Published in the 3rd European Conference on Speech Communication and Technology, Berlin, pp. 1959-1962, 1993.

DYNAMIC USE OF SYNTACTICAL KNOWLEDGE IN CONTINUOUS SPEECH RECOGNITION

Pierre Dupont

France Télécom, CNET Lannion A - TSS/RCP,
BP 40, 2 Route de Trégastel, F-22301 Lannion cedex - France

ABSTRACT

The control of continuous speech recognition by a context-free based language model requires a parsing process which may overload the acoustic decoding algorithm. We present a new approach to integrate such a language model in the search process. This approach extends the beam search Viterbi algorithm. In our case, the pruning technique not only selects the most likely acoustic hypotheses but also governs the dynamic expansion of a network structure. This algorithm is general enough to cope with the self-embedded recursivity of context-free languages and it favourably compares with other parsing techniques applied to spoken inputs.
We present results which show that the syntactical knowledge may be efficiently included at the frame level of an acoustic decoding algorithm.

Keywords: *Continuous Speech Recognition, Context-Free Language Models, Beam Search Viterbi Algorithm.*

1. INTRODUCTION

Many continuous speech understanding systems use different language models for recognition, on the one hand, and linguistic analysis, on the other hand. This is mainly due to the computational overhead required to parse spoken inputs during the acoustic decoding. Trigrams, for instance, define word probabilities given the two predecessors. As such, they are easily incorporated in the acoustic decoding. However, their major drawback comes from their inability to capture long span dependencies typical of natural language. As a consequence, the recognized word sequence may not be accepted by the subsequent linguistic analysis.

Instead of trigrams, Stochastic Context-Free Grammars (SCFGs) have been proposed. They are of great interest in virtue of their ability to represent linguistic constraints in a stochastic formalism. This offers a unified framework for acoustic and syntactic modelling of speech [1]. Moreover, the SCFGs could bridge the gap between language models used for recognition and those used for understanding. The problem of estimating the rule probabilities has been already investigated [2, 3, 4]. In this paper, we cope with the integration of such a language model into a continuous speech recognizer.

The processing of general context-free grammars requires a recursive parsing algorithm which may overload the acoustic decoding. One classical solution consists in parsing a word lattice (or N-best word sequences), obtained in a first step without accurate language modelling. Word graphs have been recently proposed as another interface between recognition and understanding [5]. Correction procedures may also be applied to recover the uttered sentence from a garbled word sequence. However, all these two-step approaches do not take full advantage of the language modelling since most of the hypotheses considered during the acoustic decoding are ruled out in view of producing an intermediate result, i.e a word lattice, N-best word sequences or a word graph, of reasonable size.

We present a new way to include the language model at the frame level of the acoustic decoding. The proposed technique relies on mixed transition networks presented in section 2. They gather the acoustic modelling based on context-dependent Hidden Markov Models (HMMs), phonological rules, the lexicon, and the context-free language model. In section 3, we describe how the initial network, associated with the grammar axiom, is recursively expanded during the Viterbi search controlled by a beam pruning technique. Section 4 concerns the Greibach normalization which ensures that the proposed technique can process any context-free language in an efficient way. Comparisons with similar approaches are drawn in section 5. The results presented in section 6 show significantly better recognition rates than those obtained by a two-step approach.

2. MIXED TRANSITION NETWORKS

A lexicon word may be represented and stored as a stochastic finite state network, possibly constructed by concatenating subword models, including various pronunciations and taking phonological rules into account. This representation embodies the acoustic models, the phonological and lexical knowledge. It corresponds to a network of HMMs being associated with each word.

In the PHIL90 continuous speech recognition system [6], the linguistic knowledge is included in the stochastic network representation. As the language model constraints (or assigns probabilities to) possible word sequences, word models are accordingly linked. The resulting structure is a single HMM. The different paths through this stochastic finite state network describe the modelled sentences. While this approach offers the advantage of a simple Viterbi decoding to find the optimal state sequence, it may become intractable whenever the lexicon size exceeds a few words or when the language model contains more than a few rules. Another drastic limitation arises when the language model presents self-embedded recursivity. In that case, the corresponding network would contain an infinite number of states and transitions. In other words, the recursive nature of context-free grammars may not be eliminated even if backward loops are added.

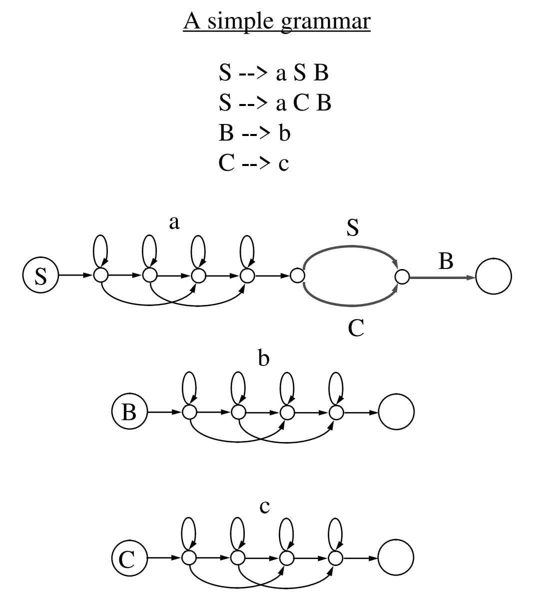

Figure 1: Mixed Transition Networks

We define a set of Mixed Transition Networks (MTNs), each of them being associated with one nonterminal symbol. These MTNs are equivalent to Recursive Transition Networks except that the terminal symbols, i.e. the lexicon words, are replaced by their assigned HMM (see figure 1). It means that the set of MTNs combines the language model and the acoustic-phonetic models. For the sake of simplicity, the rule probabilities are omitted here and we represent word models as single Markov chains. In fact, this is extended to the case of multiple Markov chains with multiple input and output states because of the context-dependent phonetic modelling. Terminals and nonterminals are respectively denoted by lower case and upper case letters.

Let us mention that the terminals usually identify the lexicon words while the definition of nonterminals is arbitrary. They may refer to syntactic or semantic categories and various definitions may cover the same language. Moreover, the underlying grammar may be rewritten in a normal form (see section 4). Therefore, the number and size of the MTNs are accordingly adapted.

3. EXPANSION OF THE INITIAL MTN

We use the beam search Viterbi algorithm to perform the acoustic decoding. The key point is that the pruning mechanism not only selects the most likely acoustic hypotheses but also governs the dynamic expansion of a particular MTN which, in turn, guides the search process. During the time-synchronous recognition, the MTN assigned to the grammar axiom, i.e. the start symbol «S», is progressively expanded.

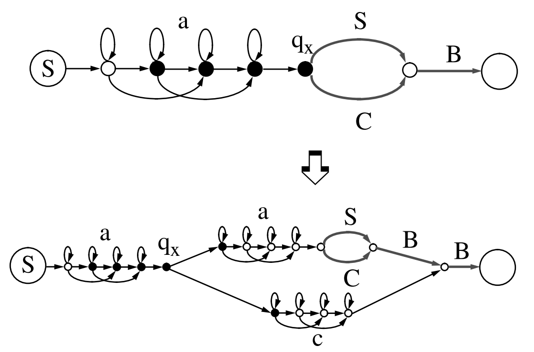

Figure 2: Dynamic expansion of the initial MTN, after state q_x became active

The first time a given nonterminal transition is encountered, its associated MTN is copied in place of this transition. From this point, the search may be pursued until it reaches other nonterminal transitions a few frames later. This dynamic expansion implies an incremental growth of the initial MTN along the time.

Figure 2 illustrates the copying of the MTNs associated with the nonterminals «S» and «C». A state is active if at least one path reaching this state has not been pruned. All active states are here denoted by black circles. The copies are performed when the end state q_x of the first word «a» becomes active.

Since a given nonterminal may appear in different contexts, its associated MTN is effectively copied several times if and only if the pruning allows the activation of the states from which the corresponding transitions arise. The potentially exponential growth of this dynamical structure is furthermore reduced by the fact that the probability of entering a network decreases with the recursion depth and is therefore bounded by the pruning threshold. The proposed approach is thus able to cope with self-embedded recursive structures.

As we use pruning, the theoretical optimality of the Viterbi search is lost. Experiments on speaker independent recognition of telephonic speech show that the search may be strongly pruned without affecting the recognition rate (see section 6). Let us quote that the size of the expanded network clearly depends on the quality of the acoustic modelling. An interesting feature is that a better acoustic modelling or a higher quality speech signal leads to a less expanded MTN, due to pruning control. The recognition process is therefore not only improved but also requires less storage.

4. GREIBACH NORMALIZATION

A left-recursive grammar would cause the dynamic expansion to fail in a recursive loop. To avoid this problem, the language model has to be translated in Greibach Normal Form.

<u>Definition [7]:</u> A CFG $G = (N, \Sigma, P, S)$ is said to be in Greibach Normal Form (GNF) if G is ε-free and each non-ε-production in P is of the form $A \dashrightarrow a\alpha$ with $a \in \Sigma$ and $\alpha \in N^*$.

<u>Theorem</u>: Any context-free language may be described by a grammar G written in GNF. The constructive proof is given in [7].

Let us note that the simple grammar mentioned in the example above was already written in GNF. The Greibach normalization guarantees that transitions starting from the initial node of any MTN are not labelled by nonterminals. Consequently, the network expansion is not recursive at a given time frame. Furthermore, the network expansion is strictly restricted to the part of the search space actually explored. In that sense, the MTN expansion based decoding algorithm described in section 3 is optimal when the underlying grammar is expressed in GNF.

5. COMPARISON WITH SIMILAR APPROACHES

Several other attempts have been made at introducing a parsing algorithm into a time-synchronous Viterbi search [8, 9, 10]. The LR parsing approach suffers from the drawback that it only covers a subclass of context-free grammars [7]. Chart parsing techniques, like the Earley algorithm [11] or the CYK algorithm [12] have a time complexity which is cubic in the length of the input sentence when handling incorrect data [9]. The beam search strategy is furthermore difficult to apply with those techniques in the case of general context-free grammars, as mentioned by Ney [9].

The proposed approach overcomes all these problems. In particular, the beam pruning controls the search space growth in a very efficient way and drastically reduces the theoretical complexity of a network expansion based parsing. The great efficiency of our method relies on the fact that to process speech inputs (as opposed to written inputs) many equivalent parses have to be treated for different time indexes. This results from the high uncertainty on word boundaries in continuous speech recognition. For the sake of search optimality, it usually requires different chart or table entries associated with those equivalent parses. In our approach, the syntactic processing is minimal since a nonterminal appearing in a given context needs to be expanded only once. Subsequent equivalent parses amount to run through the expanded MTN.

All these conclusions hold for SCFGs since the introduction of rule probabilities merely modifies some transition probabilities. It does not affect the parsing complexity and, on the contrary, may help to better prune the search.

Note that the proposed approach can be used for applying semantic constraints on spontaneous speech (see [13], for example). Indeed, the general formalism of context-free grammars is not bound down to represent purely syntactical knowledge. Here again, we would take advantage of the MTN representation and of the Greibach normalization as explained in section 4.

6. RESULTS

A first experiment has been performed on speaker independent French city name spelling recognition over the telephone. The acoustic modelling was based on HMM with unimodal gaussian density functions. In that experiment, the lexicon size was 30 (26 letters, digit «2» and 3 different silent words). Table 1 summarizes the results obtained on a test database of 1,641 city name spellings. The first line refers to unconstrained letter sequence recognition. The second line refers to a two-step approach (described in [14]) which consists in correcting the garbled letter sequence from a known list of 5,000 city names. The most likely city name is computed by using a discrete HMM, given the recognized letter sequence. The third line refers to the fully integrated approach in which the city name list is expressed as a grammar written in GNF. It offers the possibility to check the efficiency of our integrated approach on a language model composed of 5,054 rules.

Recognition type	**City name error rates**
Unconstrained	57.2%
Two steps	6.6%
Integrated	3.0%

Table 1: Test set error rates.

The low perplexity of the language model, i.e letter sequence constraints, explains why the error rates are significantly better compared to unconstrained recognition. The results show moreover that the city name error rate is divided by 2.2 as a consequence of integrating the language model in the acoustic decoding algorithm. The right hand side of each rule contained 6 nonterminals, on the average. The maximal number of MTN copies was thus about 30,000. Remark that an average number of 1,643 MTN copies were performed due to pruning control. The expanded MTN was composed of about 47,700 states and 150,000 transitions while a complete expansion would have resulted in a network of about 725,000 states and 2,300,000 transitions.

A second experiment has been performed on an air travel information task in French. The lexicon was made of 182 words. The test set consisted of 915 continuously spoken sentences. Preliminary results have been obtained in speaker-dependent mode. Without language model, the word error rate was 11.5%. The control of the acoustic decoding by a language model composed of 310 rules and 51 nonterminals reduced the word error rate to 4.2%. The size of the search space dynamically developed and explored was less than 10% of the whole.

7. CONCLUSIONS

We propose a general technique to guide a continuous speech recognition system by a context-free based language model.

By assigning a Mixed Transition Network to each nonterminal of the grammar written in GNF, we define an efficient representation to control the acoustic decoding, at the frame level. It allows us to significantly improve the recognition rates in comparison with a two-step approach.

Results show that a very small part of the potential search space may be dynamically developed and explored due to pruning control.

REFERENCES

[1] H. Ney, *Stochastic Grammars and Pattern Recognition*, Proc. of NATO ASI, 1990, pp. 319-344.

[2] K. Lari and S.Y. Young, *The Estimation of Stochastic Context-Free Grammars using the Inside-Outside Algorithm*, Computer Speech and Language, Vol. 4, No. 1, 1990, pp. 35-56.

[3] F. Jelinek, J.D. Lafferty and R.L. Mercer, *Basic Methods of Probabilistic Context Free Grammars*, Proc. of NATO ASI, 1990, pp. 345-360.

[4] A. Corazza, R. De Mori, R. Gretter and G. Satta, *Some Recent results on Stochastic Language Modelling*, Proc. of the International Workshop on Structural and Syntactic Pattern Recognition, 1992, pp. 163-183.

[5] M. Oerder and H. Ney, *Word Graphs: an Efficient Interface between Continuous Speech Recognition and Language Understanding*, Proc. ICASSP, 1993, Vol. 2, pp. 119-122.

[6] D. Jouvet, K. Bartkova and J. Monné, *On the Modelization of Allophones in an HMM based Speech Recognition System*, Proc. EUROSPEECH, 1991, pp. 923-926.

[7] A.V. Aho and J.D. Ullman, *The Theory of Parsing, Translation, and Compiling, Vol. 1: Parsing*, Prentice-Hall, 1972.

[8] K. Kita and W.H. Ward, *Incorporating LR parsing into Sphinx*, Proc. ICASSP, 1991, pp. 269-272.

[9] H. Ney, *Dynamic Programming Parsing for Context-Free Grammars in Continuous Speech Recognition*, IEEE Trans. on Signal Processing, Vol. 39, No. 2, 1991, pp. 336-340.

[10] M. Okada, *A Unification-Grammar-directed One-Pass Search Algorithm for Parsing Spoken Language*, Proc. ICASSP, 1991, pp. 721-724.

[11] J. Earley, *An Efficient Context-Free Parsing Algorithm*, Communications of the ACM, Vol. 13, No. 2, 1970, pp. 94-102.

[12] D. H. Younger, *Recognition and Parsing of Context-Free Languages in time* n^3, Information and Control, Vol. 10, No. 2, 1967, pp. 189-208.

[13] W. Ward and S. Young, *Flexible Use of Semantic Constraints in Speech Recognition*, Proc. ICASSP, 1993, Vol. 2, pp. 49-50.

[14] D. Jouvet, A. Lainé, J. Monné & C. Gagnoulet, *Speaker-Independent Spelling Recognition over the Telephone*, Proc. ICASSP, 1993, Vol. 2, pp. 235-238.