To appear in Proc. ICASSP-2000, Istanbul, Turkey

ENHANCED LANGUAGE MODELLING WITH PHONOLOGICALLY CONSTRAINED MORPHOLOGICAL ANALYSIS

A.C. Fang and M. Huckvale

Department of Phonetics and Linguistics
University College London
Gower Street WC1E 6BT, London, England

ABSTRACT

Phonologically constrained morphological analysis (PCMA) is the decomposition of words into their component morphemes conditioned by both orthography and pronunciation. This article describes PCMA and its application in large-vocabulary continuous speech recognition to enhance recognition performance in some tasks. Our experiments, based on the British National Corpus and the LOB Corpus for training data and WSJCAM0 for test data, show clearly that PCMA leads to smaller lexicon size, smaller language models, superior word lattices and a decrease in word error rates. PCMA seems to show most benefit in open-vocabulary tasks, where the productivity of a morph unit lexicon makes a substantial reduction in out-of-vocabulary rates.

1. INTRODUCTION

In this paper we present a novel approach towards the enhancement of language modelling that is achieved through *phonologically constrained morphological analysis* (PCMA). PCMA is the decomposition of word tokens into their component affixes and stems constrained by both orthography and pronunciation. In its simplest form, PCMA involves the analysis of words into a sequence of sub-word units which express the morphological structure of the word subject to the constraint that the pronunciation of the whole is derivable simply from the concatenation of the pronunciation of the parts. As an example, PCMA accepts the decomposition of *abandoned* into *abandon+ed* since the pronunciation for the whole string may be concatenated from the parts. On the other hand, *academician* is not decomposed into *academic+ian* for the reason that the parts do not allow direct derivation of the pronunciation.

This paper describes our investigations into the use of PCMA for speech recognition based mostly on the 100-million-word British National Corpus (BNC; [1]) for training and test material. In the following sections we describe the preparation of the training and test data and present baseline statistics obtained with the Abbot connectionist/HMM continuous speech recognition system ([2]; henceforward referred to as Abbot) from conventional word-based models. We then describe the PCMA approach towards language modelling in detail. We then present statistics obtained from PCMA models and discuss comparisons with word-based language models in terms of lexicons, lattice scoring, perplexity measures, and finally word accuracy rates.

2. DATA AND BASELINE STATISTICS

2.1 Text and speech data

The BNC was used as a basis on which both training and test data sets were selected. There are 4,124 files in the corpus, 3,209 written and 815 transcribed speech. The written texts were randomised and 10 chunks of 10 million words were selected for use as training sets. The remainder of the written texts were divided into 10 chunks of one million words each for use as test sets.

To establish baseline statistics, two sets of language models were trained from the first 10m-word chunk in the training set (train-01-raw) and the first and the second chunks in the training set totalling about 20 million words (train-01-02-raw). The CMU-Cambridge Toolkit [3] was used for this purpose with linear discounting. The vocabulary sizes were set at 20k, 40k, and 65k. The pronunciations were mapped from a dictionary of British English Example Pronunciations [4]. A text-to-speech system was used to generate pronunciations for lexical items from the training sets that did not have a corresponding entry in BEEP.

For the first set of recognition experiments, 100 sentences (1786 words) were randomly selected from the first test data set (BNC1). In addition, a further 100 sentences (2002 words) were randomly chosen from the Lancaster-Oslo-Bergen Corpus (LOB1). These sentences were read by a single male speaker of British English in anechoic conditions and the recording was digitally acquired at 16 kHz.

2.2 Lexicons and OOV rates

Table 1 summarises the coverage of the pronunciation lexicons constructed from the training sets. As can be noted, BNC1 has a higher out-of-vocabulary (OOV) rate than LOB1, especially at 65k level.

Lexicon	*Size*	*OOV %*	
		LOB1	*BNC1*
train-01-raw-20k	19,998	9.23	9.28
train-01-raw-40k	39,994	7.23	7.48
train-01-raw-65k	64,978	5.85	6.79

Table 1: A summary of pronunciation lexicons

2.3 Perplexity measures

Perplexities were measured for the word models trained with 10- and 20-million word training sets. These were calculated using the CMU toolkit, which by default ignores

OOV words. From Table 2, we can see that LOB1 has produced higher perplexities than BNC1. On the other hand our small test samples seem to have significantly higher perplexities than the LOB corpus taken as a whole.

	10m			*20m*		
	LOB1	*BNC1*	*LOB*	*LOB1*	*BNC1*	*LOB*
20k	434	448	277	377	404	248
40k	512	514	324	441	457	289
65k	562	538	350	480	481	311

Table 2: Perplexities for word models

2.4. Lattice scores

The Abbot recognition system (version 0.76) was used to obtain our baseline lattice and word accuracy scores. Word lattices were generated with parameters which increased the default number of hypotheses to 100. The maximal scores for the lattices were calculated by finding the best path matching the correct transcription using a dynamic programming search. The overall performance could then be computed according to the number of incorrect matches, deletions, and insertions in the best path. Table 3 summarises the word lattice scores for the two test sets with lexicons of various sizes. It is noticeable that the overall performance increases with the increase in lexicon size showing that coverage is an issue.

	LOB1 (%)	BNC1 (%)
20k	86.4	86.6
40k	88.4	88.6
65k	89.1	88.7

Table 3: Word lattice scores

2.5 Word accuracy rates

Table 4 summarises the performance of Abbot with language models trained with the 10m-word training data set (train-01). Three vocabulary sets of respective sizes 20k, 40k, and 65k were used in the training of the language models.

	LOB1 (%)	BNC1 (%)
20k	54.7	52.9
40k	56.3	54.6
65k	55.8	55.0

Table 4: Word recognition scores

Table 5 summarises the performance of Abbot with a language model trained with the 20m word training set (train-01-02).

	LOB1 (%)	BNC1 (%)
20k	56.4	54.5
40k	57.4	55.2
65k	57.6	55.3

Table 5: Word recognition scores

Across the various vocabulary sizes, Abbot performed consistently better with the larger language models. The better performance on LOB1 is probably related to its smaller perplexity and smaller OOV rate.

3. PCMA LANGUAGE MODELS

The use of morphological analyses in the construction of language models is motivated by the benefits that stem from a reduction in the size of the lexicon. In addition, a morph based pronunciation dictionary has fewer minimally different pronunciation pairs than a wordform dictionary. However, in our approach, the morphological analysis is not simply a process whereby words are decomposed into various parts according to their prefixes and suffixes. In order that morphological words or word parts may be reconstructed back into their corresponding orthographic forms, the decomposition itself has to be conditioned by phonological constraints. This ensures that a legal pronunciation may be directly generated from the decomposed parts and that the decoder used in speech recognition need not be affected. As an example, in our approach, the decomposition of abandoned into *abandon + -ed* is allowable because the pronunciation may be constructed from the parts:

```
ABANDON     = ax b ae n d ax n
-ED         = d
ABANDONED   = ax b ae n d ax n d
```

On the other hand, the decomposition of *academician* is not allowed since its pronunciation cannot be reconstructed from its parts, i.e., *academic* and *–ian*:

```
ACADEMIC      = ae k ax d eh m ih k
-IAN          = ia n
ACADEMICIAN   = ax k ae d ax m ih sh n
```

Once the decomposition is successful, the word is presented as a sequence of its component parts with a trailing hash sign (#) indicating the presence of a prefix and a leading hyphen (-) indicating a suffix. As an example, *disregarded* is decomposed into three parts: *dis# regard -ed*. PCMA is therefore a process with two sequential operations. Firstly, the word in question is decomposed into its corresponding morphological parts according to rules. Secondly, the decomposition is constrained by a pronunciation lexicon (BEEP, in our experiment) so that the system only retains those component parts that allow for the direct derivation of the pronunciation for the original word.

Based on [5], a total of 115 prefixes and suffixes were built into the morphological analyser. Table 6 lists 30 most frequent affixes together with frequency counts extracted from the LOB corpus for the 52,703 word types as a result of the morphological analysis.

4353	-s	889	de#	369	-ies
2527	-ing	849	-al	350	-en
1909	co#	786	di#	331	-or
1854	-d	767	un#	318	-ation
1693	-es	694	-ion	316	en#
1599	-ed	471	be#	278	-ment
1546	-er	432	ex#	268	-ions
1481	re#	418	pro#	264	-ity
1275	-ly	405	pre#	263	-able
1036	in#	396	dis#	253	-ness

Table 6: Top 30 affixes

The use of morph units increases the number of tokens used in language modelling by about 14%.

To appear in Proc. ICASSP-2000, Istanbul, Turkey

4. COMPARISON OF PCMA MODELS WITH WORD MODELS

Comparisons were made between PCMA and word-based models in terms of lexicon size, perplexity, model size, lattice scores, and word recognition rates.

4.1 Lexicon size

Our experiments show that morphological analysis substantially reduces the lexicon size. Take the 65k-lexicon as an example. Of the 64,978 items, 32,323 (49.7%) can be analysed by the morphological units listed in Table 7. Phonological constraints reduce this number slightly to 21,663 items (33.3%), which results in a reduction of 29.2% for the lexicon as a whole. Table 7 summarises the sizes and OOV rates of the PCMA lexicons.

Lexicon	*Size*	*Red. (%)*	*OOV (%)*	
			LOB1	*BNC1*
20k	13,370	33.2	7.03	7.08
40k	25,158	37.1	6.09	5.90
65k	46,000	29.2	5.02	5.45

Table 7: The size of morph-based lexicons.

As well as reducing the size of the lexicon, PCMA also reduces the OOV rate since many OOV words are simply different morphological inflexions of units in the lexicon. OOV rates for LOB1 and BNC1 are now comparable across the different vocabulary sizes.

4.2 Perplexity scores

Perplexity scores for the PCMA models are listed in Table 8. Morph sequence perplexities are about 55% of the word sequence perplexities. For instance, when trained with 20m-word set, the 65k PCMA model has a perplexity score of 218 with LOB1 and 207 with BNC1, a reduction of respectively 54.5% and 56.8% when compared with Table 2.

	LOB1		BNC1		LOB	
	10m	20m	10m	20m	10m	20m
20k	207	187	200	183	144	131
40k	224	201	220	200	158	144
65k	244	218	227	207	169	153

Table 8: Morph perplexities for PCMA models

However to directly compare morph-sequence perplexities to word-sequence perplexities is it necessary to compensate for the fact that there about 14% more morph-unit tokens in the test data than there are word-unit tokens. Scaling the log-probabilities accordingly shows that the morph-mapped word perplexities are actually slightly higher than word-based perplexities. Any improvements in word error rate are probably not to do with decreased perplexity alone.

To confirm that the reductions in perplexity were not simply due to the reduction in lexicon size, three additional lexicons were constructed that contain full 20k, 40k, and 65k PCMA items selected according to frequency of use. With models trained with 20m words, the LOB corpus yielded 173, 187, and 193 as perplexity scores. When compared to Table 2, the reductions are respectively 30.2%, 35.3%, and 50.8%, suggesting that the reductions are not merely due to the reduction of lexicon size.

4.3 Language model size

As a result of lexicon size reduction (c.f. Table 7), the language model size is accordingly reduced. Table 9 shows that with language models trained with 10 million words with different vocabulary sizes, as an example, the reduction rate is about 25% for bigrams and 10% for trigrams.

	No. of bigrams			*No. of trigrams*		
Size	*Word*	*PCMA*	*Red.*	*Word*	*PCMA*	*Red*
20k	1333754	1002651	24.8	4489914	4028923	10.3
40k	1542090	1129039	26.8	4798293	4243928	9.1
65k	1748975	1309768	25.1	5127667	4568349	10.9

Table 9: Reduction of model size

When training data size increases to 20 million words, the reduction rate accordingly becomes higher. At 40k level, according to Table 10, the reduction of the number of bigrams is as high as 29.2% and that of trigrams is 14.6%.

	No. of bigrams			*No. of trigrams*		
Size	*Raw*	*Morph*	*Red*	*Raw*	*Morph*	*Red*
20k	2049486	1488535	27.4	7831775	6791797	13.3
40k	2409319	1705237	29.2	8470465	7232267	14.6
65k	2708455	1961257	27.6	9017522	7753435	14.0

Table 10: Reduction of model size

4.4. Lattice scores

The results for maximal morph accuracy are listed in Table 11. It is significant that lattices generated through PCMA lexicons have achieved the maximal performance, i.e., 100% minus OOV rates. In Table 3, by contrast, the word error rate is nearly twice the OOV rate.

	LOB1 (%)	BNC1 (%)
20k	92.2	93.4
40k	94.3	95.2
65k	94.9	95.2

Table 11: Morph lattice scores

Direct comparison between morph-unit lattices and word-lattices is difficult because the average length of the units is shorter in the morph-unit lattice.

4.5. Word accuracy rates

Finally, word recognition rates were obtained with the PCMA models. According to Table 12, models trained with the 10m-word training set scored nearly 4% better than conventional word models (shown in Table 4).

	LOB1 (%)	BNC1 (%)
20k	58.6	55.7
40k	60.2	56.4
65k	59.9	57.6

Table 12: Word recognition scores

With the BNC1 test set, word recognition improvements

are slightly varied across the different vocabulary sets over the conventional models: 2.8%, 1.8%, and 2.6%. Table 13 summarises the performance of PCMA models trained from 20 million words from BNC. When compared to those listed in Table 5, the enhancement rates across different vocabulary sets were respectively 2.7%, 3.6%, and 3.8% for LOB1, and 1.0%, 1.2%, and 2.0% for BNC1.

	LOB1 (%)	BNC1 (%)
20k	59.1	55.5
40k	61.0	56.4
65k	61.4	57.3

Table 13: Word recognition scores

While the increase of training data size has resulted in better word accuracy rates for LOB1, the performance seems to have deteriorated for the other test set, BNC1. This deterioration is mainly due to the increase in the number of insertions.

4.6. Closed-vocabulary tests

The recognition tasks described in 4.5 are characterised by relatively large vocabularies and OOV rates: the materials were taken from general English corpora. To compare these results with more conventional recognition materials, PCMA modelling was also applied to the test materials of the WSJCAM0 database [6, 7]. The 1105 test sentences were divided into two groups: whether they arose from the 5k or the 20k WSJ lexicons. Table 14 summarises the recognition results for a 20k word, 65k word, 20k-equivalent morph-unit and a 65k-equivalent morph-unit lexicon.

Model	Word				PCMA			
Voc. size	20k		65k		20k		65k	
Test size	5k	20k	5k	20k	5k	20k	5k	20k
Results	58.8	61.5	64.9	66.2	60.3	63.2	63.4	65.2
Overall	60.3		65.6		61.9		64.3	

Table 14: Results from close-vocabulary tests

At 20k level, PCMA models were marginally better than the word-based models (61.9% vs 60.3%) whereas at 65k level the performance of PCMA fell below that of word-based models. This was probably due to the fact that the words in the test material were adequately covered by the larger word lexicon.

5. CONCLUSION

As described in this article, PCMA has shown through empirical tests a number of benefits:

- Reduced lexicon size – PCMA generates a much smaller lexicon for the same coverage, a reduction of 30% of the conventional pronunciation lexicon.
- Enhanced lattices – a larger proportion of correct readings are found in morph lattices compared to word lattices. In fact the morph lattices are at near-maximum performance.
- Reduced perplexities – morph sequence perplexities are only 50% of equivalent word-sequence perplexities.
- Reduced language model size – PCMA is capable of reducing word bigrams by 25% and word trigrams by about 10% .
- Increased word accuracy rates – PCMA has reduced the word error rate in absolute terms by about 2% and in relative terms by about 5% although this improvement was observed only in open-vocabulary recognition tasks.

We conclude that PCMA obtains most of its effect through the increased productivity of a fixed size lexicon. In tasks with high OOV rates, such as those derived from the BNC, the increase in coverage compensates for deficiencies arising from the use of fewer, smaller units. There seems to be no benefit with regards to language model perplexity, which might be expected since the trigram morph-unit model operates on a smaller 'window' of the sentence. The increase in morph lattice rates could be due to both a decrease in lexicon size and a decrease in the number of minimally different pronunciations in the lexicon.

It is possible that some of the deficiencies of the morph-unit model could be addressed by further work, in particular by adding phonological constraints on morph-unit combinations in a recognition post-processor, or by interpolating word and morph-unit language models.

ACKNOWLEDGEMENTS

The work was supported in part by the Engineering and Physical Science Research Council, UK, Grant No GR/L81406. We thank Tony Robinson and Steve Renals for assistance with the Abbot recognition system.

REFERENCES

[1] Burnard L. *Users Reference Guide for the British National Corpus*. Oxford University Computing Services, Oxford, 1995.

[2] Hochberg M., Renals R., and Robinson A. "ABBOT: The CUED Hybrid Connectionist-HMM Large Vocabulary Recognition System". *Proceedings of Language Technology Workshop, Austin Texas, Jan 1995*.

[3] Clarkson P. and Rosenfeld R. "Statistical Language Modeling using the CMU-Cambridge Toolkit". *Proceedings of Eurospeech*, 1997.

[4] Robinson T., Fransen J., Pye D., Foote J. and Renals S. "WSJCAM0: A British English Speech Corpus for Large Vocabulary Continuous Speech Recognition". *Proceedings of ICASSP*, 1995, pages 81-84.

[5] Quirk R., Greenbaum S., Leech G., and Svartvik J. *A Grammar of Contemporary English*. Longman, London, 1972.

[6] Fransen J., Pye D., Robinson T., Woodland P., and Young S. *WSJCAM0 Corpus and Recording Description*. Linguistic Data Consortium, 1994.

[7] Paul D.B. and Baker J.M. "The design for the Wall Street Journal-based CSR corpus". *Proceedings of Fifth DARPA Speech and Natural Language Workshop*, 1992, pages 357-362.