PROVIDING SUBLEXICAL CONSTRAINTS FOR WORD SPOTTING WITHIN THE ANGIE FRAMEWORK[1]

Raymond Lau and Stephanie Seneff

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

ABSTRACT

We describe our recent work in implementing a word-spotting system based on the ANGIE framework and the effects of varying the nature of the sublexical constraints placed upon the word-spotter's filler model. ANGIE is a framework for modelling speech where the morphological and phonological substructures of words are jointly characterized by a context-free grammar and are represented in a multi-layered hierarchical structure. In this representation, the upper layers capture syllabification, morphology, and stress, the preterminal layer represents phonemics, and the bottom terminal categories are the phones. ANGIE provides a flexible framework where we can explore the effects of sublexical constraints within a word-spotting environment. Our experiments with spotting city names in ATIS validate the intuition that increasing the constraints present in the model improves performance, from 85.3 FOM for phone bigram to 89.3 FOM for a word lexicon. They also empirically strengthens our belief that ANGIE provides a feasible framework for various speech recognition tasks, of which word-spotting is one.

1. INTRODUCTION

Previously, we proposed a new methodology for incorporating multiple subword linguistic phenomena, including phonology, syllabification, and morphology, into a single framework, which we named ANGIE, for representing speech and language ([7]). At that time, we had suggested that ANGIE would be a suitable lexical model for a variety of tasks. We had demonstrated its feasibility for bidirectional letter-to-sound/sound-to-letter generation, forced phonemic-to-phonetic alignment, and phonetic recognition. We had mentioned word-spotting and flexible speech recognition as natural tasks to tackle next. In the present work, we discuss our implementation of a word-spotter employing the ANGIE framework, including empirical evidence of feasibility and search related issues.

We also examine the effect on performance of altering the nature of the subword lexical constraints imposed upon the filler or trash model. Work by others has shown that increasing the detail of acoustic modelling and of cross-word lexical constraints tends to improve word-spotting performance, given adequate training data, to the extent that full continuous speech recognition yields the best word spotter (e.g., [3], [5], [6]). We naturally believe that a similar conclusion holds for subword lexical modelling as well. Since the ANGIE framework (discussed in detail in the next section) makes it easy to vary the types of subword lexical constraints imposed, we explore their impact in the present work.

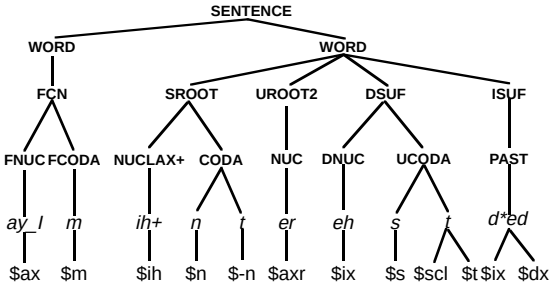

Figure 1: Sample parse tree for the phrase "I'm interested." From top to bottom, the layers shown are: sentence, word, morphology, syllabification, phonemics, and phonetics.

2. ANGIE

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. The bottom terminal layer can also be letters in the case of sound-to-letter/letter-to-sound generation. 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, described later.

The nonterminals in our grammar are carefully chosen so that lexical constraints are propagated downwards in the parse tree. We currently have 100 "phoneme" nonterminals which include stressed and unstressed versions for the vowels, onset and non-onset realizations for the

[1]This research was supported by DARPA under contract N66001-94-C-6040, monitored through Naval Command, Control and Ocean Surveillance Center, and by a fellowship from the National Science Foundation.

consonants, morph-specific phonemes (e.g., *d*ed* for the past tense morpheme "ed"), some function word specific entries (e.g., *dh_the*), and some pseudo-diphthongs (e.g., *aol*). The remaining nonterminals are generic units such as UROOT and CODA. However, we do include approximately twenty special inflexional suffices (e.g., "ism"). The terminals include sixty-four phones along with several deletable units (e.g., to handle gemination), which are marked for their left context.

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 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 ([2]) using forced alignments originally obtained from our SUMMIT ATIS recognizer ([10]) and subsequently iterated within the ANGIE forced recognition system as described in our earlier paper ([7]). 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, we feel that ANGIE permits the easy addition of new words to the vocabulary, because subword probabilities can be shared with existing words. In principle, ANGIE also provides a subword model for the *detection* of out-of-vocabulary words. The bottom-up nature of ANGIE should also facilitate the latter application. Finally, knowledge of the subword structure provided by an ANGIE parse also permits us to potentially improve our acoustic modelling. Later, we will discuss some related work involving hierarchical duration modelling and its impact on word-spotting performance.

3. ANGIE BASED WORD-SPOTTER

We previously reported in [7] on the feasibility of the ANGIE framework in two intermediate speech recognition tasks: phonemic-to-phonetic alignment and phonetic recognition. Word-spotting provides a natural next task to tackle. There is a natural continuum from word-spotting to flexible speech recognition in that in the former case, we are interested in a small set of keywords whereas in the latter case, we want to grow the set of interesting words to a full size vocabulary. Also, there is a natural analog between the filler word in a word-spotter and a new, out-of-vocabulary word in a speech recognizer that can detect the occurrence of new words. The ANGIE framework also allows for experimentation with different subword lexical models for the filler word.

3.1. Acoustic Models and Segmentation

Our word-spotter uses the same segment-based framework as SUMMIT ([9]). For the acoustic features, we compute average MFCCs for the left, middle and right third of each segment and delta MFCCs at the start and end boundaries along with segment duration. We use 10 mixture diagonal Gaussians for each phone.

3.2. Search Strategy

Each utterance is processed in a left to right search. For each partial hypothesis, we consider extending it by adding a phone to the end. We compute the new score to include the previous score, plus the acoustic score for the new phone, plus the change in the linguistic score, computed with ANGIE, for the new path. For the ANGIE score, we consider the set of possible parses for the given phone sequence, subject to pruning, and take the highest scoring parse as the representative score. To control the complexity of the ANGIE parser, we simplify the ANGIE framework at word boundaries by reducing the advancement probability to only depend on the left phone context. Replacing with a generic phone bigram start probability allows the set of hypothesized paths to be collapsed to include only the partial ANGIE parses for the current word and to merge theories which share the same word sequence and same ending phone at word boundaries. Besides reducing the search complexity greatly, this also mitigates sparse data problems which tend to occur across word boundaries.

So far, the system we have presented resembles very much the phonetic recognizer we used in [7]. However, several critical changes were needed. The best-first search strategy used by our phonetic recognizer proved to be unworkable in the context of a word-spotter. We have found that the issues involved in normalizing short vs. long theories so that their scores can be reasonably compared are extremely complex. We experimented with several normalization and pruning schemes to no avail. Instead, we decided to change our search control strategy altogether. We settled upon a variant of the *stack decoder* suggested in [4]. Our search operates as follows:

1. Initialize the stack with a null theory.
2. Of the shortest (in terms of ending time) theories, pop the best (highest scoring) theory off the stack.
3. If the theory has reached the end of the utterance, return it as a putative theory for the utterance. Stop if we have the number of theories desired.
4. Else, consider all possible phone extensions.
5. And, for each extension that meets minimum scoring thresholds, insert them into the stack. Prune the stack if needed.
6. Go to 2.

A key feature of this variant of the stack decoder lies in the way we sort the stack in step 2 and the way we do

the stack pruning in step 5. For stack pruning, we consider only up to a constant, n, number of theories that end at a given time boundary and keep the n highest scoring theories. The key feature is that theories which compete against each other for survival in our search all end at the same time boundary, i.e., they cover the same time span in the utterance. This achieves a *de facto* normalization of the theory scores, at least for the acoustic territory covered in the utterance. Short theories are never compared to long theories (where length is in terms of time). There is still the issue of normalizing for the number of hypothesized linguistic elements, but the difficulties posed by a low scoring segment are significantly mitigated.

3.3. Generating Keyword Hypotheses

We use ANGIE to model both the keywords and the filler. The precise manners of modelling the filler will be described in the next section. To generate keyword hypotheses, we allow both the keywords and the fillers to compete in the search. Keywords have additional boosts to encourage their selection over the filler. We take the best path, and output all the keywords hypothesized, along with their time alignments and scores. We compute the keyword score simply as the difference between the path score at the end of the keyword and at the beginning. Posterior scoring is not used at this time.

4. CORPUS AND EVALUATION

For acoustic training, we use a 5000 utterance subset of ATIS-3. For testing, we use the Dec '93 test set. Our word-spotting task is 39 city names. Given the keyword hypotheses generated, as described in the previous section, we plot a *receiver operating characteristics (ROC)* curve and compute a *figure of merit (FOM)* according to the procedure prescribed in [8].

5. VARYING THE FILLER

We have run a series of experiments with our ANGIE word-spotter where we varied the subword modelling of the filler. In all cases, the full ANGIE model was used to model the keywords. The subword models we explore, include:

Phone Bigram A phone bigram model constrains the probability distribution of phones in the filler space. This serves as a baseline.

ANGIE Pseudo-words ANGIE is trained with full knowledge of the ATIS lexicon, but during word-spotting, the word constraints for the non-keywords are removed. ANGIE proposes possible subword structures for the filler space and governs where these "pseudo-words" can end. We permit nonsensical syllables to occur although their scores will be low because they are not well supported by the ANGIE training data. We also allow nonsensical combinations of syllables; however, an additional constraint on syllable order is imposed so that, for example, prefix syllables are only permitted to occur in the prefix position of pseudo-words, etc.

ANGIE Syllables ANGIE proposes possible syllables to cover the filler space. Only syllables in a predefined lexicon are permitted. We also simplify the prediction probability across syllable boundaries to a left phone context advancement probability, as for normal word boundaries. As a result, nonsensical combinations of syllables are permitted without any restrictions on their ordering and cross-syllable trigram constraints are lost.

ANGIE Morph Constraints This variant is a combination of the constraints in the pseudo-words and syllables cases. We permit a set of allowable syllables (also referred to later as morphs) to generate pseudo-words for the filler space. The full ANGIE probability model is imposed within each pseudo-word. We still permit nonsensical combinations of morphs but we include the ordering constraint as in the pseudo-words case.

ANGIE Word Constraints with Unknowns Here, the filler space is modelled as a series of words from a vocabulary of approximately 1200 words. The words are modelled with the full ANGIE model. We also allow for "unknown" words which are modelled by ANGIE as in the morph constraints case. No cross-word language model is used other than the phone bigram advancement probability across word boundaries. No word unigram is used other than boosts for the keywords.

ANGIE Word Constraints w/o Unknowns As in the previous case, but "unknown" words are not permitted.

The above list is given in the order of increasing subword lexical constraints being imposed upon the fillers permitted. The exception is the pseudo-words and syllables cases, which are roughly similar in terms of level of constraints provided, although they provide a different set of constraints. We have avoided cross-word constraints to focus on the effects of subword models.

Filler Model	FOM	Speed Factor
Phone Bigram	*85.3*	-
Pseudo-Words	86.3	1.00
Syllables	87.7	1.78
Morph Constraints	88.4	1.26
Word Constraints w/Unk	88.6	1.29
Word Constraints w/o Unk	89.3	1.34
Full Recognition w/Word Bigram	*93.9*	-

Table 1: Word-spotting FOMs for various filler models. Higher speed factors indicate faster running times. Word-level statistics are excluded from all but the last system.

5.1. Experimental Results

We summarize the results of our word-spotting experiments in Table 1 and Figure 2. We also include an entry with the results of using the full SUMMIT recognizer with word bigrams for comparison purposes.

The general trend is one of increasing performance with increasingly constrained filler models. However, the same cannot be said for speed. The speed factor column in Table 1 is normalized so that the pseudo-words experiment has a speed of one. (We have omitted the speed for the phone bigram case because that system can be implemented much more efficiently via a straight forward Viterbi search.) We

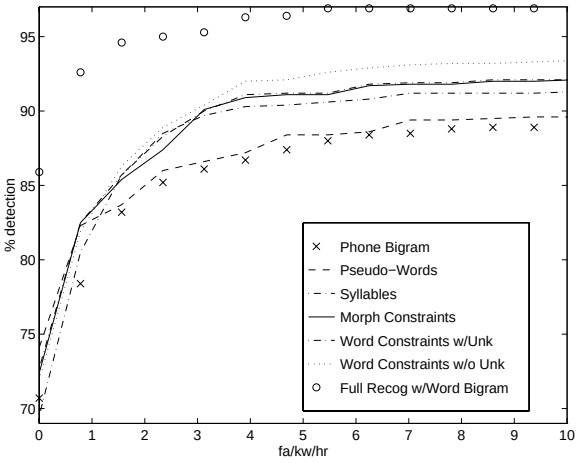

Figure 2: ROCs for various filler models.

note that the tradeoff between speed and performance is inverted. The lowest performing system happens to be the slowest. We believe this to be the case because increasing constraints also narrows down the search space that needs to be explored. Another interesting observation is the incredible speed of the syllables system. Because that system only has left phone context advancement constraints across syllable boundaries, that system permits significant sharing of theories at syllable boundaries, resulting in a favorable speed/performance operating point. Finally, we note that eliminating the possibility of unknown words in the systems with word constraints results in better performance. Our test set includes 0.57% out-of-vocabulary words. Perhaps a smaller vocabulary with lower coverage of the test set may result in a benefit from including the unknown words; however, preliminary experiments suggest little change in performance down to a vocabulary of the 400 most frequent words (in training).

6. DURATION MODELLING

Since we are using ANGIE to model the subword structure of both keywords and fillers, we also consider adding a hierarchical duration model, from Chung ([1]), beyond the simple phone duration included in the acoustic feature set. When we add hierarchical duration scores to the keywords, our performance improves between 1.4 and 2.3 FOM, depending on the subword constraint set. We also try including it for non-keywords, but performance suffers in that case. We suspect that Chung's model does not work well on the "pseudo-words" being proposed for the filler space, although other factors may be involved. The duration results are summarized in Table 2. Performance speeds degrade only slightly with Chung's model added.

7. CONCLUSION

Our experimental results validate the intuitive hypothesis that the more subword lexical constraints we impose in the model for the filler, the better the performance of the word-spotter. We find that the ANGIE syllable system provides an interesting operating point in terms of performance per unit speed. We also show that the ANGIE framework provides a viable subword model for word-spotting. Combined with our previous results in phonetic recognition ([7]), we believe that ANGIE is suitable for a variety of speech recognition tasks. ANGIE has several advantages, particularly arising out of the sharing of subword structures and also the enabling of more detail models, such as Chung's duration model, based on a detailed subword structural representation. The addition of Chung's duration model brings our word constraints system very close in performance to full continuous speech recognition. Our natural next step is to progress towards a full, flexible speech recognition system based on the ANGIE framework.

Filler Model	**FOM w/o Dur**	**FOM w/Dur**
Morph Constraints	88.4	89.8
Word Constraints w/Unk	88.6	90.0
Word Constraints w/o Unk	89.3	91.6

Table 2: Word-spotting FOMs with the addition of Chung's duration model for the keywords.

8. REFERENCES

[1] G. Chung and S. Seneff, "Hierarchical duration modeling for speech recognition using the ANGIE framework," in *Proc. Eurospeech '97*, Rhodes, Greece, Sept. 1997. These proceedings.

[2] 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.

[3] A. S. Manos, "A study on out-of-vocabulary word modelling for a segment-based keyword spotting system," Master's thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, Apr. 1996.

[4] D. B. Paul, "Efficient A^* stack decoder algorithm for continuous speech recognition with a stochastic language model," Tech. Rep. TR 930, MIT Lincoln Laboratory, July 1991.

[5] J. Rohlicek, P. Jeanrenaud, K. Ng, H. Gish, B. Musicus, and M. Shiu, "Phonetic training and language modeling for word spotting," in *Proc. ICASSP '93*, Minneapolis, MN, pp. II–459–II–462, Apr. 1993.

[6] R. C. Rose, "Definition of subword acoustic units for wordspotting," in *Proc. Eurospeech '93*, Berlin, Germany, pp. 1049–1052, Sept. 1993.

[7] 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.sls.lcs.mit.edu/raylau/icslp96_angie.pdf.

[8] "The road rally word-spotting corpora (RDRALLY1)," Sept. 1991. NIST Speech Disc 6-1.1.

[9] V. Zue, J. Glass, M. Phillips, and S. Seneff, "The MIT SUMMIT speech recognition system: A progress report," in *Proc. DARPA Speech and Natural Language Workshop Feb '89*, Philadelphia, PA, pp. 179–189, Feb. 1989.

[10] 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.