Statistical Modeling of Phonological Rules through Linguistic Hierarchies

Stephanie Seneff and Chao Wang

Affiliation:

Spoken Language Systems Group

Computer Science and Artificial Intelligence Laboratory

Massachusetts Institute of Technology

Send Correspondence to:

Stephanie Seneff

200 Technology Square, Room 643

Cambridge, Massachusetts 02139 USA

email: seneff@csail.mit.edu

or

Chao Wang

200 Technology Square, Room 639

Cambridge, Massachusetts 02139 USA

email: wangc@csail.mit.edu

Number of pages: 42

Number of tables: 2

Number of figures: 4

Keywords: pronunciation modeling, phonological modeling, speech recognition, dialogue systems.

Abstract

This paper describes our research aimed at acquiring a generalized probability model for alternative phonetic realizations in conversational speech. For all of our experiments, we utilize the SUMMIT landmark-based speech recognition framework. The approach begins with a set of formal context-dependent phonological rules, applied to the baseforms in the recognizer's lexicon. A large speech corpus is phonetically aligned using a forced recognition procedure. The probability model is acquired by observing specific realizations expressed in these alignments. A set of context-free rules is used to parse words into substructure, in order to generalize context-dependent probabilities to other words that share the same sub-word context. The model maps phones to sub-word units probabilistically in a finite state transducer framework, capturing phonetic predictions based on local phonemic, morphologic, and syllabic contexts. We experimented within two domains: the MERCURY flight reservation domain and the JUPITER weather domain. The baseline system used the same set of phonological rules for lexical expansion, but with no probabilities for the alternates. We achieved 14.4% relative reduction in concept error rate for JUPITER and 16.5% for MERCURY.

1 Introduction

In the early years of speech recognition research, it was believed that an important contribution to success would be the application of formal rules to account explicitly for predictable phonological reductions in certain contexts in conversational speech (Cohen 1989; Gauvain et al. 1993; Weintraub et al. 1989; Zue 1983). Some examples of such rules are shown in Figure 1[1] – rules which predict, for example, the contexts for flapping an alveolar stop or for palatalizing an alveolar fricative.

(insert Figure 1 here)

As the hidden Markov model (HMM) framework gained in popularity, formal rules tended to play a less prominent role. Instead, the assumption was made that context effects could be handled via context-dependent acoustic models, thus accounting for the variabilities in the Gaussian mixtures associated with these typically triphone models. In part, this shift was predicated on an admission that perhaps we do not understand the rules sufficiently well to formally characterize them.

With the recent trend towards a shift of attention from read to spontaneous speech, such as the Switchboard corpus (Godfrey et al. 1992), the issue of accounting for phonetic variability has resurfaced as a significant research problem (Strik and Cucchiarini

[1]We will use the ARPABET convention for all phone representations in this paper. See http://www.isip.msstate.edu/projects/switchboard/doc/education/phone_comparisons/ for a detailed description of the symbol set.

1999). For example, a "cheating" experiment reported in (McAllaster et al. 1998) showed that the word error rate for a Switchboard test corpus could be reduced from 40% to 8% by explicitly accounting for the phonetic pronunciations in the lexicon according to their actual realizations in the test corpus. However, many researchers have shown that an overabundance of alternative lexical pronunciations without any attempt to model their relative likelihoods can lead to a degradation in performance, due to the increased chance of erroneously matching obscure alternatives. Furthermore, the technique of simply enumerating individual variants for each word in the lexicon is tedious and shows little generality.

In the recent resurgence of interest in phonology, data-driven approaches have played a much stronger role than in the early days, when knowledge-based approaches dominated the literature. One popular technique, as exemplified by the work of (Cremelie and Martens 1999), is to allow for a generous set of confusions, including substitutions, deletions and insertions, and then use a forced recognition mode to search the expanded space for alternative, better scoring, deviations from the canonical lexical forms. A large set of "rules" with associated probabilities can then be gleaned from the observation space. Such an approach is attractive in that it is able to generalize from observed words to unobserved words with similar surrounding phonetic context, a technique that is related to the strategy we adopt here.

An interesting approach towards generalizing observed phonological expression from seen to unseen words is presented in (Tajchman, Fosler, and Jurafsky 1995). For a set of ten phonological rules accounting for vowel reduction, syllabification of nasals and liquids, alveolar stop flapping rules, and /h/ devoicing rules, they developed an iterative algorithm that could determine the probability that each rule applies. Basically, rule firings were tabulated and pooled among pronunciations of all words, subsequently normalizing scores such that each word's total probability is 1.0. They demonstrated dramatic improvements in recognition error rate on the Wall Street Journal task, if the additional critical step was taken to prune away realizations where probability was below a threshold proportional to the best scoring pronunciation of that word.

Phonologists have long been aware that syllable structure plays an important role in predicting phonological reductions (Kahn 1980). For example, consonants in syllable onset position are far less likely to be reduced than in coda position. An excellent description of the role of the syllable in the Switchboard data has been presented by (Greenberg 1999). He showed, through studies on a large corpus of hand transcribed Switchboard data, that 28% of consonants in *coda* position were deleted. Furthermore, not only syllable structure, but also morphological information, is important in characterizing phonetic expression.

In the research reported here, we address the problem of accounting for phonological variations in conversational speech through an approach that combines formal knowledge via phonological rules with automatic data driven methods. Our methodology begins with a set of formal phonological rules which are used to expand a set of lexical entries into alternative pronunciation variants. A large speech corpus is phonetically aligned, and a parsing mechanism is applied to these alignments to capture the likelihoods of the alternative pronunciations. The probabilities take into account a large number of factors, including syllable position, stress, phonetic context, and even morphology, such as function versus content word. The relevant factors are obtained by parsing each word in the training corpus using a carefully constructed context-free grammar. The probability model is superimposed on the parse tree, and is chosen so as to best capture the relevant conditioning factors while minimizing sparse data problems. It is also configured so as to specifically predict only the expressed phonetic productions, without inadvertently learning undesirable phonotactic information. This point is important because it permits generalization from a common word to a rare word with the same local sub-word context. More generally, it permits training on speech data from one domain and testing on data from another domain where available training material may be sparse or non-existent.

Our research is based on the SUMMIT *landmark*-based speech recognition system (Glass 2003). While SUMMIT's approach is quite distinct from the standard

HMM formulation, it has produced state-of-the-art performance in phonetic recognition tasks (Chang and Glass 1997; Halberstadt and Glass 1998). Our research is concerned almost exclusively with telephone-quality conversational speech, collected through interactions between users and various domain-specific conversational systems (Seneff and Polifroni 2000; Glass and Hazen 1998; Seneff et al. 2000).

This paper is an extension of the work described in (Seneff and Wang 2002). We report here on evaluations in the JUPITER weather domain in addition to the MERCURY flight domain, and based on two different phonological rule sets: a *full* set and a *reduced* set, restricted to rules that would lead to insertions or deletions in the phonetic realizations. These extensions are directly motivated by the experiments described in (Hazen et al. 2002) which showed that a *reduced* phonological rule set, along with an appropriate replacement of the acoustic models, can lead to an *improvement* in recognition performance. It therefore becomes important to assess whether the performance of the system with the reduced rule set can be further improved through the use of statistics learned from an observation space, or, alternatively, whether the addition of probabilities to the full rule expansions can close the gap between the two strategies.

In the remainder of this paper, we first describe SUMMIT, including its finite state transducer (FST) formulation and phonological modeling framework (Hetherington

2001). Section 3 describes our phonological probability model, which uses the ANGIE system (Seneff et al. 1996; Seneff 1998) to obtain sub-word linguistic hierarchies. Section 4 describes the two-step process of acquiring the trained FST mapping phones to unique sub-word contexts, and explains how the recognizer is reassembled to incorporate the acquired probability model. Section 5 provides the details of our experiments in both the MERCURY and JUPITER domains and reports on our results for both domains, for both speech recognition and understanding. Section 6 concludes with a summary and a discussion of future plans.

2 SUMMIT System

In a landmark-based approach, it is more critical to capture phonological rules than in a frame-based approach, particularly rules that would lead to the deletion or insertion of a landmark. Typical rules include, for example, epenthetic silence insertion at locations of voicing change, gemination ("fro*m* *M*aine") and palatalization ("ga*s* *sh*ortage") rules, and rules accounting for unreleased stops or even wholly deleted stops, as in "wanna" for "want to." There are also devoicing rules for fricatives and stops, and various vowel reduction rules. (Hazen et al. 2002) provide a detailed description of pronunciation variations and how they are modeled in the SUMMIT system. Example rules are shown in Figure 1.

In the SUMMIT system, landmarks are established based on spectral change. Each landmark is considered either as a boundary between two phones, or as a phone-internal event, and is scored using standard Mel-scale Cepstral coefficients describing the region surrounding the landmark. Words are entered in the lexicon according to their idealized phonemic pronunciations, and are expanded into alternative pronunciations according to an ordered set of phonological rules, accounting for phenomena such as flapping, gemination, and palatalization. The expanded lexicon is combined with language models and used in guiding lexical access during the search for words. In typical applications, the search produces an *N*-best list of hypotheses to be considered by later stages in a dialogue system.

SUMMIT uses finite state transducers (FSTs) to represent the acoustic, phonological, lexical, and grammar constraints. The search space is organized as a cascade of FSTs:

$$C \circ P \circ L \circ R \circ G \tag{1}$$

where C maps context-dependent acoustic model labels on its left to context-independent phone labels on its right, P maps phones to phonemes by applying the phonological rules, L is the lexicon mapping idealized phonemic pronunciations to spoken words (e.g., "I," "would," and "I'd"). R is a set of rewrite rules for reductions and contractions, mapping spoken words to their corresponding canonical forms (e.g., "I'd" $\rightarrow$

"I would | I had"), and G is the language model[2].

The phonological component, $P \circ L$, is a heavily shared but *unweighted* network mapping phones to words. Its FST-based implementation is described in detail in (Hetherington 2001).

3 Sub-word Probability Model

An important feature of our phonological model is that the prediction of phonetic expressions utilizes a hierarchy of sub-word units, specifying phonemic, morphologic, and syllabic contexts. In this section, we describe the details of the probability model. We start by describing how we acquire the sub-word contexts and the associated probabilities using the ANGIE system. We then show how the ANGIE model is transformed into the phonological model. Finally, we discuss issues having to do with word boundary effects, anticipatory assimilation, and non-speech events.

3.1 ANGIE Framework

Over the past several years, we have been exploring the utility of a parsing framework we call ANGIE (Seneff et al. 1996; Seneff 1998) for modeling word substructure.

[2]Please refer to (Hazen et al. 2002) for a detailed description of P and R.

The original intent was to model phonology, morphology, and syllable constraints in a shared probability framework, with the goal of modeling formal structure of the language in the absence of a known lexicon. The ANGIE utility has a wide range of application areas, including letter-to-sound/sound-to-letter systems (Chung and Seneff 2002; Seneff 1998), a high-performance phonetic recognizer as the first stage in a multi-stage recognition framework (Chung 2000), and automatic enrollment of new words into dialogue systems (Chung et al. 2003; Seneff et al. 2003). In its usage here, we are attempting to translate the ANGIE probability model into a finite state transducer mapping phones to phonemic units, in order to attach probabilities to arcs in a lexical phone graph.

In ANGIE, a parse tree is obtained for each word by expanding the rules of a carefully constructed context-free grammar, intentionally designed such that every parse tree lays out as a regular two-dimensional grid, as shown in Figure 2. Each layer is associated with a particular aspect of sub-word structure: migrating from morphemics to syllabics to phonemics to phonetics at the deepest layer.

(insert Figure 2 here)

To aid in the parsing process, ANGIE utilizes a lexical encoding that is represented in two tiers – words are entered as sequences of "morphs," essentially syllable-level units encoding the spelling in their name and marked for positional and stress in-

formation; a separate lexicon defines phonemic pronunciations for the morphs. The morph specifications in the lexicon are enforced by the parsing process, to assure accurate parsing of a corpus. We currently distinguish for English a small set of unique morph classes, namely, stressed and unstressed prefix ("`spre`", and "`pre`"), stressed and unstressed root ("`sroot`" and "`uroot`"), function word ("`fcn`"), and two kinds of suffix ("`dsuf`" and "`isuf`")[3]. The context-free rules encode positional constraints for the morph units: in our current model, a content word must contain at least one "`sroot`," and other morphs must obey the following sequential constraints relative to the "`sroot`":

```
[spre] [pre] sroot [uroot] [dsuf] [isuf]
```

Figure 3 illustrates ANGIE's two-tiered lexicon for selected words. Part (a) shows the representation of words in terms of morphs, part (b) shows the phonemic representation of the morphs, and part (c) provides the baseforms lexicon that is derived automatically from the word and morph lexicons. ANGIE's phoneme inventory is enhanced to include markers for stress (+) and onset (!) position, as well as some diphone units such as /st!/ representing an onset "st" cluster. The baseforms lexicon is used to link the ANGIE derived phonological model back into the lexicon in the FST formulation of the recognizer search space, as will be more fully described in

[3] "`Dsuf`" roughly corresponds to "derivational suffix," and "`isuf`" to "inflectional suffix." But they do not necessarily follow strict conventions for these terms, for pragmatic reasons.

Section 4.

(insert Figure 3 here)

Although the rules are context free, context dependencies are captured through a superimposed probability model. The particular choice for the probability model was motivated by the need for a balance between sufficient context constraint and potential sparse data problems from a finite observation space. We were also motivated to configure the probability model such that it would be causal, with strong locality, for practical reasons having to do with the nearly universal left-to-right search path in recognition tasks, with the ultimate goal of attaching the learned probabilities to arcs in a finite state network.

We were thus motivated to view the parse tree as a series of parse *columns*, where each column is encoded as an ordered set of tags, identifying the label of each node from the terminal phone to the root "sentence" node. At each advance in time, the goal is to predict the next column given just the previous column, but a strict column-column formulation would suffer from serious sparse data problems. Hence we decided to decompose the column into a set of locally conditioned probabilities, applied to each column entry. We restricted context to include only the left sibling and child for each entry in the parse matrix, with the exception of the terminal phone, which is conditioned on the entire preceding column.

Thus, the probability formulation we have developed for ANGIE can be written as follows:

$$\mathrm{p}(C_i|C_{i-1}) = \mathrm{p}(a_{i,0}|C_{i-1}) \prod_{j=1}^{N-1} \mathrm{p}(a_{i,j}|a_{i,j-1}, a_{i-1,j}) \tag{2}$$

where C_i is the i^{th} column in the parse tree and $C_i = \{a_{i,j}, 0 \leq j < N\}$, and $a_{i,j}$ is the label at the j^{th} row of the i^{th} column in the two-dimensional parse grid. N (equal to 6 in our experiments) is the total number of layers in the parse tree. In practice, the column probability computation ends once the column merges with its left sibling, i.e., when $a_{i,j-1}$ has the same name as $a_{i,j}$. As illustrated in Figure 2, the column index i begins at the left of the parse grid, and the row index j begins at the bottom of each column.

Each terminal symbol is predicted based on the entire preceding column, and the column probability is built bottom-up based on a trigram model, considering both the child and the left sibling in the grid. The probabilities, $\mathrm{p}(a_{i,0}|C_{i-1})$ and $\mathrm{p}(a_{i,j}|a_{i,j-1}, a_{i-1,j})$, are trained by tabulating counts in a corpus of parsed sentences, mapping words to their corresponding phonetic realizations.

3.2 Phonological Probability Model

The ANGIE model intentionally captures both phonological and phonotactic aspects of the language (e.g., how often a word begins with /f/). However, for the purpose of modeling the likelihood of the phonological variants, the phonotactic contribution to the probability model should be removed. Our intended goal is to predict the *phonetic* realization given the *known* phoneme sequence of the word. We are making the simplifying assumption that each word, in a specific phonetic realization, has a unique parse. While the original ANGIE formulation jointly predicts both the terminal phone and the column above, for this application we are interested instead in predicting the phone given *both* the entire left column and the column above. Our phonological model (PM) can thus be formulated as follows:

$$PM = \mathrm{p}(a_{i,0}|C_{i-1}, \{a_{i,j}, j > 0\}) \tag{3}$$

Using the Bayesian formulation for conditional probability, $a_{i,0}$, the *phonetic* realization associated with the right column, can be expressed as the probability of the entire right column normalized by the total probability of the upper column, given the left column:

$$PM = \frac{\mathrm{p}(a_{i,0}, \{a_{i,j}, j > 0\}|C_{i-1})}{\mathrm{p}(\{a_{i,j}, j > 0\}|C_{i-1})} = \frac{\mathrm{p}(C_i|C_{i-1})}{\mathrm{p}(\{a_{i,j}, j > 0\}|C_{i-1})} \tag{4}$$

The denominator in Equation 4 can be computed as the marginal probability of the joint probability $\mathrm{p}(a_{i,0}, \{a_{i,j}, j > 0\}|C_{i-1})$ summing over all instances of the terminal phone. Let u_k denote a phonetic realization of $a_{i,0}$, then:

$$\mathrm{p}(\{a_{i,j}, j > 0\}|C_{i-1}) = \sum_k \mathrm{p}(a_{i,0} = u_k, \{a_{i,j}, j > 0\}|C_{i-1}) \tag{5}$$

where k is indexed over all possible realizations of the terminal phone given the upper column.

Thus the ANGIE column bigram probability is summed over all observed instances of $a_{i,0}$ associated with each unique upper column, $\{a_{i,j}, j > 0\}$, to determine the upper column's *total* probability conditioned on the left column. This sum then becomes the denominator to normalize the column bigram probability.

Thus, for instance, a /t/ in "butter" realized as a flap ([dx]) and one realized as a closure ([tcl]) would both contribute to the denominator, and so would instances of these realizations in other related words such as "shuttle" and "stutter."

3.3 Modeling Word-Boundary Effects

A critical aspect in phonological modeling is the effective capturing of word-boundary effects. ANGIE's probability model retains only the last *phone* of the preceding word as the left-context condition for the first phone in the word, in order to ameliorate potential sparse-data problems. But even with this backoff condition, it is still impossible to assure that every phone-column transition possible at word onsets is observed in every phone context[4]. Furthermore, without the need to preserve language model information, it seems counterproductive to condition *all* phoneme-to-phone mappings on the left phone, when many such mappings are not particularly sensitive to left context. Only a small subset of the word-onset realizations are strongly tied to left context. For example, a gemination rule supporting deletion of the onset phone clearly needs to know context, whereas a rule mapping onset phoneme /l!/ to phone [l] is very general.

Our solution was to implement the capability to specify in the ANGIE grammar an explicit set of phoneme-to-phone mappings which, if appearing word-final, should be liaison'ed to the subsequent word's onset phone. For example, if /s/ is realized as [sh], then the following word should obligatorily start with a palatal (/sh/, /zh/, /y/, etc.) The observations control the actual set of word-boundary ties that are

[4]In principal, this problem exists word internally as well, but is much less prevalent there.

sanctioned. This mechanism effectively retains right context dependency constraints across word boundaries. All non-liaison'ed word-ending phones map to a generic word-start node, which then advances to completely context-independent realizations of word-start columns.

3.4 Anticipatory Assimilation

There is a further problem with modeling the anticipatory phonological effects that are prevalent in fluent speech. While, in the interest of preserving causality, it is customary to formulate language models to be forward predictive, humans are planning the future phonetic stream as they pronounce words, and they tend to anticipate an articulatory gesture in advance of the phoneme that requires this gesture. In these cases, a *backward* predictive model would be more appropriate. Considering the palatalization rule discussed above, it is evident that the production of /s/ as [sh] in "thi*s* ship" has little to do with the phoneme /ih/ of "th*i*s." The original probability model adopted in the ANGIE framework is inappropriate for this circumstance.

Fortunately, since we have already incorporated a mechanism to identify word-ending contexts that are tied to the observed subset of subsequent word-beginnings, it becomes feasible to model *just these explicitly identified situations* with a *backward* predictive statistic. This can be done by simply tabulating counts on appropriate

observations. The observation space is defined as exactly those right *phoneme* conditions that were observed when the liaison'ed left phone occurred. For /s/ realized as [sh], the set would likely consist of /jh/, /ch/, /y/, /sh/, and /zh/. A tabulation of *all* observed phonetic realizations of the /s/ phoneme in the context of this set of next-word onsets would define the denominator, where the numerator consists of just those cases where /s/ was realized as [sh]. This will yield a measure of the probability that /s/ is palatalized when a palatal occurs on the right.

3.5 Modeling Non-Speech Events

In speech recognition for spoken dialogue interaction, it is absolutely essential to provide some kind of explicit modeling of non-speech events. Effective modeling of the acoustic characteristics of coughs, laughter, and other kinds of noise can lead to a huge improvement in recognition accuracy (Hazen et al. 2002). In SUMMIT, five different categories of noise are explicitly identified: <laughter>, <coughing>, <hang_up>, <background>, and a catch-all category, <noise>. Several distinct acoustic models have been acquired to represent *each* of these noise classes, and the pronunciation models support a fully interconnected graph connecting the acoustic models associated with each class. Thus, for example, there are four different acoustic models for <cough>, and any particular cough event can be accounted for by any

number of these models in any sequence.

To model these noise events in the ANGIE framework, we provided a single phonemic label as the lexical pronunciation for each noise event (e.g., "cc" for "cough"). All of the sequential observations are captured by a model mapping this unique "phoneme" to its corresponding "phones," which are licensed by the rules to occur in any order. Observation counts are updated using standard procedures, and, as a result, the final statistical model accounts for their observed sequential occurrence patterns.

4 Experimental Methods

The incorporation of our phonological model into speech recognition is straightforward within the FST framework. In this section, we describe the two-step process of acquiring a trained FST mapping phones to unique sub-word contexts, and explain how the recognizer is reassembled to incorporate the acquired probability model.

4.1 Training Procedure

We have discussed the ANGIE framework for modeling word structure, and we have shown how ANGIE's probability model can be reconfigured to support prediction of

each subsequent phone, given the entire previous column and the column above the new phone. Now we will describe how a finite state transducer encoding this probability model is obtained through a cooperative interplay between SUMMIT and ANGIE.

The training procedure begins with a large corpus of orthographically transcribed utterances. These are first processed through standard SUMMIT alignment tools to produce aligned phonetic transcriptions, honoring the context-dependent phonological rules specified in SUMMIT. The aligned word and phonetic transcriptions are then used to train the probabilities in an ANGIE grammar, which is designed to support parsing of all of the variants appearing in the training corpus. In practice, the ANGIE rules need only cover all *possible* alternative realizations of each phone, without regard to surrounding context conditions. The restriction to SUMMIT's phonological space will guarantee that all *observations* honor the dependencies, and the probability model will therefore learn the context conditions from the data.

Once the ANGIE grammar has been trained, a second pass through the data computes the column-column transition probabilities given the trained grammar, and normalizes each column prediction, following Equation 5, to remove phonotactic dependencies. The resulting probability model is written out as a finite state transducer, P_A, with phones as the input symbols and phonemes in ANGIE's preterminal layer as the output symbols. In addition, at each advance to a new syllable, the syllable layer

symbol (encoding stressed root, function word, prefix, etc.) is emitted, which has the desired effect of preserving the distinct statistics of these syllable types.

4.2 Assembling the Speech Recognizer

We have described a procedure to create the finite state transducer, P_A, mapping phones to ANGIE's phoneme units, with arc probabilities reflecting the generalized observation space. Now we will describe how it is incorporated into the SUMMIT recognition framework, to be combined with a word lexicon and the n-gram language model.

As mentioned in Section 2, SUMMIT uses a phonemically based lexicon, which is then expanded into unweighted phonetic pronunciations by utilizing formal phonological rules. We replaced this lexicon with a new set of baseforms that reflect ANGIE's phoneme layer symbol set. In addition, the syllable-identity symbols are inserted at the end of each syllable, consistent with P_A. In all, there are about 150 unique phoneme units, as contrasted with 78 in the original SUMMIT configuration.

ANGIE's baseforms are converted into an FST, L_A, which maps sequences of ANGIE phonemic units to words. P_A is then composed with L_A, to produce a transducer, $P_A \circ L_A$, mapping phones to words with weights on the arcs. The rest of the recognizer

constraints, C, R, and G, are kept the same[5], and the search space is constructed in the same way as described in Equation 1.

Figure 4 illustrates a portion of the lexical network representing the alternative pronunciations along with associated probabilities for the word "Atlanta." The initial vowel prefers to be realized as [ae], but can be reduced to a schwa; the first /t/ can be either a flap or a closure with an optional release; the second /ae/ can be realized as an [aa] but with high cost; and the final /t/ can be deleted altogether. It should be noted that, although the ANGIE system and the baseline system begin with the same set of phonological rules, ANGIE generates a richer network due to generalizations of alternative pronunciations that were hard coded in SUMMIT's lexicon. Thus the two alternatives for the /ae/ in "Atlanta" can be explained through a generalization from the alternatives (/aa/ or /ae/) for "aunt" specified in SUMMIT's lexicon.

(insert Figure 4 here)

5 Evaluation Experiments and Results

In (Seneff and Wang 2002), we reported on experiments that were conducted within the MERCURY flight reservation domain, where a 12% improvement in concept error

[5]Note that the acoustic models are identical to those in the baseline system.

rate was achieved, when benchmarked against a system that utilized the same set of phonological rules, but without probability support. In those experiments, we used a set of phonological rules that had become the standard set for SUMMIT at that time, containing about 165 generalized rules. However, in parallel with our research, others in the Spoken Language Systems group were conducting research comparing this so-called "full" set of rules with a "reduced" set, containing only 64 rules, mainly involving insertions and deletions (Hazen et al. 2002). They concluded that substitutions are best accounted for in the acoustic models rather than in the phonological rules. A baseline system in their experiments used no rules[6], and obtained inferior performance to both systems with the "full" or "reduced" set of phonological rules.

They also attempted to train probabilities on the rules for both the full set and the reduced set conditions, using an EM algorithm (Shu and Hetherington 2002), but were only able to obtain significant improvements for the full set, which however was still inferior to the system with the reduced set but without probability training. These experiments motivated us to attempt to train probabilities for the "reduced" rules condition, using the hierarchical modeling techniques. We decided to explore both the JUPITER and the MERCURY domains, so that our experiments could be better benchmarked against the experiments described in (Hazen et al. 2002).

In the following, we first give a brief description of the data and the baseline

[6]except for obligatory expansion of stops to include separate closure and burst segments.

speech recognizers used in our experiments. We then report both speech recognition and understanding performances in these two domains under the two conditions, and discuss some computational issues.

5.1 Evaluation Data

To demonstrate the viability of this approach, we first prepared a lexicon of over 7,000 words that had been developed to cover a training corpus used for prior SUMMIT experiments. We represented these words in terms of component morphs, and supplied phonemic pronunciations for the morphs using ANGIE's phoneme set. It required a total of 5,254 morphs to account for all the words.

The next step was to acquire a training corpus of aligned transcriptions, for a set of over 130,000 utterances obtained through user interactions with various systems developed in our group, including JUPITER and MERCURY. We developed a context free grammar that could parse nearly all utterances in the corpus, by adding phoneme-to-phone mappings to license the observed mappings. We established two sets of terminal rules, reflecting the "full" and "reduced" phonological rule sets, respectively. We separately trained the probabilities for each grammar by parsing aligned corpora obtained by running the standard SUMMIT recognizer in forced alignment mode, but with the two distinct phonological rule sets. The results of this effort were two

different FST's representing the full and reduced rule conditions. The MERCURY and JUPITER domains share the same FST, since any domain-dependent language model effects have in theory been removed.

During evaluation, about 4500 JUPITER utterances were used as held-out data for tuning various parameters. The final results were reported on 1742 unseen test utterances in the JUPITER domain, and 1669 test utterances in MERCURY.

5.2 Baseline Recognizers

For both JUPITER and MERCURY, the baseline system and the ANGIE system differ only in the pronunciation models: they use the same set of vocabulary, class bigram and trigram language models, the diphone-to-phone mapping FST, diphone acoustic models, as well as a segment duration model (Livescu and Glass 2001)[7]. The JUPITER recognizer has about 2150 unique words in its vocabulary, and the MERCURY recognizer has about 1830. The baseline system uses unweighted pronunciation networks, while the ANGIE system has probabilities for alternative pronunciations trained using the method described in the previous sections. Various parameters for these two systems, such as word and phone transition weights, and the weight of the ANGIE

[7]It is possible that iterative training of the acoustic models would further improve the ANGIE system, but by holding them constant it becomes clearer that the phonological probability modeling is the source of any improvements observed.

pronunciation FST, P_A, were tuned on development data, and the final results are reported on unseen test data.

5.3 Evaluation Results

5.3.1 Speech Recognition

Table 1 summarizes the recognition performance of the baseline and the ANGIE systems for test sets in the JUPITER and MERCURY domains. First of all, it should be noted that the JUPITER "full" baseline has a substantially higher recognition error rate than the "reduced" baseline, consistent with the results reported in (Hazen et al. 2002). Interestingly, however, the probability training was able to improve the performance on the "full" set by much more than on the reduced set, with the consequence that the trained system's "full" performance is slightly *better* than its performance on the reduced rule system. For the MERCURY domain, substantial recognition gains were achieved for both the full and reduced sets, although the gap in performance between them was reduced somewhat.

(insert Table 1 here)

5.3.2 Speech Understanding

For spoken dialogue systems, speech understanding performance is a more significant metric than speech recognition performance. In this regard, we also evaluated the *concept error rate* when the recognizer is used with a natural language understanding system to produce a meaning representation, encoded as a set of [`key:value`] pairs. The [`key:value`] pairs obtained by parsing the N-best list are compared against those obtained by parsing the orthographic transcription, and the concept error rate was computed in a similar way as the word error rate. Out of the entire test set, we are able to parse 87.2% of the utterances (full parse or robust parse) in the JUPITER domain, and 90.4% in the MERCURY domain. The rest of the utterances failed because they are out-of-domain or incomplete, or because of gaps in the parse coverage. They are excluded from this evaluation due to the lack of reference [`key:value`] pairs. Table 2 summarizes the concept error rates on the parsed subset for the two domains under various conditions. For three out of the four test conditions, concept error rate was reduced by more than 14%. This was a substantially greater improvement than was realized for speech recognition.

(insert Table 2 here)

5.3.3 Computational Issues

One issue that arose in the EM training methods explored by (Shu and Hetherington 2002) was the explosive growth in the size of the resulting FST. The FST that resulted from training word-dependent pronunciation weights led to 50 times more transition arcs than in the baseline configuration. In contrast, by decomposing words into substructure and sharing probabilities among the same sub-word context, the ANGIE system was able to achieve a compact representation of the probability space, resulting in at worst a 50% increase in the size of the final FST.

5.3.4 Discussion

The ANGIE system consistently outperformed the baseline on both recognition and understanding. However, for all but one of our test conditions, significantly greater improvements were achieved in *understanding* than in recognition. A careful analysis of the results revealed that this enhanced gain for understanding is due to an improved quality of the recognizer N-best list. This was verified by evaluating recognition performance on the basis of the hypothesis selected by the parsing process, which gave a percentage improvement commensurate with that obtained by the concept error rate. A factor which weighed heavily in the result was a superior ability for the ANGIE system to distinguish true words from "noise words." For short utterances,

the baseline system often proposed a noise-only hypothesis (such as <laughter>) as a strong competitor in the N-best list. Such a hypothesis would often win out over the top candidate as a consequence of the NL scoring strategy. We believe that ANGIE's treatment of the noise words is more effective, since the noise models have trained probabilities associated with their many alternative realizations.

It is curious that the recognition gap between the two baseline JUPITER systems disappeared when they were evaluated for understanding performance. However the ANGIE probability model was able to improve understanding for the reduced set by a substantially greater margin than for the full set, such that the case can still be made that the reduced rules are preferable, especially considering the smaller search space.

6 Summary and Future Work

This paper describes our experiments in parsing words into their linguistic substructure, in order to obtain a probability model to account for alternative phonetic realizations of words. An FST mapping phonetic realizations to sub-word structure with associated probabilities was derived by parsing a large corpus of observed phonetic sequences, and was reinserted into the recognizer's search FST, replacing the original component FST, which had no probabilities. The technique allows words to share common structure in a generic model encoding local phonetic context conditions. We

were able to obtain a significant relative reduction in concept error rate for two distinct domains: weather information and flight reservations. This improvement was achieved with only a modest expansion in the size of the resulting FST. We found that the probability model was also effective in characterizing non-speech events. We demonstrated performance gains for both the original "full" set of phonological rules utilized by the SUMMIT speech recognizer, and for a "reduced" rule set that only accounts for deletions and insertions. The reduced-rules experiment was conducted because this rule set has been shown to yield lower recognition error rates in the absence of a phonological probability model (Hazen et al. 2002).

We feel that in future work it would be better to eliminate the extra complexity consequential to the translation from SUMMIT's lexical representation to ANGIE's. It has been a tedious exercise in this work to assure exact consistency between ANGIE's baseforms and SUMMIT's. It should be relatively straightforward to alter the phonological rule specification to directly apply to ANGIE's phonemic inventory. The framework may also benefit from iterative retraining of the acoustic models.

We believe that our approach would work well for Switchboard data. It would be interesting to test the effectiveness of our modeling approach in this challenging domain.

Acknowledgments

This research was supported by DARPA under contract number NBCH1020002 monitored through the Department of the Interior, National Business Center, Acquisition Services Division, Fort Huachuca, AZ. The authors would like to thank Dr. Timothy J. Hazen for his assistance with the SUMMIT system, and Dr. Lee Hetherington for his assistance with FST tools. We are also grateful to the two anonymous reviewers who gave us many valid suggestions for improvement.

References

Chang, J. and J. Glass (1997). Segmentation and modeling in segment-based recognition. In *Proceedings of the 5th European Conference on Speech Communication and Technology*, Rhodes, Greece, pp. 1199–1202.

Chung, G. (2000). A three-stage solution for flexible vocablary speech understanding. In *Proceedings of the 6th International Conference on Spoken Language Processing*, Beijing, China, pp. 266–269.

Chung, G. and S. Seneff (2002). Integrating speech with keypad input for automatic entry of spelling and pronunciation of new words. In *Proceedings of the 7th International Conference on Spoken Language Processing*, Denver, Colorado, pp. 2061–2064.

Chung, G., S. Seneff, and C. Wang (2003). Automatic acquisition of names using

speak and spell mode in spoken dialogue systems. In *Proceedings of the HLT-NAACL 2003*, Edmonston, Canada, pp. 32–39.

Cohen, M. H. (1989). *Phonological Structures for Speech Recognition*. Ph. D. thesis, University of California, Berkeley, CA.

Cremelie, N. and J.-P. Martens (1999). In search of better pronunciation models for speech recognition. *Speech Communication 29*, 115–136.

Gauvain, J. L., L. F. Lamel, G. Adda, and M. Adda-Decker (1993). Speaker independent continuous speech dictation. In *Proceedings of the 3rd European Conference on Speech Communication and Technology*, Berlin, Germany, pp. 125–128.

Glass, J. R. (2003). A probabilistic framework for segment-based speech recognition. *Computer Speech and Language 17*, 137–152.

Glass, J. R. and T. J. Hazen (1998). Telephone-based conversational speech recognition in the Jupiter domain. In *Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney, Australia, pp. 1327–1330.

Godfrey, J. J., E. C. Holliman, and J. McDaniel (1992). SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the 1992 International Conference on Acoustics, Speech and Signal Processing*, San Francisco, CA, pp. 517–520.

Greenberg, S. (1999). Speaking in shorthand – a syllable-centric perspective for

understanding pronunciation variation. *Speech Communication 29*, 159–176.

Halberstadt, A. and J. Glass (1998). Heterogeneous measurements and multiple classifiers for speech recognition. In *Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney, Australia, pp. 995–998.

Hazen, T. J., I. L. Hetherington, H. Shu, and K. Livescu (2002). Pronunciation modeling using a finite-state transducer representation. In *Proceedings of the ISCA Tutorial and Research Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language*, Estes Park, CO, pp. 99–104.

Hetherington, I. L. (2001). An efficient implementation of phonological rules using finite-state transducers. In *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH)*, Aalborg, Denmark, pp. 1599–1602.

Kahn, D. (1980). *Syllable-based Generalizations in English Phonology*. New York: Garland Press.

Livescu, K. and J. Glass (2001). Segment based recognition on the PhoneBook task: Initial results and observations on duration modeling. In *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH)*, Aalborg, Denmark, pp. 1437–1440.

McAllaster, D., L. Gillick, F. Scattone, and M. Newman (1998). Fabricating conver-

sational speech data with acoustic models: A program to examine model-data mismatch. In *Proceedings of the 5th International Conference on Spoken Language Processing*, pp. 1847–1850.

Seneff, S. (1998). The use of linguistic hierarchies in speech understanding. Keynote Address. In *Proceedings of the 5th International Conference on Spoken Language Processing*, Sydney, Australia, pp. 3321–3330.

Seneff, S., G. Chung, and C. Wang (2003). Empowering end users to personalize dialogue systems through spoken interaction. In *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH)*, Geneva, Switzerland, pp. 749–752.

Seneff, S., C. Chuu, and D. S. Cyphers (2000). ORION: From on-line interaction to off-line delegation. In *Proceedings of the 6th International Conference on Spoken Language Processing*, pp. 142–145.

Seneff, S., R. Lau, and H. Meng (1996). ANGIE: A new framework for speech analysis based on morpho-phonological modeling. In *Proceedings of the 4th International Conference on Spoken Language Processing*, Philadelphia, PA, pp. 110–113.

Seneff, S. and J. Polifroni (2000). Dialogue management in the MERCURY flight reservation system. In *Proc. ANLP-NAACL 2000, Satellite Workshop*, Seattle,

WA, pp. 1–6.

Seneff, S. and C. Wang (2002). Modeling phonological rules through linguistic hierarchies. In *Proceedings of the ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language*, pp. 71–76.

Shu, H. and I. L. Hetherington (2002). EM training of finite-state transducers and its application to pronunciation modeling. In *Proceedings of the 7th International Conference on Spoken Language Processing*, Denver, Colorado, pp. 1293–1296.

Strik, H. and C. Cucchiarini (1999). Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication 29*, 225–246.

Tajchman, G., E. Fosler, and D. Jurafsky (1995). Building multiple pronunciation models for novel words using exploratory computational phonology. In *Proceedings of the 4th European Conference on Speech Communication and Technology (EUROSPEECH)*, Madrid, Spain, pp. 2247–2250.

Weintraub, M., H. Murveit, M. Cohen, P. Price, J. Bernstein, G. Baldwin, and D. Bell (1989). Linguistic constraints in Hidden Markov Model based speech recognition. In *Proceedings of the 1989 International Conference on Acoustics, Speech and Signal Processing*, Glasgow, Scotland, pp. 699–702.

Zue, V. (1983). The use of phonetic rules in automatic speech recognition. *Speech*

Communication 2, 181–186.

		Full Rules		Reduced Rules	
		WER(%)	RR(%)	WER(%)	RR(%)
JUPITER	Baseline	16.5	-	15.8	-
	Angie	15.3	7.5	15.4	2.5
MERCURY	Baseline	15.2	-	14.8	-
	Angie	14.2	6.6	14.0	5.4

Table 1: Comparison of recognition error rates for two domains (weather and flights), between systems that supported phonological rules but without a probability model (Baseline) and systems that were trained using the ANGIE grammar (Angie), for two different sets of phonological rules: Full and Reduced. "WER" stands for word error rate, and "RR" stands for relative reduction.

		Full Rules		Reduced Rules	
		CER(%)	RR(%)	CER(%)	RR(%)
JUPITER	Baseline	13.40	-	13.57	-
	Angie	12.53	6.5	11.61	14.4
MERCURY	Baseline	11.12	-	11.45	-
	Angie	9.53	14.2	9.56	16.5

Table 2: Comparison of understanding error rates for two domains (weather and flights), between systems that supported phonological rules but without a probability model (Baseline) and systems that were trained using the ANGIE grammar (Angie), for two different sets of phonological rules: Full and Reduced. "CER" stands for concept error rate, and "RR" stands for relative reduction.

{**left**}	**core**	{**right**}	→ **realizations**	; comments
{vowel}	t	{schwa}	→ tcl t | dx	; flapping
{}	s	{y sh zh}	→ s | sh	; palatalization
{en n}	n	{}	→ [n]	; gemination

Figure 1: Representative phonological rules provided for lexical expansion in the SUMMIT recognition framework. The left and right conditions specify the *input* phonemic contexts in which the target phoneme should be rewritten according to the phone patterns expressed on the right side of the arrow. "|" encodes alternates, and "[ ]" means optional.

5	sentence							
4	word							
3	pre			sroot			uroot	
2	nuc	ucoda		onset	nuc_lax+	coda	uonset	nuc
1	ae	t		l!	ae+	n	t_u!	ax
0	ax	tcl	t	l	ae	n	-n	ax
	0	1	2	3	4	5	6	7

Figure 2: ANGIE parse tree for the word "Atlanta," showing phonological rules expressed in preterminal-to-terminal mappings. The i^{th} *column* corresponds to the path from the i^{th} terminal phone to the root node at the top. The notation "-n" encodes a left-context dependent deletion of the unstressed onset phoneme "t_u!" (*Note: The phoneme layer utilizes diacritics to encode onset (!) and vowel stress (+).*) *Note:* The row and column indices are illustrated in the figure.

(a) Lexical Entries

atlanta	: at- lan+ ta
either	: Ei+ ther
streets	: street+ =s
the	: the*

(b) Morph Entries

at-	: ae t
lan+	: l! ae+ n
ta	: t_u! ax
Ei+	: (iy+ | ay+)
ther	: dh! er
street+	: str! iy+ t
=s	: s_pl
the*	: dh! iy_the

(c) ANGIE Baseforms

atlanta	: ae t `pre` l! ae+ n `sroot` t_u! ax `uroot`
either	: (iy+ | ay+) `sroot` dh! er `dsuf`
streets	: str! iy+ t `sroot` s_pl `isuf`
the	: dh! iy_the `fcn`

Figure 3: Illustrations of the linguistic encodings of words in the ANGIE framework. (a) Four examples of lexical entries in terms of sub-word "morph" units; (b) corresponding morphs and their phonemic representations; (c) output baseforms that are generated automatically from (a) and (b), to be used by the SUMMIT recognizer, for these words. Note that alternate pronunciations are only provided in rare cases such as "either". In addition, there are some inflection-specific units, such as "s_pl", word-specific units, such as "iy_the", and some multi-phone units, such as "str!". Symbols such as "sroot" and "fcn" identify the syllable category. See text for further details.

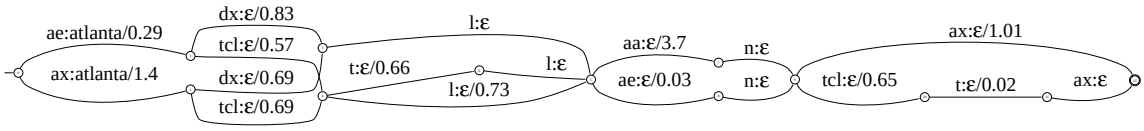

Figure 4: Example of a pronunciation graph created by the system for the word "Atlanta." Each arc is labelled with the input and output symbols, and the corresponding negative log probability. Zero or null weight corresponds to probability 1.