Open Vocabulary Speech Recognition with Flat Hybrid Models

Maximilian Bisani and Hermann Ney

Lehrstuhl für Informatik VI – Computer Science Department
RWTH Aachen University, D-52056 Aachen, Germany
{bisani,ney}@informatik.rwth-aachen.de

Abstract

Today's speech recognition systems are able to recognize arbitrary sentences over a large but finite vocabulary. However, many important speech recognition tasks feature an open, constantly changing vocabulary. (E.g. broadcast news transcription, translation of political debates, etc. Ideally, a system designed for such open vocabulary tasks would be able to recognize arbitrary, even previously unseen words. To some extent this can be achieved by using sub-lexical language models. We demonstrate that, by using a simple flat hybrid model, we can significantly improve a well-optimized state-of-the-art speech recognition system over a wide range of out-of-vocabulary rates.

1. Introduction

Large vocabulary speech recognition systems operate with a fixed large but finite vocabulary. Typical vocabulary sizes are of the order of ten to one hundred thousand word forms. This is suitable for e.g. dictation tasks for a fixed domain. In open vocabulary settings (e.g. broadcast news, political debates, etc.) the number of different words do not appear to be finite. Systems operating with a fixed vocabulary are bound to encounter so-called out-of-vocabulary (OOV) words. These are problematic for a number of reasons: 1) An OOV word will never be recognized (even if the user repeats it), but will be substituted by some in-vocabulary word. 2) Neighboring words are also often misrecognized. 3) Later processing stages (e.g. translation, understanding, document retrieval) cannot recover from OOV errors. 4) OOV words are often content words. The goal for open vocabulary scenarios is clear; the transcription system should be able to handle any spoken word without help from the user. To see how this could be achieved, let us briefly review the decision rule and knowledge sources used by a large vocabulary speech recognition system:

$$\boldsymbol{w}(\boldsymbol{x}) = \operatorname*{argmax}_{\boldsymbol{w}'} p(\boldsymbol{w}') \max_{\boldsymbol{\varphi}} p(\boldsymbol{x}|\boldsymbol{\varphi}) p(\boldsymbol{\varphi}|\boldsymbol{w}') \quad (1)$$

with

- acoustic model $p(\boldsymbol{x}|\boldsymbol{\varphi})$
  relates acoustic features $\boldsymbol{x}$ to phoneme sequences $\boldsymbol{\varphi}$, typically an HMM (vocabulary independent)
- pronunciation lexicon $p(\boldsymbol{\varphi}|\boldsymbol{w})$
  assigns one (or more) phoneme string(s) $\boldsymbol{\varphi}$ to each word $w \in V$
- language model $p(\boldsymbol{w})$
  assigns probabilities to sentences from a finite set of words $\boldsymbol{w} \in V^*$

For open vocabulary recognition, we propose to conceptually abandon the words in favor of individual letters. Unlike words, the set of different letters G in a writing system is finite. Concerning the link to the acoustic realization, the set of phonemes Φ can also be considered finite for a given language. These considerations suggest the following model:

$$\boldsymbol{g}(\boldsymbol{x}) = \operatorname*{argmax}_{\boldsymbol{g}'} p(\boldsymbol{g}') \max_{\boldsymbol{\varphi}} p(\boldsymbol{x}|\boldsymbol{\varphi}) p(\boldsymbol{\varphi}|\boldsymbol{g}') \quad (2)$$

with

- acoustic model $p(x|\boldsymbol{\varphi})$
- pronunciation model $p(\boldsymbol{\varphi}|\boldsymbol{g})$
  provides a pronunciation $\boldsymbol{\varphi} \in \Phi^*$ for any string of letters $\boldsymbol{g} \in G^*$
- sub-lexical language model $p(\boldsymbol{g})$
  assigns probabilities to character strings $\boldsymbol{g} \in G^*$

Alternatively the pronunciation model and sub-lexical language model can be combined into a

- joint "graphonemic" model $p(\boldsymbol{\varphi}, \boldsymbol{g})$

2. Grapheme-to-Phoneme Conversion

Obviously this approach to open-vocabulary recognition is strongly connected to grapheme-to-phoneme conversion (G2P), where we seek the most likely pronunciation for a given orthographic form:

$$\boldsymbol{\varphi}(\boldsymbol{g}) = \operatorname*{argmax}_{\boldsymbol{\varphi}' \in \Phi^*} p(\boldsymbol{\varphi}', \boldsymbol{g}) \quad (3)$$

In particular "graphonemic" joint sequence models have been shown to perform very well on G2P tasks [1, 2, 3, 4]. The underlying assumption of this model is that, for each word, its orthographic form and its pronunciation are generated by a common sequence of graphonemic units. Each unit is a pair $q = (\boldsymbol{g}, \boldsymbol{\varphi}) \in Q \subseteq G^* \times \Phi^*$ of a letter sequence and a phoneme sequence of possibly different

Table 1: Grapheme-to-phoneme conversion performance as a function of graphone size L and M-gram length M. All models were trained on the 20k baseline dictionary. Results are given as phoneme error rate on a disjoint subset of the 64k lexicon.

M	$L=1$	$L=2$	$L=3$	$L=4$	$L=5$	$L=6$
1	46.95	34.97	28.16	24.56	25.52	27.96
2	23.36	18.87	18.74	19.80	23.80	25.24
3	17.76	16.56	18.34	19.94	23.82	25.05
4	16.09	16.46	18.25	19.94		
5	15.22	16.49	18.01			
6	**15.01**					

length. We refer to such a unit as a "graphone".[1] The joint probability distribution $p(\boldsymbol{\varphi}, \boldsymbol{g})$ is thus reduced to a probability distribution over graphone sequences $p(\boldsymbol{q})$ which we model using a standard M-gram:

$$p(q_1^N) = \prod_{i=1}^{N+1} p(q_i|q_{i-1}, \ldots, q_{i-M+1}) \qquad (4)$$

The complexity of this model depends on two parameters: the range of the M-gram model and the allowed size of the graphones. We allow the number of letters and phonemes to vary between zero and an upper limit L: $\left|\boldsymbol{g}_q\right| = 0 \ldots L$, $\left|\boldsymbol{\varphi}_q\right| = 0 \ldots L$. Chen [4] has shown that excellent results can be obtained when restricting the segmentation to trivial chunks containing zero or one letter and zero or one phoneme ($L = 1$). We were able to verify that shorter units in combination with longer range M-gram modeling yields the best result for the grapheme-to-phoneme task. This is exemplified in table 1. A more comprehensive report on these results is currently in preparation.

3. Models for Open Vocabulary ASR

The graphone-based model integrates very easily with the standard speech recognition architecture: Any graphone can simply be added to the normal pronunciation dictionary. This means we combine the lexical entries with the (sub-lexical) graphones derived from grapheme-to-phoneme conversion to form an unified set of recognition units $U = V \cup Q$. From the perspective of OOV detection the sub-lexical units Q have been called "fragments" or "fillers" [5, 6, 7] but are typically not associated with spelling information. By treating words and fragments uniformly the decision rule becomes

$$\operatorname*{argmax}_{\boldsymbol{u} \in U^*} p(\boldsymbol{x}|\boldsymbol{u}) p(\boldsymbol{u}) \quad . \qquad (5)$$

The sequence model $p(\boldsymbol{u})$ can be characterized as "hybrid" because it contains mixed M-grams containing both words and fragments. It can also be characterized as "flat", as opposed to structured approaches that predict and model OOV words with different models. A shortcoming of this model is that it leaves undetermined were word boundaries (i.e. blanks) should be placed. The heuristic used in this study is to compose any consecutive sub-lexical units into a single word and to treat all lexical units as individual words. Galescu [8] has demonstrated that this flat hybrid model can in fact be used to perform recognition of out-of-vocabulary words, but the relative improvements he reported were rather small.

[1] Various other names have been suggested: grapheme-phoneme joint multigram, graphoneme, grapheme-to-phoneme correspondence (GPC), chunk

4. Experiments

We have tested the method discussed on the Wall Street Journal dictation task, which is a well-studied large vocabulary dictation task. Recognition results were produced on the ARPA 1993 and 1994 Hub-1 development test data. The combination of these, containing 812 sentences with 15625 words, will be referred to as "dev 93+94". We also report results on a subset called "dev rare" comprising that half of the sentences containing the rarest words (406 sentences, 8527 words).

The speech recognizer uses features derived from 16 Mel-frequency cepstral coefficients with cepstral mean normalization, and linear discriminant analysis on seven consecutive vectors with an output dimension on 32. Acoustic modeling is based on triphones with across-word context using 7001 tied states. Gaussian mixtures with a total of 405k densities were training using MCE [9]. No speaker adaptation was used.

We have three different established vocabularies with 5, 20 and 64 thousand words, each corresponding to the most frequent words in the language model training corpus. These constitute our baseline setups (cf. table 3). For each baseline pronunciation dictionary a grapheme-to-phoneme model was trained with different length constraints $L = 2 \ldots 6$ using EM training with M-gram length of 3 (cf. table 1). The recognition vocabulary was then augmented with all graphones inferred by this procedure. Next the WSJ language model corpus (10M sentences, 227M words) was modified by replacing all OOV words by their most likely (sub-lexical) graphone sequence. This modified text was used to estimate the respective hybrid language model using absolute discounting with interpolation.

Some recognition examples are shown in table 2. For quantitative analysis, we evaluated both word error rate (WER) and letter error rate (LER). Letter error rate is more favorable with respect to almost-correct words and corresponds with the correction effort in dictation applications.

4.1. Bootstrap Analysis of OOV Impact

We are particularly interested in the effect of OOV words on the recognition error rate. This effect could be studied by varying the system's vocabulary. However, changing the recognition system in this way might

Table 2: Examples of recognized OOV words: The columns show the recognition result of the respective system for the word shown on the left. Incorrect results are slanted. The vertical lines in the FH result indicate the fragment boundaries and are not part of the actual system output.

correct	baseline 20k	baseline 64k	FH 20k L=4
opportunistic	*opportunity stick*	opportunistic	opp|or|tun|ist|ic
cellar	*seller*	cellar	cell|ar
overblown	*the reply*	overblown	over|blow|n
convulsed	*can false*	convulsed	*conv*|*uls*
disenfranchised	*anderson franchise*	*disenfranchise*	dis|en|fran|chis|ed
Litvack	*slipped back*	litvack	lit|v|ack
Margulies	*marti leaves*	margulies	*mar*|*u*|*li*|*as*
Murtagh	*murray todd*	*humor tad*	*m*|*ur*|*tag*
Betty Percival	betty *personal*	*that a person will*	betty *pers*|*ible*
Noriyuki Matsushima	*or you keep matsushita*	*subsidiary p. matsushita*	n|ori|y|uk|i|ma|t|su|shi|ma
Du Liban	*do rebound*	*duly bond*	*du*|*ly bond*

introduce secondary effects such as increased confusability between vocabulary entries. Alternatively we can alter the test set. By extending the bootstrap technique proposed in [10], we create an ensemble of virtual test corpora with a varying number of OOV words, and respective WER. This distribution allows us to study the correlation between OOV rate and word error rate without changing the recognition system. This procedure is detailed in the following: For each sentence $i = 1 \ldots s$ we record the number of words n_i, the number of OOV words o_i and the number of recognition errors e_i:

$$X = (n_1, o_1, e_1), \ldots, (n_s, o_s, e_s) \tag{6}$$

For $b = 1 \ldots B$ (typically $B = 10^3$) we randomly select with replacement s tuples from X to generate a bootstrap sample

$$X^{*b} = (n_1^{*b}, o_1^{*b}, e_1^{*b}), \ldots, (n_s^{*b}, o_s^{*b}, e_s^{*b}) \tag{7}$$

Then we calculate the OOV rate and word error rate on this sample

$$OOV^{*b} := \sum_{i=1}^{s} o_i^{*b} / \sum_{i=1}^{s} n_i^{*b} \tag{8}$$

$$WER^{*b} := \sum_{i=1}^{s} e_i^{*b} / \sum_{i=1}^{s} n_i^{*b} \tag{9}$$

The bootstrap replications OOV^{*b} and WER^{*b} can be visualized by a scatter plot (see fig. 1). We quantify the observed linear relation between OOV rate and WER by a linear lest squares fit. The slope of the fitted line reflects the number of word errors per OOV word. For this reason we call this quantity "OOV impact".

5. Discussion

The recognition results are listed in table 3. First of all we note that the flat-hybrid model performs better than the corresponding baseline in all tested circumstances. Obviously the improvement in error rate depends strongly on the OOV rate: For very high OOV rates above 10%, error rate reductions of over 30% relative are possible. For the moderate OOV rates of the 20k system (2.6%), the improvement is still 15% relative. Even for very low OOV rates the performance does not deteriorate as one might have expected, but is still slightly better than baseline.

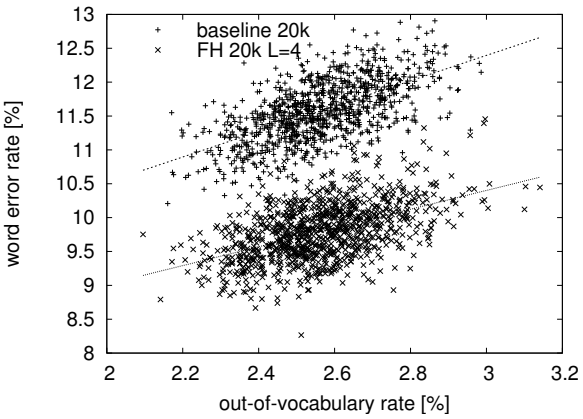

Figure 1: Word errors vs. out-of-vocabulary rate: Each data point represents one bootstrap replication of the dev 93+94 test set. The straight lines represent the linear least-squares fit.

Is is interesting to compare the OOV impact factor (word errors per OOV word): The baseline systems have values between 1.7 and 2, supporting the common wisdom that each OOV word causes two word errors. The flat-hybrid models are superior in this respect by at least 0.5, which means, for each OOV word, they make one error less.

Concerning the optimal choice of fragment size L, we note that there are two counteracting effects: Larger L values increase the size of the graphone inventory, which in turn causes data sparseness problems, leading to worse grapheme-to-phoneme performance. Smaller values for L cause the unit inventory to contain many very short words with high probabilities, leading to spurious insertions in the recognition result. The present experiments suggest that the best trade-off is at $L = 4$.

Table 3: Recognition results for different models on both test sets. "FH" stands for flat hybrid model, followed by base vocabulary and maximum fragment size. Out-of-vocabulary words (OOV) given in number of occurrences and percentage per running words. Word error rate (WER) and letter error rate (LER) given in percent.

	vocabulary		dev 93+94			dev rare			OOV
model	words	fragments	OOV	WER	LER	OOV	WER	LER	impact
baseline 5k	4986		1743	24.26	11.23	1344	32.12	14.72	1.72
FH 5k L=2	5k +	1090	(11.2%)	17.56	7.69	(15.6%)	22.92	9.96	1.20
FH 5k L=3	5k +	3016		16.67	7.26		21.80	9.42	1.14
FH 5k L=4	5k +	4085		**16.54**	7.37		**21.43**	9.59	1.15
FH 5k L=5	5k +	4381		17.36	7.94		22.76	10.34	1.18
FH 5k L=6	5k +	4474		18.41	8.57		24.31	11.31	1.27
baseline 20k	19977		400	11.58	5.06	396	15.58	6.75	1.88
FH 20k L=2	20k +	1721	(2.6%)	10.43	4.45	(4.6%)	13.55	5.74	1.52
FH 20k L=3	20k +	6700		9.95	4.26		12.56	5.31	1.32
FH 20k L=4	20k +	11622		**9.79**	4.19		**12.26**	5.19	1.27
FH 20k L=5	20k +	13708		9.88	4.22		12.28	5.18	1.17
FH 20k L=6	20k +	14858		10.09	4.39		12.58	5.43	1.22
baseline 64k	64735		76	8.92	3.81	72	10.62	4.51	1.99
FH 64k L=2	64k +	3175	(0.5%)	8.93	3.84	(0.8%)	10.45	4.48	1.63
FH 64k L=3	64k +	14346		**8.87**	3.81		**10.34**	4.43	1.46
FH 64k L=4	64k +	29335		8.90	3.80		10.40	4.40	1.46

6. Conclusion

We have shown that we can significantly improve a well-optimized state-of-the-art recognition system by using a simple flat hybrid sub-lexical model. The improvement was observed on a wide range of out-of-vocabulary rates. Even for very low OOV rates, no deterioration occurred. We found that using fragments of up to four letters or phonemes yielded optimal recognition results, while using non-trivial chunks is detrimental to grapheme-to-phoneme conversion.

7. Acknowledgements

This work was partially funded by the European Commission under integrated project TC-STAR (Technology and Corpora for Speech to Speech Translation, IST-2002-FP6-506738, http://www.tc-star.org).

8. References

[1] S. Deligne, F. Yvon, and F. Bimbot, "Variable-length sequence matching for phonetic transcription using joint multigrams," in *Proc. European Conf. on Speech Communication and Technology*, Madrid, Spain, Sep. 1995, pp. 2243 – 2246.

[2] M. Bisani and H. Ney, "Investigations on joint-multigram models for grapheme-to-phoneme conversion," in *Proc. Int. Conf. on Spoken Language Processing*, Denver (CO), USA, Sep. 2002, vol. 1, pp. 105 – 108.

[3] L. Galescu and J. F. Allen, "Pronunciation of proper names with a joint n-gram model for bi-directional grapheme-to-phoneme conversion," in *Proc. Int. Conf. on Spoken Language Processing*, Denver (CO), USA, Sep. 2002, vol. 1, pp. 109 – 112.

[4] S. F. Chen, "Conditional and joint models for grapheme-to-phoneme conversion," in *Proc. European Conf. on Speech Communication and Technology*, Geneva, Switzerland, Sep. 2003, pp. 2033 – 2036.

[5] D. Klakow, G. Rose, and X. Aubert, "OOV-detection in large vocabulary system using automatically defined word-fragments as fillers," in *Proc. European Conf. on Speech Communication and Technology*, Budapest, Hungary, Sep. 1999, vol. 1, pp. 49 – 52.

[6] I. Bazzi, *Modelling Out-of-Vocabulary Words for Robust Speech Recognition*, Ph.D. thesis, MIT Department of Electrical Engineering and Computer Science, 2002.

[7] A. Yazgan and M. Saraclar, "Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Montreal, Canada, May 2004, vol. 1, pp. 745 – 748.

[8] L. Galescu, "Recognition of out-of-vocabulary words with sub-lexical language models," in *Proc. European Conf. on Speech Communication and Technology*, Geneva, Switzerland, Sep. 2003, pp. 249 – 252.

[9] W. Macherey, L. Haferkamp, R. Schlüter, and H. Ney, "Investigations on error minimizing training criteria for discriminative training in automatic speech recognition," in *Proc. European Conf. on Speech Communication and Technology*, Lisbon, Portugal, Sep. 2005.

[10] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in ASR performance evaluation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Montreal, Canada, May 2004, vol. 1, pp. 409 – 411.