ANGIE: A NEW FRAMEWORK FOR SPEECH ANALYSIS BASED ON MORPHO-PHONOLOGICAL MODELLING[1]

Stephanie Seneff, Raymond Lau, and Helen Meng

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
http://www.sls.lcs.mit.edu

ABSTRACT

This paper describes a new system for speech analysis, ANGIE, which characterizes word substructure in terms of a trainable grammar. ANGIE capture morpho-phonemic and phonological phenomena through a hierarchical framework. The terminal categories can be alternately letters or phone units, yielding a reversible letter-to-sound/sound-to-letter system. In conjunction with a segment network and acoustic phone models, the system can produce phonemic-to-phonetic alignments for speech waveforms. For speech recognition, ANGIE uses a one-pass bottom-up best-first search strategy. Evaluated in the ATIS domain, ANGIE achieved a phone error rate of 36%, as compared with 40% achieved with a baseline phone-bigram based recognizer under similar conditions. ANGIE potentially offers many attractive features, including dynamic vocabulary adaptation, as well as a framework for handling unknown words.

1. OVERVIEW

In this paper we propose a methodology for incorporating multiple sublexical linguistic phonemena (including phonology, syllabification and morphology), into a single framework for representing speech and language. Together with a trainable probabilistic parser, this unified framework provides a viable paradigm for *multiple* tasks – letter-to-sound/sound-to-letter generation, phoneme-to-phone alignment, and speech recognition. We hope that such a unified framework promotes shared usage of the sublexical constraints amongst the different applications, which should facilitate the search processes and also make it easier to deal with out-of-vocabulary words and to add new words dynamically.

A preliminary system based on this paradigm, which we call ANGIE, has been under development in our group over the past year. Context-free rules are written by hand to generate a hierarchical tree representation, as illustrated in Figure 1. These trees are used to train the probabilities of the parser, which are later used in each of our three applications. The structure consists of five regular layers below the root SENTENCE node. Each word in the sentence is represented by a WORD node in the second layer. The remaining layers capture, in order, morphology, syllabification, phonemes, and phones/letters. Throughout this paper, we use *italics* to distinguish entries in the preterminal phoneme layer, and a preceding "$" symbolizes terminal labels, as indicated in the figure. Stress markings are distributed in multiple layers (e.g., SROOT stands for "stressed root," *ow+* is a "stressed *ow*," LNUC+ is a stressed nucleus with a long vowel, etc.). The bottom-most layer consists of letter terminals in sound-to-letter/letter-to-sound generation, changing to phone terminals for phoneme-to-phone alignment and speech recognition.

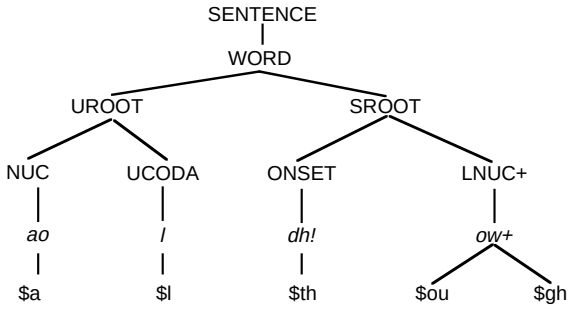

Figure 1: Example parse tree with letter terminals for the word "although."

2. LINGUISTIC MODEL

ANGIE's approach is similar to that reported in [6]. A significant difference is that we have attempted to reduce the parameter space, both by decreasing the total number of layers in the hierarchy, and by restricting the context conditions for column [2] building. Thus there is greater sharing in the probability space, which should improve parse coverage, but may sacrifice performance. We have eliminated the broad class layer,[3] and have replaced the stress layer with distributed stress markings at all nonterminal layers. The letter-terminal and the phone-terminal grammars share an identical rule set for all but the terminal layer.

A parse proceeds bottom-up and left-to-right. Each column is built from bottom to top based on spacio-temporal trigram probabilities.

[1] This research was supported by DARPA under contract N66001-94-C-6040, monitored though Naval Command, Control and Ocean Surveillance Center."

[2] The term *column* refers to the sequence of nodes along a path in a parse tree from a terminal node to the root node.

[3] Previous experiments have yielded improved pronunciation accuracy without this layer.

The terminal category is first predicted based on the *entire* left column. The prediction of a parent is conditioned on its child and the parent's immediate left sibling, without regard to the column above the left sibling.[4] The linguistic score for a full-column advance is the sum of the log probability for the terminal phone/letter and the log probabilities for the bottom up prediction scores for each column node up to the point where the parse tree merges with the left column.

The categories in our grammar are carefully chosen so that lexical constraints are propagated downwards in the parse tree to reduce perplexity and improve prediction accuracy. There are currently 100 unique preterminal "phonemes," including stressed (marked by "+") and unstressed versions for the vowels, onset (marked by "!") and non-onset realizations for the consonants, morph-specific phonemes (e.g., *d*ed* for the past tense morpheme "ed", *s*pl* for the plural morpheme, etc.), phonemes that are unique to particular function words (*dh_the*), and pseudo-diphthongs such as *aar* and *aol*. The "null" terminal in [6] has been replaced with deletable units marked explicitly for their left context. Most of the remaining nonterminal categories are generic units such as UROOT and CODA, but we do include over twenty special inflexional suffixes.

3. LETTER-TO-SOUND GENERATION

ANGIE's letter-to-sound generation system bears close resemblance to that described in [6]. A probabilistic parsing algorithm is used to parse the letters of an input word, and the pronunciation is derived from the phoneme sequence at the preterminal layer in the parse tree. One enhancement is the introduction of a new preprocessing stage to preselect the preferred training parse when ambiguity exists, to avoid having to select parse trees manually.

A breadth-first search strategy is adopted for left-to-right, bottom-up parsing – each column in the parse tree is advanced exhaustively to all the legitimate right columns. Partial theories (or partial parse trees) are scored and ranked probabilistically, and the top 20 theories with the same letter terminals are chained together, while the rest are pruned. The theories in the chain serve to provide multiple pronunciation outputs if necessary.

Training parse trees are generated from letter sequences that have first been processed through a set of "meta rules," which provide a preliminary marking/grouping of the letters before parsing takes place. The meta rules are based only on letter context (left and right). When a meta rule proposes a doubleton, the corresponding sequence of two singletons is disallowed. Without the meta rule, both options are possible. Other meta rules specially mark certain letters, for example, labelling the second "l" in "pillow" as $l2, or marking the "v" in "paving" as a $v_e, to help encode the long vowel. There are a total of 187 possible terminal categories, including the 26 standard letters, 125 doubleton letters, 6 letter contexts for deletion, and 30 specially marked letters. Such an approach is less time-consuming and more portable than would be the process of hand-selecting the appropriate training parse tree.

Our experiments were conducted with the same high frequency words in the Brown Corpus that were used in [6]. About 8,000 words are used for training, and a disjoint set of 872 utterances for testing. We have counted as correct any confusions between a consonant and the same consonant marked for onset position, since ambisyllabicity makes this distinction difficult, and since onset position is not important as a phonemic distinction. As summarized in Table 1, we have achieved a training accuracy of 75.3% per word and 91.9% per phoneme, and a testing accuracy of 68.9% per word and 89.7% per phoneme. The accuracy score takes into account substitutions, deletions, and insertions. None of the test-set words failed to parse, indicating that our more generalized grammar was effective.

Data Set	# words	# phmes	Word Acc	Phme Acc
training: l-t-s	7868	75	75.3%	91.9%
test: l-t-s	872	75	68.7%	89.7%
collapsed	872	69	69.4%	91.5%
Data Set	# words	# letters	Word Acc	Letter Acc
test: s-t-l	872	26	53.2%	89.2%

Table 1: Results for letter-to-sound (l-t-s) and sound-to-letter (s-t-l) generation.

Word	Target Phonemes	Hypothesized Phonemes
dough	*d! ou+*	*d! ah+ f*
familiar	*f! ah m! ih+ l iy er*	*f! ae+ m l! iy er*
envelope	*eh+ n v! el ow+ p*	*eh n v! el+ ah p*
exploited	*eh k s p! l oy+ t d*ed*	*eh k s p! l ow+ ih t d*ed*

Table 2: Examples of pronunciation generation errors

While these results appear to be somewhat inferior to those reported in [6], it should be noted that the latter experiment used a smaller set of phonemic distinctions (55 categories as against our 75). Test performance improves to 69.4% per word, 91.5% per phoneme accuracy if we add the following forgivable confusions: *eh*/*ih* (as a combined front schwa); *eh*/*eh+*, *ih*/*ih+*, *er*/*er+*, *el*/*l*, and *en*/*n*.[5] This result is slightly better than the 69.2%, 91.3% result reported in [6], obtained with full coverage after a backoff algorithm recovered parse failures. Table 2 shows some examples of letter-to-sound errors. An incorrect stress pattern is the source of many of the errors, which perhaps calls for special treatment of the stress pattern.

For sound-to-letter generation, we used a best-first search (see Section 5), proposing all letters bottom-up and filtering on the known sequence at the phoneme layer. The best scoring hypothesis was selected from the first five completed theories. Our results (53.2% per word, 89.2% per phoneme) are comparable to those reported in [6] (53.5% per word, 88.5% per letter).

4. PHONOLOGICAL RULES

When the terminals are *phones* instead of letters, the grammar defines a set of probabilistic phonological rules. Phonological rules are written without specifying context explicitly. Contexts for which the rules apply are learned, along with corrresponding probabilities,

[4] In [6] we conditioned on the *entire* column above the left-sibling.

[5] The last two are syllabic versus non-syllabic distinctions.

from a large body of acoustic training data. For example, the following rule[6] states that a *t* in onset position can be realized with an optional closure interval ($tcl), an obligatory release that could optionally be rounded ($tr), and an optional aspiration interval ($hh).

$$t! \rightarrow [\$tcl]\ (\$t\ \$tr)\ [\$hh]$$

The 100 phonemic units in the preterminal layer map to 65 unique phonetic labels. The vowel phonemes are marked for stress, and consonants are marked for onset position. These distinctions are not maintained at the phonetic level; instead they are manifested in distinct distributions of the probability of mapping to particular phones. Thus, *unstressed* vowels are much more likely to be reduced to schwa, and stops in *onset* position are much more likely to be released. Some of the phonemes for function words are word-specific: they map to generic phones but with a different distribution than do their non-word-specific counterparts. Thus, for example, the *ay* in the word "I" is much more likely to be reduced than is *ay* in general.

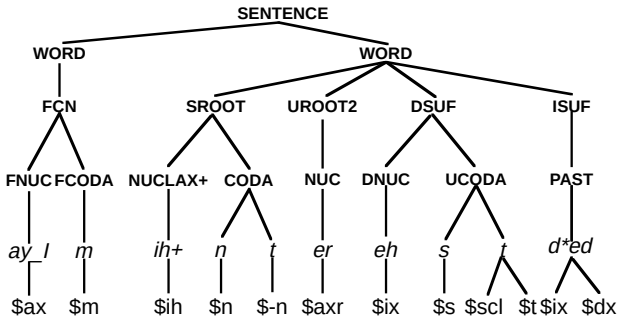

Figure 2: Example parse tree with phone terminals for the phrase "I'm interested."

Our phone set has evolved over time, and is likely to change. The choices were made empirically by examining phonemic-to-phonetic alignments using our new SAPPHIRE tool [4]. A few context-dependent units were chosen for cases where context effects were strong and there was sufficient context-specific data to build a robust model. Thus, both alveolar stops can be realized as a $dx (flap) phone, usually within intervocalic environments. A special $scl phone represents a noisy alveolar closure interval folllowing fricatives. The phoneme *h* has been merged with the aspiration interval of unvoiced stops, thus substantially alleviating the sparse data problems for this rare phoneme. We allow three distinct schwas – retroflex ($axr), front ($ix), and back ($ax); some unstressed vowels can be realized as both front and back schwa, with probabilities that adjust with context. An example parse tree with phone terminals is shown in Figure 2.

Probabilities are currently trained on some 10,000 utterances from the ATIS corpus [1, 2], seeded on forced alignments obtained using our SUMMIT recognizer [7], and subsequently iterated. Per-phone perplexity for training/test conditions is around 5.7 and 7.0 respectively, which is substantially lower than corresponding bigram phone perplexities.

[6] Brackets indicate *optional* and parentheses enclose alternates.

5. SPEECH RECOGNITION

Many current speech recognition systems handle phonological variation either by generating a pronunciation graph for each word (such as MIT's SUMMIT system) [7] or by implicitly absorbing the variations into a hidden Markov model. The former has the disadvantage of not sharing common subword structure, hence splitting training data. The latter makes it difficult to control and improve upon phonological modelling. For example, in the ATIS domain, the words "connect," "connecting," "connects," and "connection" all share a common initial phoneme sequence. Phonological variations affecting this sequence can be better learned if examples from all four words are pooled together.

By pursuing merged common subword theories during the search, we can mitigate the combinatorial explosion of the search tree, making large vocabulary recognition more manageable. Because we expect new words to share much common subword structure with existing words, we can easily add new words dynamically. In principle, we can even detect the occurrence of out-of-vocabulary words by recognizing as much of the subword structure as possible in a bottom up manner. Since the same rules support a letter-to-sound system, proposed pronunciations for new words can be generated automatically from their orthographic transcription, and entered into a pre-existing right-branching lexical tree at the phoneme layer.

At the present time, recognition in ANGIE utilizes a one-pass best-first search, with no future estimate. Words are built bottom-up from the rules, while tracking the lexicon along the phoneme layer. Whenever the parser proposes a possible word termination, the system can confirm the existence of this word in the lexicon, and could apply higher level language models at this time (either word n-gram or natural language models).

Each unique theory in the stack is associated with a single boundary in time and a partial linguistic hypothesis. The linguistic hypothesis retains the unique *word* sequence up to the last proposed word boundary, along with the unique *phone* sequence for the partial word theory under construction. The phone sequence is associated with a set of linguistic theories representing all word patterns that can begin with this phone sequence, pruned to a maximum of 15. These theories are rank ordered, and the score of the most probable one is taken as the representative linguistic score for the phone sequence. The linguistic score is computed as described previously in Section 2, with the exception that across word boundaries the terminal category score is conditioned only on the preceeding phone, instead of the entire left column. The rationale for this is twofold. One, it mitigates the sparse data problem across word boundaries, and two, it allows us to merge theories at word boundaries and thus improve the tractability of the search. The acoustic score is the sum of the acoustic scores for the individual phones. Linguistic theories, which have no knowledge of time, are shared among equivalent hypotheses.

Because the search is a best-first algorithm without a future estimate, it is important to normalize scores so that short theories and long theories are balanced. Our feeling is that this can be accomplished in part by targeting scores towards a mean value of zero. To this end,

we adjusted the acoustic models for our training data such that on average each corect phone score would realize a zero mean and a unity variance distribution. To normalize the linguistic scores, we adjusted the log probability by offsetting the mean entropy. We experimented with several different ways to define natural groupings, and found experimentally that the best algorithm was to normalize each unique phone category so that it appeared to advance *on average* with probability 1.0. In this way, paths that are "better than average" get a positive score. In addition, we introduced a fading scheme so that probabilities from the past eventually decay to zero.

In spite of the above normalization schemes, it was still the case that for some long sentences computation became unwieldy. We experimented with several different pruning algorithms, and found that the two most effective were to limit the *total* number of theories that could cross a given boundary, and to limit the *total stack size* to some fixed maximum length. It is likely that we are sometimes losing the best theory, and certainly the search is inadmissible. We may eventually decide to incorporate some sort of future estimate, although we find the idea of no look-ahead appealing.

5.1. Recognition Experiments

To date, all of our recognition experiments have been conducted in the ATIS [2] domain, and we have limited our training and test data to a subset of the ATIS3 corpus [1]. Thus far, we have been conducting experiments at the level of *phonetic* recognition, where the lexicon is only used *implicitly* to train the ANGIE subword models, i.e., the ANGIE probabilities are trained on a set of phonetic sequences associated with orthographic transcriptions for ATIS sentences. Comparisons are made with a baseline system that uses a simple phone bigram for a language model and a standard Viterbi search strategy.

Our main question was whether the sophisticated linguistic model in ANGIE would lead to better measured performance in a phonetic recognition task. Since hand-labeled phonetic transcriptions for ATIS data are not available, we felt that it would be appropriate to score each recognizer against its *own* phonetic labels, as obtained during forced alignment of the orthographic transcription. After all, this is the phonetic transcription the system would need to produce in order to perform correct lexical access. For the baseline system, we used a predefined set of ATIS phone units and phonological rules that had been developed for our existing ATIS recognizer to produce the forced alignments.

We used the same acoustic data for training and testing each system, as well as the same segmentation algorithm and model parameters. The models were Gaussian mixtures averaged over left, middle and right thirds of each segment, as well as delta Gauassian mixtures over the left and right boundaries. The two systems had somewhat different phone sets, but in both cases we collapsed down to the standard 39 "CMU" phone set [5]. In place of ANGIE's subword language model, the baseline system used a phone bigram, derived from its own training phonetic alignments. Neither system had a lexicon, but ANGIE proposed word-ends periodically, segmenting the phones into a sequence of pseudo-words. Although the results are preliminary, ANGIE achieved a phone error rate of 36%, as against the baseline's 40%. We are encouraged by this initial positive result.

6. FUTURE PLANS

ANGIE is still in a preliminary stage, and we feel there are many research directions the work could take. We plan to extend the recognizer in many directions. Our first attempt at something beyond phonetic recognition will probably be word-spotting in the ATIS domain, using ANGIE's subword models for both the known words and the surround. We will then move on to continuous speech recognition, focusing on issues of search, computation, and memory.

Ultimately, we hope to use ANGIE in a conversational system, such as our GALAXY system [3], enabling us to tag and discard unknown words, as well as adjusting the vocabulary transparently to reflect dialogue context. For example, if the user asks for bookstores in Cambridge, and the system retrieves a set of bookstores from an on-line Yellow-Pages database, it would be convenient if the system could, at the same time, update its recognizer vocabulary to include the names of these bookstores, licensing them in the language model under a generic category such as "store_name." We hope to use ANGIE as a letter-to-sound system to generate proposed phonemic pronunciations from the spellings, and then enter these new words along the phoneme layer to achieve lexicalization. Phonological rules would be embedded in the existing structure, leveraged from patterns acquired for other similar words. We see critical future needs for such capabilities for flexible vocabulary, if conversational systems are ever to become practical.

We are beginning to explore the possibility of using subword parse trees provided by ANGIE to construct complex duration models for sublexical units. For example, we suspect that the ratio of the duration of a constituent to that of its parent (e.g., onset duration relative to syllable duration) may form an interesting self-normalized duration parameter.

7. REFERENCES

1. Dahl, D. et al., "Expanding the Scope of the ATIS Task: the ATIS-3 Corpus," *Proc. ARPA Human Language Technology Workshop*, pp. 43-48, Plainsboro, NJ, March, 1994.
2. Glass, J. et al., "The MIT ATIS System: December 1994 Progress Report", *Proc. ARPA Spoken Language Technology Workshop*, pp. 252-256, Austin, TX, January 1995.
3. Goddeau, D. et al, "GALAXY: A Human Language Interface to Online Travel Information," *Proc. ICSLP-94*, pp. 707-710, Yokohama, Japan, 1994.
4. Hetherington, L and M. McCandless, "SAPPHIRE: An Extensible Speech Analysis and Recognition tool base on Tcl/Tk," *These Proceedings*, 1996.
5. Lee, K., *Large Vocabulary Speaker-independent Continuous Speech Recognition: The Sphinx System*, Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, 1988.
6. Meng, H. M., *Phonological Parsing for Bi-directional Letter-to-Sound / Sound-to-Letter Generation*, Ph.D. Thesis, MIT Department of Electrical Engineering and Computer Science, Cambridge, Massachusetts, 1995.
7. Zue, V. et al., "The MIT SUMMIT Speech Recognition System: Phonological Modelling and Lexical Access," *Proc. ICASSP-90*, pp. 49-52, Albuquerque, NM, May, 1990.