IMPROVEMENTS IN SPEECH UNDERSTANDING ACCURACY THROUGH THE INTEGRATION OF HIERARCHICAL LINGUISTIC, PROSODIC, AND PHONOLOGICAL CONSTRAINTS IN THE JUPITER DOMAIN[1]

Grace Chung 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:{graceyc, seneff}@mit.edu

ABSTRACT

This paper explores some issues in designing conversational systems with integrated higher level constraints. We experiment with a configuration that combines a context-dependent acoustic front-end, using MIT's SUMMIT recognizer, with ANGIE, a hierarchical framework that models word substructure and phonological processes, and with TINA, a trainable probabilistic natural language (NL) model. Working in the Jupiter weather domain, we develop a computationally tractable system which incorporates higher level linguistic, prosodic and phonological constraints together in the second of a two-pass strategy. Experiments are evaluated using a new understanding performance metric, and the new integrated system achieves up to 17.1% relative reduction in understanding error and 15.4% reduction in word error. In addition, we investigate the possibilities of a two-pass system which relies on the first stage for pruning based on syllable-level constraint, and applies linguistic and prosodic knowledge largely at the second stage.

1. INTRODUCTION

One critical area in developing speech understanding systems is the intelligent integration of structured linguistic knowledge. Our research is concerned with exploring various strategies for incorporating multiple linguistic knowledge sources with speech recognition in a tightly integrated manner, that is, to apply these constraints early in the recognition search process. In a tightly coupled control strategy, partial hypotheses are advanced according to high scores attributed to constraints such as prosody, semantics and syntax as well acoustics and phonology. This approach ultimately encourages final hypotheses that are linguistically meaningful, and leads to improvement in understanding as well as recognition. The difficulty in such an integrated approach lies in the expensive computational requirements, involving multiple searches.

This paper will describe some exploratory system designs for integrated conversational systems with near real-time constraints. This work follows on the research described in [5]. Our focus is mainly on the real-time issues demanded of conversational systems. To achieve efficiency, we are integrating multiple-level linguistic constraints into a high-quality acoustic-phonetic network derived from the N-best outputs of the MIT SUMMIT recognizer [2]. We are also interested in achieving domain-independence in the first stage; towards this goal, we report on results when the SUMMIT system is restricted to subword units as its lexicon. We are more concerned with *understanding* than with *recognition* accuracy. Thus we developed and utilized an understanding metric for evaluation.

All experiments are conducted in the Jupiter weather domain. Developed at MIT, this consists of spontaneous, telephone-based inquiries into weather-related information [7]. For the linguistic constraint, we have combined our existing ANGIE and TINA systems into a single framework. ANGIE and TINA are both trainable probabilistic hierarchical models based on context-free grammars [6]. ANGIE models morpho-phonemic and phonological phenomena in a bottom-up multi-layered representation, whereas TINA is the top-down NL framework used in our conversational systems. We have designed a configuration that combines the high quality acoustic modelling of the SUMMIT recognizer with both the bottom-up sublexical constraints of ANGIE and the top-down NL constraints of TINA. Our experiments involve a two-pass system where the N-best outputs of a SUMMIT recognizer are used to construct a compact acoustic-phonetic network. A second search through this network uses a parallel ANGIE-TINA control strategy, as developed in [4].

While this first experiment aims to show the benefits of the combined systems, it is only a preliminary step towards a more flexible system design. We are interested in further exploring a design where the first pass network is restricted to *low level* linguistic and phonological knowledge, while the second pass implements a parallel application of higher order linguistic knowledge and NL. We have begun to move towards this goal in a second pilot study where recognition in the initial pass requires only a lexicon of syllable-like units while word-level linguistic constraints are exclusively enforced in the second ANGIE-TINA pass. We will present some encouraging results from this system.

2. SYSTEM DESIGN

Our ultimate goal is to design a two-pass system where the first pass produces a phone graph based only on acoustic and syllable-level information and the second pass incorporates domain-specific higher level constraints via TINA and ANGIE. In an initial pilot experiment, we used the existing SUMMIT Jupiter recognizer as the first pass, decomposing the resulting N-best list back into a phone graph with associated context-dependent acoustic scores. Unless otherwise indicated, all experiments are based on the *10* best SUMMIT hypotheses, yielding a very com-

[1]This material is based upon work supported by the National Science Foundation under Grant No. IRI-96187321.

pact and high quality acoustic-phonetic network, which becomes the input to a search incorporating ANGIE and TINA (which we refer to as ANGIE-TINA).

2.1. The SUMMIT Recognizer

SUMMIT is a segment-based recognizer, and in the Jupiter domain, it utilizes context-dependent diphone boundary models. In total, there are 68 phonetic units and 631 diphones, including transition and internal units. For acoustic features, 8 different averages of 14 Mel-scale cepstral coefficients measurements are computed from regions within a 150 msec window surrounding each boundary at every 5 msec frame interval. The resultant 112 dimensional vector models the boundary of each diphone. A diagonal Gaussian mixture (using a maximum of 50 Gaussian kernels) is created for each model.

To model phonological variations among words, the system utilizes a pronunciation network generated from the phonetic baseform, and expanded through the application of hand-written phonological rules. The system also incorporates a bigram in a forward Viterbi search that yields the best scoring hypothesis, and a reversed trigram in a backward A^* search, which generates N-best outputs.

2.2. The ANGIE Framework

ANGIE is a system for speech analysis which characterizes word substructure via a multi-layered hierarchical representation. It combines a trainable probabilistic framework with a hand-written context-free grammar. From bottom to top, the layers capture phonetics, phonemics, syllabification and morphology. Also, stress information is embedded explicitly throughout sublexical nodes of the hierarchy, and the phoneme to phone layers govern phonological events.

ANGIE's parser proceeds in a bottom-up, left-to-right manner, advancing column[2] by column. Upon the completion of a word parse, ANGIE yields a linguistic score that comprises log probabilities of each column which, in themselves, are sums of trigram bottom-up probabilities and conditional probabilities for advancing columns. Training is conducted on automatically generated phonetic alignments for a large set of training utterances. ANGIE's phonological rules have been adjusted to match with SUMMIT's rules and phonetic inventory. It has been shown that ANGIE's phone perplexity is lower than that of a phone trigram (see [4]), and we have successfully employed ANGIE's linguistic model in aid of a variety of recognition tasks.

In addition, we have shown that the flexibility of ANGIE extends towards modelling prosodic events. Based on the same underlying paradigm is a statistical hierarchical duration model that accounts for rate of speech effects on durational relationships among sublexical units [1]. The model yields a word duration score that sums log probabilities of node durations throughout the structure.

2.3. The TINA Framework

Like ANGIE, TINA is based on a hand-written context-free grammar but it is augmented by (1) a set of features that enforce syntactic and semantic constraints, and (2) a trace mechanism that handles movement phenomena. Within the TINA parse tree, the probabilities depend on sibling-sibling transitions conditioned on the parent context. The TINA control strategy is implemented in a top-down manner, and an NL score can be generated for the next word candidate given the preceeding partial parse tree. TINA also supports a "robust parse" mechanism, where, in case of failure during full parse, it backs off to retain a partial parse that carries a sentence fragment. In this way, we can achieve meaning representations for agrammatical constructions and errorful recognition hypotheses.

2.4. The Integrated ANGIE-TINA System

The integrated ANGIE-TINA system manages the top-down design of TINA and the bottom-up approach of ANGIE in one top-level procedure which keeps track of partial parses, corresponding with the stack of partial paths, for each component. This configuration grew out of the development of the ANGIE word recognizer [4], which utilizes a stack decoder, an approach well-suited for the application of multiple higher order, long distance language constraints. In this one-pass left-to-right algorithm, the total score, computed at each newly extended partial path, is the sum of the previous path score, an acoustic score for the new phone candidate, and an ANGIE linguistic score for the partial word. When a word ending is hypothesized, a bigram score and word duration score, derived from the ANGIE hierarchical duration scheme, is also added.

With the addition of the NL component, the ANGIE word recognizer hypothesizes a word candidate, and calls upon the TINA parser, which stores a stack of partial parses corresponding to the current path. TINA extends these parses with the word candidate, at which point paths may be eliminated if failure is encountered at every possible parse, or else the mostly likely parse yields a score that augments the total path score. Moreover, an alternative robust parse strategy is implemented, which differs in the original in its reduced computational cost, but similarly handles spontaneous speech phenomena and recognition errors. Details for this implementation are given in [4].

2.5. Morph-based Recognition

In our second set of experiments, we consider relaxing the linguistic constraints in the first-pass system by restricting the lexical information in the SUMMIT front-end to a set of morphological units. By stripping away word-level information in the first pass recognizer, the burden of utilizing linguistic constraints is now shifted towards the second stage. We are interested in discovering the degradation in performance affected in the first pass, and the ability of the ANGIE-TINA strategy to recover that loss. Our ultimate goal is to remove domain dependencies from the first stage, and we recognize that this is only a step in that direction.

The system remains identical, except that the word-based language models of the original system are replaced by newly-trained *morph*-based[3] models. The first pass outputs a 10-best list with *morphs* instead of words. The 1603-sized extended morph lexicon is based on the original 1341 words in Jupiter.

[2] This refers to the nodes along a given path from the root to the terminal.

[3] Morphs are syllable-sized units encoded with linguistic meaning.

Sentence	Key-Value Pair
What is the temperature in boston tomorrow	WEATHER: temperature DATE: tomorrow CITY: Boston
That's all, thanks	CLOSE-OFF: yes

Table 1: *Examples of Key-Value Pairs used in the Understanding Evaluation.*

System	Word Error Rate (%)	Understanding Error Rate (%)
1. SUMMIT Top 1	12.3	19.4
2. SUMMIT N-Best	13.4	17.0
3. ANGIE only	10.4	16.2
4. ANGIE-TINA	11.1	14.1

Table 2: *Comparing recognition and understanding performance among various systems described in Section 4.*

They are the same morphological units that are embedded in the ANGIE parsing mechanism, and are therefore well-matched with the ANGIE probability models. The acoustic-phonetic networks are constructed in the same way as in the preceding experiment, with an identical ANGIE-TINA second pass.

3. UNDERSTANDING EVALUATION

The ultimate goal in building conversational systems is to improve upon overall understanding. This requires the availability of an effective method for drawing comparisons in terms of understanding performance. In our work, we have devised an evaluation measure based on the semantic representation, afforded by the TINA module.

Given a recognition hypothesis as input, TINA generates a parse tree which can be automatically translated to a semantic frame representation. From this, we employ GENESIS [3], our language generation module, to paraphrase the frame into a set of predefined key-value pairs. This set is empirically determined by judging which information in the semantic frame is important in completing the Jupiter-based inquiry. As a result, we have a simpler, collapsed meaning representation that captures only the essential information required by the system to process the inquiry in our domain. Examples of key-value pairs are given in Table 1.

To compute the final understanding error of a test set, we precompute the key-value pairs corresponding to the original orthographies of the set as reference. In cases of parse failures in TINA, this may be due to TINA's incomplete coverage, in which case the transcription is manually rephrased such that a parse can be generated while preserving the original meaning. In other cases, this may not be possible, because a percentage of the utterances lie outside the domain; that is, the spoken requests cannot be handled by the system, and no alternative phrasing would be interpretable by the dialog component. These reference key-values are deemed missing. The final understanding error is a percentage calculated from the total number of mismatches, deletions and insertions against the reference key-values. For missing key-values, in either the reference or hyptheses, deletions are counted.

Because this TINA module is identical to the NL module deployed in the real-time system, we believe that the evaluation method is a fair reflection of overall understanding performance. It simulates the situation where the system in evaluation is integrated with the dialog module; utterances with mismatched key-values would be interpreted erroneously by the dialog component of the real system; that is, a different action would result.

4. INTEGRATION EXPERIMENTS

For baseline comparison, we use the SUMMIT system, which outputs an N-best list for our subsequent experiments. We consider the system performance under two modes of operation: (1) SUMMIT Top 1: the best scoring candidate is chosen and (2) SUMMIT N-Best: a rudimentary algorithm is used to choose, from the N-best list, the most likely utterance where a meaning representation can be obtained. The latter mode is used in our real-time system, and is implemented with our TINA NL parser as a post-processor. We report on the successive performance gains of the SUMMIT system from augmenting with (3) ANGIE alone, and with (4) ANGIE-TINA fully deployed. Systems (1*m*), (3*m*), and (4*m*) are the morph-based counterparts of (1), (3), and (4). All experiments are evaluated on an unseen test set of 352 utterances.

The Jupiter system utilizes a 1341-sized word lexicon. Within this lexicon, some commonly occurring adjacent words are treated as a single word, e.g., "what is," for added constraint[4]. The ANGIE probabilistic grammar and the hierarchical duration model are both trained on 11677 utterances. The TINA word grammar is separately trained on 6531 utterances.

5. RESULTS AND ANALYSIS

We will begin by reporting results for integration experiments using the word-based SUMMIT system, followed by results for the morph-based experiments. Recognition and understanding errors for the Systems 1–4, mentioned above are reported in Table 2. When ANGIE is applied, the word error rate reduces by 15.4% (from 12.3% to 10.4%) compared with the baseline System 1. This system (without any NL) achieves an understanding error of 16.2% which improves upon the NL processing of System 2 (17.0%). When ANGIE-TINA is fully integrated, the word error rate of 11.1%, improves upon both System 1 and 2, and outperforms each one in terms of understanding error, with a value of 14.1%. This is a 17.1% error reduction relative to the NL post-processing of System 2.

It is clear from these results that, firstly, system performance benefits significantly from the combined probabilistic sublexical models of ANGIE and its duration model. The inclusion of compound words enables ANGIE to incorporate the inter-word phonological effects and pronunciation variations probabilistically in the sublexical parse structure, and we believe this has contributed to enhancing performance. Secondly, an integrated ANGIE-TINA achieves superior understanding performance via a search strategy that enables meaningful partial paths to proceed. It is also apparent that word error rate does not necessarily fall

[4]These units are retained in our morph experiments.

System	Word Error Rate (%)	Understanding Error Rate (%)
3*m*. ANGIE only	11.8	18.1
4*m*. ANGIE-TINA	13.9	17.3

Table 3: *Comparing recognition and understanding performance among various systems which use a morph lexicon in the* SUMMIT *front-end.*

System	Morph Error Rate (%)
1. SUMMIT Top 1	10.8
1*m*. Morph SUMMIT Top 1	12.8
3*m*. ANGIE only	10.9

Table 4: *Morph error rates for selected systems.*

with understanding error, and this is particularly relevant in considering the underlying goal of improving understanding in the design of conversational systems.

When N is raised to 100, the ANGIE System 3 does not improve significantly, although for the ANGIE-TINA System 4, the understanding error improves to 13.6%. We can conclude that the ANGIE-TINA guided search retrieves a greater number of correct paths from the deeper network.

Final results for the morph-based experiments are tabulated in Table 3. It can be observed that the word error rate (11.8%) in System 3*m* outperforms that of System 1 (12.3%). Similarly, for System 4*m*, understanding performance of ANGIE-TINA (17.3%) is comparable to that of System 2 (17.0%). From this, we infer that the sophisticated language models of ANGIE and ANGIE-TINA recover most of the loss in performance incurred by the morph lexicon.

We gauge the drop in performance from switching to morphs by comparing the morph error rates for the morph-based best-scoring SUMMIT output (1*m*) to the original best-scoring word-based SUMMIT (1) and the final integrated ANGIE only word output (3*m*), when given in terms of morph accuracy, as shown in Table 4. There is an 18.5% degradation in using morphs (from 10.8% to 12.8% error) but this is largely recovered even when using ANGIE alone with 10.9% error.

It should be noted that in comparing the word-based systems against their morph-based counterparts, the former utilize a word trigram within SUMMIT whereas the latter do not employ word trigrams, and instead rely solely on a bigram and ANGIE-TINA, at the word level. We claim that performance would further improve if a trigram were incorporated.

6. SUMMARY AND FUTURE WORK

The above results have shown that (1) an integrated SUMMIT with ANGIE-TINA benefits both recognition and understanding, (2) replacing words by morphs in SUMMIT still achieves a workable performance, and (3) ANGIE-TINA models are sufficiently powerful to recover those losses, demonstrating the potential for shifting the application of higher level language constraints towards the integrated ANGIE-TINA approach. However, these experiments remain preliminary, and much work remains to be completed. Some of these are ideas are outlined below.

Currently, the morph lexicon retains much word-specific information which contributes to a strong performance in the front-end. However, we envision a future system that relies more on the second-pass search for domain-specific constraint. It is conceivable that the high quality acoustic pruning of the first pass could operate on a domain-independent syllable lexicon trained on a generic English corpus, and the second pass would incorporate a vast array of domain-dependent linguistic information in a fast and intelligent search. At this stage, we have repeated the above morph experiment with a further reduced 1250-sized syllable lexicon, and results similarly indicate that, while the reduced constraints in SUMMIT produce some marginal degradation, ANGIE-TINA recovers much of the degradation incurred.

We believe that reconstructing an acoustic-phonetic network provides a much richer search space than the alternative of simply processing the top N hypotheses. Through cross-pollination effects, the second pass search may potentially traverse new and improved paths which are favored by ANGIE-TINA scores. However, from a design standpoint, this configuration is ultimately suboptimal. It is also particularly slow when the value of N is as large as 100. Therefore, we hope to entirely replace the N-best paradigm with a more efficient and direct methodology for achieving a highly pruned space that enables real-time computation associated with the complex parallel control strategies required by ANGIE-TINA, and where the size of the network can be varied with flexibility.

7. REFERENCES

1. G. Chung, *Hierarchical Duration Modelling for a Speech Recognition System*, S.M. thesis, MIT Department of Electrical Engineering and Computer Science, 1997.

2. J. Glass and T. J. Hazen, "Telephone-based Conversational Speech Recognition in the Jupiter Domain," *These proceedings*.

3. J. Glass, J. Polifroni and S. Seneff, "Multilingual Language Generation Across Multiple Domains," *Proc. ICSLP '94*, pp. 983–986, Yokohama, Japan, Sept. 1994.

4. R. Lau, *Subword Lexical Modelling for Speech Recognition*, PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institue of Technology, Cambridge, MA, May 1998.

5. R. Lau and S. Seneff, "A Unified System for Sublexical and Linguistic Modelling Using ANGIE and TINA," These proceedings.

6. S. Seneff, "The Use of Linguistic Hierarchies in Speech Understanding," *These proceedings.*

7. V. Zue, S. Seneff, J. Glass, L. Hetherington, E. Hurley, H. Meng, C. Pao, J. Polifroni, R. Schloming, P. Schmid, "From interface to content: translingual access and delivery of on-line information," in *Proc. Eurospeech '97*, Rhodes, Greece, pp. 2227–2230, Sept. 1997.