Decoder Issues in Unlimited Finnish Speech Recognition

Teemu Hirsimäki and Mikko Kurimo

Helsinki University of Technology
Neural Networks Research Centre
P.O. Box 5400, FI-02015 HUT
FINLAND
Tel. +358-9-451 3284, Fax: +358-9-451 3277
E-mail: teemu.hirsimaki@hut.fi, mikko.kurimo@hut.fi

ABSTRACT

In contrast to continuous speech recognition systems which utilize a fixed vocabulary to limit the search, practically unlimited vocabulary recognition can be achieved by constructing the recognition result from sub-word units. This paper discusses some important things to consider in sub-word based decoders, especially when recognizing languages with heavy use of inflections and compound words. Also, a decoder design implemented to achieve unlimited vocabulary Finnish speech recognition is described.

1. INTRODUCTION

Majority of the modern speech recognition systems use Hidden Markov Model (HMM) to model the acoustics of phonemes. In such systems, the central part, called decoder, is responsible for finding the most probable HMM state sequence, given the observed speech, acoustic model, and language model. In order to limit the search space, which contains all possible state sequences, the decoders usually consider only words in a fixed lexicon. This approach has been quite successful, especially in languages like English, which do not have too many inflectional forms or compound words. However, in many European languages like Finnish, Turkish, Czech or German, the inflections and compound words are so common, that covering all possible word forms in a simple lexicon is not very effective. For example, a lexicon of two million words may be needed in Finnish to get the same coverage as a 64 000-word lexicon gives in English [1].

To improve the recognition of these languages, there has been active research on approaches which abandon the word-based lexicon. In [2], grammatical morphemes for Czech were generated with a morphological analyzer. The out-of-vocabulary rate decreased clearly, but the word error rates were similar to a word-based approach. In [1], four different rule-based methods to split words into sub-word units were experimented for Finnish and German tasks with good results. The unsupervised data-driven word splitting algorithm presented in [3] has been found to be very effective in Finnish [4] and Turkish [5] speech recognition.

In contrast to splitting words into smaller units, other methods have also been proposed to tackle the problem of vocabulary growth. In [6], a dynamical lexicon adaptation was studied in German recognition task. A traditional word lattice generated in the first pass recognition was expanded by adding inflectional forms for each word in the lattice. This expanded lattice was then used in the second pass with better results. In [7], the error rates were reduced by growing the lexicon sizes up to 500 000 words in German broadcast news task.

The main contribution of this paper is to discuss how the use of sub-word units instead of whole words affects the other parts of a HMM based recognition system. The most important issues are how word breaks can be modeled, and how the choice of the sub-word units affect the search space in which the decoder tries to find the most probable state sequence. Also, with respect to these issues, the paper presents the decoder of the speech recognition system used in the Neural Networks Research Centre in Helsinki University of Technology. The system has been used to experiment different sub-word units in order to achieve practically unlimited vocabulary speech recognition for continuous Finnish speech. The emphasis of the paper is in the decoding process, so other important parts of a typical speech recognition system as feature extraction, acoustic modeling and language modeling are not discussed in detail.

2. DECODER ISSUES

2.1 Search Problem

Strictly speaking, the task of the decoder in the HMM-based recognition systems is to find the most probable state sequence given the observed speech signal, acoustic model, and language model. If the decoder had infinite amount of time, the decoding would be very simple indeed: The decoder could compute the probability of every possible state sequence in turn and pick the most probable one. However, the number of possible state sequences is astronomical even for a short segment of speech, which is the reason why there exists so many different decoding and pruning strategies. A good overview of the fundamental ideas is presented in [8].

2.2 Morphs

As a lexicon of words can be used to limit the search to the most frequent words, it is also an option to have a collection of sub-word units that are considered during the search. From now on, the term *morph* is used to refer one of these sub-word units. A morph can be a whole word, syllable, grammatical morpheme, or something else as long as it is found in the *morph lexicon*.

2.3 Word Breaks

There are a few things that need special attention when morphs are used instead of whole words. With a word lexicon, it is clear that there is a word break after each lexical unit (i.e. word). But now that several morphs can be concatenated into a single word, the breaks must be inferred in some other way. Acoustically, long silences obviously mark word breaks, but unfortunately in fluent speech, there are seldom clear silences between words. On the other hand, humans seem to conclude the word breaks mostly by understanding the words and meaning. Thus, a simple solution is to have a word break morph in the morph lexicon, and let the decoder consider a break between each morph even without acoustical evidence. Then it is left for the language model to get the word breaks right. This approach was taken in our decoder.

2.4 Choosing the Optimal Morphs

2.4.1 Limiting the Search Space

In principle, the decoder algorithms do not care what kind of morphs are used, but nevertheless, they are an important factor in the recognition process. Firstly, it is good to note that the lexicon limits the search to some state sequences, so it is a strict pruning method in contrast to acoustic and language models, which prune sequences softly by assigning probabilities. Thus, the selection of the morph set controls the trade-off between the two following two issues: How well the words of the language can be built from the morphs, and how much the morph set is able to limit the search. If the morphs are short, they can cover the words of the language well, but they allow as well many words that are not grammatically correct words but still very close to the correct ones acoustically. Also, it is very hard to get good language models over very short morphs, e.g. just phonemes. On the other extreme, there is the word lexicon, which can not cover all possible words, but limits the search effectively.

2.4.2 Pronunciation

In addition to limiting the search, the lexicon also provides pronunciation of the morphs, i.e. the sequence of the HMM states that form the morph. In Finnish, for example, the pronunciation of a morph can be derived quite accurately from the written form, while in the English, it is often necessary to know the whole word to deduce the pronunciation. Thus, in some languages the pronunciation may dictate what kind of morphs are useful.

2.4.3 Context-Sensitive Acoustic Models

Also, the choice of the phoneme models is related to the morphs. In context-sensitive acoustic modeling, several separate models are trained for each phoneme, depending on the context of the phoneme. Then the length of the morphs makes a difference, especially if the decoder is not able to model the phoneme context across lexical units. In this case, it is clear that longer morphs benefit more from the context-models, because shorter morphs induce more morph boundaries that prevent the use of the context. However, even if the decoder can take the cross-morph contexts into account, the computational overhead may depend quite much on the length of the morphs. Typically, expanding the context over the boundaries of the lexical units involves using several copies of the lexical tree, and if there are more boundaries where the context needs to be taken into account, the computational cost may be higher. For interested readers, a detailed description of using context-sensitive models across words is presented in [9].

All in all, it seems to be an open question what kind of morphs are optimal for speech recognition, if very large vocabularies are desired, but most likely, the chosen decoding approach has also an effect. This should always be kept in mind when different splitting methods are compared in speech recognition tasks.

3. DECODER DESIGN

3.1 Overview of the Decoder

The stack decoding principle used in the decoder is based on the stack decoder of the Duisburg University, Ducoder [10]. During the recognition process, the decoder maintains a set morph sequences that have most probably generated the observations up to a certain frame. These most probable morph sequences are called *hypotheses*, and they are organized in stacks according to their ending times. Each time frame has a stack of the most probable hypotheses ending in the frame in question.

The basic idea of the decoding process is as follows. Initially, there is only one empty hypothesis in the stack of the first frame. At each step, the decoder moves forward to the first frame containing hypotheses (this is also the earliest frame containing hypotheses). Then the decoder uses the morph lexicon and the acoustic models to make a local search to the near future (1–2 seconds, for example) for the acoustically best morphs starting from the current frame. Then the decoder takes each hypothesis from the current stack in turn, and makes several copies of the hypothesis by appending each of the most promising morphs at time. The newly created hypotheses are put in the future stacks according to the best ending times of the morphs that were used to expand the hypotheses (see Fig. 1). After expanding each hypothesis with each of the most promising morphs, the hypotheses in the current stack are discarded, and the decoder can proceed to the next frame that contains hypotheses.

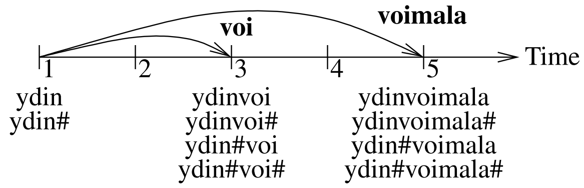

Fig. 1. The hypotheses in the stack 1 are expanded by finding the acoustically best morphs, and the expanded hypotheses are placed in stacks 3 and 5 according to the best ending times of the morphs. The hypothesized word breaks are marked with #.

3.2 Local Acoustic Search

The acoustically best morphs are searched using a morph lexicon tree. All legal morphs are combined into a big branching HMM, and common prefixes of the morphs are merged. Because the language model is not used in the local search, a traditional Viterbi search can be used to compute the best paths for each morphs effectively. As a result, the computation provide the cumulative log-probability of the best path for each morph end state for each time frame. Then the best ending times can be selected for each morph, and the best morphs (on average) can be chosen.

3.3 Using Language Model Probabilities

As mentioned above, the local search, i.e. finding the best morphs starting from a given frame, is done using only the acoustic models. In this decoding approach, it is not easy to use the language model information in acoustic search, because the morph history is ambiguous: The acoustically promising morphs can follow different morphs (in different hypotheses). Thus, whenever a new hypothesis is created by appending a previously found potential morph to a hypothesis in the current stack, the language model probability of the hypothesis can be updated, because for each hypothesis the morph history is unique.

3.4 Pruning Hypotheses

Of course, the basic idea described above would drown the decoder in the exponentially growing number of hypotheses unless some pruning method is used to discard the most improbable ones.

Firstly, it is important to prune hypotheses that can not end up in the recognition result. For example, if two hypotheses ending on the same frame contain the same morph sequence, but have the morph boundaries on different frames, it is known for sure that the more probable hypothesis will be more probable in the end, whatever happens. This is due to the fact that the future acoustic and language model probabilities do not depend on the alignment of the history. Actually, the language model probabilities depend only on $n-1$ most recent morphs, if an n-gram model is used, so whenever a hypothesis is put in a stack, the decoder checks if there exists a hypothesis with the same $n-1$ most recent morphs, and stores only the more probable one.

The above mentioned pruning is exact in the sense that it never introduces more errors. However, in practice, additional prunings are needed to make the decoding computationally feasible. The most obvious pruning is to limit the size of hypothesis list on each frame, so that a hypothesis is stored only if it is among the n best hypotheses, and its log-probability is not worse than a fixed threshold when compared to the most probable hypothesis in the stack.

All of the above prunings consider only hypotheses inside a single stack. Comparing hypotheses on different frames is not so straightforward, because it is hard to estimate how much the log-probability of the shorter hypothesis will change when it reaches the other hypothesis. Some approaches to compare hypotheses on different frames are discussed in [10].

4. DISCUSSION AND CONCLUSIONS

In this paper, we have discussed the relevant decoding and modeling issues when sub-word units are used instead of words. We have as well presented our stack decoder design for unlimited vocabulary continuous speech recognizer, which utilizes language models and lexicon based on sub-word units. We have earlier used the system to compare a word-based lexicon with different sub-word units. With the morpheme-like units instead of words, the word error rate decreased from 56% to 32% in a very large vocabulary Finnish recognition task. Corresponding letter error rates were 14% and 7.3% [4].

The decoder design is by no means optimized for sub-word recognition. Originally, the decoder approach was chosen so that it would be useful for studying language models which may also use much wider contexts than traditional n-gram-models. In a different decoding approach, the lexicon, language model and possible context-sensitive phoneme models are combined into a big HMM before the recognition. Even if language models exploiting long contexts may produce vast number of states in the HMM, minimization techniques can be used to allow the use of even 6-gram models [11]. However, other than n-gram models might be more difficult to incorporate in this approach.

As such, replacing the word lexicon with a morph lexicon should be quite straightforward in other decoding approaches too. It just has to be kept in mind, that the choice of the morphs affects generating the pronunciation lexicon, modeling the word boundaries, and handling the context-sensitive phoneme models. Especially, when comparing different word splitting algorithms, one has to be careful that comparisons are fair.

REFERENCES

[1] J. Kneissler and D. Klakow, "Speech recognition for huge vocabularies by using optimized sub-word units," in *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg, Denmark, 2001, pp. 69–72.

[2] W. Byrne, J. Hacič, P. Ircing, F. Jelinek, S. Khudanpur, P. Krbec, and J. Psutka, "On large vocabulary continuous speech recognition of highly inflectional language — Czech," in *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg, Denmark, 2001, pp. 487–489.

[3] M. Creutz, "Unsupervised discovery of morphemes," in *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, Philadelphia, Pennsylvania, July 2002, pp. 21–30.

[4] V. Siivola, T. Hirsimäki, M. Creutz, and M. Kurimo, "Unlimited vocabulary speech recognition based on morphs discovered in an unsupervised manner," in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, Sept. 2003, pp. 2293–2296.

[5] K. Hacioglu, B. Pellom, T. Ciloglu, O. Ozturk, M. Kurimo, and M. Creutz, "On lexicon creation for Turkish LVCSR," in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, Sept. 2003, pp. 1165–1168.

[6] P. Geutner, M. Finke, and P. Scheytt, "Adaptive vocabularies for transcribing multilingual broadcast news," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seattle, Washington, May 1998.

[7] K. McTait and M. Adda-Decker, "The 300k LIMSI German broadcast news transcription system." in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, Sept. 2003, pp. 213–216.

[8] X. L. Aubert, "An overview of decoding techniques for large vocabulary continuous speech recognition," *Computer Speech and Language*, vol. 16, no. 1, pp. 88–114, Jan. 2002.

[9] A. Sixtus and H. Ney, "From within-word model search to across-word model search in large vocabulary continuous speech recognition," *Computer Speech and Language*, vol. 16, no. 2, pp. 245–271, Apr. 2002.

[10] D. Willett, C. Neukirchen, and G. Rigoll, "Ducoder - the Duisburg University LVSCR stackdecoder," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Istanbul, Turkey, 2000, pp. 1555–1558.

[11] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Computer Speech and Language*, vol. 16, no. 1, pp. 69–88, Jan. 2002.