A THREE-STAGE SOLUTION FOR FLEXIBLE VOCABULARY SPEECH UNDERSTANDING[1]

Grace Chung

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

ABSTRACT

This paper discusses our three-stage approach to a flexible vocabulary speech understanding system, which can detect out-of-vocabulary (OOV) words, and hypothesize their phonetic and orthographic transcriptions. In the first stage, we introduce the column-bigram finite-state transducer (FST) which, while embedding ANGIE sublexical models, also supports previously unseen data from unknown words. Secondly, the ANGIE models utilize grapheme information, providing tighter linguistic constraint as well as instantaneous sound-to-letter capability during recognition. Thirdly, the syllable-level lexical units of the first stage are automatically derived via an iterative procedure to optimize performance. The second-stage recognizer employs ANGIE to output a word network which is parsed by TINA, our natural language (NL) processor, in stage three. Experiments with a JUPITER implementation of this system are described in [1].

1. INTRODUCTION

In the future, we foresee conversational systems capable of supporting flexible vocabularies, that is, unknown words are automatically detected at the spoken input, and corresponding acoustic, phonological and linguistic properties are inferred. From these, the system would hypothesize letter spellings, and in this way, the lexicon is dynamically extended with new words spoken at recognition time. At present, state-of-the-art systems have limited capabilities in coping with unknown words. In fact, as quoted in [5], test sentences containing OOV words can suffer a five-fold degradation in recognition performance. One challenge is to narrow this large performance gap.

Previously, we envisioned a *two*-stage architecture where the front-end consisted of a domain-independent recognizer, and this was interfaced via a subword network to a back-end which incorporated constraints from higher-order knowledge sources, such as NL, tailored for the specific domain. In [2], we implemented a preliminary system where the first stage only utilizes syllable-level linguistic information, combining ANGIE [4], the sublexical model, together with a morph[2] trigram. This first stage produces a phonetic network which guides the search in stage two. Thereafter, TINA [3], our NL module, and ANGIE models are tightly integrated within a single search, to produce final sentence hypotheses. Here, we extend the previous architecture with some novel enhancements in a *three*-stage implementation. These are designed to enable the system to detect unknown words, hypothesize their phonetic and orthographic transcriptions, and potentially incorporate them without additional training.

Our foremost objective is to devise a framework where phonetic sequences of unknown words that have never occurred in training can be supported by linguistic models, and to accomplish this without compromising in-vocabulary recognition accuracy or causing an explosion in the search space. In light of this, the underlying philosophy behind our first stage is to use only low-level linguistic constraints that are shared within all general English vocabulary, and to delay the application of domain-specific lexical knowledge until the second stage, after much of the search space has already been pruned. In principle, such an initial stage could encompass a simple phone or syllable recognizer. But in order to boost performance, we have, through novel means, embedded additional knowledge sources that do not require an explicit domain-dependent word lexicon.

In the first stage, our FST-based recognizer now utilizes a new method for generating an FST which better encapsulates the hierarchical probabilistic models of ANGIE. Called the *column-bigram* method, the resultant FST can assign non-zero probabilities for previously unobserved phonetic sequences. Secondly, we introduce the notion of simultaneously modeling grapheme and pronunciation information within ANGIE via spelling-based phoneme-level units. With an ANGIE grammar converted to an efficient FST representation, a recognizer can potentially derive letter spellings directly during recognition upon encountering unknown words. The third feature of the first stage is the ability to optimally exploit syllable-level constraints together with spelling information by automatically generating a morph-based lexicon. We will describe a procedure which begins with a column-bigram FST, and iteratively builds novel syllable-sized units by concatenating grapheme-related sequences.

A particularly novel aspect of this design is the integration of low-level linguistic knowledge to promote tighter constraint. These are (1) implicit word substructure constraints captured by the column-bigram ANGIE FST (2) grapheme information embodied in spelling-based units and (3) morph-level trigram constraints. More importantly, because we are not tied to a fixed lexicon, the morphs and word substructure information stem from training our grammar on novel automatically-derived units. The algorithm used to build these units aims to encapsulate information more efficiently by creating alternative word and morph units.

During the second stage, OOV words are allowed at restricted locations, and the ANGIE parse mechanism is used to guide phonetic hypotheses during the search. ANGIE may also be used to

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

[2] Morphs are syllable units with distinct spellings augmented with positional markers.

word			word			
stressed root		suff	stressed root			
onset	nuc+	plural	onset		nuc+	coda
d!	ey+	s	p!	l	ey+	s
d	ey	z	p	l	ey	s

Figure 1: Tabular schematic of ANGIE parse trees for words *"days"* and *"place."* Each entry in the table depicts a tree node. The rows represent from the bottom up, phonetic, phonemics, syllabification and morphology. "!" denotes onset position and "+" marks lexical stress.

propose spellings. The third stage involves using TINA to parse and re-score word graphs, computed from the second stage. This method also computes a meaning representation for further processing in the dialog system.

In the next sections, we elaborate on details of our first stage, and how the component FSTs are assembled together. Steps include generating FSTs using the column-bigram method, incorporating spelling information within an ANGIE grammar, and iterating the automatic procedure for generating novel lexical units. We then examine the roles of the second and the third stages. This system has been implemented in the JUPITER domain, and evaluated with test sentences containing unknown city names. Details and results are reported in [1].

2. STAGE ONE RECOGNIZER

As in [2], the first stage is an FST-based recognizer that uses context-dependent diphone acoustic models and proposes novel morph units. Although the actual novel morphs are hypothesized as outputs, they are subsequently decomposed into their phonetic constituents in an optimized lattice so that the later stages are not confined to the set of morphs proposed in the initial stage. Defining all language search constraints is a single FST that maps the context-dependent units to sequences of morph units. This FST is computed *a priori* via the composition: $C \circ P \circ L \circ G$, where C transduces context-dependent to context-independent labels, P applies sublexical modeling via an ANGIE-derived FST, L maps phoneme baseforms to lexical units and G is a trigram model. In Section 2.1, we elucidate the column-bigram FST which maps phones to phonemes with ANGIE probabilities. The phoneme set has been redesigned to be the spelling-enriched letter-phonemes described in Section 2.2. Both the ANGIE grammar and the trigram are trained on the novel units that are generated in an algorithm detailed in Section 2.3.

2.1. THE COLUMN-BIGRAM FST

Introduced in [4], ANGIE is a sublexical model that combines a trainable probabilistic framework with a hand-written context-free grammar. It produces a parse tree which captures phonetics, phonemics, syllabification and morphology. In [2], we converted ANGIE's hierarchical models to a single flattened FST representation, enabling us to utilize its probabilities within the recognition search. However, the algorithm relied on memorizing training data instances, and assigning pre-computed probabilities on the FST arc weights. Hence, allowable paths in the FST were limited to entire phonetic sequences that had been instantiated in the training data, and moreover, previously unseen phone sequences belonging to unknown words or rare pronunciations were disallowed. Fundamentally, this was at variance with the ANGIE parse mechanism which can generalize models across words with common substructures. These considerations have motivated us to develop the column-bigram FST.

In the column-bigram method, an entire ANGIE parse tree is viewed in a tabular format (Figure 1) which depicts a sequence of columns. Each *column* denotes the parse tree nodes along a given path from the root node to the terminal node. ANGIE parse tree probabilities can then be decomposed into probabilities from one column to another, proceeding left to right. The probability of a column, given the previous, is computed by summing component probabilities that depend on nodes from the left context. These are (1) trigram bottom-up probabilities: the probability of a column node given its left sibling node and its child node, and (2) the advancement probability: the probability for advancing to the next phone terminal given the left-context column. Consequently, the probability generated, when we proceed from one column to the next, can be seen as a bigram probability for adjacent column pairs, and consequently our column-bigram FST is similar in structure to a general bigram model FST.

During the training phase, the ANGIE grammar and its probabilities are initially trained up. Then, all unique ANGIE columns (with distinct tree nodes), that have occurred in training data, are enumerated. For all adjacent column pairs that are observed, column bigram probabilities are computed by summing the above-mentioned component probabilities, and recorded. When constructing the FST, every unique node corresponds with a unique ANGIE column. For a particular column, the outgoing arcs of that node represent transitions to other columns, and the bigram probabilities reside on the arc weights. In our design, we choose to emit phoneme units of the pre-terminal tree layer, given phone terminals as input labels. Also at a morph boundary, we emit the morph class[3] of the left column. This provides additional higher-level or long-distance linguistic information about the left column at the output which may be beneficial at a later stage.

In this scheme, a consideration has been to reproduce as much of the parse mechanism's flexibility as possible while keeping the size of the search space, as determined by the FST size, manageable. Thus, restricting the FST to recording only the column pairs that are observed in training serves to omit a portion of ANGIE's probability designated by its over-generalizing ability. We believe that, in most cases, this space yields low probability estimates, because those column pairs in question did not occur in training. This effectively limits the number of arcs in the resulting FST. However from our preliminary investigation of the design, many novel morph transitions may be disallowed due to lack of training observations. Therefore as a measure to overcome this and control any sparse data problems that would prevail at morph boundaries, we have implemented a simple back-off or smoothing mechanism.

A back-off node corresponding with every morph class is constructed. At any morph-final column, the probability of transitioning to a back-off node from a column C_i is computed as follows:

$$P(Backoff|C_i) = 1 - \sum_j P(C_j|C_i) \qquad (1)$$

where $P(C_j|C_i)$ is the probability assigned to column C_j given

[3] Examples of morph classes include prefix, inflexional suffix, stressed root, and so forth.

its left context column C_i. In practice, this is estimated by summing the total probability of arcs exiting a column node, C_i, and assigning the remaining probability space to the transition towards the back-off node. This space is a direct consequence of allocating probability towards unseen data by the ANGIE parse mechanism. Then, the probability estimate for exiting the back-off node is computed as the maximum likelihood estimate for the probability of the next column given by left morph class context.

Compared with our previous strategy, the column bigram captures a larger portion of the ANGIE probability space by replicating some of the parse mechanism's ability to generalize through sharing models of common substructures. In fact, while the ANGIE grammar is trained from a fixed two-tiered lexicon containing words and morphs, the resulting FST is not confined to these, but instead only captures an implicit knowledge of word substructures. Novel paths through the FST which license previously unseen sequences are possible because the algorithm only observes adjacent column pairs rather than entire sequences of phones. For example in Figure 1, by observing in training the words "days" and "place," the FST now contains a path for the unobserved word "plays," because the relevant adjacent column pairs have all been observed in the two training words. Even more morph transitions can be supported when the back-off mechanism is included, and this ameliorates further sparse data problems. In addition, unlike our previous method, the output symbols are at the phoneme level, rather than morphs from a fixed lexicon. Hence, concatenating novel phoneme sequence outputs yields new morphs.

2.2. SPELLING-BASED UNITS

We are driven to incorporate grapheme information in the first stage for two reasons. First by equipping models with spelling knowledge, we could directly deduce spellings from phonetic hypotheses at unknown words. Spellings can be accessed within the recognizer models, and this obviates the need for a separate sound-to-letter module. Secondly, we may exploit spelling information as another form of low-level domain-independent linguistic constraint that can be used in our first-stage recognizer.

With the availability of ANGIE's hierarchy and an efficient FST representation for it, we propose to encode grapheme information within the grammar. In the past, for sound-to-letter/letter-to-sound applications, letter units were used in lieu of phones at the terminal layer in ANGIE. Here, our approach is to embody spelling knowledge within the phoneme pre-terminal layer of the ANGIE grammar. This is done by replacing the phoneme set of the pre-terminal layer by an expanded set of *letter-phoneme* units. We conceive of units that codify spelling and pronunciation information simultaneously; that is, the letter-phoneme to phone layer of an ANGIE parse tree has the dual purpose of modeling both phonological processes and letter-to-sound/sound-to-letter conversion. The inventory of letter-phonemes is constructed by annotating letter units with markers designed empirically to signify phonemic pronunciation, stress and so on. For example, the letter-phoneme *a_l+* represents a long stressed vowel spelled as an "a" and pronounced as the /ey/ phone. From the letter-phoneme baseform of a word, one can deduce both its spelling and phonemic sequence. More examples and greater detail can be found in [1] where this has been implemented in the JUPITER domain.

The major benefit of this novelty is that the new ANGIE grammar is conveniently converted to a column-bigram FST where, instead of emitting phonemes, the FST emits letter-phoneme sequences. And so phone sequences can be mapped to spelling hypotheses directly by a single FST with a likelihood score that reflects the combination of pronunciation and letter-to-sound modeling. This is particularly desirable for novel letter-phoneme sequences belonging to unknown words, during recognition. Moreover, in expanding the functionality of the pre-terminal layer of the parse tree, we have enriched the probability models with greater contextual information, thereby imposing tighter constraints. This may deliver improved recognition performance in the first stage.

2.3. AUTOMATIC LEXICAL GENERATION

As mentioned earlier, the system composes the column-bigram FST P (built from the spelling grammar) with an intermediate lexical FST L before combining with a trigram FST G. The role of the intermediate FST is to map the letter-phonemes to the larger-sized units of the trigram. The resultant FST should combine tight linguistic constraint in conjunction with flexibility to support OOV sequences. In the past, lexical units in the first stage are morphs. With the more cumbersome spelling ANGIE grammar, we are faced with concerns for the FST size that would impact computational efficiency. Our remedy here is to design a new set of lexical units which is automatically derived and optimizes on the probability models. Ideally, this would lead to smaller FSTs and an optimized set of morph-level units where sparse data problems are minimized while tight constraints are maintained.

The core idea relies on the fundamental insight that, in the first stage, the morph units are not required to directly correspond with a single syllabification of the word lexicon of the second stage; in fact, we are not tied to any fixed lexicon specific to our domain. The reason for this is that the second stage uploads a phonetic network from the first. Therefore, we are required simply to improve phonetic accuracy, whereas the actual underlying morph or word hypotheses are only relevant in the second stage. We posit that our chances for producing correct phonetic sequences for both known and unknown data may improve upon re-optimizing lexical space. And our optimized lexical organization should enable ANGIE to capture sublexical information more compactly.

Using the same forced aligned training data, the procedure is an iterative algorithm which hinges on the following properties of our column-bigram FST: (1) by concatenating the output labels, we can hypothesize the spelling underlying the word of a phonetic sequence and the respective morph boundary locations, and (2) given any phonetic sequence, multiple paths exist within the FST, emitting distinct and novel letter-phoneme sequences with probabilities. In fact, within a sentence, even for phonetic sequences of known words, the highest scoring output sequence may not be the one asserted in the original ANGIE morph lexicon. This suggests that an alternative set of morph-level units may yield better results by way of higher probabilities and thus reduced perplexity numbers. With each iteration of our algorithm, a new morph and word lexicon is proposed. The morphs are a concatenation of the novel letter-phoneme sequences that were discovered by finding highest scoring paths in the FST, and the words are constructed from the morphs, using a simple set of concatenation rules. The morph and word lexicons are used to train up a new ANGIE grammar for the next iteration.

The iterative procedure is set out in the following steps:

1. *Initialization*: Begin with an initial set of rules that incorporate letter-phoneme units.

2. *Train grammar*: Use the forced aligned set of orthographic and phonetic transcriptions to train an ANGIE grammar.

3. *FST Generation*: Use the forced alignments and trained ANGIE grammar to generate a column-bigram FST.

4. *Search*: For the phonetic sequence of each training utterance, search for the highest scoring path through the column-bigram FST, and output a corresponding letter-phoneme sequence along with morph class labels at morph boundaries.

5. *Construct morphs*: For each morph, infer the spelling by concatenating the letter-phoneme sequence, after removing contextual markers. The morph class is also deduced. For example, the letter-phoneme sequence, *d! ay+*, can be concatenated to form a stressed morph, *day+*. If the morph has not been previously encountered, add it to the lexicon.

6. *Construct words*: Construct the underlying "word" by concatenating morph sequences using some simple rules. Each word must contain only one stressed morph, and word boundaries are inserted whenever permissible according to ANGIE. If the word has not been previously encountered, add it to the lexicon.

7. *Go to step 2*: Upon completion of the new lexicons, begin with a new grammar and orthographic transcriptions for training data that are derived from the new word lexicon. Return to Step 2.

The above procedure has been implemented successfully in [1] with a reduction in both perplexity and the FST size. After several iterations, the final morph lexicon is used for generating the lexical FST. A further step to increase ability to support novel sequences is to decompose all stressed roots into their respective onsets and rhymes. In doing this, the number of unique lexical units will be greatly reduced, and the trigram model will now be based on a set of *sub-morphs*: a collection of unstressed morphs combined with onsets and rhymes of stressed morphs. On the one hand, this represents a relaxation on constraints. But splitting the stressed roots permits novel combinations of onsets and rhymes to form new stressed morphs, where the most serious sparse data problems exist. And this will be supported by both ANGIE probabilities in the column-bigram FST as well as in the trigram model FST. Note that both these FSTs are derived from the new morph units.

3. STAGE TWO: ANGIE-BASED SEARCH

In the second stage, the search space is constrained by optimized phonetic networks output from the first. The arc weights on these networks consist of acoustic scores and language model scores with reduced weighting. Similar to that described in [2], the control strategy integrates together a word bigram, the ANGIE sublexical models, and optionally, TINA. This strategy has been altered to a best-first search augmented with future estimates given by the potentials on the FST nodes.

The recognizer allows the incidence of unknown words at certain restricted locations. In general, the ANGIE parse mechanism only licenses phoneme sequences expressed in the word lexicon. However, where unknown words are permissible, we override this constraint, and phone sequences will be proposed whenever an ANGIE parse succeeds. In the case where a novel phonetic sequence is hypothesized, the associated probability score is returned with an OOV flag. A hypothesized spelling may also be accessed from ANGIE, if the letter-phoneme grammar is used.

4. STAGE THREE: TINA Parsing

In the final stage, the N-best output of the second stage is converted into a word network via an algorithm that computes goodness scores computed from ranking hypotheses by their frequencies in the list. A viterbi search with beam pruning sweeps the network, applying TINA parsing and finding the highest scoring sentence according to the TINA score combined with the goodness scores. TINA is specially trained to support unknown words under specific categories such as proper names. A meaning representation is obtained directly during this stage and will be of further use for the dialog system.

5. CONCLUSIONS

This paper has described our three-stage speech understanding system capable of dynamically extensible vocabulary. The most novel contribution is the design of our first stage which relies exclusively on low-level knowledge such that previously unseen phonetic sequences belonging to new words can be recognized. Our objective has been to dually maximize flexibility and constraint. This has entailed combining sublexical models, embedded with grapheme information together with syllable-level costraints in an FST paradigm. In particular these linguistic models are based on automatically generated novel units which are guaranteed to improve probability likelihoods. Application of word-level constraints are delayed until stage two where the ANGIE parse mechanism can detect unknown word occurrences and hypothesize their phone and spelling sequences. NL constraints are imposed in the third stage on a compact network, containing hypotheses of known and unknown words. Encouraging results reported in [1] point to the potential of this design for a future system that not only copes with unknown words but adopts them immediately, enlarging the recognizer vocabulary each time.

6. ACKNOWLEDGEMENT

I would like to thank Dr. Stephanie Seneff for supervising this research, and for providing valuable comments on this paper.

7. REFERENCES

1. G. Chung, "Automatically Incorporating Unknown Words in JUPITER," in these Proceedings.

2. G. Chung and S. Seneff, "Towards Multi-Domain Speech Understanding Using a Two-Stage Recognizer,", in *Proc. Eurospeech '99*, Budapest, Hungary, pp. 2655-2658, Sep. 1999.

3. S. Seneff, "TINA: A Natural Language System for Spoken Language Applications," in *Computational Linguistics*, 18(1), pp. 61-86, 1992.

4. S. Seneff et al., "ANGIE: A New Framework for Speech Analysis Based on Morph-Phonological Modeling," in *Proc. ICSLP '96*, Philadelphia, PA, Oct. 1996.

5. V. Zue et al., "JUPITER: A Teleophone-Based Conversational Interface for Weather Information," in *IEEE Trans. Speech and Audio Processing*, 8(1), pp.85–96, 2000.