Proceedings of the Institute of Acoustics

FINDING THE N-BEST PHRASES IN A CONTINUOUS SPEECH RECOGNITION SYSTEM WITH PARTIAL TRACEBACK

M. Kadirkamanathan

Speech Research Unit, Defence Research Agency, St. Andrews Rd, Malvern, WR14 3PS, England.

1 INTRODUCTION

Spoken language recognition involves many sources of information including acoustic, phonetic and language knowledge. It is also viewed as a multi-level task since acoustic pattern matching is performed on much shorter durations of time while the language parsing is performed over the identified words of much longer durations. Attempts to incorporate some language processing within the acoustic matching algorithms have been expensive and inelegant. The communication between the acoustic matching engine and the language parser imposes a large amount of communication between the two and extra computing. This is alleviated to some extend by using an open loop system involving an N-best words/sentence recogniser and a language parser. The acoustic matching engine produces a number of candidate choices for each word or sentence thus providing the language parser with enough alternatives and flexibility to parse the spoken words. It is hoped that the correct word in each case would be included among the top choices.

The N-best sentence recognition paradigm [3] [1] [4] [5] obtains the top N sentences which have the highest scores. It is an extension of the viterbi word search algorithm. In Viterbi search algorithms each hypothesis is uniquely identified not only by the words but also their found locations along the time axis. Consequently some of the sentences among top N may may contain the same words with different time alignments. There are many approximate versions of this algorithm which obtain different sentences among the top choices [2]. All implementations of the N-best sentences algorithm are bidirectional where the best sentence is obtained during the forward pass and the next best sentences are obtained during the backward pass, once the end of input is reached.

Most of the top N sentences in the sentence hypotheses generation algorithm differ from another choice only by a few words due to the nature of the dynamic programming optimization. It therefore seems efficient to use the raw lattice representation to include the top N sentences as alternative phrases to words or phrases of the best sentence.

This article proposes an implementation of the N-best paradigm which can be used with a continuously running speech recogniser. Where the N-best findings do not have to wait until the end of the speech input. When implementing the N-best paradigm in continuously running mode it is no longer possible obtain ranked sentences as the data is received continuously all the time. Hence the next best solutions can only be described in a more raw form; for example as an alternative choice of phrases to the best recognised words. The term N may now be used to refer to the number of segment alternatives for each word or phrase. Hence the implementation is termed the N-best phrases hypotheses algorithm.

2 PARTIAL-TRACEBACK AND THE N-MULTIPLE HYPOTHESIS

In a 1-best continuous speech recogniser with partial trace-back, the partial theory (or recognised words) could be published as soon as all active paths at the current time instant are identified as having a common ancestor. [6] [7]. This is based on the finding that all active paths could be mapped into a tree spanning out with time. A path is uniquely identified by the sequence of states at each time instance in the past. A branch of a tree is formed when two paths share a common sequence of ancestors from the start of input up to an instant in the past. All active paths mapped in this fashion constitute to the tree. The location of the oldest branch in time, *i.e.* the youngest ancestor state that is common to all currently active paths, marks the end of the resolved region. The words corresponding to the path up to this point in time could be published as recognised words. The unresolved region occupies the time between the resolved region and the current instant. This region will grow with each input frame but will be truncated whenever a new or a more recent location of the oldest branch is detected. The partial trace-back is merely the mechanism which implements this. The words corresponding to the newly discovered common path are published as more recognised words.

In the N-best algorithm there will be N paths at each state corresponding to a partial word hypothesis at a time instant. These paths may be sorted as a ranked list in terms of decreasing scores. This is how it is done in the time synchronous forward search in [1]. The tree of partial paths will now contain more branches which have a layered structure in terms of rank. A branch, in addition to the properties mentioned above, will either maintain its rank or drop ranks with time. It will become dead when it falls out of the ranking.

At word boundaries where a ranked list of word end candidates are obtained only the best word candidate theory is extended to form new branches. The others are frozen only to be used during the trace-back. These branches of alternatives will meet the optimum path at branch point. This meeting point could be as far back as the beginning of the input data but usually only a few word durations away.

Although many paths are formed with N times as many hypotheses present in the system, most of the best alternative (2nd and lower ranked) paths will become frozen at various time instances to allow only the potential candidates for the optimum path to be kept alive. The identification of the optimum sequence of words may therefore be obtained in the usual way.

3 THE N-BEST PHRASES ALGORITHM

In order to obtain the N-best phrases in a time synchronous manner the forward search needs to be interleaved with the partial trace-back and the N-best phrases extraction scheme.

For each input frame the forward search updates the scores associated with each state in each word. Then the partial trace-back mechanism may be activated to locate if more resolved regions could be identified. If so more N-best phrases could be extracted along with the optimum words.

To extract the best segment alternatives, steps similar to those described in [2] could be carried out. This done by computing the extra cost associated with chosing an alternative segment instead of the corresponding optimum one.

1. Initialize stack of alternate phrases.
2. Initialize accumulated excess cost score s to zero. Initialize word position t to the end in time of the newly resolved region.
3. perform trace-back computation from t with score s, chaining back through word ends to produce the next highest scoring phrase. Stop as soon as the trace-back it hits the optimum path. Output

the words.

4. At each word boundary along the trace-back, add the excess cost, s, to the difference between a word lower down in the word end list and the current word. This is the lowest cost of choosing a segment containing this word instead of the optimum segment. Insert the word and the score in the stack of alternate phrases sorted by increasing s.

5. Pick the next entry or the first entry if the cycle is executed for the first time. It must be noted that any further entries into the stack will only be made lower than the current segment.

6. Perform steps 3 - 5 recursively until the desired number of alternatives have been obtained. A threshold may also be imposed on the extra cost.

The fig. 1 shows a screen dump an N-best phrases output. Below the menu is a portion of speech data and the first line below that shows the correct annotation of the speech. The second line depicts the top recognition words with their start and end times marked with vertical bars. The boxes below show the alternative phrases to the best recognised words. The right and left hand edges coincide with best word boundaries. The shades indicate the extra cost of choosing the segment printed inside the box instead of the best words. The boxes are not placed in the rank order owing to need to plot as many alternatives as possible into the display area.

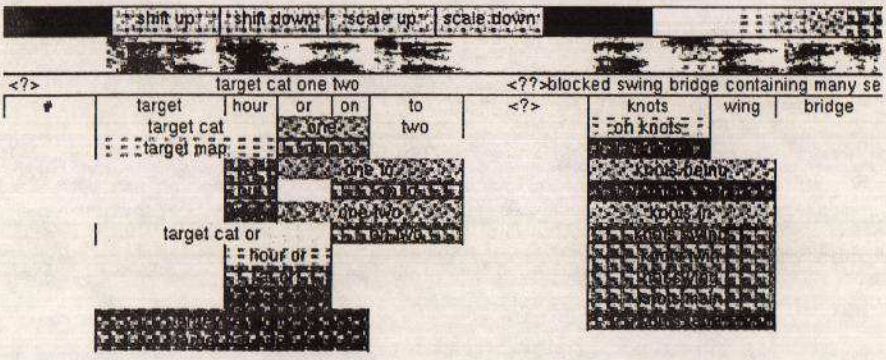

Figure 1: Output of the N-best phrases recogniser

3.1 Implementation issues and unresolved trace-back

The partial trace-back may in exceptional circumstances have a large unresolved region extending a long time into the past. This implies that the two or more live paths do not share a common ancestor until far back into the past. This is a theoretical possibility which exists even with the 1-best version which may result in a system to run out of storage memory. With N many times more active paths present at any one instant, the possibility of a large unresolved region seems to be greater in the N-best phrases generation algorithm. This however has not happened in the experiment conducted in this article. Should it occur, some of the oldest branches of the tree could be pruned out on the basis of its rank until the unresolved region is reduced to a manageable size.

4 THE ARM EXPERIMENT

A subword word HMM based word recognition experiment is carried out in order to evaluate the N-Best phrases obtained with the algorithm. The task is the Airborne Reconnaisance Mission and the ARM database consisted of three reports for each speaker. Speaker independent triphone subword models were used in the recognition. Each model had 3 states with unimodal gaussian distributions. These models were trained with 61 speakers. The test set consisted of another 10 speakers. A word penalty of 30 was imposed (in natural log probabilities).

With a number of word or phrase alternatives presented in the output, an alternative was chosen as correct only if it matched correctly all the words that it would replace. Figure 2. summarises the results.

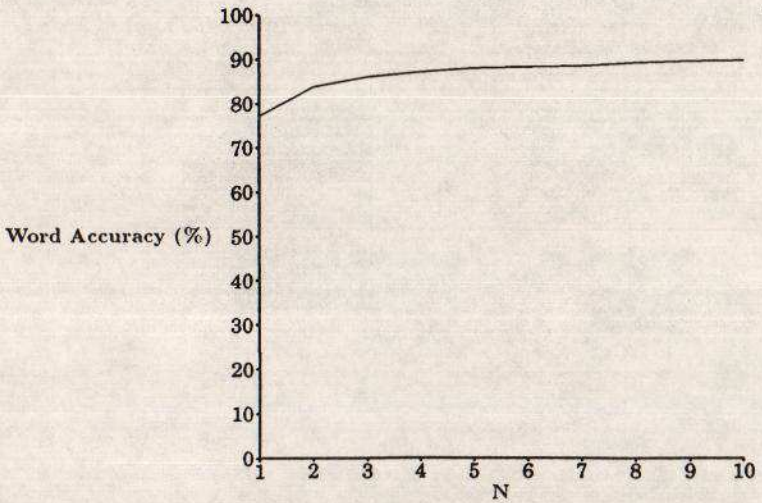

Figure 2: Word recognition accuracy of the ARM database.

5 DISCUSSION

In the proposed N-best phrases hypotheses generation algorithm the results do not have to wait until the end of the input data to be published. The scheme also works in a continuous mode producing the alternative phrases in a time synchronous way. The memory requirements therefore do not increase with longer sentences which may be a limitation with the bidirectional implementations. In this implementation a maximum duration had been imposed on the alternative phrases in order to limit the maximum size of memory used.

The proportion of the time spent in the partial trace-back is negligible compared with the cost of the forward search.

The recognition experiment performed on the ARM database illustrates that the word errors may be reduced at most by a half with a postprocesser which chooses appropriately from the list of words and phrases. The algorithm runs about ten times slower than realtime on a 32.4 SPECmark processor.

References

[1] Schwartz R., Chow Y. L., *The N-Best algorithm: An efficient and exact procedure for finding the N most likely sentence hypotheses*, Proc. ICASSP90, pp.81-84, Albuquerque 1990.

[2] Schwartz R., Austin S., *A comparison of several approximate algorithms for finding multiple (N-Best) sentence hypotheses*, Proc. ICASSP90, pp.701-704, Toronto 1991.

[3] Young S. J., *Generating multiple solutions from Connected word DP recognition algorithms*, Proc. Institute of Acoustics, Vol 6-4 pp.351-354, 1984.

[4] Steinbiss V., *Sentence hypothesis generation in a continuous speech recognition system*, Proc. European Conf. Speech Comm. and Tech., pp.51-54, Paris 1989.

[5] Marino J., Monte E., *Generation of multiple hypothesis in connected phonetic-unit recognition by a modified one-stage dynamic programming algorithm* Proc. Europen Conf. Speech Comm. and Tech., pp.408-411, Paris 1989.

[6] Bridle J. S., Brown M., D., Chamberlain R. M., *An algorithm for connected word recognition* Proc. ICASSP82, pp.899-902, Paris 1982.

[7] Brown P., Spohrer J., Hochschild P., Baker J., *Partial traceback and dynamic programming* Proc. ICASSP82, pp.1629-1632, Paris 1982.