Context-dependent Probabilistic Hierarchical Sub-lexical Modelling Using Finite State Transducers[1]

Xiaolong Mou, Stephanie Seneff, Victor Zue

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

Abstract

This paper describes a unified architecture for integrating sub-lexical models with speech recognition, and a layered framework for context-dependent probabilistic hierarchical sub-lexical modelling. Previous work [1, 2, 3] has demonstrated the effectiveness of sub-lexical modelling using a core context-free grammar (CFG) augmented with context-dependent probabilistic models. Our major motivation for designing a unified architecture is to provide a framework such that probabilistic sub-lexical components can be integrated with other speech recognition components without sacrificing the flexibilities of their independent developments and configurations. At the same time, we are able to obtain a tightly coupled interface between recognizers and sub-lexical linguistic components. We also present a view of using layered probabilistic models to augment CFGs. It captures context-dependent probabilistic information beyond the standard CFG formalism, and provides the flexibility of developing suitable probabilistic models independently for each sub-lexical layer. Experimental results show that the context-dependent probabilistic hierarchical sub-lexical modelling approach can achieve comparable performance to pronunciation network approaches on utterances that contain only in-vocabulary words, while being able to substantially reduce errors on utterances with previously unseen words.

1. Introduction

The goal of sub-lexical modelling in speech recognition is to model the construction of words robustly using sub-lexical units, which are supported by acoustic models of a speech recognizer. For example, when acoustic models are established at the phone level, commonly used sub-lexical modelling approaches include explicit modelling using pronunciation networks and implicit modelling using Hidden Markov Models (HMMs). In these cases, the mapping from phones to words is mainly based on local context-dependent constraints such as phonological rules. Further studies have revealed more detailed hierarchical linguistic structures of sub-lexical units, and researchers have recently demonstrated that incorporating hierarchical sub-lexical linguistic information can improve the robustness and flexibility of sub-lexical support for speech recognition [1, 2, 3].

One important problem that influences the effective use of sub-lexical linguistic constraints is the lack of a general framework to express and apply hierarchical linguistic knowledge. Most speech systems use separately implemented sub-lexical linguistic components, which are specifically designed for some particular recognizer architecture, and are difficult to re-configure or extend to suit a wider variety of application requirements. Another factor in using such sub-lexical linguistic knowledge is the integration interface with a speech recognizer. Possible interfaces include *N*-best lists and phone graphs produced by a recognizer. Further linguistic constraints are applied to the *N*-best lists or phone graphs at a subsequent stage. While this approach can reduce the complexity at later stages, early integration into the recognition search phase may be desirable because the search can then be guided from the available hierarchical sub-lexical linguistic constraints. Incorrect hypotheses could be pruned early in a tightly coupled interface.

In this paper, we present a unified architecture based on finite state transducers (FSTs), which seamlessly integrates context-dependent probabilistic hierarchical sub-lexical modelling with speech recognition. FSTs have been widely used recently [4] as an effective and flexible framework for speech recognition. In this formalism, a speech recognizer is represented as the composition of a series of FSTs combining various knowledge sources across the linguistic hierarchy. In our approach, we construct the hierarchical sub-lexical model by composing a CFG parser and a series of probabilistic FSTs, each of which is designed to capture the layer-specific context-dependent probabilities. Such context-dependent information is beyond the standard CFG formalism. The resulting FST is then composed with other FSTs in the recognizer, such as the acoustic modelling and language modelling FSTs. The final search space is a unified network that encodes knowledge at both sub-lexical and language levels. This unified architecture allows independent developments of recognition and linguistic components, while maintaining tight integration between the two.

2. Sub-lexical modelling using probabilistic hierarchical approaches

Most modern speech recognizers use sub-lexical units for acoustic modelling to increase the recognizer's flexibility and alleviate the sparse data problem. When units smaller than words are used as basic units for speech recognition, it is necessary to establish a mapping from sub-lexical units to words, since words are natural units for applying higher level linguistic knowledge. One approach is to use pronunciation networks when basic acoustic models are based on phones. The pronunciation network is usually generated by converting a dictionary to a baseform network; then context-dependent phonological rules

[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.

To be presented at the 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, September, 2001

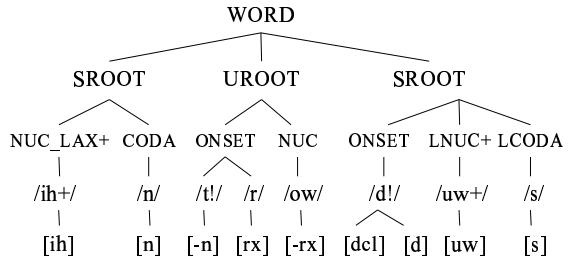

Figure 1: *The* ANGIE *hierarchical structure of the word "introduce". The bottom layer shows the phone realizations. [-n] is a phone deletion following by an [n] phone.*

are applied to the baseform network to allow different phone realizations. The phonological variations can also be modelled implicitly in an HMM acoustic model.

Another approach [1, 2, 3] demonstrated by researchers to be effective is to use a core context-free grammar to describe the sub-lexical hierarchical structure of words and augment the CFG with context-dependent probabilities to capture different phonological variations. While the pronunciation network is constructed from explicit rules for different local contexts, the probabilistic CFG approach relies on a large amount of training data to learn the context-dependent knowledge, especially at lower levels of the sub-lexical hierarchy. At the same time, it preserves strong structural constraints at higher levels. Moreover, it has the potential to provide detailed sub-lexical structural analysis for words not limited to a specific vocabulary.

2.1. Sub-lexical linguistic hierarchy

Generic structural constraints for sub-lexical units can be obtained by defining a CFG which encodes sub-lexical grammatical information. The constraint is tighter at higher levels of the sub-lexical hierarchy. At lower levels, a CFG usually does not provide effective constraints by itself because realization variations tend to be more context-dependent.

Figure 1 shows an example parse tree of sub-lexical units using the ANGIE [1] sub-lexical rules. ANGIE is an hierarchical framework developed in our group, which models multiple sub-lexical linguistic phenomena, including phonetics, phonemics, syllabification and morphology. The hierarchical representation has a regular, layered structure. The categories in the grammar are grouped into separate sets for each of the layers. For example, the morphology layer contains categories such as stressed and unstressed roots (SROOTs and UROOTs), and the syllabification layer contains categories of onsets and nucleus (ONSETs and NUCs), etc. Stress and onset markers(+ and !) are preserved in the phonemic layer, since both of them can provide additional information for phonological effects at the phonetic layer. Such a regulated grammar is able to provide generic word structural constraints, while maintaining simplicity.

2.2. Context-free rule-production probabilistic models

We have conducted some research to examine the structural constraints above the phoneme layer provided by hierarchical sub-lexical models using context-free grammars [5]. Context-free rule-production probabilities are also used to augment the grammars. We found that context-free grammars with rule-production probabilities are effective in describing the generic word structures above the phonemic level. However, it is not sufficient to account for context-dependent information, especially at lower sub-lexical levels.

2.3. Context-dependent probabilistic hierarchical models

It is important to realize that at lower levels of the sub-lexical hierarchy, especially at the phonetic level, the phonological realization of phones is highly context-dependent. For example, the word "keep" is phonemically /kip/, and the /k/ is aspirated (followed by a burst of air). In contrast, the word "ski" is phonemically /ski/ but the /k/ is not aspirated, because the /k/ is preceded by an /s/. In order to model such context sensitive knowledge, researchers have proposed probabilistic frameworks with CFGs that account for context information through training. In our work, we applied the probabilistic models of ANGIE. There are two types of probabilities in ANGIE: the phone advancement probability, which is the probability of the current phone (terminal) given the path from the top node to the previous phone in the parse tree (called a column in [1]); and the trigram bottom-up probability, which is the probability of an internal parse tree node (non-terminal) given its first child and left sibling. A back-off mechanism is also used to alleviate the sparse training data problem. The probability of a parse tree is given by the product of the column probabilities conditioned on its left column, assuming that they are independent. The conditional probability of a column is further defined as the product of the phone advancement probability and trigram bottom-up probabilities of the internal nodes in the column. For example, the conditional probability of the second column in Figure 1 is defined as:

$$\begin{aligned} P\,(\mathrm{C}_2 \mid \mathrm{C}_1) = {} & P\,(\mathrm{[n]} \mid \mathrm{[ih]},\ \mathrm{/ih+/},\ \mathrm{nuc\_lax+},\ \mathrm{sroot},\ \mathrm{word}) \\ & * P\,(\mathrm{/n/} \mid \mathrm{[n]},\ \mathrm{/ih+/}) \\ & * P\,(\mathrm{coda} \mid \mathrm{/n/},\ \mathrm{nuc\_lax+}) \end{aligned} \quad (1)$$

3. FST based context-dependent probabilistic hierarchical modelling

In this work, we propose a unified architecture based on FSTs that seamlessly integrates context-dependent probabilistic hierarchical sub-lexical modelling with speech recognition. The hierarchical sub-lexical model is designed by composing a Recursive Transition Network (RTN) parser for CFGs and a series of context-dependent probabilistic FSTs.

3.1. Motivation and related work

The major motivation for designing a unified architecture is to provide a framework for integrating probabilistic sub-lexical components with other speech recognition components without sacrificing the flexibilities of their independent developments and configurations. At the same time, we are able to obtain a tightly coupled interface between a recognizer and sub-lexical linguistic components. Using probabilistic hierarchical sub-lexical models also provides linguistic support for previously unseen words through structural and probabilistic information generalized from training data. Furthermore, it can provide detailed sub-lexical analysis obtained from the parse tree for further use with higher level components.

Previous research toward integrating sub-lexical linguistic knowledge into speech recognition includes using a separate sub-lexical parser by Lau [2], and a three-state recognition architecture by Chung [3]. In Chung's work, ANGIE based sub-lexical modelling is used in the first and second stage. The ANGIE sub-lexical model is also represented by an FST, and the FST is constructed by enumerating all parsing columns occurred in training, then connecting them using precomputed ANGIE probabilities. This approach is effective and is able to

capture a large portion of the sub-lexical probabilistic space when phonetic observations are well supported by training data. A full probability space without limiting the generalization power of probabilistic sub-lexical modelling on previously unseen words may be desirable when sub-lexical parse trees contain columns not connected or not seen in the training data.

3.2. FST based sub-lexical models using a layered approach

It is important to notice that, while an input phone sequence can be parsed by an RTN according to the corresponding CFG, it is difficult to apply the context dependent probabilities directly in the RTN. When traversing an RTN, a stack is usually used to remember the returning state when entering a sub-network of a non-terminal. No sibling context information is remembered. One possible solution is to build a state machine with a second stack to record the context information. In this work, however, we adopt a layered approach to capture the context dependent probabilities. One convenient factor is that the CFG we use is a regulated one and the parse tree has a fixed number of layers. The RTN constructed from the sub-lexical CFG will output a tagged parse string first. Then a series of probabilistic FSTs are applied, each of which is designed to capture the context-dependent probabilities for a particular layer, and filter out the irrelevant parse tags. This design provides a view of using layered probabilistic models to augment CFGs, and facilitates independently choosing a suitable probabilistic model for each layer. In this work, we use trigram bottom-up probabilities for intermediate layers, and the phone advancement probabilities for the phone layer, as described in section 2.3. Other models are also possible to use with this framework.

3.2.1. The skip phone and parsing FSTs

For an input phone sequence, a skip phone FST S illustrated in Figure 2 is applied first. It encodes the left context of possible deleted phones. The skip phone FST is then composed with an RTN R, illustrated in Figure 3, constructed from the sub-lexical CFG. It outputs the tagged parse string representing the parse tree. For example, the first "SROOT" sub-tree in Figure 1 is represented by "<SROOT> <NUC_LAX+> <ih+> ih </ih+> </NUC_LAX+> <CODA> <n> n </n> </CODA> </SROOT>." This tagged parse string is used to apply context-dependent probabilistic models for each parse tree layer.

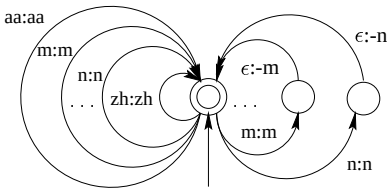

Figure 2: *The skip phone FST diagram. An arc labeled "ϵ:-n" represents that a phone can be deleted if following an [n] phone. In this case, it outputs a "-n" marker for sub-lexical parsing.*

3.2.2. The intermediate trigram probabilistic layers

In our work, the probabilities of intermediate layers are modeled by trigram bottom-up probabilities as illustrated in Equation 1. The probability of the parent is conditioned on its left sibling and first child. The FST L_i is designed to capture this context-dependent probability for the intermediate layer i above the phone level. It takes in the tagged parse string, ignores irrelevant parse tags through filter arcs, and applies trigram probabilities of layer i. Figure 4 shows the diagram of the state transitions from the current parent P and its left sibling L to its right sibling R and the parent P. The probability Prob (P | L, K) is applied during the transitions, where K is the first child of P. It could be further simplified given that the categories we use are grouped into separate sets for each layer.

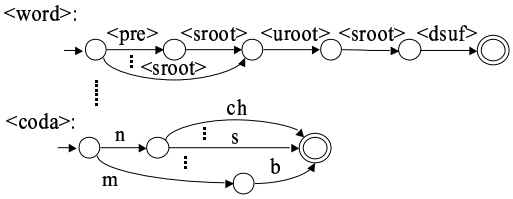

Figure 3: *The RTN diagram for the sub-lexical CFG parser. It parses the input phone sequence and gives a tagged parse string representing the parse tree.*

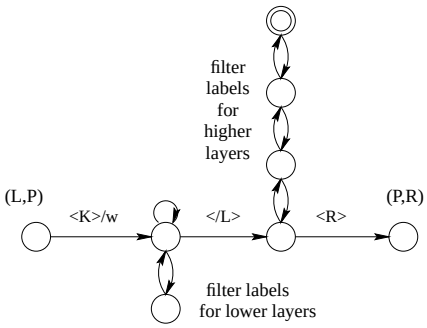

Figure 4: *The state diagram in an intermediate layer probabilistic FST. It shows the transitions from a state (L,P) to state (P,R), where P, L, and R are the current parent, its left sibling and its right sibling in the parse tree, respectively. "w" is the probability Prob (P | L, K), where K is the first child of P.*

3.2.3. The phone advancement probabilistic layer

The probabilities of the phone layer are defined by the phone advancement probabilities also illustrated in Equation 1. The probability of the next phone is conditioned on its left column. The phone layer FST P encodes such phone transition probabilities. Context-dependent phone bi-gram probabilities across word boundaries are also applied. Figure 5 shows the state transitions from the left column to the right column. Back-off states are added to apply phone advancement probabilities for unseen columns during recognition.

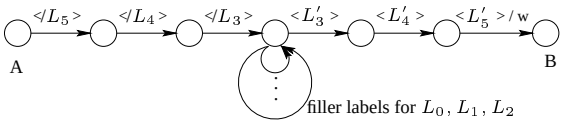

Figure 5: *The state diagram in a phone advancement probabilistic FST. It shows the transition from the left column A ($[L_0, L_1, L_2, L_3, L_4, L_5]$) to the right column B ($[L_0, L_1, L_2, L'_3, L'_4, L'_5]$). L_0 is the top layer category. "w" is the probability of the right phone (L'_5) given the left column.*

3.2.4. Definition of the complete sub-lexical model

Given the skip phone FST, parsing RTN and the probabilistic models described above, the complete context-dependent probabilistic hierarchical sub-lexical model is then defined by the following FST M:

$$M = S \circ R \circ L_1 \circ L_2 \circ \ldots \circ L_{N-1} \circ P \quad (2)$$

where N is the number of parse tree layers, S is the skip phone FST, R is the tagged parsing RTN, L_1 through L_{N-1} are the probabilistic FSTs for intermediate layers above the phone level, and P is the phone advancement probabilistic FST.

M is precomputed and optimized. It maps a phone sequence to a sub-lexical parse tree with context-dependent probabilistic support. We can choose the output of M for further processing. In our work, we output the phonemic layer categories which can be used for further mapping to morphs and words by recognizers.

3.3. Integrating sub-lexical models with speech recognizers

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:

$$S \circ A \circ C \circ M \circ L \circ G \quad (3)$$

where S is the acoustic segmentation; A is the acoustic model, C is the context-dependent relabelling, M is the sub-lexical model defined by Equation 2, L is the lexicon, and G is the language model. Such a unified architecture allows seamless integration of the sub-lexical models with speech recognizers, while preserving great flexibility for each component.

4. Experimental results

The sub-lexical models described above are trained and evaluated in the JUPITER [7] English weather domain. The training set consists of 98,121 utterances, and the independent test set consists of 2,321 utterances, of which 1,631 utterances contain in-vocabulary words only.

Two probabilistic hierarchical models are constructed for evaluation. The first one uses a sub-lexical CFG up to the word level, and the second one uses a simpler CFG up to the syllable level, which yields a smaller sub-lexical FST M. Since the context-dependent probabilistic hierarchical sub-lexical models have the potential of providing both structural and probabilistic constraints for previously unseen words, the recognizer can be configured to output an unknown word tag for a phoneme sequence that is not in the vocabulary. We have evaluated the sub-lexical models in terms of phone perplexity and recognition word error rate (WER) on the full test set and its subset which contains only in-vocabulary words. The baseline system is a pronunciation network sub-lexical model with phone bigram probabilistic support, and it does not detect unseen words. Table 1 shows the phone perplexity and the recognition WER results. Note that we map all unseen words to an unknown word tag when calculating WERs for the hierarchical models. We see that the word-level context-dependent probabilistic hierarchical sub-lexical model has a comparable performance to the pronunciation network sub-lexical model on utterances that contain only in-vocabulary words, while being able to substantially reduce the errors on the full test set. The word-level model yields a better performance than the syllable-level model, since it has stronger constraints at higher levels.

Sub-lexical Models	*Full Test Set*		*In-vocabulary Test Set*	
	Perp.	*WER*	*Perp.*	*WER*
Baseline	12.3	23.3 %	6.2	12.0 %
Word-level Sub-lexical CFG	6.6	17.2 %	5.1	13.1 %
Syllable-level Sub-lexical CFG	7.2	20.5 %	6.3	14.6 %

Table 1: *Phone perplexity and WER results of different sub-lexical models on the full test set and its in-vocabulary subset.*

5. Conclusions and future work

The work presented in this paper demonstrates the feasibility of constructing probabilistic hierarchical sub-lexical models in a novel layered framework and integrating such a sub-lexical model with speech recognition in a unified architecture. Furthermore, the proposed sub-lexical models are also able to perform detailed sub-lexical analysis, and output relevant information for use with components in a higher linguistic hierarchy.

The probabilistic models for each layer are not restricted to the trigram and phone advancement models used in this work. Future work includes evaluations on the probabilistic models for each sub-lexical layer, and improvements through better modelling of layer-specific information. It is convenient to conduct such work under this unified architecture. We are also interested in incorporating such a layered framework at higher linguistic levels. However, the grammars used at language levels are usually much more complicated compared to the regulated sub-lexical grammars. An alternative approach is to directly construct state transducers that are able to remember the context-dependent information. Although such a device may not be finite state, it can be composed dynamically during recognition.

Acknowledgments Lee Hetherington offered great help on the FST libraries which made this work possible. Grace Chung and Issam Bazzi kindly shared their experiences on ANGIE and out-of-vocabulary modeling.

6. References

[1] Seneff, S., Lau, R., and Meng, H. "ANGIE: A new framework for speech analysis based on morpho-phonological modelling", Proc. ICSLP'96, Philadelphia, PA, 1996.

[2] Lau, R. and Seneff, S., "A unified framework for sub-lexical and linguistic modelling supporting flexible vocabulary speech understanding", Proc. ICSLP'98, Sydney, Australia, 1998.

[3] Chung, G., "A three-stage solution for flexible vocabulary speech understanding", Proc. ICSLP'00, Beijing, China, 2000.

[4] Mohri, M., Pereira, F., and Riley, M., "Weighted finite-state transducers in speech recognition," Proc. of ISCAASR'00, Paris, 2000.

[5] Mou, X. and Zue, V., "Sub-lexical modelling using a finite state transducer framework", To appear in Proc. ICASSP'01, Salt Lake City, Utah, 2001.

[6] Glass, J., Chang, J. and McCandless, M., "A probabilistic framework for feature-based speech recognition", Proc. of ICSLP'96, Philadelphia, PA, 1996.

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