Large Vocabulary Speech Recognition of Slovenian Language Using Data-Driven Morphological Models

Tomaž Rotovnik, Mirjam Sepesy Maučec, Bogomir Horvat, and Zdravko Kačič

Faculty of Electrical Engineering and Computer Science, University of Maribor
Smetanova 17, 2000 Maribor, Slovenia
{tomaz.rotovnik, mirjam.sepesy, bogo.horvat, kacic}@uni-mb.si
http://www.dsplab.uni-mb.si

Abstract. A system for large vocabulary continuous speech recognition of Slovenian language is described. Two types of modelling units are examined: words and sub-words. The data-driven algorithm is used to automatically obtain word decompositions. The performances of one-pass and two-pass decoding strategies were compared. The new models gave promising results. The recognition accuracy was improved by 2.5% absolute at the same recognition time. On the other hand we achieved 30% increase in real time performance at the same recognition error.

1 Introduction

Slovenian language is highly inflected language, like other languages in the Slavic family. It possesses significant properties, which make it stand out as a potentially problematic language for automatic speech recognition. Slavic languages have complex morphological structure. It is possible to produce many different word forms from the same root with suffixes. The experiment has shown that the Slovenian corpus requires a vocabulary of 600K words in order to achieve a 99% training-set coverage. The vocabulary size for the Slovenian language must therefore be an order of magnitude larger than for the English language. Currently the most advanced speech recognition systems can handle vocabularies of 20K to up to 60K words. Considering this we have to restrict the vocabulary size in the case of Slovenian language. This would result in a high Out Of Vocabulary (OOV) rate. To solve this problem we propose the use of smaller lexical units.

2 Vocabulary, pronunciation dictionary and language models

Two different types of language models were built: word-based models and models at a sub-word level (named morphological models). For training them, the corpus of 60M words was used, obtained from the archives of a Slovenian newspaper VEČER, spanning the period from 1998 through 2000.

Vocabulary of word-based models was chosen to be the most frequent 20K words from the training corpus.

Inflectional change of the word mostly affects word ending, whereas stem remains unchanged. Because of this we split words in two smaller lexical units: stems and endings. Some words (non-inflectional) can not be decomposed. We leave them unchanged. The decomposition is determined automatically using the longest match principle (scanning the list of endings). The list of endings was defined by iterative algorithm, which searches for the minimum number of different units in a training-set. Decomposing vocabulary words resulted in 8497 different basic units. Vocabulary size was reduced by 58%.

Phonetic transcriptions of words were made automatically under basic grammatical principles using 30 phones. The number of phones was smaller than usual, because we did not differ between long and short vowels and we also excluded some rare phones. Sub-word vocabulary contained 134 homographs from stems and 31 homographs from endings (see table 1). Homographs are words or parts of words that have the same orthographic transcription, but different phonetic transcription. Some examples of word decomposition are shown in table 2. The first two words in the table have the same orthographic and phonetic transcriptions but different meaning (homographs). They differ in accentuation of vowels in pronunciation. The third word has the same ending as the first two words. The last four words have the same stem and different endings, except fourth and fifth words, which do not have endings at all.

Table 1. Vocabulary of stems and endings

	Unique	Homographs	Σ
Stems	6836	134	6970
Endings	1661	31	1692
Σ	8497	1665	8662

Table 2. Examples of word decomposition into stem-ending

Word	Decomposition		Transcription	Translation	Possible new biphon
	stem	ending			
kamen	kam	-en	kam - En	Stony	a-m
kamen	kam	-en	kam - @n	Stone	a-m
česen	čes	-en	CEs - @n	Garlic	E-s
pol	pol	-0	pOL	pole(celestial)	/
pol	pol	-0	pOw	Half	/
poleg	pol	-eg	pOl - Eg	Beside	O-l
polčas	pol	-čas	pOw - Cas	half-time	O-l

Bigram, trigram and fourgram backoff language models were built [4]. The trigram and fourgram models were used only for rescoring, whereas bigram language model was used for generating word graphs.

3 Acoustic models

As acoustic models word internal triphone models with 16 Gaussians mixtures for each state and for each model were built. The basic models had to be expanded with all unseen biphones for the sub-word based models. When decomposing words into stems and endings, and using word internal triphones (biphones and monophones included) from word-based vocabulary, it is possible to get new biphones as shown in table 2. Thus all missing biphones were created and added to the basic triphone models. The total number of models was 5983. The number of states was reduced from 10K to 3K after performing tree-based state tying. 4090 tied-state triphone models were obtained.

Acoustic models were trained on SNABI speech database [1]. It contains speeches of 52 speakers where each speaker read in average more than 200 sentences and 21 speakers read also the text passage. The total database consists of approx. 14 hours of speech.

4 LVCS Recognition

Two different decoding strategies were used. The first strategy included a standard time-synchronous Viterbi beam search decoder [2]. The second strategy is called two-pass decoding and includes two recognition stages [3]. A word graph is achieved as a result of keeping more then 1-best hypothesis. It defines possible word strings which can be used as grammar constraints in the second pass decoding. Although the process of word graph generation is very time-consuming, more complex acoustic and language models are used in order to obtain an overall better performance in accuracy.

5 Results

The results were evaluated on the SNABI test-set. It consisted of 779 pronunciations spoken by 7 speakers and contained phonetically rich sentences.

Baseline system uses words as basic units, Viterbi beam search decoder and bigram language models. Results are shown in table 3. The recognition accuracy of only 41.6% was obtained. Using words as basic units provide an OOV rate of 22,26%. For the second baseline system 2-pass decoder with the same vocabulary and trigram language model were used. Better recognition accuracy was achieved against real time degradation (57% increase).

In the next experiments morphological models were used. OOV rate was improved by 70%. The recognition accuracy was slightly better, (0.5% absolute) by using a one-pass decoder, but real time performance increased by 29% because

of a smaller search space. The second experiment included a two-pass decoder with trigram language models for rescoring. Comparing the recognition accuracy with the second baseline system, it can be seen that we got slightly worse results (about 2% absolute), while real time performance increased by 32%. The best results were obtained in the last experiment, when fourgram language model was used. The fourgram language model attempts to capture the correlation between the stems and endings of two neighbouring words. It can be seen as a counterpart of the word-based bigram model. Comparing the results of the two-pass decoder using fourgrams of stems and endings with one pass decoder using word bigrams, the recognition accuracy improved by 3.41% absolute.

Table 3. Results of experiments

Basic units	**words**			**stems and endings**		
Vocabulary size	20,000			8,662		
OOV [%]	22.26			6.51		
Decoding Strategy	ACC [%]	Mem [MB]	Speed [RT]	ACC [%]	Mem [MB]	Speed [RT]
1. Pass (bigram)	41.62	70	9.66	42.05	45	6.90
2. Pass (trigram)	44.16	200	15.18	42.24	155	10.26
2. Pass (fourgram)	-	-	-	45.03	530	10.40

6 Conclusion

Recognition accuracy at sub-word level is as good as the word based recognition, while real time recognition performance increases. The greatest benefit of modelling inflected languages at sub-word level is OOV rate improvement. The proposed technique for word decomposition is language-independent and can easily be applied to other Slavic languages as well.

References

1. Kačič, Z., Horvat, B., Zögling, A.: Isues in design and collection of large telephone speech corpus for Slovenian language, LREC 2000.
2. Young, S., Odell, J., Ollason, D., Kershaw, D., Valtcheva, V., Woodland, P.: The HTK Book, Entropic Inc., 2000.
3. Zhao, J., Hamaker, J., Deshmukh, N., Ganapathiraju, A., Picone, J.: Fast Recognition Techniques for Large Vocabulary Speech Recognition, Texas Instruments Incorporated, August 15, 1999.
4. P. Clarkson, R. Rosenfeld: Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings of Eurospeech*, 1997.