Detection of OOV Words Using Generalized Word Models and a Semantic Class Language Model

Thomas Schaaf

Interactive Systems Labs
University of Karlsruhe, Germany
tschaaf@ira.uka.de

Abstract

This paper describes an approach to detect out-of-vocabulary words in spontaneous speech using a language model built on semantic categories and a new type of generalized word models consisting of a mixture of specific and general acoustic units. We demonstrate the construction of the generalized word models as replacements for surnames in a German spontaneous travel planning task GSST [1]. We show that the use of our generalized word models improves recognition accuracy in cases where out-of-vocabulary words appear and does not lead to a degradation of the overall recognition accuracy. In our experiments we measured recall and precision rates of OOV-detection which are close to their theoretic optimum. Furthermore, we compared the effect of using cross-word-triphones vs. using context-independent cross-word models. We show that when using generalized word models with cross-word-triphones, the expected number of consequential errors following an OOV word can be reduced significantly by 37%.

1. Introduction

For many languages, especially highly inflected languages and languages that allow compounding of words (such as German) we will in the foreseeable future not be able to use recognizer vocabularies that cover all of the potentially spoken words (> 2 million in one year of a newspaper). While out-of-vocabulary (OOV) words resulting from inflections of known base forms often cause recognition errors that can be recovered in a post-processing step, possibly using linguistic knowledge [2], errors resulting from unknown proper names (persons, places, companies, etc.) are usually difficult if not impossible to detect and repair automatically. For instance, the number of surnames can be very large. Figures 1 and show the coverage of 20 million entries of German surnames taken from a telephone directory. It consists of more than one million unique entries with half a million occurring at least twice.

In many applications, we can not provide all the words to be recognized by a speech recognition system in advance (e.g. in dictation). In such cases, we want to be able to extent the vocabulary of the recognizer to include potential missing words. In dictation systems, the user typically provides the system with unknown words to be included in the recognizer's vocabulary. However, in applications where such an approach is not feasible, we have to detect OOV words automatically. For example in a speech-to-speech translation system, the German utterance "guten Tag ich bin Herr Schaaf" ("Good day I am Mister Schaaf") which contains the OOV word "Schaaf" might be misrecognized as "guten Tag ich bin haarscharf" and translated as "hello I am precise".

The above example also brings up what we call a consequential OOV error. The word "Herr" is being misrecognized although it is part of the vocabulary. This is a side-effect of the misrecognition of the following OOV word. In translation systems, misrecognized proper names often result in absolutely incomprehensible outputs and errors from missing proper names are very annoying to the user.

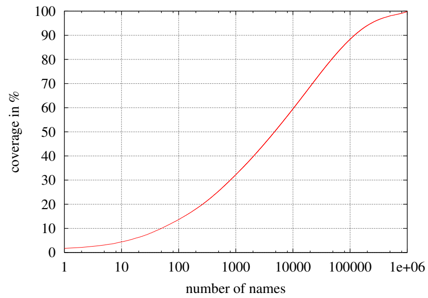

Figure 1: *coverage of German surnames.*

In an ideal scenario, a speech recognizer would detect the fact that an OOV word was spoken and trigger a clarification dialog in which the user can provide acoustic and orthographic representations of the new word. Then the new word can be incorporated into the vocabulary and into the language model [3][1].

2. Head-Tail Structure of Generalized Word Models

A generalized phone is an acoustic unit that models all or at least a large subset of the phone inventory of a language. A generalized word is a word that contains one or more generalized phones. Some speech recognition systems use a sequence of one or more generalized phones (here called a "mumble" word) to model unknown words. Such recognizers expect that in an OOV situation, the likelihood of a mumble word is higher than the likelihood of any well trained word from the regular vocabulary. For low quality speech, especially in spontaneous human-to-human dialogs, we have often observed that a general mumble word eats up too many regular words, often introducing more errors than detecting OOV words.

The most common approach to model OOV is to use a phoneme loop with a duration model [3][4].

In [5] a set of 15 generic whole word models where trained which differ in the number of states. Each whole word model is trained with a set of regular words of a certain number of phonemes. These words are used in combination with a language model class to model OOV words. These small number of whole word models found 21% of the OOV words in a spontaneous speech task with a precision of 49%.

In a more specific approach [6], more than 10.000 syllables are added to the dictionary and treated like normal words in the recognition vocabulary except that they were modeled with context independent phoneme models and the language model probabilities were trained to capture the language specifics. The handicap was that two levels of linguistic models are mixed together with the effect that the language model had to be trained on two different linguistic levels and merged together. Also they try to detect all kind of OOV words which lead will lead to an unspecific language model so that the many of the syllables will always be active and might reduces the recognizer's efficiency dramatically. These approach found on a spontaneous task 24% of the OOV words with a precision of 69%.

Our approach, presented now, uses a set of generalized word models consisting of a specifically modeled head and a general tail (see figure 2). A head consists of a sequence of regular phones and models a prefix of one or more potentially unknown words. A tail consists of a sequence of generalized "mumble" phones and models the remainder of such words. In a time synchronous decoder that proceeds from left to right, it is advantageous to model the head of a generalized word more specifically because this produces competitive likelihoods for the first part of the word such that better pruning decisions can be made. We call this type of generalized word models "Head-Tail-Generalized-Words" (HTGW).

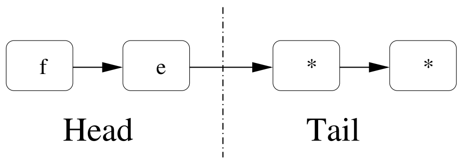

Figure 2: *head-tail-generalized-word.*

One important property of HTGW models is their length. While the general tail has little impact quality, the length of the head is determining the quality of the acoustic match. While longer heads allow to model the words more specifically, the corresponding number of different generalized word models grows exponentially with the sequence length. For the experiments reported in this paper, we used an inventory of 38 phones. We only investigated head sequences of length 0, 1 and 2, since we didn't want to artificially inflate the recognizer's vocabulary. Even for a head length of 2, many of the theoretically possible head sequences ($38^2 = 1444$) are irrelevant or highly unlikely to be observed in any language. Motivated by German phonotactics, we reduced the number of legal heads to the sequences consisting of alternating vowels and consonants. The German phones set consists of 18 vowels and 20 consonants. Many other languages also possess an alternating vowel-consonant phonotactic. For a head sequence length of 2 the set of legal heads contained 780 elements (about half of all possible).

In our experiments, the tail part of the generalized words is modeled using a sequence of one or more instances of one single generalized speech phone. Thus, the number of possible tails is always one for any given tail length. Note that the length of the tail realizes a minimum duration model. Variation of the length of the tail should have a direct effect on the number of consequential errors. Ideally, the length of the tail should reflect the mean duration of OOV words. In practice, one has to experiment with different tail lengths.

As far as the acoustic model is concerned, we investigated the effect of different types of phone models in the head sequences. The question is whether we can use context independent (CI) acoustic models or whether we should use context dependent (CD) models (resulting in cross word triphones (XWT) at word boundaries. Table 1 gives an overview on the investigated combinations of context dependent and context independent acoustic models.

Experiment	length	first phone	second phone
GW1	0	-	-
GW38-CI	1	CI	-
GW38-XWT	1	XWT	-
GW780-2CI	2	CI	CI
GW780-XWT-CI	2	XWT	CI
GW780-XWT-CD	2	XWT	CD

Table 1: *used acoustic models in the head.*

In certain cases, it is possible to refine the head-tail model using additional information about the set of OOV words that should be covered. For example, German street names (mostly compound words) typically end in 'platz' or 'strasse'. Such words might be modeled by Head-Middle-Tail generalized words, with a specific head and tail and a general middle part.

To incorporate our OOV word model into a recognition system, one can either put each generalized word as a separate entry into the recognizer's vocabulary or use a more complex word model (phone graph) that reduces the required number of states as shown in figure 3 (three-state phone models with cross word triphones at the beginning).

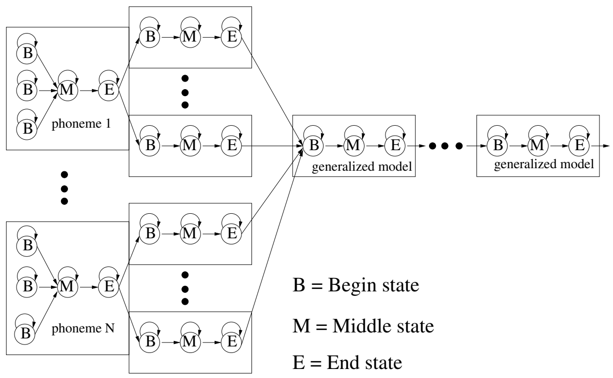

Figure 3: *possible realization of a generalized word.*

3. Semantic Class Language Model

Our application of the Head-Tail-Generalized-Word models is to detect unknown surnames in spontaneous speech. Since the Head-Tail-Generalized-Word models only the acoustic realization of OOV words but not their semantic context, we augment the language model to predict certain classes of OOV words. We use a trigram language model over a vocabulary of 11k words which is trained on 637k word of the GSST [1] task. A semantically motivated class of surnames is introduced and the 436 most frequent surnames are mapped to this class. The top ten names yield 33% of the probability mass and the majority of surnames in the training corpus is not in the recognition vocabulary (OOV words). The resulting language model has a test-set perplexity of 61. The advantage of introducing a surname class is that it has clear semantics and that the number of occurrences allows for robust estimation of the respective n-grams.

If the recognizer's vocabulary is (very) large, an OOV word is often recognized as a sequence of similar sounding in-vocabulary words. In such cases, a semantic class language model should help the generalized-word model in the competition with similar sounding words. It is a straight-forward extension to use the same generalized-words in different semantic classes for example to predict unknown hotel names, towns and surnames.

4. Experiments

4.1. Used data bases for testing

The used recognizer is trained on the GSST domain of time scheduling and travel planing. To measure the influence of the head-tail-generalized-words in spontaneous speech the evaluation set eval99.1 of GSST [1] is used (table 2).

	Time	Spk.	Words	#OOV surname	OOV rate
eval99.1	54 Min.	8	8577	46	1.2%

Table 2: spontaneous evaluation database

A second test-set enriched with surnames was collected (table 3). Each speaker had to read a set of 150 sentences. This set was constructed by using template sentences and a set of top 30 surnames from Hannover's telephone directory.

	Time	Spk.	Words	OOV	#OOV
top30HH	57 Min.	8	5760	20.8%	1200

Table 3: read evaluation database

4.2. Evaluation criteria

As quality measure, we used word error rate (WE) and word + UNK error rate (WCE). The difference to WE is that in cases of an unknown word in the reference the word is mapped to the symbol <UNK> and in the hypotheses every time an HTGW word occurs it is mapped to <UNK>. After this mapping the word accuracy is computed.

To measure how accurately surnames are detected and how many are found, we used precision (PRC) and recall (RCL). All unknown surnames in the reference are mapped to the symbol <UNK:Surname> and all the hypothese's generalized-words of this surname class were mapped to the same symbol. The reference is aligned against the hypotheses and matching symbols are counted as correctly assigned tags. The precision and recall are then computed with the following formulas.

$$PRC_X = \frac{\text{Number of correctly assigned tags for class X}}{\text{Number of total tags for class X}} \quad (1)$$

$$RCL_X = \frac{\text{Number of correctly assigned tags for class X}}{\text{total number of elements in class X}} \quad (2)$$

where $X \in \{surname\}$.

4.3. The baseline recognizer

The baseline recognizer used for the following experiments was built using our Janus-III Speech Recognition Toolkit [7]. The baseline system is an 11k vocabulary fully continuous speech recognizer. For speech extraction, we derived 13 MEL-scale cepstral coefficients (MFCC) with their first and second order derivatives. The dimension is reduced from 39 to 32 by performing a linear discriminant analysis. We use 2500 context dependent subtriphones as acoustic models with 12 Gaussians each, trained with around 52 hours of spontaneous speech from the GSST corpus.

In the decoding phase, incremental adaptation using constraint MLLR [8][9] is used to adapt to speaker and channel condition. This adaption is done after the processing of an utterance. To speed up the recognition, phoneme lookahead [10] and BBI-trees [11] are used. The recognizer runs with 2 real-time on a 300MHz Ultra-Sparc with a baseline result of 22.6% word error rate on the eval99.1 data.

4.4. Details on acoustic modeling of the generalized words

Using the top30HH data, the most important numbers are the WCE and REC. Because every utterance containes an unknown surname, PRC shows only that we hypothesize more than one unknown word per OOV and not that we hypothesize an OOV at a completely wrong place. We tested the baseline system that did not contain the required names (BASELINE). Then we did a cheating experiment where all needed words (surnames) where modeled like a generalized-word but none of the head-tail-generalized-words are in the vocabulary (CHEAT).

First, we can see in the results from table 4 that the cheating experiment has a 0.4% error rate. The baseline is with 38.9% nearly a hundred times greater. This shows that having the right word in the vocabulary makes a big difference in usability. In detail, the baseline system made 2240 errors and the cheating experiment only 195 and identified 1159 surnames correctly. This lead to the result that every unknown word produced 1.8 errors in average, one by itself and 0.8 as consequential errors.

If we take a look on the experiments using the generalized-words we can see in REC and WCE that, as expected, heads of length two are better than of length zero or one. There is one exception between the pure context independent two phoneme head GW720-2CI and the context dependent GW38-XWT with a head length of one, these result and the mutch better result of GW780-XWT-CI compared with GW720-2CI shows the strong importance of using crossword tri-phones at the beginning of a general word. The best results are achieved using a head of length two with context dependent phonemes, which reduced the error rate by more than 50% of the cheating experiment's reduction. In detail the GW780-XWT-CD system made 1215 and found 707 correct surnames, this lead to the result that a correctly placed general word reduced consequential errors by 37% (0.3 per instance) in average per unknown word. This is comparable to the reduction of consequential errors reported in [5].

Next we took a look on the effect of the length of the tail. As shown in table 5, the optimal length seems to be at length two with respect to WCE. Looking on the (plain) word error rate, we see that a tail of length 2 produce 0.8% more error than a tail of length 3, the reason is that more consequential errors remain. This indicates that the tail is to short. The system with a tail of length 3 outperforms also a tail of length 4, because

System	head lengt	WCE	REC	PRC
BASELINE	-	38.9%	-	-
GW1	0	38.7%	2.2%	100.0%
GW38-CI	1	36.9%	8.7%	100.0%
GW38-XWT	1	34.2%	19.2%	100.0%
GW780-2CI	2	35.2%	12.8%	99.4%
GW780-XWT-CI	2	27.0%	38.8%	99.8%
GW780-XWT-CD	2	21.1%	58.9%	100.0%
CHEAT	-	0.4%	96.6%	100.0%

Table 4: results on top30HH data

it has a 1.4% better WCE and only 0.1% more consequential errors remain.

tail	WCE	WE
0	26.3%	42.6%
1	21.0%	35.5%
2	20.7%	34.2%
3	21.1%	33.4%
4	22.5%	33.3%

Table 5: different length of the tail with GW780-XWT-CD

It should be noted that adaption was not used in the experiments on this database.

4.5. Generalized words and spontaneous data

With this experiment we show that there is no drawback in using generalized-words on real spontaneous data. Because the number of unknown surnames in the eval99.1 data is small, 46 instances of nine unique words, the results in overall error reduction are so small that they are in the range of 'noise' but what we can see 6 is that there is a small reduction of the error rates compared to the baseline and no increase in the real time factor (RTF), using the best system from the last experiment (GW780-XWT-CD).

We also did a cheating experiment which we named 'CHEAT2'. In the cheating experiment the nine missing names were added to the recognition vocabulary, but we did not add the head-tail-generalized-word. Table 6 summarizes the results. The generalized words achieve more than 50% of the gain that the cheating experiment has. Reduction of consequential errors perform equal to the last experiment.

System	WCE	REC	PRC	WE	RTF
BASELINE	22.6%	-	-	22.6%	2.0
GW780-XWT-CD	22.2%	57%	77%	22.5%	2.0
CHEAT2	21.9%	74%	100%	21.9%	2.0

Table 6: results on top30HH data

5. Conclusion

We introduced a method of building general words using a specifically modeled head and a general modeled tail. The generalized word models have the task to model many different words. The semantic class language model has the job to allow only a specific class of unknown words be predicted or to select among different classes of a generalized word. Our experiments showed the high importance using cross word triphone acoustic models for head-tail-generalized-words, like for the rest of the vocabulary. Modeling the head part with context dependent models achieved the best results. We were able to reduce the number of consequential errors by about 37%. The proposed approach identifies unknown surnames with a high recall in the range of 57% to 59% and a precision of about 77% on real spontaneous speech using an LVCS recognizer with a vocabulary of 11k words.

6. Acknowledgments

I want to thank my colleagues of the Interactive System Labs for their useful discussions and interest in my work. A very special thank to Jürgen Fritsch and Ivica Rogina for their mental and active support that this paper became reality.

7. References

[1] Wolfgang Wahlster (Ed.), *Verbmobil: Foundations of Speech-to-Speech Translation*, Springer, 2000.

[2] P. Geutner, *Adaptive Vocabularies in Large Vocabulary Conversational Speech Recognition*, Ph.D. thesis, University of Karlsruhe, 1999.

[3] A. Asadi, R. Schwartz, and J. Makhoul, "Automatic modeling for adding new words to a large-vocabulary continuous speech recognition system," in *Proc. ICASSP '91*, Toronto, Canada, 1991, pp. 305–308.

[4] B. Suhm, M. Woszczyna, and A. Waibel, "Detection and transcription of new words," in *Proc. Eurospeech '93*, Berlin, Germany, 1993, pp. 2179–2182.

[5] P. Fetter, *Detection and Transkription of OOV Words*, Ph.D. thesis, Daimler-Bens AG, 1998.

[6] T. Kemp A. Jusek, "Modeling unknown words in spontaneous speech," in *Proc. ICASSP '96*, Philadelphia, USA, 1996, pp. 530–533.

[7] Michael Finke Juergen Fritsch Petra Geutner Klaus Ries and Torsten Zeppenfeld, "The JanusRTk switchboard/callhome 1997 evaluation system," in *Proc. of LVCSR Hub5 Workshop May, 13-15.*, Baltimore, Maryland, 1997.

[8] M. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," 1997.

[9] V. Digalakis and L. Neumeyer, "Speaker adaptation using combined transformation and bayesian methods," in *Proc. ICASSP '95*, Detroit, MI, 1995, pp. 680–683.

[10] M. Woszczyna, *Fast Speaker Independent Large Vocabulary Continuous Speech Recognition*, Ph.D. thesis, University of Karlsruhe, Germany, 1998.

[11] J. Fritsch and I. Rogina, "The bucket box intersection (BBI) algorithm for fast approximative evaluation of diagonal mixture gaussians," in *Proc. ICASSP '96*, Atlanta, GA, 1996, pp. 837–840.