A Scalable Syllable Speech Recognizer

C. Rico García[1], O. Schreiner[2] and W. Minker[1]
[1] *University of Ulm, Germany*
[2] *University of Göttingen, Germany / DaimlerChrysler Research and Technology, Ulm, Germany*

Abstract

Modularity is widely accepted as a dear advantage in any system. This is especially true for speech recognition systems, which must be scaled to different task demands. In order to reach this goal, each layer in a speech recognition system can be incorporated as a finite state machine, including the traditional HMM-Layer, lexicon and language model. Here we present a method of construction for a pure finite state speech recognition system. This includes the conversion of traditional HMMs into a finite state transducer and the related problems. The methid is first proved for a lexicon based whole wod recognizer and then extended to a subword unit approach using syllables. The syllables in this approach are phoneticaly motivated in order to minimize cross unit dependencies. Syllable level pruning is used for improved footrin an runtime.

1 Introduction

One of the main constraints in state of the art recognizers is the difficulties given when handling large vocabulary tasks. Natural language is limited to a few thousand words. However, specialized applications manages very large vocabulary size, increasing the number of items the recognizer must deal with to the order of hundred of thousand of words. We investigate a novel recognition strategy in order to deal with large vocabulary tasks. In particular we present an alternative to pronunciation lexicon based speech recognized, and relying on the modularity property associated to the use of finite state machines (FSM), we insert a syllabic recognition level. This allows us to perform a lexicon independent main recognition task.

The use of finite state machines (FSM) [2] in speech decoding has been shown to be an attractive alternative. FSMs generalize the levels involved in speech decoding, and therefore simplifies considerably the decoding strategy and converts the ancient rigid, hard to modify system in a modular one. A set of key operations [5] are performed over the WFSM, they are integrated through the composition operation [1], optimization is performed by means of epsilon removal [3], determinization [2] and minimization [4].

We build and compare two recognizers in which the main recognition task is performed in different levels. In the first recognizer we bring up the idea of modularity introduced by the WFSM's. Here, we perform the "classical" recognition approach based on a pronunciation lexicon dictionary making use of WFST. This transducer contains the acoustic model information, lexical knowledge, and a simple grammar, which could easily be substituted by a more sophisticated one. Clearly, the main recognition task is performed over the lexicon level, outputting the decoded words. The second system results of an increment of the first optimized recognizer with a syllable layer. The newly created recognizer is based on a syllable dictionary, and allows to characterize the lexicon in a phonetically independent manner. At this point we split the recognizer into two parts, the first part performs the main recognition task, and contains the acoustic model information and syllable knowledge, the result of this recognizer are phonemes. The second part is composed by a phoneme described word lexicon and a grammar.

In Section 2, we briefly introduce the parts involved in the two recognizers reporting its results in Section 3. Finally in Section 4 we give future directions and summarize our findings.

2 System Components

Acoustic Model C: The acoustic models implemented as HMMs can be compiled to WFSTs with a similar topology. For every input edge to each state, a set of edges containing the emitted symbols in the target state are created. The input label is the emitted symbol, while the associated weight to each edge is the product of the probability of the emitted symbol and the probability of the transition. The output is epsilon. Every path that traverses the transducer must output a single model unit.

Lexicon transducer L: The lexicon contains a description of the lexica in terms of the units modeled by the HMM, called pronunciation.

The pronunciation can be modelled as the union and closure of trivial linear transducers with one edge for each unit in the sequence. Each edge has as input label the corresponding unit, and has unitary weight, i.e. no weight. One of the edges in the sequence must have the word as output, all the other edges should have epsilon output.

Grammar transducer G: A simple grammar is introduced in the system. It outputs "sentences" containing isolated words and allowing before and after the word all necessary pauses and noises. The grammar transducer G can be easily substituted by a more sophisticated one.

Syllable acceptor: Our recognizer system is incremented with a syllable layer. The aim of this layer is to perform the main recognition task in a limited alphabet, and make the lexicon phonetically independent. The main recognizer is therefore composed by an HMM transducer, a syllable acceptor, and a set of rules mapped to

a transducer giving a phonetically independent output.

$$R_1 = C_1^* \circ min\left(det\left(S\right)\right)^* \circ det\left(P\right)^* \qquad (1)$$

The syllable output of this level can be used with various word lexica afterwards.

The syllable level S accepts certain phones combinations, corresponding each of them to one syllable. This layer groups the phones in the possible syllables, avoiding all the groups that do not produce a syllable.

Phonetic independence rules The goal, is to make a robust primary recognizer, which will constitute a first recognition level. Over this level further recognition improvement can be computed. Therefore our primary recognizer must have a suitable output alphabet. We find the phonemic alphabet more suitable than the phonetic one becuase it is smaller and not context dependant. Thus a transducer P performing this task is inserted in our system.

Decoding Strategy Having implemented the previous components as transducers the complete recognizer can be regarded as a unique transducer integrated by the composition of all the components. The problem of estimating the uttered sequence is reduced to search for the best path on the decoded graph. The decoded graph is obtained by the composition of the data evidence acceptor and the recognizer. A Viterbi search is performed in order to find the best path.

For the experiments to follow, we used a German cities navigation recognition task. The data corpus contains 1950 utterances pre-processed as explained in next Section. The data is uttered by 564 speakers of different conditions. The HMMs model 1693 units, i.e. common words, part of words, thriphones, diphones and monophones. The lexicon contains approximately 1000 city names. It makes use of all the potential the HMMs give and model the words in term of all kind of units mainly with triphones. The syllable level is composed by 8118 syllables which we model only in terms of phones mainly with triphones.

The experiments were performed on a variety of Intel Pentium IV systems running Linux, including four with 2GB RAM.

3 Results

Basic Recognizer In a first experiment, we evaluate the recognition rate given by the system. The recognition for the first system is therefore performed on

$$T_1 = C^* \circ (L^* \circ G) \qquad (2)$$

The system is tested with the data corpus providing it a recognition rate of a 85.685%.

Syllabic Recognizer Since we want to compare the performance of the primary recognizer R_1 1, it was tested with and additional transducer R_2. This transducer diagrammed in figure 1 is composed by a lexicon transducer

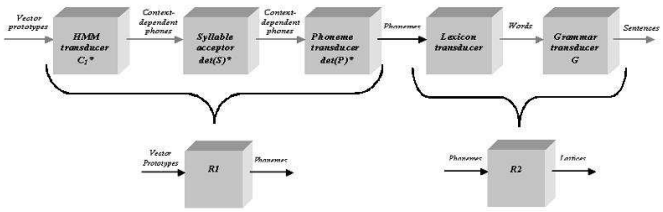

Figure 1: Chain of transducers used in the recognition test of the syllabic recognizer.

which lexica is described in terms of phonemes and the grammar used in the first system.

The results evidenced a slightly lower maximum recognition rate of 79.5270%, due to the loss of cross syllabic information. However we posses now a powerful basic recognizer which does not depend on the vocabulary size.

4 Conclusions

We have presented a syllabic recognizer based on weighted finite state machines, as a way to combat large vocabularies tasks. We introduced a "classical" recognition approach based on a pronunciation lexicon. We presented the problems generated by the use of semi soft vector quantization with a weighted finite state transducers approach, and optimize the system following several strategies and introducing variables that allowed us to optimize the recognition rate. A primary recognizer based on syllabic pronunciation was built. This recognizer allows a flexible substitutions of different recognizers in a second decoding pass. Finally both systems were compared evidencing the perfect suitability of the proposed solution for large vocabulary tasks.

This work was partly funded by the German Ministry of Education and Research (BMBF) in the framework of the SmartWeb project under grant 01 IMD01 D. The responsibility for the content lies with the authors.

References

[1] M. Riley M. Mohri, F. Pereira. Weighted automata in text and speech processing. In *Proceedings of the ECAI 96 Workshop*, 1996.

[2] Mehryar Mohri. Finite-sate transducers in language and speech processing. *Computational Linguistics*, 23:269–311, 1997.

[3] Mehryar Mohri. Generic epsilon-removal and input epsilon-normalization algorithms for weighted transducers. *International Jouranl of Foundations of Computer Science*, 13:129–143, 2002.

[4] Mehryar Mohri. Minimization algorithms for sequential transducers. *Theoretial Computer Science*, 234:177–201, March 2000.

[5] Mehryar Mohri, Fernando C.N. Pereira, and Micheal Riley. General-purpose finite-state machine software tools, http://www.research.attt.com/sw/tools/fsm, vol. at&t labs-research, 1997. http://www.research.attt.com/sw/tools/fsm, vol. AT&T Labs-Research, 1997.