemann, "A category based approach for recognition of out-of-vocabulary words", *Proc. ICSLP*, (1997) 228.

[6] M. Boros, M. Aretoulaki, F. Gallwitz, E. Nöth, H. Niemann, "Semantic Processing of Out-of-Vocabulary Words in a Spoken Dialogue System", *Proc. EUROSPEECH*, (1997) 1887.

[7] T. Kuhn, P. Fetter, A. Kaltenmeier and P. Regel-Brietzmann, "Improved modeling of OOV words in spontaneous speech", *Proc. ICASSP*, (1996) 534.

[8] D. Klakow, "Language-Model Optimization by Mapping of Corpora",*Proc. ICASSP*, (1998) 701.

[9] P. Beyerlein, X. Aubert, R. Haeb-Umbach, D. Klakow, M. Ullrich, A. Wendemuth, P. Wilcox, "Automatic Transcription of English Broadcast News",*Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, Virginia*, (1998).

[10] X. Aubert "One-pass cross-word decoding for large vocabularies based on a lexical tree search organization", elsewhere in these Proceedings.