I. Troch, F. Breitenecker, eds. ISBN 978-3-901608-35-3

A Method for Minimization of Pronunciation-lexicon Finite-state Model for Automatic Speech Recognition

Simon Dobrišek, France Mihelič
University of Ljubljana, Slovenia

Corresponding author: Simon Dobrišek
University of Ljubljana, Faculty of Electrical Engineering
SI-1000 Ljubljana, Tržaška 25, Slovenia
Email: simon.dobrisek@fe.uni-lj.si

Abstract. The paper presents work-in-progress on a large-vocabulary automatic speech recognition (LV-ASR) system that is being developed for the Slovenian language. The concept of a single-pass token-passing algorithm for fast speech decoding that can be used with the designed multi-level system structure is discussed. State-of-the-art LV-ASR systems are mostly based on a finite-state network (FSN) models and implementations of such systems require their optimization in terms of their network size and algorithmic complexity. The concept of weighted finite-state transducers (WFSTs) provide an algorithmic framework for such optimizations. Within this framework, LV-ASR systems are normally represented as unified weighted FSNs. This very general concept has many advantages, however it is also not very flexible if one wants to experiment with the mathematical models of LV-ASR sub-components that are not strictly based on the finite-state automata theory. Our system structure does not strictly follow the WFST concept, and consequently we decided to developed some *ad-hoc* methods for system minimisation. The paper present one such method that can be used for minimization of the FSNs that can be used for modelling pronunciation lexicons. The experimental results show that the proposed algorithm achieves the minimization that is at least as good as the one achieved by the algorithms that are proposed within the WFST framework.

1 Introduction

The common mathematical model of the human speech communication process comprises two major components. The low-level component is called the acoustic model and is related to the acoustic realization of human spoken language. This model incorporates general phonetic and psycho-acoustic knowledge about the human speech production and perception system. It is used to model the sequences of acoustic speech observations and is normally defined as a set of sub-word phonetic models. The well-known hidden Markov models (HMMs) [5] and artificial neural networks (ANNs) [1] are usually used for this purpose. The phonetic models are often considered to be context dependent and this dependency is modelled using decision trees and rule-based knowledge representation models [6].

The high-level component of the whole speech model is called the language model. This model comprises several sub-components, like pronunciation lexicons, context-free and/or stochastic grammar models, and semantic models. This part of the joint model is related more to the phonological and linguistic aspects of human spoken language. The language model is commonly represented as a finite-state machine, or more precisely, as deterministic or non-deterministic FSNs [3].

When HMMs are used for the acoustic model, the joint speech model can be represented as a unified multi-level probabilistic FSN (directed graph). This approach reduces the problem of ASR to the problem of searching for the most probable path through such a network, given an acoustic speech observation sequence [4]. The search algorithms are often called speech decoding algorithms.

The performance of LV-ASR systems, which are based on the above modelling concept, critically depends on the size of their FSNs. Due to the limited computational performance of current computer systems, any implementation of a LV-ASR system requires certain optimization of the network in terms of its size and algorithmic complexity. The concept of WFSTs provide a general representation and algorithmic framework for such an optimization that can be achieved using general algorithms for model composition, weighted determinisation and minimization, and weight pushing [4].

Many state-of-the-art LV-ASR systems are based on multi-pass recognition strategy [3]. The first pass is performed using smaller ASR models and simple speech decoding algorithms. The output of the first pass is the most probable word FSN network that is then rescored using more complex models in the subsequent passes. This approach is often used for the off-line LV-ASR.

One of our main research goals is the development of our own state-of-the art LV-ASR system. The main motivation for this is our research commitment to develop such a system for our mother tongue, the Slovenian language. Our language is richly inflected with a grammar that is similar to that of other Slavic languages. It has certain distinctive characteristics, like dual grammatical number, two accentual norms, and abundant inflection. Although

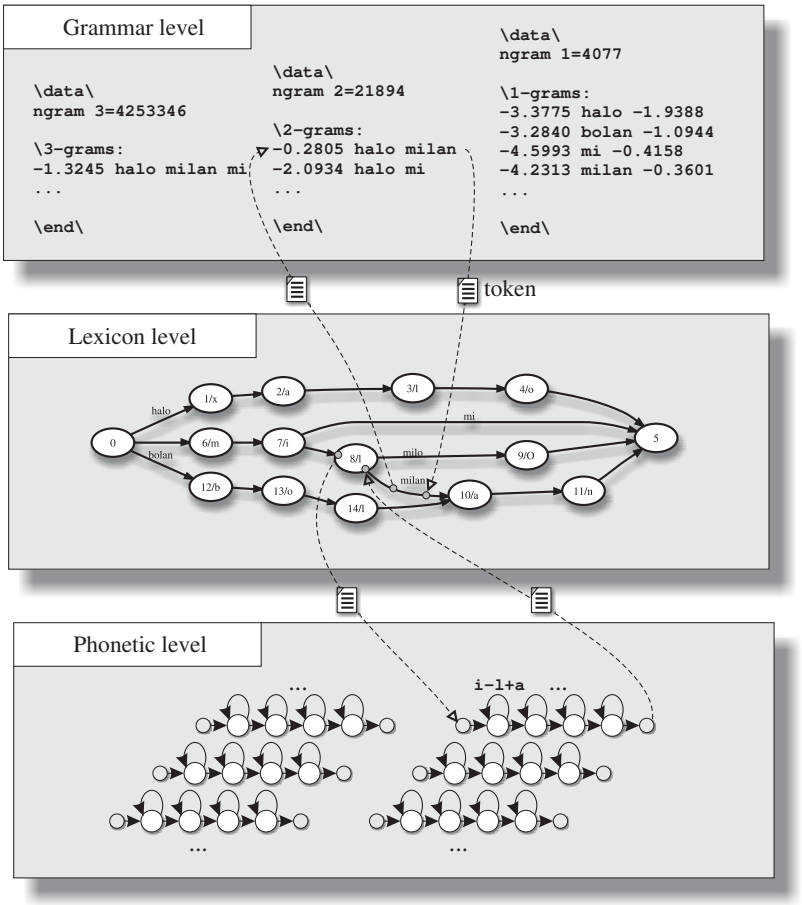

Figure 1: The three-level LV-ASR system structure and a symbolic illustration of a multi-level token-passing speech decoding.

Slovene is basically a subject-verb-object language, word order is very flexible, often adjusted for emphasis or stylistic reasons. These characteristics provide a good reason, why we should not strictly follow the concepts that are used for the less inflected languages with less flexible word order, like English or German. We also have ambition to develop a single-pass LV-ASR system that requires certain adaptations of the system structure.

In order to provide more flexibility, we divided our LV-ASR system into three major components that can be developed and optimized separately. In this paper we briefly present work-in-progress on the system. Especially, we focus on the lexicon level, embedded between the other two levels, as shown in the following section. We also present the concept of a single-pass token-passing algorithm that performs multi-level search for the most probable sequences of words, given an input acoustic speech observation sequence. Finally, we present a simple method for dynamic minimization of a deterministic FSN that is used for the modelling of a pronunciation lexicon.

2 Multi-level LV-ASR System

Our LV-ASR system is designed as a three-level system that follows the general paradigm of spoken language modelling. The three levels are defined as follows:

- The grammar level with context-free and/or stochastic grammar models;
- The lexicon level with a pronunciation-lexicon FSN;
- The phonetic level with a set of context-dependent phonetic models.

From the algorithmic point of view the main level is the lexicon level. In our system pronunciation lexicons are strictly modelled as FSNs. The other two levels are not necessary modelled as such. We believe that this multi-level structure enables more flexible research on new acoustic and grammar models that might be required due to the mentioned distinct characteristics of our language.

The system structure is illustrated in Fig 1. The speech decoding algorithm is based on a simple token-passing paradigm [7]. Tokens are small data structures from which one can reconstruct their path through FSNs. Each token has its own cost value and the costs increase while the tokens pass through the network. Costs values are normally assigned using some pattern matching similarity or probability measures. The token passing mechanism is triggered by the incoming sequence of speech observation vectors. Tokens with the lowest cost are those that

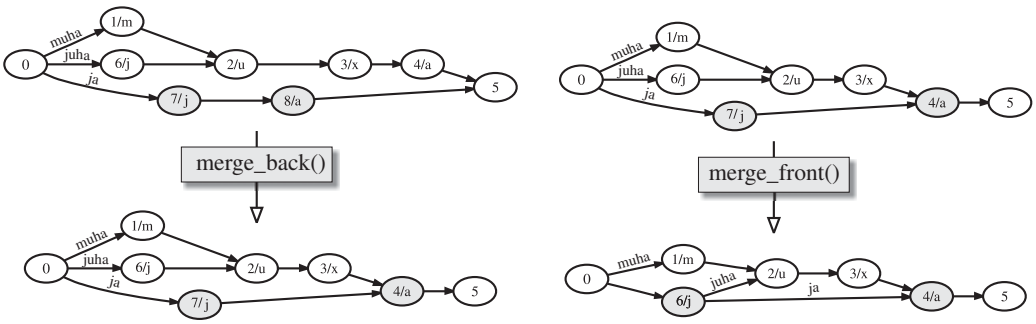

Figure 2: An illustration of the proposed algorithm that optimally merges a new word pronunciation into the exisiting pronunciation-lexicon FSN.

passed the most probable paths through the network, given the incoming observation sequence. LV-ASR systems that are based on this paradigm require also some pruning mechanism that simultaneously eliminates the less perspective tokens that are passing through the network.

The multi-level structure of our system requires a bit more complex token-passing algorithm that was originally proposed by Young et al [7]. For instance, when tokens pass into the acoustic or into the grammar level, they are tagged with the additional back reference that enables their proper return to the main FSN in the lexicon level. This mechanism is illustrated in Fig. 1. When a token passes an allophone node in the main FSN, it enters the phonetic model in the acoustic level that is selected according to the token's phonetic context. This is possible as all the tokens are tagged with their phonetic contexts, i.e. with the information about which node in the main FSN they have already passed. When the token exits the phonetic model, it is returned back to proper position in the main FSN using the back reference that was added to the token.

Tokens are also tagged with their word contexts. This information is used when tokens pass into the grammar level. This is triggered by their pass through the word transition in the main FSN. The main task of the grammar level is to assign (as fast as possible) the cost of passing the particular token through its current word transition in the main FSN considering its previous word context. The grammar model can thus be developed and optimized separately from the lexicon level. The same holds for the acoustic level which task is to assign (also as fast as possible) the costs of passing the particular token through the phonetic model, considering the input speech observation vectors.

As mentioned, LV-ASR systems require optimization of all the three system levels in terms of their size and algorithmic complexity (i.e., speed). In this paper we focus just on the size optimization of the main FSN that is used in the lexicon level.

3 Pronunciation-lexicon FSN Minimization

The size of the pronunciation-lexicon FSN has critical impact to the overall performance of the proposed LV-ASR system. The size of such FSNs can be minimized using the general algorithms that are proposed with the concept of WFSTs. Due to their generality these algorithms are rather complex and require some considerable effort to be implemented. We decided to develop and implement our own optimisation algorithm that is suitable for the presented problem. The proposed algorithm is simple and is based on the concept of merging new words into the pronunciation-lexicon FSN using the backward-forward algorithm. Fig/ 2 illustrates this algorithm on the word "*ja*" with the pronunciation `[j a]` that is merged with the existing pronunciation-lexicon FSN.

The back-merge algorithm iterates backward from the exit node of the FSN and merges nodes of a newly added word pronunciation with the existing nodes in the FSN. This merge is performed if some conditions are satisfied. These conditions are as follows:

- The new and existing node are both predecessors of the current node;
- The existing node is not the entry node of the FSN;
- The existing node has only one successor node;
- The two nodes have the same allophone tag;
- The transition from the existing node to the current node has no word tag.

If the above conditions are satisfied, the new and existing nodes are merged, and the current node becomes the previously merged node. The iteration is repeated while the above conditions are satisfied or the beginning of the newly added pronunciation is reached.

The front-merge algorithm is performed after the back merge one. This algorithm performs merging from the opposite side of the FSN. It also performs word tag pushing, if necessary. The front merge and possible the word label pushing is performed if the following conditions are satisfied:

- The new and existing node are both siblings of the current node;
- The existing node is not the last node that was processed by the back-merge algorithm;
- The existing node has only one predecessor node;
- The two nodes have the same allophone tag;
- The transition from the current node to the existing node has no word tag, or if so, the existing node has only one successor.

If the above conditions are satisfied, the new and existing nodes are merged, and the current node becomes the previously merged node. If the transition from the current node to the merged node had a word tag, it is pushed forward to the transition from the merged node to its single successor node. The iteration is repeated while the above conditions are satisfied.

4 Experimental Results

The above simple FSN optimisation algorithms, denoted as BFMA, proved to be at least as good as the general optimization algorithms that are defined with the concept of WFSTs and are implemented in the AT&T FSM library. In the experiment, we constructed several pronunciation lexicon FSNs from a list of 35k word pronunciations using different algorithms. The results are given in Tab. 1. For the optimization that was achieved using the AT&T

Method	Number of FSN nodes	Number of FSN arcs
NOPT	237206	272650
WFST	22715	58146
BFMA	19332	54747

Table 1: The size of the pronunciation lexicon FSNs achieved by the compared FSN minimization algorithms. The label NOPT denote the FSN that was composed without without any optimization.

FSM library, we had to run a sequence of algorithms. The non-optimal transducer, which was initially generated form the list of word pronunciations, was first determinized and minimized, then the output word symbols were pushed forward and the transducer was again determinized and minimized. On the other hand, the presented algorithm runs in a single backward-forward pass for each added word pronunciation that is optimally merged with the existing pronunciation lexicon FSN.

5 Conclusions

The LV-ASR system for the Slovenian langauge is still being developed and this paper briefly presents our work-on-progress. Many state-of-the-art concepts have already been implemented into the presented multi-level system. For instance, a speaker adaptive training (SAT) method based on the constrained MLLR adaptation was used for building the context-dependent phonetic HMMs that are currently used for the acoustic model [2]. For speaker-independent mid-size vocabulary ASR, we have already achieved the word-recognition rate that exceeds 90 %, and our next goal is to achieve similar real-time word recognition rates for the vocabularies that exceed 100k words.

6 References

[1] Bourlard H., Adali T., Bengio S., Larsen J., and Douglas S. (Eds.): *Neural Networks for Signal Processing*. IEEE Press, 2002.

[2] Dobrišek S., Vesnicer. B, Žganec G. J., Mihelič F.: *Adaptation of Acoustic Feature Space Using Canonical Acoustic Model* In: Proceedings of the 9th International Multiconference Information Society, 2006, Ljubljana, Slovenia, 89 – 92.

[3] Jelinek F.: *Statistical Methods for Speech Recognition*. The MIT Press, 1998.

[4] Mohri M., Pereira F., and Riley M.: *Weighted finite-state transducers in speech recognition*. Computer Speech and Language, 16 (2002), 69–88.

[5] Lawrence Rabiner R., and Juang B. H.: *Fundamentals of Speech Recognition*. Prentice Hall, New Jersey, 1993.

[6] Nock H. J., Gales M. J. F., and Young S. J.: *A comparative study of methods for phonetic decision-tree state clustering*. In: Proceedings of the European Conference on Speech Communication and Technology, 1997, Rhodes, Greece, 111–114.

[7] Young S. J., Russel N. H., and Thornton J. H. S.: *Token passing: A simple conceptual model for connected speech recognition systems*. Cambridge University Engineering Department Technical Report, F/INFENG/TR.38, 1989.