A UNIFIED FRAMEWORK FOR SUBLEXICAL AND LINGUISTIC MODELLING SUPPORTING FLEXIBLE VOCABULARY SPEECH UNDERSTANDING[1]

Raymond Lau and Stephanie Seneff

Spoken Language Systems Group
MIT Laboratory for Computer Science
Cambridge, Massachusetts 02139 USA
http://www.sls.lcs.mit.edu, email:{raylau, seneff}@mit.edu

ABSTRACT

In [9], we introduced the ANGIE framework for modelling speech where morphological and phonological substructures of words are jointly characterized by a context-free grammar and represented in a multi-layered hierarchical structure. In [6], we demonstrated a competitive word-spotter based on the ANGIE framework and presented several results comparing the performance of various sublexical filler models. In the present work, completed as a part of [5], we extend the ANGIE framework to a competitive full continuous speech recognition system. Furthermore, given that ANGIE is based on a context-free framework, we have decided to combine ANGIE with TINA ([8]), a context-free based framework for natural language understanding, into an integrated system. The integrated system led to a 21.7% reduction in word error rate compared to a baseline word bigram recognizer on ATIS. Numerous issues relating to the construction of the combined system were explored. We have also examined the addition of new words to the recognizer vocabulary, one of the areas which we believe will benefit from the ANGIE framework and also from the ANGIE-plus-TINA integration. Our combined system achieved an error rate reduction of 20.8% over the baseline system and outperformed several other configurations we tested not involving an integrated ANGIE-plus-TINA.

1. INTRODUCTION

Many spoken language systems employ a higher level language understanding component in addition to a speech recognition component. The interface between the two components is best characterized as a feed forward only process, with either an N-best list (the top N full sentence hypotheses from the recognizer) or a word graph (a graph representation of the top scoring hypotheses from the recognizer, as in [10]) being passed from the recognizer to the understanding component. The understanding component then either rescores the hypotheses or chooses the highest scoring one that parses. Little progress has been made in terms of feeding knowledge in the reverse direction, from the understanding component to the recognition component. An understanding component is needed to obtain useful results from a spoken language system, where honoring the user's request rather than recognition is the aim. However, early attempts at leveraging the understanding component for better recognition have met with only limited success (e.g., [7], [11]). As a result, the recognizer consumes time pursuing hypotheses which may clearly be eliminated by the understanding component, perhaps even at the expense of pruning away more promising, from the understanding component's point of view, hypotheses.

Our ANGIE subword model, discussed in greater detail in the next section, is based on an underlying context-free framework. Context-free grammars also underly numerous natural language understanding systems, including the TINA system from MIT ([8]). The goal of the work discussed in this paper is to explore whether our subword framework can be integrated with a natural language understanding system more tightly, so that knowledge feeds in both directions, allowing the NL system to help filter unpromising hypotheses early.

We also believe that the combination of ANGIE and TINA should yield a system to which new words can be easily added without requiring extensive sublexical or linguistic retraining. ANGIE's shared hierarchical subword model provides a framework whereby subword structural information can be shared between words in the recognizer vocabulary and new words to be added to the vocabulary.

2. ANGIE AND TINA FRAMEWORKS

ANGIE is a framework for subword lexical modelling which we introduced in [9] and which is discussed more fully in [5]. In ANGIE, word substructure is characterized by a set of context-free rules and a set of trained probabilities. The context-free rules are written by hand and generate a very regular, layered, hierarchical structure, as illustrated by the example parse shown in Figure 1. The subword structure is represented by four layers beneath the WORD node. The layers are, from bottom to top, phonetics, phonemics, syllabification and morphology. Stress markings are distributed through several layers, so for example, SROOT stands for "stressed root" and /ih+/ stands for "stressed /ih/." The rules governing the phonemics to phonetics layer are particularly noteworthy because they govern which phonological processes are permitted. Typically, such rules are captured in a context dependent manner, but since ANGIE uses context-free rules, any context dependency will be captured by our choice of rules and nonterminals along with the trained probability model.

The ANGIE probability model consists of two types of probabilities, computed based on a bottom-up, left-to-right parse: *advancement probabilities* and *trigram bottom-up probabilities*. The former are the conditional probabilities of a leaf node in the parse tree given its immediate left column, where a column is defined as the nodes along the path from the root to a leaf. The trigram bottom-up probabilities are the conditional probabilities of an internal node given its left sibling and its child. The full

[1]This material is based upon work supported by the National Science Foundation under Grant No. IRI-9618731.

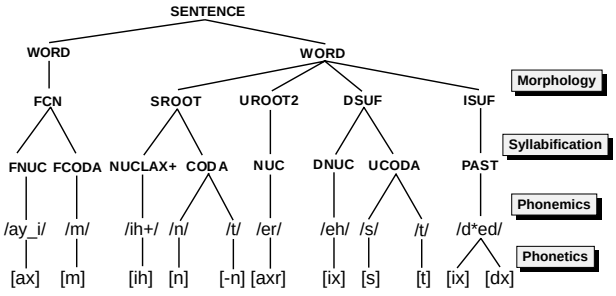

Figure 1: Sample parse tree for the phrase "I'm interested."

column probability is the sum of the log advancement probability and the log trigram bottom-up probabilities for the nodes up to the point where the current column merges with its left column. The linguistic score for an entire parse is the sum of all the column log probabilities.

Our ANGIE probabilities are trained on approximately 10,000 utterances from the ATIS corpus ([3]) using forced alignments originally obtained from our SUMMIT ATIS recognizer ([13]) and subsequently iterated within the ANGIE forced recognition system as described in our earlier paper ([9]). ANGIE has a phone perplexity of 7.15 on test data as compared to 14.91 for a phone bigram and 9.20 for a phone trigram.

We believe ANGIE offers several advantages for speech recognition tasks. Because of the hierarchical structure, different words which share common word substructures will share common subtrees in an ANGIE parse. This permits pooling of training examples across all words with a given substructure. Further, as mentioned earlier, we believe that ANGIE permits the easy addition of new words to the vocabulary by sharing subword structural information between existing in-vocabulary and new words. In principle, ANGIE also provides a subword model for the detection of new words. The bottom-up nature of ANGIE should facilitate the latter application. Finally, knowledge of the subword structure provided by an ANGIE parse also permits us to use the information for prosodic modelling, as in [1].

For our natural language integration work, we used MIT's TINA system, which is describe in greater detail in [8]. The TINA NL processing system shares many similarities with ANGIE. It also makes use of a context-free grammar, designed by hand, and a probability model trained automatically from data. TINA has two other features worth mentioning. TINA uses constraints to enforce feature matching, such as number and verb tense agreement, and also to handle gaps, which occur fairly frequently in English wh-queries. TINA has a robust parsing mechanism to handle sentences which are not completely well formed due to either poor user verbalization or recognition errors.

Unlike ANGIE's bottom-up parsing strategy, TINA uses a top-down methodology. The practical ramification for our task is that we cannot implement both ANGIE and TINA in a single parser. Instead, we will need to integrate the two parsing strategies during our search process. We considered converting TINA to a bottom-up strategy, but rejected the option due to the difficulty of supporting gap phenomena when going bottom up. We also considered using a top-down strategy for ANGIE, but we felt that subword structures are of an inherently bottom-up organization and we wanted to retain the bottom-up sharing, for example, of syllables, across words.

3. RECOGNITION WITH ANGIE

In previous conference presentations, we had reported on the success of implementing phonetic recognition ([9]) and word-spotting ([6]) systems based on the ANGIE framework. Before discussing our integration work, we will briefly report results at implementing a basic continuous speech recognition system using ANGIE. Our implementation of a continuous speech recognizer uses a stack-decoder strategy similar to what we used in our word-spotter. For word-level statistics, we incorporate a word bigram score when our search algorithm reaches a word boundary. We compared our recognizer using context-independent acoutic models trained on a 5000 utterance subset of ATIS with a similarly configured baseline system using the MIT SUMMIT recognizer. On the December 1993 test set, the ANGIE system achived an 18.8% error rate, comparable to the baseline's 18.9%. Further details on our recognizer implementation can be found in [5].

4. INTEGRATING SUBLEXICAL AND LINGUISTIC MODELLING

Recall from our previous discussion that our sublexical ANGIE model is primarily of a bottom-up design whereas our supralexical TINA linguistic model is of a top-down design. A natural organizational point to combine these two models into a single search is at the lexical level. Our stack decoder consults the ANGIE subword model as it attempts to construct a word hypothesis bottom-up. At each putative word ending, the decoder consults the TINA NL component to obtain a score for extending the sentence hypothesis with the proposed word. The decoder then combines the scores from the two sources and the search proceeds. A graphical illustration of this process can be found in Figure 2. Although our organization is not as tightly integrated compared to combining the sublexical and linguistic modelling into a single parser, it still provides a much quicker feedback cycle from the NL component to the recognition. In particular, the NL component is consulted at the end of each putative word as opposed to at the end of each utterance, as would be the case with either word graphs or N-best resorting.

Our integration efforts encountered one serious difficulty. The robust parsing mechanism within TINA proved computationally expensive. This mechanism implements an effect similar to the following top level rule:[2]:

$$\text{sentence} \Rightarrow \text{skip}^* \; [\text{full\_parse}] \; (\text{skip} \mid \text{partial\_parse})^*$$

Naturally, the expense of hypothesizing the insertion of skip words at all possible points in the sentence leads to a combinatorially explosive search space. Our solution is to remove the robust parsing mechanism from TINA and implement our own greedy strategy in the decoder. Our strategy is as follows: For a particular hypothesis, parse as many words as possible, allowing

[2]Here, the * refers to zero or more, the | refers to alternatives, and [] means optional.

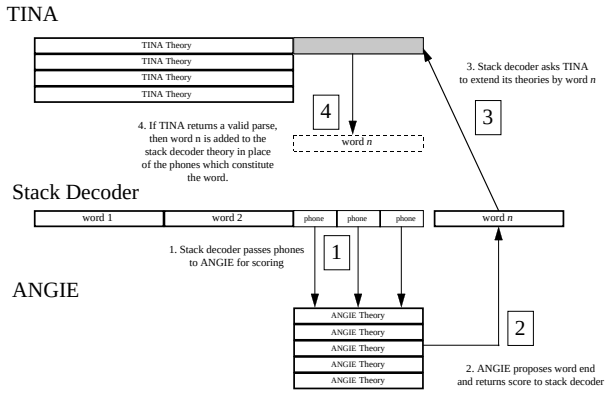

Figure 2: Integration of ANGIE and TINA into the stack decoder search process.

Recognizer	Total	Sub	Del	Ins
SUMMIT w/Word Bigram	18.9	11.7	4.9	2.3
ANGIE w/Word Bigram	18.8	11.7	4.2	2.9
SUMMIT w/Word Bigram and TINA 100-best Resorting	18.2	10.8	5.5	1.9
ANGIE-plus-TINA Integrated	14.8	8.7	4.5	1.6

Table 1: Comparison of word error rate percentages for different recognizers.

only full parses. When the NL parser fails, it retraces backwards until it finds a point where the NL grammar permits a sentence end, and it starts a new parse at that point. This strategy is obviously greedy, in that we restrict the locations where a parse breaking point is inserted to the point of least backward retracement following a parse failure. However, our greedy strategy ran two orders of magnitude faster than the original robust parse mechanism, and resulted in a tractable system.

A comparison of our integrated ANGIE-plus-TINA recognizer and several baseline recognizers is shown in Table 1. As can be seen from the table, the integrated ANGIE-plus-TINA system results in a 21.7% error rate reduction from the SUMMIT word bigram system. Moreover, we see that the tight integration also results in an 18.7% error rate reduction from the TINA 100-best resorting system, illustrating the value of the tight integration. In the resorting system, we linearly interpolated TINA and SUMMIT scores, using a set of optimized weights.

An examination of the error rate reductions show that much of the improvement is in substitutions. An examination of the top substitution errors corrected by the integrated ANGIE-plus-TINA system suggests a large reduction of number agreement errors, as in the confusion between "flights" and "flight." Many of the remaining top substitution errors in the ANGIE-plus-TINA system are ones which are gramatically correct in our rules, for example, "New York" being substituted for "Newark" and "a" being substituted for "an."[3]

[3]We do not enforce a feature agreement constraint for "a" and "an" in our TINA rules.

5. ADDING WORDS TO VOCABULARY

Since introducing the ANGIE framework in [9], we have been suggesting that one of the advantages of the framework is the potential to support flexible vocabulary changes, such as the addition of new words to the vocabulary. We believe that ANGIE's ability to share word substructures between existing and newly added words will provide better lexical support for the added words. Further, with the integration of an NL processing system, the combined system should have better linguistic support than, for instance, a class n-gram.

For our new word study, we envision the following scenario. The recognizer is part of a conversational system. The system may retrieve information from a database in response to a user query. For example, the user may inquire about flights to California and the system may retrieve a list of cities there. At this point, we would want to increase the recognizer's vocabulary to include all the cities. The salient features of this class of scenarios are that we know the category of the new words to be added to the system vocabulary, the addition must be done online without extensive lexical retraining, and the number of words to be added is small relative to the size of the total vocabulary. The ATIS data set consists of ATIS-2 and ATIS-3 subsets with 34 additional city names present in the ATIS-3 subset. We chose to use these additional city names as new words to be added in our experiment.

Because we want to be able to compare systems, with and without the new city names, which are identical except for the availability of lexical and linguistic training for the new city names, we adopted the following experimental methodology. We start with a recognizer trained on the full training data, with the new city names. Then we artificially removed the benefits of the additional lexical training by setting the lexical arc weights of the new city names to zero in the SUMMIT baseline case and by not using those examples to update the probability model in the case of ANGIE. This allows both systems to be trained on exactly the same set of acoustic data. Our approach is similar to that used in the new word work in [4]. For the linguistic training, we allow both the class bigram baseline and TINA to see the new city names, but as unknown cities.

For our experiment, we compared adding the new city names to three systems: a baseline SUMMIT recognizer with a pronunciation graph for sublexical modelling and a class bigram for linguistic modelling, an ANGIE recognizer with a class bigram, and an ANGIE-plus-TINA recognizer. Adding new words involves primarily three steps. Adding the baseforms, adding sublexical support and adding linguistic support. We assume that we know the correct baseform for all the new city names in an effort to limit our study to the effects of the sublexical and linguistic models. In an actual system, a dictionary, or a letter-to-sound system, perhaps using the ANGIE framework, can be used. To provide sublexical support, in the SUMMIT pronunciation graph case, we add the baseforms, expanded by phonological rules, to the graph with zero lexical arc weights, which corresponds to neutral weights. In ANGIE, we allow the parser to share probabilities and structures with existing vocabulary words. However, we discovered that there were several cases where ANGIE assigned zero probabilities to some of the structures in the new city names because such structures did not occur in training data. These structures were licensed by the context-free rules, but had

	SUMMIT	ANGIE	ANGIE-plus-TINA
Reduced	34.2%	31.2%	32.8%
Augmented	19.2%	19.2%	15.2%
Full	18.9%	18.8%	14.8%

Table 2: Error rates of different systems in the presence of simulated new word additions to the active vocabulary.

no support from the probability model. There are several potential solutions. The one we pursued was to operate ANGIE in phoneme-to-phone generation mode, which generated likely phone sequences based on probabilities for similar structures in training data, and then use the resulting phone sequences as supplemental training data to prime the probability model, eliminating the zero probabilities. This can be done very quickly, as the number of new words to be added is assumed small. For linguistic support, we added the new words to the city name category in both the class bigram and the TINA grammar, assigning probabilities within the category uniformly over both previously existing and newly added words.

The results of our new word experiment are summarized in Table 2. In this table, "full" vocabulary refers to a system trained with knowledge of all city names, "reduced" refers to the artificially reduced lexical and linguistic training described earlier, and "augmented" refers to the simulated addition of the new city names to the reduced vocabulary configurations. We make several observations from the table. Most disappointing is that ANGIE does not surpass the baseline SUMMIT pronunciation graph if we consider the augmented vocabulary configuration. However, a closer evaluation shows that a possible explanation for this is that the loss of lexical training data impacted performance only slightly, increasing error rate from 18.9% to 19.2% and 18.8% to 19.2% in the SUMMIT and ANGIE cases, respectively. This observation is consistent with that made in [4] on the same set of simulated new word additions. However, other authors in the literature have noticed more dramatic detriments from lack of lexical training with other choices of word additions (e.g., [12]). We conclude that more work is needed to conclusively determine how ANGIE compares with a pronunciation graph when new words are added. Another point to observe is that with the reduced configurations, ANGIE outperforms the pronunciation graph, suggesting that ANGIE is better able to deal with a large number of unknown words. The ANGIE-plus-TINA configuration shows a large benefit in the augmented vocabulary setups as compared to SUMMIT or ANGIE without TINA. We examined more closely the error rate increase for ANGIE-plus-TINA going from full vocabulary to augmented vocabulary (14.8% to 15.2%) by considering an augmented ANGIE configuration with a fully-trained TINA configuration and vice-versa. In both cases, we achieve an error rate of 15.0%, suggesting that the degradations due to limited TINA training and due to limited ANGIE training are roughly equal.

6. SUMMARY AND FUTURE WORK

In this paper, we described the successful integration of our ANGIE sublexical modelling framework with the TINA natural language processing system. The combination resulted in a 21.7% error rate reduction as compared to a baseline system with a word bigram. We also explored the addition of new words to both an ANGIE-based recognizer and a combined ANGIE-plus-TINA system. In the ANGIE only case, the results were comparable to that of the baseline SUMMIT pronunciation graph. The combined system performed better than the other tested augmented vocabulary configurations.

The development of the ANGIE framework is an ongoing process. In terms of future work, we intend to pursue use of our integrated ANGIE-plus-TINA system in actual conversational domains, instead of the ATIS data we used for the reported study. (Some of this work is reported by our colleague in [2]). Also, we would like to switch to context-dependent acoustic models, now that the basic recognition infrastructure is operational.

7. REFERENCES

1. G. Chung and S. Seneff, "Hierarchical duration modelling for speech recognition using the ANGIE framework," in *Proc. Eurospeech '97*, Rhodes, Greece, pp. 1475–1478, Sept. 1997.
2. G. Chung and S. Seneff, "Improvements in speech understanding accuracy through the integration of hierarchical linguistic, prosodic, and phonological constraints in the Jupiter domain," in *Proc. ICSLP '98*, Sydney, Australia, Nov. 1998. These Proceedings.
3. D. A. Dahl, M. Bates, M. Brown, W. Fisher, K. Hunicke-Smith, D. Pallett, C. Pao, A. Rudnicky, and E. Shriberg, "Expanding the scope of the ATIS task: The ATIS-3 corpus," in *Proc. ARPA Human Language Technology Workshop '92*, Plainsboro, NJ, pp. 45–50, Mar. 1992.
4. I. L. Hetherington, *The Problem of New, Out-of-Vocabulary Words in Spoken Language Systems*. Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, Oct. 1994.
5. R. Lau, *Subword Lexical Modelling for Speech Recognition*. Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, May 1998. URL http://www.raylau.com/thesis/thesis.pdf.
6. R. Lau and S. Seneff, "Providing sublexical constraints for word spotting within the angie framework," in *Proc. Eurospeech '97*, Rhodes, Greece, pp. 263–266, Sept. 1997. URL http://www.raylau.com/angie/eurospeech97/main.pdf.
7. R. Moore, M. Cohen, V. Abrash, D. Appelt, H. Bratt, J. Butzberger, L. Cherny, J. Dowding, H. Franco, J. M. Gawron, and D. Moran, "SRI's recent progress on the ATIS task," in *Proc. Spoken Language Systems Technology Workshop '94*, Plainsboro, NJ, pp. 72–75, Mar 1994.
8. S. Seneff, "TINA: A natural language system for spoken language applications," *Computational Linguistics*, vol. 18, no. 1, pp. 61–86, Mar. 1992.
9. S. Seneff, R. Lau, and H. Meng, "ANGIE: A new framework for speech analysis based on morpho-phonological modelling," in *Proc. ICSLP '96*, Philadelphia, PA, vol. 1, pp. 110–113, Oct. 1996. URL http://www.raylau.com/icslp96_angie.pdf.
10. S. Seneff, M. McCandless, and V. Zue, "Integrating natural language into the word graph search for simultaneous speech recognition and understanding," in *Proc. Eurospeech '95*, Madrid, Spain, pp. 1781–1784, Sept. 1995.
11. V. Zue, J. Glass, D. Goodine, H. Leung, M. Phillips, J. Polifroni, and S. Seneff, "Integration of speech recognition and natural language processing in the MIT VOYAGER system," in *Proc. ICASSP '91*, Toronto, Canada, pp. 713–716, May 1991.
12. V. Zue, J. Glass, D. Goodine, M. Phillips, and S. Seneff, "The SUMMIT speech recognition system: Phonological modelling and lexical access," in *Proc. ICASSP '90*, Alburquerque, NM, pp. 49–52, Apr. 1990.
13. V. Zue, S. Seneff, J. Polifroni, M. Phillips, C. Pao, D. Goddeau, J. Glass, and E. Brill, "The MIT ATIS system: December 1993 progress report," in *Proc. ARPA Spoken Language Technology Workshop '94*, Plainsboro, NJ, pp. 66–71, Mar. 1994.