SPECOM 1320
1 December 2003 Disk used
ARTICLE IN PRESS
No. of Pages 18, DTD = 4.3.1

Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Speech Communication xxx (2003) xxx–xxx

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

The use of subword linguistic modeling for multiple tasks in speech recognition ☆

Stephanie Seneff *

Laboratory for Computer Science, Massachusetts Institute of Technology, Spoken Languages System Group, 200 Technology Square, Room 643, Cambridge, MA 02139, USA

Abstract

Over the past several years, I have been conducting research on subword modeling in speech recognition. The research is most specifically aimed at the difficult task of identifying and characterizing unknown words, although the proposed framework also has utility in other recognition tasks such as phonological and prosodic modeling. The approach exploits the linguistic substructure of words by describing graphemic, phonemic, phonological, syllabic, and morphemic constraints through a set of context-free rules, and supporting the resulting parse trees with a corpus-trained probability model. A derived finite state transducer representation forms a natural means for integrating the trained model into a recognizer search. This paper describes several research projects I have been engaged in, together with my students and associates, aimed at exploring ways in which recognition tasks can benefit from such formal modeling of word substructure. These include phonological modeling, hierarchical duration modeling, sound-to-letter and letter-to-sound mapping, and automatic acquisition of unknown words in a speech understanding system. Results of several experiments in these areas are summarized here.

1. Introduction

1.1. Background

Speech is first and foremost a *communicative* signal. It is a complex encoding of linguistic messages for the purpose of conveying information among humans who share the code. Speech scientists have been studying various aspects of the speech code for many decades, and engineers have been involved in designing computer systems that attain a certain degree of competence in understanding and rendering the code.

At the core of human communication is the notion of "words" as the fundamental units. Above the word level, it is apparent that words group into phrases, and phrases group into higher level units such as clauses and sentences. In addition to studies of how words are organized into meaning, studies of the *substructure* of words in multiple languages have revealed a number of organizational principles (Scalise, 1986). The exact specification of that substructure still eludes us,

☆ This research was supported by DARPA under contracts N66001-94-C-6040 and N66001-96-C-8526, monitored through Naval Command, Control and Ocean Surveillance Center, and contract NBCH1020002, monitored through the Department of the Interior, National Business Center, Acquisition Services Division, Fort Huachuca, AZ, and by the National Science Foundation under grant no. IRI-96187321.

* Tel.: +1-617-253-0451; fax: +1-617-258-8642.
E-mail address: seneff@sls.lcs.mit.edu (S. Seneff).

0167-6393/$ - see front matter © 2003 Published by Elsevier B.V.
doi:10.1016/j.specom.2003.11.001

however, particularly for languages such as English with a rich borrowing from other languages. The inventory of *phonemes* for any particular language can usually be enumerated quite specifically. We are now also reasonably confident that the syllable exists as an intermediate layer between words and phonemes, although most speech recognition systems make little or no use of this syllable layer. There is also the possibility of breaking words down into *meaning* units (i.e., morphemes), which may not necessarily align precisely with syllable units based strictly on phonology and sonority. The difficulty of defining exactly how the phonemes of a word might group themselves into natural subunits has been a major hurdle to the design of systems that utilize this intermediate structure.

1.2. Historical context

My interest in developing hierarchical structures to represent word substructure was inspired by research conducted in the 1970s and 1980s by a team of researchers and students at MIT, dating back originally to Chomsky and Halle's theory of generative phonology (Chomsky and Halle, 1968). A doctoral thesis by Kahn led to the formalization of phonological phenomena with respect to syllable structure (Kahn, 1976). In the 1980s, a team of researchers led by Allen developed a sophisticated letter-to-sound generation system called MITalk, based on a decomposition of words into meaning units called morphs (Allen et al., 1987). At the same time, Randolph developed formal rules to parse words into syllables, with the aim of formally encoding a distinctive-feature formalism (Randolph, 1989). Zue was also codifying his acoustic phonetic knowledge into formal, ordered, context-sensitive rules that could be utilized to expand lexical pronunciations for a speech recognition task (Zue, 1983). Church's doctoral thesis (1983) proposed applying context-*free* rules [1] to parse syllables, in order to capture phonological effects such as flapping and palatalization, arguing that conditions for phonological phenomena could be encoded effectively in category names, thus avoiding explicit context dependencies.

1.3. Motivation

While researchers have made significant progress in speech recognition in the last two decades (Bacchiani and Ostendorf, 1998; Cook et al., 1998; Nguyen et al., 1995; Woodland and Young, 1993), there are still many remaining problems which could be addressed through the use of formal representation of the substructure of words. Nearly all speech recognition systems today are based on a simple model in which words are represented explicitly in a lexicon encoding their phonemic realizations, and class *n*-gram language models provide linguistic constraint. One shortcoming of such a representation is that unknown words are not formally represented, and therefore will be recognized as an acoustically similar substitution of a known word, often adversely affecting the recognition of neighboring words as well (Hetherington, 1994; Hetherington and Zue, 1993). Another limitation is the lack of a syllable-based framework for characterizing phonological rules, as well as the difficulty to capture such rules in an appropriate probability formulation. [2] The durational aspects of phonemes depend on their position within the syllable and the word, but this information is usually not available to a recognizer. Finally, the task of modeling an association between letters and their pronunciations is likely to benefit from knowledge of the linguistic context of each letter.

1.4. Overview

Through my earlier and ongoing work in parsing words into meaning, via the TINA natural understanding system (Seneff, 1992), I have come to believe that a similar approach can be used

[1] A context free rule is a rule that rewrites a symbol generally into a sequence of zero or more symbols. A context-*sensitive* rule attaches conditions under which the symbol is permitted to be rewritten.

[2] Although one could argue that the use of triphone modeling does provide some *implicit* information about phonetic context.

effectively below the word level, yielding a parsimonious and trainable hierarchical representation of word substructure (Seneff, 1998). It is my belief that such representations may have significant advantages over a flatter structure, in that they should be capable of generalizing knowledge across similar contexts. I have subsequently investigated various ways in which such linguistic substructure can be utilized in speech recognition tasks. These investigations are predicated on a common theme that involves parsing words into their underlying linguistic constituents via a formal grammar, expressed through context-free rewrite rules. The resulting structural information is then augmented with a probability framework, where probabilities are determined by tabulating counts in parse trees obtained by parsing a large corpus of representative speech materials. A final optional step is to reformulate the trained parse trees as a finite state transducer (Hetherington, 2001), typically with inputs and outputs associated with the terminals and preterminals of the parse tree, respectively. This step then enables a straightforward mechanism for incorporating the linguistic models directly into a recognizer search.

The ideas discussed above have been formalized into a framework, called Angie, and several different topics of research have been investigated by members of the Spoken Language Systems group within this framework. These include phonetic recognition (Chung, 2001; Chung and Seneff, 1998; Lau, 1998; Lau and Seneff, 1998), hierarchical duration modeling (Chung, 1997; Chung and Seneff, 1997), sound-to-letter and letter-to-sound generation systems (Meng, 1995; Meng et al., 1996; Chung et al., in press; Seneff et al., 1996), unknown word detection and modeling (Chung, 2001; Mou et al., 2001; Parmar, 1997), and phonological modeling (Seneff and Wang, 2002). This paper will provide motivation for the approach we have taken, and will describe instances of all of the above applications within a common thread. While some of the investigations are on-going, it seems appropriate at this time to provide a detailed accounting of this research, partly in the hope that others might be inspired to pursue similar avenues of research.

2. Angies subword linguistic representation

Most of the work in speech recognition to date has been focused on the task of correctly producing the sequence of words that were spoken. The notion of characterizing any information beyond the word sequences is usually not treated as part of the explicit goal, although some amount of phonological and semantic knowledge is generally viewed as a necessary adjunct to success. Usually, each word is represented in the lexicon as a sequence of phonemes, and in some systems a phonological rule framework permits the expansion of lexical entries to explicitly account for phonological effects like flapping or devoicing (Cohen, 1989; Gauvain et al., 1993; Glass and Hazen, 1998; Weintraub et al., 1989). Typically the rules are precompiled into the lexicon, yielding an expanded lexicon of alternate pronunciations.

In order to address the issue of out-of-vocabulary (OOV) words, some recognition systems have included a generic model for unknown words as part of the recognizer's phonetic model. The approach typically adopts a generalized probabilistic subword model as a pronunciation model, such as a phone bigram, for the "word" OOV (Bazzi and Glass, xxxx). The word OOV then competes with known words, and the goal is that it would score better than known words for spoken out-of-vocabulary words, preventing the system from erroneously substituting a vocabulary entry with a similar pronunciation. An interesting example of a more sophisticated use of this technique is the work by Onishi et al. (2001), which developed distinct subword models for two different *classes* of unknown words: city-name and surname. In an evaluation experiment, they were able to achieve perfect disambiguation of the unknown word class whenever an unknown was detected, and with a slight improvement in overall recognition performance, when compared with a baseline that had no unknown word model.

In the Angie framework, we are interested in building a single subword linguistic model that can be effective for both the known and the unknown words. Thus, the purpose for building hierarchical structure below the word level is multifold. One main goal is to predict phone sequences of the language without explicit ties to a particular vocabulary. A bottom-up parsing procedure has the important property that it supports significant structure sharing among both in-vocabulary and OOV words that begin with the same phone sequence. If words are further decomposed into syllables, which then form the basic recognition unit, even greater sharing is possible, since words such as "retention" and "contention" can share everything except their prefix in common syllable nodes.

Exactly what linguistic knowledge should be encoded in the Angie parse trees is open for debate. In all of the experiments we have conducted, syllable structure plays a critical role. At this point, we have developed several grammars with distinctive symbol sets at the preterminal and terminal layers of the parse tree, but we have consistently distinguished stressed and unstressed syllables, which are further decomposed into onset, nucleus, and coda, according to standard syllable theory (Selkirk, 1982; Kahn, 1976). This choice of decomposition reflects in part our belief that the position of a consonant within its syllable plays a critical role in its phonetic expression. For example, a /t/ in a syllable onset /st/ cluster is unaspirated, whereas a /t/ would normally be aspirated in onset position. Greenberg (1999) has shown, through studies on a large corpus of hand transcribed Switchboard data (Godfrey et al., 1992), that 28% of consonants in *coda* position were deleted, a rate that is substantially higher than the rate for onset position.

The substructure that is captured in Angie's grammar rules includes morphology, stress, syllable structure, and phonological variants. Probabilities are trained automatically from a parsed corpus. We have used the approach of seeding on phonetic transcriptions provided by automatic alignment of training data using our segment-based summit speech recognizer (Glass et al., 1996; Glass and Hazen, 1998), which expands idealized phonemic baseforms into phonetic alternatives via formal phonological rules.

The shared probability model is important for generalizing phenomena over similar contexts. Rare words can benefit from observations of common words that have the same local phonetic environment. And words that are completely unknown to the recognizer can be generated with a non-zero probability by following the parse tree fragments of words with localized equivalent patterns. For example, "queen" can be decomposed into the onset of "quick" and the rhyme of "seen."

Parse trees in Angie are further characterized by structural units that encode positional roles of the syllables in a word. Thus, unstressed syllables are identified as "prefix" if they begin a word, and as "suffix" if they are a terminal unit carrying syntactic information, such as "-ing" for present participle or "-ness" marking a nominalization. This additional information is beneficial not only for further constraints but also because the position of a syllable within a word impacts other aspects, such as the prosodic characteristics. For example, in prepausal lengthening a final unstressed syllable is affected much more strongly than an initial one (Chung and Seneff, 1997).

In addition to the parse framework, a set of explicit subword lexical units can offer further constraint in phonetic recognition tasks. One possibility is to augment a phonetic recognizer with a lexicon containing the inventory of all the unique syllables present in a corresponding word lexicon for the task. A syllable *n*-gram will then provide additional language model support to improve the quality of the phone or phoneme graph being proposed as outputs of the recognizer. In addition to simple syllable units, we have also investigated the use of more detailed units which we call "morphs," [3] essentially syllables marked for both their spelling and positional information within the word. Another unit above the syllable is the metrical foot (Hayes, 1995), which consists of a stressed syllable and zero or more adjacent unstressed syllables. This unit of recognition, which is intermediate betweeen phonemes and words, provides a convenient compromise in yielding fairly strong constraint while still supporting substantial coverage of novel words and disfluencies in conversational speech.

[3] This follows roughly the definition given in (Allen et al., 1987, p. 24), which is a representation of morphological units such as prefix and root that is also tied to the word's spelling.

In ANGIE, we currently represent our lexicon in two tiers—words are entered as sequences of morphs, and morphs are in turn entered as sequences of phonemes. We currently distinguish for English five different possible morph positions: prefix, stressed root, unstressed root, "dsuf" and "isuf". [4] Context-free rules encode positional constraints for the morph units—for example, unstressed root always follows immediately after stressed root, and isuf's are always terminal.

As mentioned previously, it is often not obvious where to place syllable boundaries in English words. There are many cases of ambisyllabicity, as in the word "co*nn*ect" where it is not clear whether the intermediate consonant belongs with the preceding or following syllable. Placement of the boundary can also be influenced by the underlying morphology—when there is a clear inflectional ending our policy has been *not* to shift the terminal consonant of the root into onset position, even though this would be in accord with a maximal-onset rule. Hence "dancing" becomes "danc-ing" rather than "dan-cing". Often we introduce a double consonant phonemically as a means of implementing explicit ambisyllabicity, which reduces via a gemination rule to a single phonetic realization. Hence, "connect" becomes "con-nect" with two /n/ phonemes at the phonemic layer reducing to one at the phonetic layer. This makes the boundary between the word-internal syllables behave analogously to boundaries between word sequences, as in "on next" or "seven nine." Such lexicalized geminations are nearly always associated with a spelling that includes a doubleton letter, such as the "nn" in "connect."

2.1. Example parse tree

One of the main goals of ANGIE's modeling is to provide letter-to-sound and sound-to-letter mappings, and, particularly for this purpose, we have found it beneficial to provide a pair of grammars with a shared superstructure but two distinct sets of rules mapping preterminals to terminals: one expecting phonetic units as the terminals and the other expecting graphemics. The preterminal layer contains the phonemic sequence exactly matched to the entries in the morphs of the two-tiered lexicon. The terminals are either the letters of the spelling of the word or the phones of the particular spoken realization. Thus letter-to-sound and phonological rules are licensed on the preterminal-to-terminal mappings. The upper layers capture syllabification, morphology, and stress.

Example parse trees in ANGIE for the word "commission" are given in Figs. 1 (letter terminals) and 2 (phone terminals). The lexical representation of the word consists of a prefix (com−) a stressed root (mis+) and an inflectional suffix (= sion). Phonemically, there are both a final /m/ for the prefix and an onset /m!/ for the root. These geminate in the phonetic realization into a single [m]. [5] Similarly, the "mis+" unit ends phonemically with an /s/. The /s/ is palatalized to a [sh] at the phonetic level, with the onset /sh!/ of the "= sion" marked as deleted. Fig. 3 illustrates how sharing of subword units can be achieved, using the examples "mis+" and "= sion."

2.2. Lexicon creation

ANGIE relies heavily on the availability of a specifically prepared two-tiered lexicon, in which words are represented in terms of their underlying morphs. We first obtained, through careful hand-editing, a seed lexicon of some 10,000 words, derived from the common words of the Brown corpus (Kucera and Francis, 1967) augmented with words from some of our conversational domains such as ATIS (Zue et al., 1991) and Jupiter (Glass and Hazen, 1998). We have since converted all the common words of Pronlex [6] into ANGIE's lexical format (Parmar, 1997). We have utilized a semi-automatic process which first parses the letters of

[4] "dsuf" roughly corresponds to "derivational suffix," and "isuf" to "inflectional suffix," but we sometimes violate strict conventions for pragmatic reasons.

[5] [−m] is a code for "deleted in the context of preceding [m]".

[6] A pronunciation lexicon for the words in the Comlex lexicon, produced and distributed by the Proteus Project at New York University, under the auspices of the Linguistic Data Consortium (see http://www.ldc.upenn.edu).

word									
pre			sroot			isuf			
uonset	nuc		onset	nuc_lax+	coda	uonset		nuc	
k!	em		m!	ih+	s	sh!		en	
c	o	m	m2	i	s	s2	i	o	n
com-			mis+			=sion			

Fig. 1. ANGIE parse tree for the word "commission," with letters as the terminals. An aligned sequence of morphs is shown below the parse tree. *Note:* "!" denotes onset position and "+" marks stress. The second letter in a doubleton is specially tagged for additional constraint (m2, s2).

word									
pre				sroot			isuf		
uonset		nuc		onset	nuc_lax+	coda	uonset	nuc	
k!		**em**		m!	ih+	s	sh!	en	
kcl	k	**ax**	m	-m	ih	sh	-sh	ax	n
com-				mis+			=sion		

Fig. 2. ANGIE parse tree for the word "commission," with phones as the terminals. An aligned sequence of morphs is shown below the parse tree. The highlighted entries illustrate units involved in the trigram language model as applied to the bottom-up prediction of the preterminal layer.

Word Lexicon

commission	com- mis+ =sion
mister	mis+ ter
mansion	man+ =sion

Morph Lexicon

com-	k! em
man+	m! ae+ n
mis+	m! ih+ s
sion	sh! en
ter	t! er

Fig. 3. Selected entries from a word and morph lexicon for ANGIE.

each word into a set of hypothesized phonemic alternatives, and then parses the phonetic units as provided by Pronlex into phonemes, constrained by the choices produced by the letter-parsing step. Of course the automatic procedures are not error-free, so extensive hand correction is required to perfect the lexicon.

We hope to use the resulting morph lexicon as a basis for a generic morph-based recognizer for general English. A phonological model can then be trained on any large corpus of spoken utterances. There would still be some possibility of unseen morphs in new material, but these would likely be covered generatively by the rule base. Such a lexicon is also useful for training a reversible letter-to-sound system. Ultimately, we would like to augment it with additional information such as part-of-speech, and perhaps add a feature propagation mechanism to ANGIE's framework to utilize such features, similar to the one developed for the TINA natural language understanding system (Seneff, 1992).

2.3. Probability model

In ANGIE, a parse tree is obtained for each word by expanding the rules of a carefully constructed context-free grammar. The grammar is intentionally arranged such that every parse tree lays out as a regular two-dimensional grid, as shown in Fig. 4. Each layer is associated with a particular aspect of subword structure: migrating from morphemics to syllabics to phonemics to phonetics at the deepest layer. 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, as well as the convenience of providing arc

sentence								
word								
sroot		uroot			sroot2			
nuc_lax+	coda	uonset		nuc	onset		lnuc+	lcoda
ih+	n	t!	r	ow	d!		uw+	s
ih	n	-n	rx	-rx	dcl	d	uw	s
in+		tro			duce+			

Fig. 4. ANGIE parse tree for the word "introduce," showing phonological rules expressed in preterminal-to-terminal mappings. The morph sequence is shown below the terminal phones.

probabilities for a finite state transducer (FST) representation (Hetherington, 2001).

Given these considerations, the probability formulation we have developed for Angie can be written as follows:

$$P(C_i|C_{i-1}) = P(a_{i,0}|C_{i-1}) \prod_{j=1}^{N-1} P(a_{i,j}|a_{i,j-1}, a_{i-1,j}) \quad (1)$$

where C_i is the ith column in the parse tree and $C_i = \{a_{i,j}, 0 \leqslant j \leqslant N\}$, and $a_{i,j}$ is the label at the jth row of the ith column in the two-dimensional parse grid. [7] In words, each phone 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. These probabilities are acquired by tabulating counts in a large corpus of parsed sentences, mapping words to their corresponding phonetic realizations. This process will become clearer when we give an explicit example in the next section.

Angie's language model, while restricted to phone-to-phone transitions, is very powerful, and captures generic linguistic knowledge of English while a partial word is under construction. We have determined empirically that, within the ATIS flight information domain, Angie is able to achieve a significantly lower perplexity on unseen data than a phone trigram similarly trained (Lau and Seneff, 1997). Once a word is completed, higher level language models can be incorporated as well (e.g., syllable/word n-grams).

2.4. Phonological rule expression

Angie's ability to encode and generalize phonological rules is best illustrated through an example. Consider the parse tree shown in Fig. 4 for the word "introduce" pronounced casually as "innerduce." The two special phones [−n] and [−rx] are "deletion" phones, meaning that they occupy no temporal space and have no acoustic model. The deletion category is tied to the preceding phone's identity. The grammar developer would specify that /t!/ can be realized as "[−n]", meaning "/t/ in onset position can be deleted after [n]." The probability model captures the important context conditions—falling stress and following schwa. The deletion of the /ow/ is predicated on the realization of the preceding /r/ as a retroflexed schwa ([rx]).

Fig. 5 illustrates the context conditions that are learned, with regard to this t-deletion rule. The column above the [n] encodes coda position in a stressed syllable. It predicts a deletion after [n] with no awareness of which phoneme actually follows. The trigram column-building step decides which phoneme was deleted. Other possibilities would be /t/, /d/, /d!/, and /n!/. The training procedure would collapse together the /t/ deletion here with other similar environments, such as "in*te*grate," "can*ta*loupe," "en*ter*tain," "San*ta* Clause," "hun*ter*," and "pan*ty*hose." The column above the [−n] would learn through training that it is rarely followed by anything other than [ax], [rx], and [ix]. The system would thus learn from examples that the right context must be a schwa, but it could be front, back or retroflexed. This "fact" was not informed by any rule, but rather discovered from observation of training data.

2.5. Spellnemes

In the original grammars we developed for Angie, we adopted the point of view that there would be two parallel grammars with identical

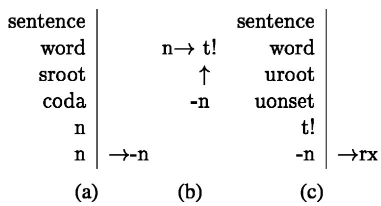

Fig. 5. Schematic of probability model in Angie, and its accounting of the context conditions for t-deletion in words such as "introduce." In (a) and (c) are the two column contexts for the predictions of the phones "−n," symbolizing deletion after /n/, and "rx," a retroflexed schwa. (b) Illustrates the bottom-up trigram prediction of the deleted phone's parent phoneme, which in the example is "t!," a /t/ in onset position (see text for discussion).

[7] j indexing begins at the bottom of the column.

parse tree superstructure, but terminating in phonetic units on the one hand and in graphemic units (spellings) on the other hand. The phonetic terminals would characterize possible phonological rules, such as alveolar flapping, vowel reduction, anticipatory palatalization, etc. (Zue, 1983). The graphemic units would encode letter-to-sound rules of English, with the terminals consisting of the letters of the English alphabet, sometimes grouped to form natural paired units such as "th" or "ng."

In more recent work, however, (Chung, 2000a,b; Chung, 2001) we have begun to explore potential benefits of introducing a new unit type which we refer to as a letter-phoneme, or "spellneme." These are units which tie together letters or letter sequences with their corresponding pronunciation in a single symbol string. For example, the spellneme unit, "a_x+" symbolizes a short /a/, pronounced as in the word "cat," whereas the "a_l+" symbol string refers to a long /a/ as in "cape." Other symbols distinguish "soft" and "hard," for consonants such as 'c' (compare "cent" with "car"), etc. In general, we expect a sequence of spellnemes to encode both the spelling and the pronunciation of the word.

An advantage of these spellneme units is that they provide greater constraint, in that they are more specific than either letters or phonemes. We expect this to translate into better performance, not only on the sound-to-letter task, but also on other tasks such as phonetic recognition, due to the richer language model, which has in fact been borne out by experiments (Chung and Seneff, 1998). There is also the notational convenience that both the spelling and pronunciation can be derived through simple string manipulations. The "commission" example, with spellnemes as preterminals and phones as terminals, is illustrated in Fig. 6.

<table>
<tr><td colspan="10">word</td></tr>
<tr><td colspan="4">pre</td><td colspan="3">sroot</td><td colspan="3">isuf</td></tr>
<tr><td colspan="2">uonset</td><td colspan="2">nuc</td><td>onset</td><td>nuc_lax+</td><td>coda</td><td>uonset</td><td colspan="2">nuc</td></tr>
<tr><td colspan="2">c_k!</td><td colspan="2">om</td><td>m!</td><td>i_x+</td><td>s</td><td>si!</td><td colspan="2">on</td></tr>
<tr><td>kcl</td><td>k</td><td>ax</td><td>m</td><td>-m</td><td>ih</td><td>sh</td><td>-sh</td><td>ax</td><td>n</td></tr>
<tr><td colspan="4">com-</td><td colspan="3">mis+</td><td colspan="3">=sion</td></tr>
</table>

Fig. 6. ANGIE parse tree for the word "commission," with spellnemes as the preterminals (highlighted), and phones as the terminal units. This is to be compared with the corresponding parse tree in Fig. 2.

3. Incorporating ANGIE into recognition tasks

The ANGIE parsing framework provides powerful mechanisms for learning important aspects of subword structure. However, parsing is in general a computationally expensive process, and so it becomes important to consider ways of encapsulating the results of ANGIE parse training into a speech recognition task while still preserving close to real-time performance. Fortunately, the SUMMIT segment-based recognizer has been formulated in terms of a FST framework (Hetherington, 2001), which provides the opportunity to encode complex linguistic knowledge and embed it in the core recognizer search engine. The dominant approach we have taken is to reconfigure ANGIE parse trees as a "column bigram" (Chung, 2000b), where a "column" is a unique path from a terminal node up to the root node. An assignment of the probability associated with a transition from one column to the next can be computed directly from the ANGIE parse framework. In most of our experiments, we have preserved only a portion of the information in the ANGIE parse tree via the input and output symbols associated with the FST. These simplifications were thought to be necessary both because of the complexity inherent in an FST representing the full ANGIE parse space, and because the full parse space would likely lead to an impractical FST size (but see (Mou et al., 2001) for a more general FST solution).

For many applications, the full linguistic model obtained from ANGIE is correlated with other linguistic knowledge encoded in higher level language models introduced either in the same stage or in later stages of a multi-stage recognizer. An intended goal is to influence the search through language model constraints before the lexical entries have been retrieved. Once a proposed word is known, and at the point where the word n-gram score is being introduced, the linguistic component of the ANGIE contribution to the word can be subtracted out, leaving behind only the phonological component. In other configurations, a more precise formulation using Bayes' formula to remove from each column–column score all contributions except the phoneme-to-phone probability assignment, where the intent is to use Angie strictly as a phonological rule model. The latter technique is particularly effective when the speech corpus used for phonology training is not well matched to the application domain, as is often the case when a new domain is being launched. In such cases, the upper layers of Angie's parse tree would obtain probability training from an inappropriate linguistic model, which can be eliminated by the normalization procedure. The likelihoods of the phoneme-to-phone mappings should be independent of the domain, and thus can be used to capture phonological rule probabilities generically (Seneff and Wang, 2002).

In the remainder of this section we will touch on several applications where the Angie linguistic hierarchy has been found to be useful for improving speech recognition performance. We begin with the task of duration modeling in recognition, evaluated in the ATIS flight information domain (Chung and Seneff, 1997). This is followed by an experiment in acquiring probabilities on phonological rule productions, conducted in the Mercury flight reservation domain (Seneff and Polifroni, 2000), but making use of a large training corpus from the Jupiter weather domain (Seneff and Wang, 2002). Next, we address the issue of detecting and accounting for unknown words, through experiments conducted in the Jupiter domain (Chung, 2000a,b). We conclude with a discussion of our research in the highly related topic of automatic new-word acquisition (Chung and Seneff, 2002). Due to space restrictions, each of the topics is described only briefly. The interested reader is referred to the literature for details of the experiments.

3.1. Duration modeling

Angie's parse trees can provide access to intermediate structures within words, which can be useful for characterizing prosodic information. Thus far we have only attempted to characterize prosody through *timing* measures. However, we have found that significant improvements in both phonetic recognition and word spotting can be gained through the use of relative duration models relating parents to children at all layers of an Angie parse tree (Chung, 1997; Chung and Seneff, 1997). The approach adopted involved normalizing the duration of each constituent in the parse tree with respect to its particular *children*, and then measuring the portion it occupies of its *parent's* total duration. The procedure propagates to the top of the tree to yield a word-by-word speaking rate parameter, which can then be folded back into the phonemic layer to tighten the distributions on absolute phoneme duration. This too leads to improved overall recognition.

To quantitatively assess the effectiveness of the hierarchical duration model, an experiment was conducted on phonetic recognition in the ATIS flight information domain, where the sophisticated duration models were benchmarked against a phonetic recognizer configuration which made use of the raw phone duration as a feature in the phone-based acoustic models. The enhanced system augmented the standard system with two sets of Gaussian models, as suggested above: relative duration models, across the entire parse tree, and models for absolute phoneme duration, normalized with respect to the estimated word-by-word speaking rate parameter. It was found that the performance gains attributable to the hierarchical duration models were stronger when the linguistic models included a richer knowledge base, with the best gains yielding improvements in phonetic error rate from 29.7% to 27.4%. [8]

The word spotting experiments (Lau, 1998) were also conducted in the ATIS domain, where the task was to detect all city names in the user utterances, treated as keywords. Results were reported in terms of a "figure of merit" (FOM), derived by integrating over a receiver operator characteristic (ROC) curve, which gives detection rate as a function of false alarm rate. The addition of the hierarchical duration model to the scores for the keywords yields performance gains on the

[8] For the details of these experiments, please see Chung (1997).

FOM from 89.3% to 91.6%. Details of this experiment can be found in (Chung, 1997; Lau, 1998).

We believe that this direction of research has many as yet unexplored branches, both in terms of incorporating hierarchies above the word level and in incorporating other prosodic measures such as fundamental frequency and energy.

3.2. Acquiring phonological rule probabilities

This section describes a set of experiments we have conducted, aimed at acquiring a probability model to support phonological rules describing the mappings from the phonemic baseforms of a lexicon to the actual phonetic realizations in spontaneous speech. In this case, we are only interested in the component of ANGIE's probability model that predicts the terminal phone unit of each subsequent column. It is not enough to simply discard the predictions of the chain of parents moving up the right hand column, but rather they must be considered as contributing to the conditioning context for the terminal phone.

The procedures we have adopted appear complex, but are relatively straightforward to execute, given a SUMMIT recognizer with an associated phonological rule set, an ANGIE grammar with ANGIE's phonemes in the preterminal layer and SUMMIT's phones in the terminal layer, a lexicon of words in the domain, with baseforms available in both SUMMIT's phonemic units and in ANGIE's phonemic units, and a large corpus of utterances for training.

The approach we have taken, then, is to start with a SUMMIT recognizer, complete with its standard set of phonological rules, which, when applied to the SUMMIT baseforms, yields a finite state transducer specifying all the phonetic variants possible for each word in the lexicon, but in the process, losing the mapping from phones to phonemes (Hetherington, 2001) (this transducer inputs phones and outputs words). We then insert a column bigram FST mapping phones to ANGIE phonemes, along with an ANGIE baseforms FST to map from ANGIE phonemes to words. The column bigram FST will contain probabilities computed by training ANGIE on an observation space obtained by parsing the phonetically aligned corpus. The process can be iterated.

The ANGIE model intentionally captures both phonological and linguistic aspects of the language, such as the frequency of different syllable onset patterns. However, for the purpose of modeling the likelihood of the phonological variants, the linguistic contribution to the probability model needs to be removed. Specifically, our phonological model is designed to predict each subsequent phone, using the entire previous column and the column above the new phone as the context. This can be achieved by essentially inverting the probability model of the right column such that the predictor focuses totally on the prediction of $a_{i,0}$, the *phonetic* realization associated with the right column. In practice, this means summing over all observed instances of $a_{i,0}$ following C_{i-1} to compute a total probability for each particular set of $\{a_{i,j}, j > 0\}$, i.e., each unique upper column. This sum then becomes the denominator in a normalization step. Thus, the probability of the right column's phone is modelled as the probability of the phone and the upper column, normalized by the total probability of the upper column, given the left column:

$$
\begin{aligned}
P(a_{i,0}|C_{i-1},\{a_{i,j},j>0\})
&= \frac{P(a_{i,0},\{a_{i,j},j>0\}|C_{i-1})}{P(\{a_{i,j},j>0\}|C_{i-1})} \\
&= \frac{P(C_i|C_{i-1})}{\sum_{a_{i,0}} P(a_{i,0},\{a_{i,j},j>0\}|C_{i-1})} \\
&= \frac{P(C_i|C_{i-1})}{\sum_{a_{i,0}} P(C_i|C_{i-1})}
\end{aligned}
\tag{2}
$$

$$
P(a_{i,0}|C_{i-1},\{a_{i,j},j>0\}) = \frac{P(a_{i,0}|C_{i-1})\prod_{j=1}^{N-1} P(a_{i,j}|a_{i,j-1},a_{i-1,j})}{\sum_{a_{i,0}} P(a_{i,0}|C_{i-1})\prod_{j=1}^{N-1} P(a_{i,j}|a_{i,j-1},a_{i-1,j})}
\tag{3}
$$

To acquire the probability model for the column bigram, the corpus is first processed through forced alignment using standard methods available in SUMMIT, to yield a phonetic transcription associated with each utterance. The ANGIE grammar is then trained on parse trees associated with the corpus. Next, the corpus is reparsed, but this time using the trained grammar, and with the intent of producing a column bigram mapping phones to ANGIE phonemes, removing the linguistic predictions through the procedures described above. An attractive aspect of this approach is that the corpus for training does not have to be restricted to the intended application domain, since the language model component of the column bigram probability space has been completely removed. A summary of the steps in this procedure is given in Fig. 7.

To demonstrate the viability of this approach, we have trained the system on a corpus consisting of a mixed set of 80,700 utterances [9] from the Jupiter weather domain and 13,800 utterances from the Mercury flight reservation domain. The trained model was then tested on an independent test set of 848 utterances exclusively from the Mercury domain. Results are summarized in Table 1. We were able to realize a significant reduction in word error rate, compared with the SUMMIT baseline system, when training on a trigram language model. Perhaps more significantly, when we evaluated on *understanding* error rates, the performance improvement was even greater: concept error rate dropped from 11.9% in the baseline system to 10.4% with the phonological probability modeling, suggesting that the probabilities are differentially helping the content words. For further details concerning these experiments, please see Seneff and Wang (2002).

3.3. Modeling unknown words

One of the most significant applications for ANGIE in subword modeling is both the detection and the characterization of new words. Our approach to this problem is predicated on the notion that the known words can serve as a model for the unknown words: by decomposing words into their linguistic constituents, novel combinations of these constituents can yield representations for the unknown words. To fully characterize new words, one needs both their phonemic and their graphemic representations. Thus, if a subword hierarchy can capture both of these aspects, then it has utility to provide both constraint and valuable linguistic knowledge.

In our early work on sound-to-letter and letter-to-sound tasks (Seneff et al., 1996) we formulated a grammar whose terminals were the letters of the spelled form, with the preterminals encoding phonemic information. We conducted experiments using the Brown corpus and obtained competitive results on the letter-to-sound task (91.5% phoneme accuracy on an unseen test set). For the sound-to-letter task, we modified the search such that the preterminal phonemes were provided as inputs and the parsing process then predicted the most likely letter sequence corresponding to these phonemic specifications. This strategy gave a reasonable performance (89.2% letter accuracy on an independent test set) but it assumed a perfect phonemic transcription as the input sequence, and it still falls far short of the performance level necessary for new word enrollment.

In later experiments conducted by Chung (2000a), we attempted the much more ambitious *waveform*-to-letters task. These experiments were conducted within the Jupiter weather domain (Glass and Hazen, 1998), and we selected as a test set a set of utterances that contained unknown city names. The task therefore involved first identifying the presence of the portion of the speech waveform associated with the unknown city, subsequently proposing a possible spelling for that city.

We attacked this problem through a two-stage procedure, where the first stage utilized subword structure mainly as a language model in support of phonetic recognition, and the second stage involved parsing the resulting phone graph into a sequence of proposed known and unknown words. For both stages, we utilized a grammar that contained SUMMIT phonetic units as the terminals and *spellnemes* as the preterminals. Thus we mapped phones directly to units that encode both the phonemic and the graphemic information. This approach is distinguishable from the approaches addressed by Bazzi and Glass (xxxx) and Onishi et al. (2001) in that a detailed sound-to-letter system is embedded in the linguistic model characterizing the unknown words. We anticipate that linguistic constraint achieved as a side-effect of the

[9] Since we have available a much larger corpus from Jupiter.

1. Obtain corpus of phonetic alignments using SUMMIT.
2. Train ANGIE grammar on phonetic alignments.
3. Obtain column bigram FST by parsing corpus using trained grammar and inverting right-column production probabilities.
4. Compose diphone FST from SUMMIT with ANGIE column bigram to yield FST mapping phonetic models to ANGIE phonemes.
5. Compose with ANGIE's lexicon.
6. Add standard word n-grams as language models.

Fig. 7. Steps in training phonological rule probabilities using ANGIE and SUMMIT.

Table 1
Speech recognition (WER) and understanding (CER) performance for telephone quality speech collected within the Mercury flight reservation domain

No. of Utterances	WER (%)		CER (%)	
	Baseline	+ Angie PM	Baseline	+ Angie PM
848	17.3	16.3	11.9	10.4

The system which utilized an ANGIE pronunciation model (+ Angie PM) is contrasted with a baseline system that utilized the same set of phonological rules but had no probability model for the alternate pronunciations.

statistical sound-to-letter mappings will provide a richer linguistic model for the unknown words, with the additional benefit (and goal) of providing a full characterization of the unknown words in terms of both their spellings and pronunciations.

Since all of the linguistic information in the first stage was ultimately discarded, we were not required to represent this information perfectly. In fact, for certain words, we felt that the training process would generalize better if we discarded rare forms, allowing their more common homomorphs to stand in for them. A good example to clarify this point is the word "champagne," whose unique spelling is very difficult to predict from observations of other words in English. An English sound-to-letter system not explicitly trained on this word would likely produce something like "shampain." We decided to formalize such "mistakes" by introducing these odd spellings intentionally in order to reduce the perplexity of the task and better generalize the models. We realized further that even the boundaries of the words were not necessary to preserve in the first stage of our system. We therefore decided to license a realignment of word boundaries by reorganizing the syllables into foot-like units, each of which contained a single stressed syllable and zero or more unstressed syllables on either side. Furthermore, we developed an iterative procedure which realigned these foot-like units with each iteration. Each realignment would support the same phonetic sequence as the original but with a reduced perplexity. The result of all of this training was a significant net reduction in the size of the FST, along with a reduction in the perplexity of the task, both of which are positive outcomes. The foot-like pseudo-words were supported by a standard class trigram, to yield additional constraint in the search.

The second stage of this system parsed the phone graph into ANGIE parse trees, this time using a grammar that had been trained on the Jupiter word lexicon, and allowing unknown words to compete with known words in the search. Table 2 reports some recognition and under-

Table 2
Recognition performance in terms of word error rate (WER) and concept error rate (CER) in the Jupiter weather domain, for utterances containing out-of-vocabulary city names

	WER (%)	CER (%)
Baseline	24.6	67.0
Two-stage ANGIE	15.6	31.3
Three-stage ANGIE/TINA	17.4	21.8

The percentages indicate error rates for *all* the words. Unknown words were counted as correct if they were identified as such. The baseline system had no capability to handle unknown words.

standing error rates for three different systems, where words are considered as correct if they are unknown to the recognizer and correctly identified as unknown cities. Every utterance in this selected test set contained an unknown city, although the systems were unaware that this was the case. All systems used the same set of context-dependent acoustic models. The baseline system was a standard SUMMIT recognizer configuration utilizing a word trigram language model, but with no capability to deal with unknown words. The two-stage system utilized the foot-based ANGIE grammar reconfigured as an FST in the first stage, supported by a foot-trigram language model. The output of this stage was a phone graph that was subsequently searched in a second stage, parsing with an ANGIE grammar that mapped to the recognizer vocabulary, supporting novel unknown generations as unknown words. An optional third stage parsed the word graph proposed by the second stage using our TINA natural language system (Seneff, 1992), which sometimes favored a solution that was suboptimal in the second stage scores. The two-stage system yielded a 36.6% reduction in word error rate (WER), and a 53.3% reduction in concept error rate (CER), compared to the baseline system. The reason that CER is not 100% for the baseline is that other concepts, such as dates, state names, and topic of inquiry (e.g., "will it rain?") are also counted. With the addition of NL support, the concept error rate improved further to 22%, a net reduction of 67.5%, although this was accompanied by an increase in recognition error.

3.4. Automatic acquisition of new words

In addition to proposing unknown words, the system described above is also capable of proposing spellings for these words. Some examples of proposed spellings are given in Fig. 8. A recognition evaluation of the proposed spellings in terms of letter substitutions, insertions, and deletions, was computed for the unknown words that were correctly tagged as such. The result was a 57.8% letter error rate, which, while quite high, is still substantially better than chance performance. The significant result is that we have formulated a procedure for modeling unknown words by a technique of generalizing from the known words, and have been able to locate the unknown words in a user utterance and propose a set of alternate spellings and pronunciations for these words.

alameda	$\longrightarrow$	alumida
hanover	$\longrightarrow$	anover
hatteras	$\longrightarrow$	sateras
madagascar	$\longrightarrow$	madigasgar
mapleton	$\longrightarrow$	mapelton
mountainview	$\longrightarrow$	mountonvue
youngstown	$\longrightarrow$	janston

Fig. 8. Some examples of unknown cities and their proposed spellings, produced by the ANGIE two-stage recognizer. Spellings were extracted from the letter-phonemes at the preterminal layer of the ANGIE parse tree.

A further experiment aimed at improving the sound-to-letter performance, conducted by Gabovich (2002), utilized the PhoneBook (Dupont et al., 1990) isolated word corpus as the acoustic data. PhoneBook is a set of approximately 92,000 isolated word utterances spoken over the telephone by a large number of native speakers of American English. It utilizes a vocabulary of about 8000 words, and the data have been assigned to speaker-disjoint and vocabulary-disjoint training and test sets.

Our interest was in understanding how reliably we can spell the unseen words from PhoneBook if we choose to represent them only by the column bigram FST obtained by training an ANGIE grammar on the corpus. We compared performance on the training set with that on the test set, to measure how well the models are able to generalize to unseen data. We are aware that sparse data problems will cause a certain percentage of the words to fail, even if given a perfect phonetic transcription, which would force these words to choose a suboptimal solution. But a more serious source of error is the difficult recognition task of producing a phonetic transcription from a waveform without explicit knowledge of a lexicon.

We began with a lexicon of words in PhoneBook represented in terms of morph sequences whose pronunciations were in turn represented as phonemic units.[10] We developed a procedure to convert the phonemic baseforms into *spellnemic* baseforms by utilizing a letter-to-phoneme ANGIE grammar and then inferring the correct spellneme by associating the terminal letter sequence with the corresponding preterminal phoneme. We could then derive an ANGIE grammar mapping *phonemes* to spellnemes semi-automatically. Furthermore, we took advantage of the research on phone-to-phoneme modeling to create a statistical representation of the phonological rules accounting for the variations in pronunciation of the individual words. We trained the phone-to-phoneme grammar on aligned phonetic transcriptions for the training corpus, and derived a corresponding column bigram FST. In parallel, the phoneme-to-spellneme grammar was trained on phonemic representations automatically obtained by parsing the letters of the training corpus using a letter-to-phoneme grammar, verified at the morph level against the lexicon. The language model was then just a composition of the two resulting column bigram FSTs. Weights were optimized on an independent development set.

This system was tested on PhoneBook data, with the main goal of observing how well training would generalize to words that the system had never observed in training. The task is more difficult than phonetic recognition, in that a sound-to-letter system is embedded in the overall task. For example, a recognition of "fragmental" as "fragmittle" has only a single phonetic error (missing /n/), but gets a 50% letter error rate.

Results are summarized in Table 3. Overall letter error rate (LER) increased from 34.1% to 41.0% when comparing the training set with the test set, which we feel reflects fairly good generalization capabilities. We were also interested in assessing how well the system would perform on the test set if the phonetic transcription were perfect. Notice that this is different from and more difficult than the phoneme-to-letter task discussed in Section 3.3, since it is mapping from *phones* to letters. This experiment will measure the capabilities of the sound-to-letter system independent of the phonetic recognition subtask. We obtained a LER of 12.7% using as inputs the forced phonetic alignments, for the subset (70.4%) of the test set that had any solution at all through the FST space. Nearly 30% failed to parse, clearly indicating that we need to add a back-off mechanism to support them. However, overall performance on the set that parsed versus the set that failed drops by less than 6%.

Table 3
Recognition performance (letter error rate) on the training and test sets for the task of automatically proposing a spelling of an unknown spoken word, for the PhoneBook telephone-quality isolated word corpus

	Sub (%)	Del (%)	Ins (%)	LER (%)
Training Corpus	16.6	7.5	10.0	34.1
Test Corpus	21.2	9.3	10.5	41.0

See text for details.

3.4.1. Integrating pronunciation and spelling information

In the context of an interactive dialogue system, there are further options available to help with the task of unknown word acquisition. Having detected that there may be an unknown city, the system can solicit from the user a spoken spelling form for the word. An ANGIE grammar with letters as terminals can be used to parse a letter graph produced by the SUMMIT recognizer. This independent source can be matched against the proposed solutions from the word pronunciation in order to select something that is consistent with both sources. A final resource that is available with telephone input is a keypad entry of the unknown word. This provides the interesting constraint that each key disambiguates into one of three possible letters. This can be formulated as a strict language model and provide further constraint to the problem.

We have been pursuing the above ideas in a joint research project with Chung.[11] For the experiments described below, we have created an

[10] This lexicon had been prepared by Livescu as part of her research on duration modeling.

[11] Now at CNRI in Washington, DC.

ANGIE lexicon of about 100,000 personal names, originally obtained from the Web, and have trained ANGIE parse trees on this corpus to produce a model mapping phonemes to letters.

In an initial experiment (Chung and Seneff, 2002), we developed a recognition system that is able to integrate information from a keypad input of the spelling of the word with information culled from a spoken pronunciation of the word, as schematized in Fig. 9. We defined the search space by composing an FST mapping phonetics to graphemics with an FST specifying all possible pronunciations obtainable from the keypad inputs. We also incorporated a morph bigram for increased linguistic support, where the possible organizations of letters into morph units are determined by the parsing grammar.

To evaluate this idea, we conducted experiments on both the OGI name corpus (Cole et al., 1992) and a set of enrollment data obtained from our Mercury system (Seneff and Polifroni, 2000). In both cases, about 16% of the names were not present in our lexicon, an indication that the unknown word problem would be unavoidable in a personal name recognition task. The OGI set contains isolated first and last names, whereas the Mercury data are utterances containing both first and last name spoken sequentially.

Results are summarized in Table 4. The system performs very well on letter accuracy for the in-vocabulary portion of both sets, as might be expected. However, it should be pointed out that this system has no explicit knowledge of the vocabulary that it was trained on, such as a word lexicon. It is encouraging that the system was able to obtain a perfect spelling for nearly half of the unknown words. If the search were restricted by a word lexicon, clearly none of the OOV words would have obtained a correct spelling. For further information on this topic, please see Chung and Seneff (2002).

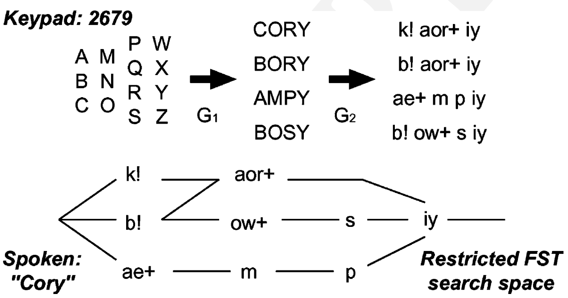

Fig. 9. A schematic for integrating keypad input with phonetic recognition, to produce a hypothesized spelling and pronunciation for the name "Cory." Entered at the keypad is the sequence "2679," producing a total of 144 possible four-letter sequences. Using a subword language model, FST G_1 sets out probable names, and FST G_2 maps the letters to phonemes probabilistically, based on a grammar encoding letter-to-sound rules.

Table 4
Performance results for an experiment integrating telephone keypad inputs with spoken names, to produce hypothesized spellings for the names

Test set	IV subset (84%)		OOV subset (16%)	
	LER (%)	WER (%)	LER (%)	WER (%)
Mercury	1.7	8.1	12.0	43.2
OGI	1.8	8.1	13.3	57.3

Letter error rates (LER) and word error rates (WER) are reported for the in-vocabulary (IV) and out-of-vocabulary (OOV) portions of the Mercury and OGI test sets. Both sets have about a 16% OOV rate.

An extension of this work resulted in a system that can recognize a *spoken* spelling of a word jointly with the corresponding *pronounced* word, using an integrated solution that improves the recognition of the spelled letters by incorporating the constraints of a sound-to-letter model applied to the pronounced word. We have integrated this technology into a dialogue system that can learn new words by prompting a user to speak and spell the word in a single turn (Chung et al., in press). We have thus far only incorporated this capability into a user enrollment phase in the Orion task delegation system (Seneff et al., 2000), but we expect it to be much more generally applicable. Table 5 gives its letter and word error rates on a corpus of telephone quality "speak-and-spell" data, divided into in-vocabulary and OOV subsets.

3.5. Novel FST configurations 1104

All of the research efforts described thus far involving an FST formulation of ANGIE's parse structure have taken the point of view that the parse trees are decomposed into a simple column–

Table 5
Letter error rates (LER) and word error rates (WER) for speak-and-spell utterances for peoples' first and/or last names

Set	No. of Utterances	LER (%)	WER (%)
In-Voc	416	8.4	27.4
OOV	219	12.4	46.1

Results are reported separately for words that were in the 100,000 word training lexicon (In-Voc) and words that were not part of the training lexicon (OOV). See (Chung et al., in press) for details.

column transition matrix. This approach depends upon direct observation in the training set of every unique column pattern in every appropriate column context, and therefore can suffer from sparse data problems. Recent research by Mou et al. (2001) has been able to successfully encode the *entire parsing mechanism* into an FST formulation, thus retaining the generality that the rules achieve directly in the FST representation. His strategy is to use a recursive transition network formulation to encode the context free rule component, which can produce as outputs a detailed parse of the input phone sequence. Subsequently, each layer can be separately modelled so as to ignore all of the elements that are irrelevant to that layer, inserting just the portion of the probability model that is provided for that layer. By jointly composing the FSTs representing all of the layers, the entire probability model can be inserted into the resulting FST composition. While the resulting FST is substantially larger than the FSTs obtained from the column bigram approach, this formulation is attractive because it provides a detailed parse of each word, and because it permits us to explore a variety of different probability formulations to help identify which aspects of Angie's probability model are most crucial.

4. Summary and future work

This paper describes a framework for acquiring subword linguistic knowledge by parsing letters and/or spoken pronunciations into words, via a context-free grammar, and acquiring a supporting probability model from a corpus of observations within the domain of interest. We have identified several different ways in which such a framework has utility in tasks related mainly to speech recognition. These include letter-to-sound and sound-to-letter modeling, acquiring a probability model for phonological realizations of words in fluent speech, developing a duration model that takes into account the hierarchy and also produces as a side-effect a word-by-word speaking rate estimate, acquiring a model for unknown words by generalizing from the observed known words, and obtaining a high quality phonetic graph in the first stage of a two-stage large-vocabulary recognition task. The ideas described here encompass research that I have been conducting over the last several years, collaboratively with both students and researchers in the SLS group.

We are encouraged by the results of the phonological modeling, which demonstrated significant reductions in understanding error rates in our Mercury flight reservation domain. Probability modeling for phonology might have even higher payoff in recognition involving human–human dialogues, such as the Switchboard corpus (Godfrey et al., 1992), where the speaking style is likely to be considerably more casual than that used in spoken interactions with a computer dialogue system.

We anticipate that the Angie hierarchical representations can play a role in subword modeling for speech synthesis. For instance, a correct durational model is more critical in speech synthesis, and it is known that phoneme durations depend significantly on the position of the phoneme in the syllable and of the syllable in the word. Furthermore, Angie's hierarchical framework might provide a convenient mechanism to aid in unit selection for concatenative speech synthesis.

The main original motivation for characterizing word substructure was to be able to model unknown words as derivative from the substructure of known words. The ability to support the automatic acquisition of new words to both the recognition and understanding components of a spoken conversational system will likely lead to a breakthrough in dialogue system design. A system's ability to immediately augment its working vocabulary with a list of names obtained from a Web page being presented to the user will greatly enhance the set of services it can offer to its user population. Furthermore, if the user can simply speak and spell a word they would like to see added, they are empowered to configure the system in ways that will be of much greater use to them in the future. Ongoing research is aimed at developing conversational systems with flexible vocabularies, where proper nouns presented to the user in Web pages are automatically added to the system's working vocabulary, and the user is empowered to personalize the system to their own favored information sources through natural spoken interaction.

Acknowledgements

The research described in this paper would not have been possible without the contributions made by many former and current members of the Spoken Language Systems Group. They include, in alphabetical order, Grace Chung, Vladislav Gabovich, Raymond Lau, Karen Livescu, Helen Meng, Xiaolong Mou, Aarati Parmar, and Chao Wang. I am especially indebted to Grace Chung, whose doctoral thesis made significant advances towards the goals that were originally conceived for the ANGIE framework. Victor Zue read drafts of the manuscript and made many helpful suggestions.

References

Allen, J., Hunnicutt, M.S., Klatt, D., 1987. From Text to Speech: The MITalk System. Cambridge Studies in Speech and Science Communication. Cambridge University Press, Cambridge.

Bacchiani, M., Ostendorf, M., 1998. Using automatically derived acoustic subword units in large vocabulary speech recognition. In: Proc. ICSLP'98, Vol. IV, Sydney, Australia, November 1998, pp. 1319–1322.

Bazzi, I., Glass, J., xxxx. Learning units for domain-independent out-of-vocabulary word modeling. In: Proc. Eurospeech, Aalborg, Denmark, pp. 61–64.

Chomsky, N., Halle, M., 1968. The Sound Pattern of English. Harper & Row, New York, NY (Republished in Paperback, MIT Press, Cambridge, MA, 1991).

Chung, G., 1997. Hierarchical Duration Modeling for a Speech Recognition System. S.M. Thesis, MIT Department of Electrical Engineering and Computer Science.

Chung, G., 2000a. Automatically incorporating unknown words in Jupiter. In: Proc. ICSLP, Beijing, China, October 2000, pp. 520–523.

Chung, G., 2000b. A three-stage solution for flexible vocablary speech understanding. In: Proc. ICSLP 2000, Beijing, China, October 2000, pp. 266–269.

Chung, G., 2001. Towards multi-domain speech understanding with flexible and dynamic vocabulary. Ph.D. Dissertation, MIT Department of Electrical Engineering and Computer Science, June 2001.

Chung, G., Seneff, S., 1997. Hierarchical duration modeling for speech recognition using the ANGIE framework. In: Proc. EUROSPEECH'97, Rhodes, Greece, September 1997, pp. 1475–1478.

Chung, G., Seneff, S., 1998. Improvements in speech understanding accuracy through the integration of hierarchical linguistic, prosodic, and phonological constraints in the Jupiter domain. In: ICSLP'98, Sydney, Australia, December 1998, pp. 935–939.

Chung, G., Seneff, S., 2002. Integrating speech with keypad input for automatic entry of spelling and pronunciation of new words. In: Proc. ICSLP '02, Vol. III, Denver, CO, September 2002, pp. 2061–2064.

Chung, G., Seneff, S., Wang, C., in press. Automatic acquisition of names using speak and spell mode in spoken dialogue systems. In: Proc. HLT-NAACL, Edmonton, Canada, May 2003.

Church, K.W., 1983. Phrase-structure parsing: a method for taking advantage of allophonic constraints. Ph.D. Thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA.

Cohen, M.H., 1989. Phonological structures for speech recognition. Ph.D. Dissertation, University of California, Berkeley, CA.

Cole, R. et al., 1992. A telephone speech database of spelled and spoken names. In: Proc. ICSLP'92, Banff, Canada, October 1992.

Cook, G., Robinson, T., Christie, J., 1998. Real-time recognition of broadcast news. In: Proc. ICSLP'98, Vol. IV, Sydney, Australia, November 1998, pp. 1319–1322.

Dupont, S., Bourlard, H., Deroo, O., Fontaine, V., Boite, J.-M., 1990. Hybrid HMM/ANN systems for training independent tasks: experiments on phonebook and related improvements. In: Proc. ICASSP, Albuquerque, NM.

Gabovich, V.Y., 2002. A multi-stage sound-to-letter recognizer. M.Eng Thesis, MIT, May 2002.

Gauvain, J.L., Lamel, L.F., Adda, G., Adda-Decker, M., 1993. Speaker independent continuous speech dictation. In: Proc. EUROSPEECH'93, Berlin, Germany, September 1993, pp. 125–128.

Glass, J., Chang, J., McCandless, M., 1996. A probabilistic framework for feature-based speech recognition. In: Proc. ICSLP'96, Philadelphia, PA, October 1996, pp. 2277–2280.

Glass, J.R., Hazen, T.J., Telephone-based conversational speech recognition in the Jupiter domain. In: ICSLP'98, Sydney, Australia, December 1998, pp. 1327–1330.

Godfrey, J., Holliman, E., McDaniel, J., 1992. SWITCHBOARD: telephone speech corpus for research and development. In: Proc. ICASSP-92, San Francisco, pp. 517–520.

Greenberg, S., 1999. Speaking in shorthand—a syllable-centric perspective for understanding pronunciation variation. Speech Commun. 29, 159–176.

Hayes, B., 1995. Metrical Stress Theory: Principles and Case Studies. University of Chicago Press, Chicago.

Hetherington, I.L., 1994. The problem of new, out-of-vocabulary words in spoken language systems. Ph.D. Thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, October 1994.

Hetherington, I.L., 2001. An efficient implementation of phonological rules using finite state transducers. In: Proc. Eurospeech 2001, Aalborg, Denmark, September 2001, pp. 1599–1602.

Hetherington, I.L., Zue, V., 1993. New words: implications for continuous speech recognition. In: Proc. Third European Conf. on Speech Communication and Technology, Berlin, Germany, September 1993.

Kahn, D., 1976. Syllable-based generalizations in English phonology. Ph.D. Thesis, Department of Linguistics and Philosophy, MIT, Cambridge, MA.

Kucera, H., Francis, W., 1967. Computational Analysis of Present-Day American English. Brown University Press.

Lau, R., 1998. Subword lexical modeling for speech recognition. Ph.D. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, May 1998.

Lau, R., Seneff, S., 1997. Providing sublexical constraints for word spotting within the ANGIE framework. In: Proc. EUROSPEECH'97, Rhodes, Greece, September 22–25, 1997, pp. 263–266.

Lau, R., Seneff, S., 1998. A unified system for sublexical and linguistic modeling using ANGIE and TINA. In: ICSLP'98, Sydney, Australia, December 1998, pp. 2443–2446.

Meng, H., 1995. Phonological parsing for bi-directional letter-to-sound/sound-to-letter generation. Ph.D. Thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, June 1995.

Meng, H., Hunnicutt, S., Seneff, S., Zue, V., 1996. Reversible letter-to-sound/sound-to-letter generation based on parsing word morphology. Speech Commun. 18, 47–63.

Mou, X., Seneff, S., Zue, V., 2001. Context-dependent probabilistic hierarchical sub-lexical modeling using finite state transducers. In: Proc. EUROSPEECH 2001, Aalborg, Denmark, September 2001, pp. 451–454.

Nguyen, L., Anastasakos, T., Kubala, F., LaPre, C., Makhoul, J., Schwartz, R., Yuan, N., Zavaliagkos, G., Zhao, Y., 1995. The 1994 BBN/BYBLOS speech recognition system. In: Proc. ARPA Spoken Language Systems Technology Workshop'95, Austin, TX, January 1995, pp. 693–696.

Onishi, S., Yamamoto, H., Sagisaka, Y., 2001. Structured language model for class identification of out-of-vocabulary words arising from multiple word classes. In: EUROSPEECH 2001, AAlborg, Denmark, September 2001, pp. 693–696.

Parmar, A.D., 1997. A semi-automatic system for the syllabification and stress assignment of large lexicons. M.Eng Thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, June 1997.

Randolph, M.A., 1989. Syllable-based constraints on properties of English sounds. Ph.D. Thesis, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, September 1989.

Scalise, S., 1986. Generative Morphology. Foris Publications, Dordrecht, Netherlands.

Seneff, S., 1992. TINA: a natural language system for spoken language applications. Comput. Linguist. 18 (1), 61–86.

Seneff, S., 1998. The use of linguistic hierarchies in speech understanding. Keynote Address. In: ICSLP'98, Sydney, Australia, December 1998, pp. 3321–3330.

Seneff, S., Polifroni, J., 2000. Dialogue management in the mercury flight reservation system. In: Proc. ANLP-NAACL 2000, Satellite Workshop, Seattle, Washington, May 2000, pp. 1–6.

Seneff, S., Wang, C., 2002. Modelling phonological rules through corpus-trained linguistic hierarchies. In: ISCA Tutorial and Research Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language, Estes Park, CO, September 2002, pp. 71–76.

Seneff, S., Lau, R., Meng, H., 1996. ANGIE: A new framework for speech analysis based on morpho-phonological modeling. In: Proc. ICSLP'96, Vol. 1, Philadelphia, PA, October 1996, pp. 110–113. Available from <http://www.sls.lcs.mit.edu/raylau/icslp96_angie.pdf>.

Seneff, S., Chuu, C., Cyphers, D.S., 2000. ORION: from on-line interaction to off-line delegation. In: Proc. ICSLP '00, Vol. II, Beijing, China, October 2000, pp. 142–145.

Selkirk, E., 1982. Syllables. In: van der Hulst, Harry, Smith, Norval (Eds.), The Structure of Phonological Representations. Foris, Dordrecht, pp. 337–383.

Weintraub, M., Murveit, H., Cohen, M., Price, P., Bernstein, J., Baldwin, G., Bell, D., 1989. Linguistic constraints in hidden Markov model based speech recognition. In: Proc. ICASSP '89, Glasgow, Scotland, May 1989, pp. 699–702.

Woodland, P.C., Young, S.J., 1993. The HTK tied-state continuous speech recognizer. In: Proc. European Conf. on Speech Commun. and Technology, Vol. 3, pp. 2207–2210.

Zue, V., 1983. The use of Phonetic rules in automatic speech recognition. Speech Commun. 2, 181–186.

Zue, V., Glass, J., Goodine, D., Hirschman, L., Leung, H., Phillips, M., Polifroni, J., Seneff, S., 1991. The MIT ATIS system: preliminary development, spontaneous speech data collection and performance evaluation. In: Proc. European Conf. on Speech Communication and Technology, Vol. 2, Genoa, Italy, September 1991, pp. 537–541.