OUT-OF-VOCABULARY WORD MODELING USING MULTIPLE LEXICAL FILLERS

Gilles Boulianne, Pierre Dumouchel

Centre de recherche informatique de Montréal
550 Sherbrooke St. W.
Montréal, Québec
Canada H3A 1B9
{gboulian, pdumouch}@crim.ca

ABSTRACT

In large vocabulary speech recognition, out-of-vocabulary words are an important cause of errors. We describe a lexical filler model that can be used in a single pass recognition system to detect out-of-vocabulary words and reduce the error rate. When rescoring word graphs with better acoustic models, word fillers cause a combinatorial explosion. We introduce a new technique, using several thousand lexical fillers, which produces word graphs that can be rescored efficiently. On a large French vocabulary continuous speech recognition task, lexical fillers achieved an OOV detection rate of 44% and allowed a 23% reduction in errors due to OOV words.

1. INTRODUCTION

Out-of-vocabulary words are an important problem in many speech recognition applications. Increasing vocabulary size helps alleviate the problem, but will never solve it since language evolves, and new words constantly appear. Reliable detection of OOV words would be a better solution to improve performance and robustness of the technology.

The problem is worse in continuous speech, where a single out-of-vocabulary word will often cause several recognition errors. Estimates range from 1.2 errors/OOV word on the English language Wall Street Journal task to 2.2 errors/OOV word on a similar French task [1, 2]. In French, particularly, typical OOV rate may range over 5-6% for a 20,000 word vocabulary [2]. In this situation, detecting OOV words to reduce errors to one per OOV word would allow a significant reduction of the overall error rate.

The most common approach to OOV word detection consists in explicitly modeling out-of-vocabulary words by providing some form of pronunciation and a representation in the language model. Proposed models range from "fillers" or "generic word models" that allow any phonetic sequence, to detailed sub-lexical models. The main objective of these models is to cover all possible OOV word pronunciations, while providing some constraints in the form of phone N-grams [3] or morpho - phonemic constraints [4].

In this work, we propose a simpler model that does not aim to model all possible OOV pronunciations, and avoids modeling in-vocabulary pronunciations. This approach keeps tight lexical constraints both on in- and out- of- vocabulary words. The model is a filler word provided with the subset of pronunciations, taken from a larger dictionary, which do not occur in the task vocabulary. Thus the model is constrained both by the task lexicon and the larger lexicon.

Our experiments with larger lexicons show that, contrary to expectation, reasonable sizes will provide good OOV models. We first present results on a large vocabulary, continuous speech recognition task in French, that show the effectiveness of lexical constraints when applied in a single pass recognition system. In addition, we examine a problem that arises when OOV models are used in multi-pass systems. In these systems, when large acoustic models are used for rescoring, OOV models that do not impose tight constraints on pronunciation will require very large recognition networks and a costly search. We address this problem in more detail in Section 2.2 and we propose a solution, involving multiple lexical fillers, that allows word graphs to be rescored efficiently.

2. LEXICAL FILLERS

The idea of using a larger dictionary for OOV words is not new. It has been used with success in large vocabulary tasks, for example in [2]. In this system, 64K word dictionary and language model were used, and the recognizer output was mapped to the task vocabulary of 20K words.

In our approach, we also use a larger dictionary, but only to provide pronunciations to a filler word model. The recognition vocabulary is augmented by one word, namely the filler word, both in the dictionary and the language model.

Our OOV model also differs from a generic filler in that it only allows a restricted number of pronunciations, instead of all possible pronunciations, but most importantly it does not model in-vocabulary words. As will be shown in the experimental results (section 3), this characteristic limits the degradation of in-vocabulary word recognition observed for generic filler models [3]. In addition, strong constraints reduce network size and search effort during recognition.

2.1. Single lexical filler

The closed vocabulary is augmented with one additional filler word. Possible pronunciations for this word are obtained from all the pronunciations present in a larger dictionary but not present in the task dictionary. A typical lexical filler could have between 40K and 100K pronunciations.

The filler word is also added to the language model, with a unigram probability chosen to reflect the probability of OOV words in the training corpus. Higher N-gram probabilities are set to 0 for the filler word; otherwise, the search network becomes unmanageably large, since several tens of thousands of pronunciation are associated with each appearance of the filler word in the language model. Contrary to generic filler models, where the unigram probability is chosen manually and determines the number of correct detections and false alarms, constraints on pronunciation make the lexical filler behavior mostly insensitive to the particular value used for its unigram probability (see section 3.2).

2.2. Multiple lexical fillers

Today's laboratory speech recognition systems are almost always multi-pass systems. The goal of the first pass in these systems is to provide the highest inclusion rate and lowest search effort for the second pass. Hypotheses are recorded in the form of a word graph or a N-best list of word hypotheses. Using words as the intermediate unit is usually a good compromise: smaller units such as phones would not retain much more information, since most words have very few pronunciations.

When a filler is used (generic or lexical), however, the pronunciation that has been determined at great cost by the first pass is not recorded in the word graph. The information about the best pronunciation is lost. When the word graph is used for rescoring with different acoustic models, the whole pronunciation network corresponding to each filler present in the graph has to be searched again, with even costlier models.

The idea behind multiple lexical fillers is to retain some of the pronunciation information in the filler identity itself, by merging pronunciations into classes and using a different filler for each class. In the experiments reported below, we defined pronunciation classes as follow:

> Two pronunciations are in the same class if, when mapping all vowels to V and all consonants to C, they have the same sequence of V and C symbols.

Thus the words *saltimbanque* et *dégringole* will be represented by the same filler, because their pronunciations belong in the same class (since they both correspond to the CVCCVCVC symbol sequence). That particular way of grouping pronunciations yields a large number of filler words, typically several thousand, each one having a small number of possible pronunciations.

In the language model, all filler words are assigned the same unigram probability as the single lexical filler, and no higher order N-gram is added. The first pass search effort remains essentially the same for single or multiple fillers, since the set of pronunciations to be searched is the same. But the second pass search effort for multiple fillers is reduced in proportion to the number of pronunciations per filler, which is much smaller when thousands of lexical fillers are used.

3. EXPERIMENTS

The following experiments were performed on a standard French language, large vocabulary dictation task from the AUPELF'97 evaluation [5, 6], which has a vocabulary of 20,000 words. Acoustic models were trained on BREF-80 and a subset of BREF-Total, containing 53 hours of speech from 100 speakers. Acoustic parameters were 12 mel- frequency cepstral coefficients plus energy, and their derivatives. Cross-word triphone acoustic models were trained with 3981 output distributions, each being a Gaussian mixture with 32 components and sharing a single global full covariance matrix. For rescoring experiments the same models with 64 components per mixture were used. Note that models were not gender-dependent, and that no speaker adaptation (such as VTLN or MLLR) was used.

The speech recognition system [7] was transducer-based, with rules for *liaison* both in training and recognition. Transducers were built and manipulated using the FSM library tools [8].

Language models were trained on 168M words of *Le Monde*; their size and perplexity are shown in Table 1. A small homophone class trigram model was used to generate the word graphs (more exactly the homophone class graphs). When rescoring, a larger word trigram language model was used, and homophone class graphs were mapped to words by composition with a homophone-class-to-word transducer.

The test set contains 576 sentences, with a frequency weighted OOV word rate of 3.78%, and over 40% of the sentences contain OOV words. Only 16.5% of OOV words are proper names.

LM	2-grams	3-grams	Voc. size	Ppx
Graph gen.	247K	78.5K	13.5K	128
Rescoring	2.8M	10.6M	20.0K	87

Table 1. Size, vocabulary and development set perplexity of language models used.

3.1. Baseline

We first need a proper measurement of the effect of OOV words in the case of continuous speech. As mentioned in the introduction, each OOV word tends to cause several errors, and this is precisely the main problem that we try to solve. In order to estimate how many errors are caused by OOV words, we measure the accuracy separately on a subset of 344 test sentences which contain only in-vocabulary words. This gives an estimate of the error rate if there were no OOV words (first line of Table 2). Assuming that other aspects of this subset are representative of the whole corpus, we can subtract this error rate from the whole corpus error rate (second line of table) to get an estimate of the error rate due to OOV words (third line of table).

Naturally, the in-vocabulary subset is not exactly representative of the whole. In particular, the fact that it contains no OOV words may be associated with a better perplexity (see [2]); in that case the subset error rate will underestimate the real in-vocabulary error rate, leading to an overestimate of the error rate due to OOV. So this estimate must not be taken as an absolute number. Nevertheless, we will use the estimate throughout the rest of the paper for comparison purposes, since relative improvements on the same OOV subset are still meaningful.

Sentences	Words	Error rate
In-vocabulary only	4913	17.5 %
All sentences	8647	24.7 %
Errors due to OOV words		7.2 %

Table 2. Estimated OOV error rate, baseline system.

The OOV rate being 3.78%, we estimate that each OOV causes on average 1.9 errors (7.2%/3.8%). This is consistent with results on this task reported elsewhere [2].

3.2. Single filler experiments

The pronunciations for the single filler experiments came either from a 64K word dictionary or a 600K word dictionary. The 64K dictionary was generated mostly automatically from grapheme-to-phoneme rules, with a small percentage selected automatically for correction by hand [7]. The 600K dictionary pronunciations were generated with the same rules applied to the 600K most frequent words found in the LM training corpus. For the 64K dictionary, the filler had 40K pronunciations, while for the 600K dictionary, the filler had 107K pronunciations. None of these pronunciations occurred in the 20K task dictionary.

Condition	Overall WER	Due to OOV	OOV detect	IV false
No filler	24.7%	7.2%	0%	0%
64K dict filler	23.0%	5.48%	44.0%	0.52%
600K dict filler	22.8%	5.46%	41.0%	0.28%

Table 3. Overall word error rate, errors due to OOV words, OOV correct detection and in-vocabulary false alarm for single lexical filler.

Table 3 shows the word error rate obtained for the no filler, 64K dictionary and 600K dictionary conditions. The errors due to OOV were estimated as for Table 2. The last two columns show the OOV detection rate, which measures how many OOV words in the utterances were correctly recognized as a filler word, and the IV false alarm rate, which measures how many in-vocabulary words were incorrectly recognized as filler words. Note that deletions of OOV words are counted and contribute to decrease the OOV detection rate; similarly, insertions of filler words contribute to increase the IV false alarm.

A good fraction of OOV words are correctly detected (44%), even given the rather tight constraints on their pronunciation. We also observe a very low false alarm rate for in-vocabulary words. OOV detection allows the overall error rate to go from 24.7% to 23.0%, mainly by reducing insertions. Looking only at the errors due to OOV, we see a drop from 7.2% to 5.5%, which represents a relative reduction of more than 23%.

Interestingly, using a much larger dictionary of 600K words did not significantly decrease the overall error rate compared to the 64K dictionary. It seems that, contrary to expectation, it is not necessary to use a very large number of pronunciations to capture the essence of OOV words.

For generic fillers, the operating point (OOV detection rate vs. IV false alarms) has to be adjusted by tuning a filler penalty. With lexical fillers, we found that changes to the unigram probability of the filler word had little effect. Changing the unigram probability by 9 orders of magnitude shifted the OOV detection rate from 33% to 44% while the IV false alarm rate only went from 0.13% to 0.53%. This insensitivity to the unigram probability makes a system using lexical fillers more robust to changes in the language model.

Our results are not directly comparable with those in [3] about generic fillers, which were obtained on a smaller English vocabulary of 2000 words. Nevertheless, it is interesting to note that a generic filler model, at a similar OOV detection rate of 46.8%, produced a higher in-vocabulary false alarm rate of 1.3%.

Note that a larger task vocabulary will increase the overlap between in-vocabulary pronunciations and a generic filler. Conversely, a smaller task vocabulary will result in fewer constraints for a lexical filler. Accordingly, we can probably expect generic fillers to perform better on small vocabularies and lexical fillers to perform better on large vocabularies, although this would require experimental verification.

3.3. Multiple filler experiments

We generated word graphs with the models used in the single filler experiments. To rescore them using new acoustic and language models, we built a recognition network using each sentence word graph as a constraint on the language model. We found out that although word graphs with fillers are about the same size as word graphs without fillers, the recognition network, which includes the pronunciation and acoustic model information, increases by a factor so large that we could generate it for a few sentences only, due to the large number of pronunciations per filler.

Analysis revealed that each occurrence of a filler word in the graph required the inclusion of 40K pronunciations in the recognition network. Even though filler words represent only a few percent of all word graph arcs, and network optimization reduces any unnecessary duplication, networks could still be potentially a thousand times larger than those without fillers.

When multiple fillers are used, however, the number of pronunciations for each filler word that appears in the word graph can be reduced arbitrarily. After grouping the 64K dictionary pronunciations according to their CV patterns (see section 2.2), we obtained 2492 filler words, each one having 16 pronunciations on average. The recognition networks obtained from the multiple filler word graphs are manageable, being 4 times larger than without fillers.

We estimate the graph inclusion from the error rate of the best hypothesis contained in the graph. For the word graphs generated with the baseline system (no fillers), this error rate was 5.8% compared to 5.5% for the word graphs with fillers. Table 4 shows the results obtained after rescoring these word graphs with acoustic models of 64 components per mixture.

Condition	**Overall WER**	**Due to OOV**	**OOV detect**	**IV false**
No fillers	19.4%	7.4%	0%	0%
2492 fillers	18.0%	6.0%	31.8%	0.11%

Table 4. Error rates for word graph rescoring.

Error rate reductions are not as large as in the single pass system. We must note, however, that we experimented only with word graphs generated at an operating point with a rather small OOV identification rate (35.5% with 0.17% IV false alarm). Determining the best operating point for the first pass to produce the best results for the second pass will require further experimentation.

4. CONCLUSION

We described a lexical filler for out-of-vocabulary word modeling. It does not account for all possible pronunciations but explicitly avoids modeling in-vocabulary pronunciations. We also introduced multiple lexical fillers in order to produce word graphs that can be efficiently rescored in a second pass with more complex acoustic models. Our experiments on a 20K word French task show that lexical fillers, achieve an OOV detection rate of 44% (at a false alarm rate of 0.5%), allowing a 23% relative reduction of errors due to OOV words.

5. REFERENCES

[1] J.L. Gauvain, L.F. Lamel, G. Adda, and M. Adda-Decker, "Speaker-independent continuous speech dictation," *Speech Communication*, vol. 15, no. 1, pp. 21–27, 1994.

[2] G. Adda, M. Adda-Decker, J.L. Gauvain, and L. Lamel, "Text normalization and speech recognition in french," in *Proc. Eurospeech'97*, September 1997, Rhodes.

[3] I. Bazzi and J.R. Glass, "Modeling out-of-vocabulary words for robust speech recognition," in *Proc. ICSLP 2000*, October 2000, Beijing.

[4] X. Mou and V. Zue, "Sublexical modelling using a finite state transducer framework," in *Proc. ICASSP 2001*, May 2001, Salt Lake City.

[5] J.M. Dolmazon, F. Bimbot, G. Adda, J.C. Caerou, J. Zeiliger, and M. Adda-Decker, "Arc b1 - organisation de la première campagne aupelf pour l'évalua tion des systèmes de dictée vocale," in *JST97 FRANCIL*, April 1997, Avignon.

[6] J.M. Dolmazon, F. Bimbot, G. Adda, J.C. Caerou, J. Zeiliger, and M. Adda-Decker, "Première campagne aupelf d'evaluation des systèmes de dictée vocale: organisation et resultats," in *preparation*, 2000.

[7] G. Boulianne, J. Brousseau, P. Ouellet, and P. Dumouchel, "French large vocabulary recognition with cross-word phonology transducers," in *Proc. ICASSP 2000*, June 2000, Istanbul.

[8] M. Mohri, F. Pereira, and M. Riley, "A rational design for a weighted finite-state transducer library," *Lecture notes in Computer Science*, p. 1436, 1998.