SUB-LEXICAL MODELLING USING A FINITE STATE TRANSDUCER FRAMEWORK[1]

Xiaolong Mou and Victor Zue

Spoken Language Systems Group
MIT Laboratory for Computer Science
200 Technology Square, Cambridge, Massachusetts 02139, USA
{mou,zue}@sls.lcs.mit.edu

ABSTRACT

The finite state transducer (FST) approach [1] has been widely used recently as an effective and flexible framework for speech systems. In this framework, a speech recognizer is represented as the composition of a series of FSTs combining various knowledge sources across sub-lexical and high-level linguistic layers. In this paper, we use this FST framework to explore some sub-lexical modelling approaches, and propose a hybrid model that combines an ANGIE [2] morpho-phonemic model with a lexicon-based phoneme network model. These sub-lexical models are converted to FST representations and can be conveniently composed to build the recognizer. Our preliminary perplexity experiments show that the proposed hybrid model has the advantage of imposing strong constraints to the in-vocabulary words as well as providing detailed sub-lexical syllabification and morphology analysis of the out-of-vocabulary (OOV) words. Thus it has the potential of offering good performance and can better handle the OOV problem in speech recognition.

1. INTRODUCTION

Currently, typical conversational systems are built for specific domains, with a predefined vocabulary for the domain. Usually the recognizer is constrained by a strict lexical network generated from the vocabulary. Each word in the vocabulary is represented by a pronunciation network and these networks are combined into a single lexical network. While such a scheme provides strong sub-lexical constraints for in-vocabulary words, the recognizer usually suffers great performance degradation when the utterances contain OOV words. This problem demands a better balanced sub-lexical modelling approach that can account for both the in-vocabulary and OOV words. In this paper, we will try to integrate sub-lexical morpho-phonemic structure described by the ANGIE [2] system into the recognizer, and compare it with several other models. Similar architecture can also be used to incorporate higher-level semantic knowledge into the recognizer, resulting in a uniform representation across different linguistic hierarchical layers.

In this work, we will mainly focus on the word to phoneme sub-lexical structure mappings. In order to facilitate the construction and exploration of different sub-lexical models, we use the FST recognizer framework. It can integrate acoustic segmentation, application of acoustic models, context-dependent relabelling, application of phonological rules, lexicon, language model and potentially high-level linguistic knowledge etc. into a single weighted FST by composing a series of FSTs. By constructing different sub-lexical model FSTs and composing them with the rest of the FSTs in the recognizer, one can easily build the recognizer with different sub-lexical models.

We have implemented an ANGIE [2] morpho-phonemic model and a novel hybrid model which combines the ANGIE model with a lexicon-based phoneme network model by constructing an FST with an in-vocabulary branch and an ANGIE OOV branch. The same topology for the hybrid model was used in [4] except that the OOV branch is now modeled by ANGIE morpho-phonemic rules. We have compared the ANGIE-based models with some other models including a simple lexicon-based phoneme network model, a phoneme network model with fillers and a phoneme *n*-gram model. We also demonstrated the feasibility of using this flexible FST framework to construct different sub-lexical models.

In the next sections, we will describe the FST framework, and the concepts and implementations of the different sub-lexical models using such a framework. Perplexity results of these models are then given. Finally, conclusions and future work are presented.

2. THE FINITE STATE TRANSDUCER FRAMEWORK

In this section, we will first introduce the FST framework for the complete recognizer, and then elaborate on the FST representation of sub-lexical models.

2.1. Recognizer Architecture

The speech recognizer we use is the MIT SUMMIT [6] segment based recognition system. The recognizer's search space is defined as the following cascade of FSTs:

$$S \circ A \circ C \circ P \circ L \circ G \tag{1}$$

where S is the acoustic segmentation; A is the application of acoustic models, C is the context-dependent relabelling, P represents the phonological rules, L is the lexicon, and G is the language model. The compositions $S \circ A$ and $C \circ P \circ L \circ G$ are usually precomputed and optimized, and the composition of $S \circ A$ with $C \circ P \circ L \circ G$ is computed on-the-fly by the decoder. Thus the decoder only sees a *single* composed $C \circ P \circ L \circ G$ FST, allowing very flexible construction and manipulation of both sub-lexical modelling and language modelling.

[1]This research was supported by a contract from the Industrial Technology Research Institute, and by DARPA under contract N66001-99-1-8904 monitored through Naval Command, Control and Ocean Surveillance Center.

2.2. Word to Phoneme Level Sub-lexical Modelling

To more clearly explain the word to phoneme level sub-lexical modelling, we can further decompose the lexicon FST L described above into the following two FSTs:

$$L = M \circ V \tag{2}$$

where M is the phoneme level sub-lexical model, which defines the phoneme level sub-lexical structures, and V is the vocabulary FST. The phoneme level sub-lexical model could be a simple lexicon-based phoneme network model, a phoneme network model with fillers, a phoneme *n*-gram model, or an ANGIE morpho-phonemic model, for example. No matter how the sub-lexical models are constructed, they will be represented by a single FST M. V is constructed from the recognizer's vocabulary, which maps sequences of phonemes to words. It has two branches, the in-vocabulary branch and the OOV branch, which allow the phoneme to word mapping for any arbitrary phoneme sequences. The weights for these two branches are assigned to reflect an OOV penalty. Throughout the work in this paper, the weight (λ, shown in figure 1) for in-vocabulary branch is chosen to be 0.95, and the weight for OOV branch is assigned to 0.05. These weights define the operation point of the recognizer (false alarm rate and OOV detection rate). Figure 1 gives the topology of the vocabulary FST V.

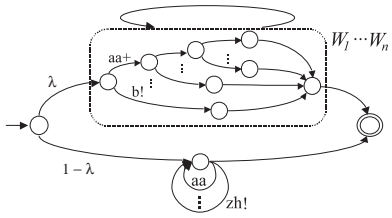

Fig. 1. The topology of the vocabulary FST V.

3. SUB-LEXICAL MODELLING

In this section, we will introduce several different sub-lexical models. As we see, the need to model OOV words is one driving factor in exploring these models. Another key point here is the need to derive sub-lexical structures of the new words. This information is crucial for the effort of automatically incorporating new words into the recognizer.

3.1. Lexicon-based Phoneme Network Model

This is the simplest model, and is the typical model for most domain dependent speech recognition systems with a predefined vocabulary. It provides the strongest restriction on the acceptable phoneme sequences for the recognizer. It performs well if the user uses in-vocabulary words only. However, the recognizer performs significantly worse when OOV words are included.

3.2. Phoneme Network Model With Fillers

In order to allow OOV words in the recognizer, one approach is to use phoneme fillers to model and detect the OOV and partial words, with unique filler path for each phoneme. This is similar to another model for OOV words [4], in which a bigram model is used in the OOV branch. This model accepts any arbitrary phoneme sequence through fillers, which are also used in a standard keyword spotting system. The operation point (the false alarm and OOV detection rate) can be controlled by a penalty for detecting OOV words. It can also provide phoneme hypothesis sequences for OOV words, thus allowing a subsequent post processor to further hypothesize the sub-lexical structure of the OOV word. Although this approach maintains tight constraints over in-vocabulary words, the phoneme fillers do not represent any sub-lexical morpho-phonemic knowledge by themselves, and the constraints for OOV words are generally loose.

3.3. Phoneme *N*-gram Model

Another feasible compromise is to build the sub-lexical models using solely statistical knowledge. For example, we can use phoneme *n*-gram models to model both the in-vocabulary and OOV words. With a large amount of training data, statistical models can capture the underlying sub-lexical morpho-phonemic knowledge by learning the probabilities of different phoneme connections. Compared with the phoneme filler approach, this method can better model the OOV words and partial words, because it learns statistically about the legitimate phoneme sequences, and assigns different probabilities for different connections based on the training data. The disadvantage of this approach compared to the previous filler model is that it also relaxes the constraints for in-vocabulary words at the same time.

3.4. ANGIE Morpho-phonemic Model

In this work, we will focus on the solution where the morpho-phonemic knowledge is encoded explicitly into the sub-lexical models. we use ANGIE hierarchical rules to model the sub-lexical structures of words. ANGIE is a stand-alone application developed in our group, which incorporates multiple sub-lexical linguistic phenomena (including phonology, syllabification and morphology) into a single framework for representing speech and language. It has recently been used to support flexible vocabulary speech understanding [3]. Figure 2 illustrates the word to phoneme part of the ANGIE sub-lexical hierarchy. As we can see, the word "introduce" is comprised of a stressed root, an unstressed root followed by another stressed root. The lower layers show the syllabification and the phonemes. To incorporate ANGIE into our FST framework, we use an FST representation of the morpho-phonemic rules. Thus this sub-lexical model itself knows about the sub-lexical word-to-phoneme hierarchy. The FST representation of morpho-phonemic rules is trained using the same standard FST training tool used to train other types of sub-lexical models. Here we can see the uniform FST framework provides great flexibility of constructing, training and evaluating different sub-lexical models. Compared to the phoneme *n*-gram models, this ANGIE model provides stronger constraints for both in-vocabulary words and OOV words, due to the combination of low level linguistic knowledge and statistical learning from large amounts of training data. For in vocabulary words, this model is still more relaxed than the lexicon-based phoneme network model. However, it is a better balance for in-vocabulary and OOV words.

3.5. Lexicon and ANGIE Hybrid Models

We also investigated a novel idea of combining the ANGIE sub-lexical model with the lexicon-based phoneme network model, which has the potential to maintain the strong constraint provided

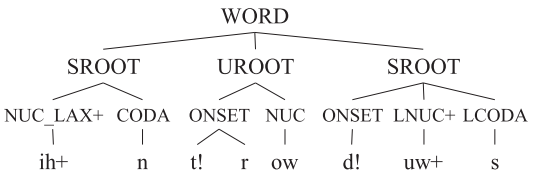

Fig. 2. The ANGIE hierarchical structure of the word "introduce". The bottom layer shows the phoneme labels.

by the lexicon phoneme network, as well as to impose relatively tight constraint over the OOV words. The details of the model implementations are presented in the next section.

4. FST IMPLEMENTATION OF SUB-LEXICAL MODELS

Now we will give the detailed construction of FST M mentioned in equation (2) for different phoneme level sub-lexical models. After constructing M, we can compose it with the rest of the FSTs to build a recognizer conveniently. There may be computational issues, however, because some settings of M may result in non-linear increase in the size of the composed FST. We will then need to compromise the complexity of M accordingly. The FSTs are trained using a straightforward EM algorithm for simple finite state networks, or an inside-outside algorithm for recursive transition networks (RTNs), such as FSTs built from rules written in the form of a context free grammar.

4.1. Lexicon-based Phoneme Network Model

The lexicon-based phoneme network model is equivalent to our current SUMMIT baseline system. Only phoneme sequences that form legitimate in-vocabulary words are allowed. Figure 3 gives the topology of FST M for this model.

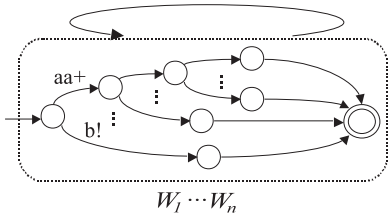

Fig. 3. The topology of FST M for the lexicon-based phoneme network model.

4.2. Phoneme Network Model With Fillers

The phoneme network model with fillers is our first attempt to address the problem of OOV words. It has an OOV branch that contains phoneme loops to accept arbitrary phoneme sequences. This branch is essentially equivalent to a phoneme uni-gram model after training. The other branch is the same as the phoneme network model mentioned above, which accepts in-vocabulary words only. Figure 4 gives FST M's topology for this model. Note that it is similar to the topology of FST V shown in figure 1. The difference is that M is a trained network, and it models the sub-lexical phoneme structures rather than the phoneme-to-word mapping. In practice, they can be directly combined instead of composing.

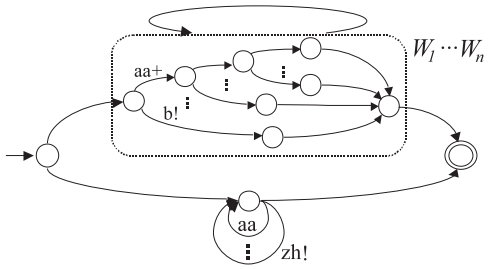

Fig. 4. The topology of FST M for the phoneme network model with fillers.

4.3. Phoneme *N*-gram Model

The phoneme *n*-gram model tries to model both in-vocabulary words and OOV words by learning the phoneme connection probabilities within a short history context. This is the equivalent FST representation of the widely used *n*-gram model. Note that smoothing is represented by proper weighted back-off arcs, which are used to alleviate the sparse data problem. Figure 5 gives an example of the topology of FST M for the phoneme bi-gram model.

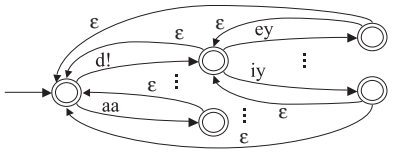

Fig. 5. The topology of FST M for the phoneme *n*-gram model.

4.4. ANGIE Morpho-phonemic Model

The ANGIE mopho-phonemic hierarchy knowledge is written in the form of a context free grammar. However, the underlying language specified by the grammar is actually a regular language in this case. The context free grammar is compiled into an RTN, which is a natural representation for context free grammars. Compiling into RTNs also makes it easier to deal with real context free languages when necessary. The compiled RTNs are then trained, and composed with other FSTs of the recognizer. Figure 6 illustrates this model.

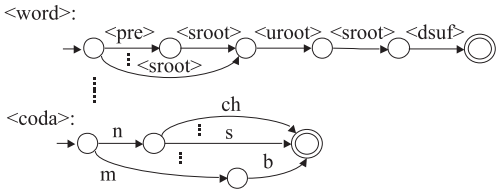

Fig. 6. The topology of FST(RTN) M for the ANGIE morpho-phonemic model.

4.5. Lexicon and ANGIE Hybrid Models

We also present here a novel approach of combining the lexicon-based phoneme network model with the ANGIE morpho-phonemic model. Since the lexicon model has the strongest constraint over

in-vocabulary words, and ANGIE model can better handle OOV words, we construct an in-vocabulary-only branch using the lexicon model and an OOV branch using the ANGIE model. This is similar to the phoneme filler model setting mentioned above, expect that the OOV words are now modeled by ANGIE morpho-phonemic rules rather than the phoneme uni-gram fillers. Figure 7 shows this model's topology.

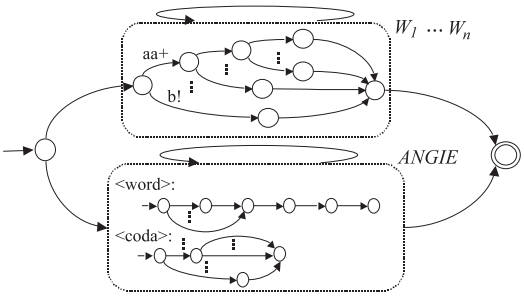

Fig. 7. The topology of FST(partially RTN) M for the lexicon and ANGIE hybrid model.

5. CORPUS AND EXPERIMENTAL RESULTS

The sub-lexical models described above are trained and evaluated in the JUPITER [5] English weather domain. The training set consists of phoneme transcriptions of 99,062 utterances, and the independent test set consists of phoneme transcriptions of 2,443 utterances, of which 2,297 utterances do not contain OOV words.

We have evaluated the sub-lexical models in terms of perplexity results on the full test set and its subset which contains only in-vocabulary words. The perplexity numbers are obtained using an FST-based tool which essentially composes the input phoneme strings with different sub-lexical FSTs, searches the most likely path, and then computes the average log probability per phoneme. Table1 shows the results.

Sub-lexical Models	*Perplexity on Test Set with OOV Words*	*Perplexity on Test Set without OOV Words*
Lexicon-based Phoneme Network	∞	2.638
Phoneme Network with Fillers	9.344	2.643
Phoneme bi-gram	6.334	5.955
ANGIE Morpho-phonemic	3.733	3.654
Lexicon and ANGIE hybrid	3.602	2.843

Table 1. Perplexity results of different sub-lexical models on the full test set and its in-vocabulary-only subset.

From the results we see that the lexicon-based phoneme network model has the lowest perplexity number on the in-vocabulary-only test set. However, it fails to model any OOV word. Thus, on the test set containing OOV words, its perplexity is infinite. The phoneme network with fillers model has a high perplexity on the OOV test set, due to its inadequate ability to make use of sub-lexical structural information. The phoneme bi-gram models can model both in-vocabulary and OOV words, but it has a significantly higher perplexity on the in-vocabulary test set than the previous two models. Our ANGIE model successfully reduced the perplexity on both test sets with or without OOV words. Finally the proposed lexicon and ANGIE hybrid model is able to combine the benefits and has a better overall perplexity result.

6. CONCLUSIONS AND FUTURE WORK

This work described in this paper shows the feasibility of incorporating ANGIE sub-lexical linguistic knowledge into speech recognition using the FST framework. The advantages of using ANGIE sub-lexical linguistic knowledge include better constraint over OOV words and the ability to analyze the sub-lexical hierarchy of OOV words, which is absent for phoneme fill or phoneme *n*-gram models. This ability is quite useful in many ways. For example, it can lead to easy hypotheses of new word spellings according the sub-lexical analysis, and help automatically incorporate new words. We also see that the lexicon and ANGIE hybrid model has an overall better performance than other settings.

In this paper, we showed preliminary perplexity results for the proposed ANGIE-based sub-lexical models. Future work include the evaluation of their speech recognition performance, along with their receiver operating characteristics.

It is also very interesting to explore the use of similar FST architectures at higher levels of the language processing hierarchy. For example, we can try to incorporate some natural language processing procedures directly into the recognizer, rather than interfacing the speech recognizer and a separate natural language processing module with an *N*-best list. This results in a tightly coupled speech recognition and natural language processing system, where the high-level linguistic knowledge is incorporated at very early stages of speech recognition. However, since the linguistic phenomena at higher levels are much more complicated than at the sub-lexical level, proper adaptations may be necessary for using the FST framework.

Acknowledgments Lee Hetherington offered great help on the FST tools used in this work. Grace Chung and Issam Bazzi kindly shared their experiences on ANGIE and OOV modeling, respectively. Stephanie Seneff also provided many insights on the initial system construction.

7. REFERENCES

[1] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," *Proc. of ISCA ASR2000*, Paris, 2000.

[2] S. Seneff, R. Lau, and H. Meng, "ANGIE: A new framework for speech analysis based on morpho-phonological modelling," *Proc. of ICSLP*, Philadelphia, 1996.

[3] G. Chung, "A three-stage solution for flexible vocabulary speech understanding," *Proc. of ICSLP*, Beijing, 2000.

[4] I. Bazzi and J. Glass, "Modeling out-of-vocabulary words for robust speech recognition," *Proc. of ICSLP*, Beijing, 2000.

[5] V. Zue, *et al*, "JUPITER: A telephone-based conversational interface for weather information," *IEEE Trans. on Speech and Audio Processing*, vol. 8, no. 1, January 2000.

[6] J. Glass, J. Chang, and M. McCandless, "A probabilistic framework for feature-based speech recognition," *Proc. of ICSLP*, Philadelphia, 1996.