MODELLING PHONOLOGICAL RULES THROUGH LINGUISTIC HIERARCHIES[1]

Stephanie Seneff and Chao Wang

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA

ABSTRACT

This paper describes our research aimed at acquiring a generalized probability model for alternative phonetic realizations in conversational speech. The approach begins with the application of a set of ordered context-dependent phonological rules, applied to the baseforms in the recognizer's lexicon. The probability model is acquired by observing specific realizations expressed in a large training corpus. A set of context-free rules represents words in terms of a substructure that can then generalize context-dependent probabilities to other words that share the same sub-word context. The model is designed to capture phonetic predictions based on local phonemic, morphologic, and syllabic contexts, thus permitting training on corpora whose lexicon is divergent from that of the intended application. The training corpus consisted of a large set of Jupiter weather-domain speech data [9] augmented with a much smaller set of Mercury flight-domain data [20]. The baseline system utilized the same set of phonological rules for lexical expansion, but with no probability modelling for alternate pronunciations. We evaluated on a test set of utterances exclusively from the flight domain. Using this approach, we achieved a 12.6% reduction in speech understanding error rate on the test set.

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 [2, 7, 23, 24]. Some examples of such rules are shown in Figure 1 – rules which predict, for example, the contexts for flapping an alveolar stop or for palatalizing an alveolar fricative.

As the hidden Markov model (HMM) framework gained in popularity, such formal rules tended to play a less prominent role. Instead, the assumption was made that context effects could be handled by simply defining 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 [10], the issue of accounting for phonetic variability has resurfaced as a significant research problem [17]. For example, a "cheating" experiment conducted by McAllister et al. [16] 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 Mateus [5]) 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" can then be gleaned from the observation space, and frequency counts can be tabulated to yield probability estimates for those rules. Thus, one does not rely on a linguist to define the formal rules, but rather lets the data dictate which rules are most productive. Such an approach is attractive in that it is able to generalize from observed words to unobserved words with similar surrounding phonetic context.

Phonologists have long been aware that syllable structure plays an important role in predicting phonological reductions [15]. For example, consonants in syllable onset position are far less likely to be reduced than in coda position. It is therefore perhaps surprising that the speech recognition community has typically ignored the syllable in characterizing word lexical entries and in modelling phonological effects.

An excellent description of the role of the syllable in the switchboard data has been presented by Greenberg [11]. 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. For example, the words "no" and "know" have identical syllable structure but significant differences in their phonological expression due in all likelihood to their distinct functional roles in the language. Similarly, the "ing" inflexional suffix is

[1]This work was supported by DARPA under contract N66001-99-1-8904 monitored through Naval Command, Control and Ocean Surveillance Center.

{**left**}	**core**	{**right**}	→ **realizations**
{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.

much more likely to be reduced to "in"' than other instances of "ing" (compare "Redding" with "reading").

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 separate parsing mechanism is applied to a large corpus, 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 language model information. This point is important because it permits generalization from a common word to a rare word with the same local syllable 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 [8]. While SUMMIT's approach is quite distinct from the standard HMM formulation, it has been shown to produce state-of-the-art performance in phonetic recognition tasks [1, 12]. In our research, we are concerned almost exclusively with telephone-quality conversational speech, collected through interactions between users and various domain-specific conversational systems [20, 9, 21].

In the remainder of this paper, we first describe SUMMIT, including its finite state transducer (FST) formulation and phonological modelling framework [14]. We then describe our phonological probability model, which uses the ANGIE system [18, 19] to obtain sub-word linguistic hierarchies. We describe the two-step process of acquiring the trained FST mapping phones to unique sub-word contexts, and explain how the recognizer is reassembled to incorporate the acquired probability model. In Section 4, we report performance on unseen test data from the Mercury flight domain, giving results from both speech recognition and speech understanding experiments. Finally, we conclude with a summary and a look to the future.

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. These include, for example, epenthetic silence insertion at locations of voicing change, gemination ("fr*om M*aine") and palatalization ("gas *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. In all, there are about 250 generalized rules accounting for these various phenomena, similar to those shown in Figure 1. Hazen et al. [13] provide a detailed description of pronunciation variations on different levels and how they are modelled in the SUMMIT system.

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 according to an ordered set of phonological rules into alternative pronunciations. 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 then organized as a composition of these FSTs:

$$C \circ (P \circ L) \circ G \tag{1}$$

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 an ordered set of phonological rules, L is the lexicon mapping idealized phonemic pronunciations to words, and G is the language model.

The resulting pronunciation model, $P \circ L$, is a heavily shared but *unweighted* network mapping phones to words. Its FST-based implementation was first described in detail in [14]. Recently, an EM training algorithm was developed to learn FST weights and applied to the problem of pronunciation modelling [22]. In this paper, we describe an alternative technique based on parsing words into a linguistic hierarchy that encodes syllable context, utilizing the ANGIE framework to generalize probabilities among similar substructure.

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 context, including phonemic, morphologic, and syllabic contexts. In this section, we describe the details of the probability model. We start by providing a description of how to acquire the sub-word context and the associated probabilities using the ANGIE system. We then describe how the ANGIE model is transformed into the phonological model and incorporated into the recognizer.

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

Figure 2: ANGIE parse tree for the word "introduce," 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 phoneme "t!" (*Note: The phoneme layer utilizes diacritics to encode onset (!) and stress (+).*)

3.1. ANGIE Framework

Over the past several years, we have been exploring the utility of a parsing framework we call ANGIE [18, 19] for modelling word substructure. The original intent was to model phonology, morphology, and syllable constraints in a shared probability framework, with the goal of modelling 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 [6, 19], an explicit accounting of unknown words in speech recognition tasks [3], and strong linguistic support for a high-performance phonetic recognizer as the first stage in a multi-stage recognition framework [4]. In recent experiments we have explored the possibility of encoding the ANGIE probability model as a finite state transducer mapping phones to phonemes. It is this mechanism that can be used 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. The grammar is intentionally arranged 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 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, with the ultimate goal of attaching the learned probabilities to arcs in a finite state network.

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

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

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 is the total number of layers in the parse tree. The column index i begins at the left of the parse grid, and the row index j begins at the bottom of each column. 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. The probabilities, $P(a_{i,0}|C_{i-1})$ and $P(a_{i,j}|a_{i,j-1}, a_{i-1,j})$, are trained by tabulating counts in a corpus of parsed sentences, mapping words to their corresponding phonetic realizations.

3.2. Phonological Probability Model

As was mentioned earlier, 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 modelling the likelihood of the phonological variants, the linguistic contribution to the probability model needs to be removed. Our goal is to produce the probability of the phone sequence, given the word. We are making the simplifying assumption that each word, in a specific phonetic realization, has a unique parse into a sequence of columns. Furthermore, to cope with sparse data problems and to assure generalization, the context conditioning is restricted to column pairs. Specifically, our phonological model (PM) will predict each subsequent phone, using the entire previous column and the column above the new phone as the context:

$$PM = P(a_{i,0}|C_{i-1}, \{a_{i,j}, j > 0\}) \quad (3)$$

This can be computed by essentially inverting the ANGIE column probability model such that the predictor focuses totally on the prediction of $a_{i,0}$, the *phonetic* realization associated with the right column. Using a Bayesian formulation, this probability can be expressed as the probability of the phone *and* the upper column (i.e., the entire right column), normalized by the total probability of the upper column, given the left column:

$$PM = \frac{P(a_{i,0}, \{a_{i,j}, j > 0\}|C_{i-1})}{P(\{a_{i,j}, j > 0\}|C_{i-1})} \quad (4)$$

The denominator in Equation 4 can be computed as the marginal probability of the joint probability $P(a_{i,0}, \{a_{i,j}, j > 0\}|C_{i-1})$ summing over all instances of $a_{i,0}$, i.e.,

$$P(\{a_{i,j}, j > 0\}|C_{i-1}) = \sum_{a_{i,0}} P(a_{i,0}, \{a_{i,j}, j > 0\}|C_{i-1}) \quad (5)$$

Substituting Equation 5 into Equation 4 and recognizing that $\{a_{i,0}, \{a_{i,j}, j > 0\}\}$ is simply the column C_i, we can obtain:

$$PM = \frac{P(C_i|C_{i-1})}{\sum_{a_{i,0}} P(C_i|C_{i-1})} \quad (6)$$

$P(C_i|C_{i-1})$ is the ANGIE column bigram probability, which can be computed according to Equation 2.

In practice, Equation 6 means that we first sum the ANGIE column bigram probability over all observed instances of $a_{i,0}$ to compute a total conditional probability for each particular set of $\{a_{i,j}, j > 0\}$, i.e., each unique upper column. This sum then becomes the denominator to normalize the column bigram probability.

3.3. Modelling Word-Boundary Effects

A critical aspect in phonological modelling is the effective capturing of word-boundary effects. ANGIE's probability model retains only the last *phone* of the preceding word as context 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. 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 "l!" to /l/ is very general.

Our solution was to implement the capability to specify in the ANGIE rules an explicit set of phoneme-to-phone mappings which, if appearing word-final, should be liaisoned 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 rule only states that "s" realized as /sh/ is liaisoned, and the observations control the actual set of word-boundary ties that are sanctioned. This mechanism effectively retains right context dependency constraints (with probabilities trained from observations) across word boundaries. Except in the liaisoned condition, all other word-ending phones map to a generic word-start node, which then advances to completely context-independent realizations of word-start columns, as well as to a generic *pause* model.

3.4. Training Procedure

We have discussed the ANGIE framework for modelling word structure, and we showed 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 the SUMMIT system and the ANGIE framework.

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 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, as described previously, to remove linguistic dependencies. The resulting column-bigram model is written out as a finite state transducer, 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, it emits the syllable layer symbol (encoding stressed root, function word, prefix, etc.), which has the desired effect of preserving the distinct statistics of these syllable types.

above	: ah pre b! ah+ v sroot
airlines	: ehr+ sroot l! ay+ n sroot s_pl isuf
either	: (iy+ , ay+) sroot dh! er dsuf
in	: en_in fcn
the	: dh! iy_the fcn
west	: w! eh+ st sroot

Figure 3: Representative baseforms from ANGIE's word lexicon. These include a small set of alternate pronunciations ("either"), as well as some inflection-specific phonemes, ("s_pl"), some word-specific phonemes, such as "iy_the", and some diphone units, such as "st." Symbols such as "sroot" and "fcn" identify the syllable category. See text for further details.

3.5. Assembling the Speech Recognizer

We have described a procedure to create a finite state transducer mapping phones to ANGIE's sub-word units, with probabilities attached to the arcs 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, which is enhanced to include markers for stress (+) and onset (!) position, as well as some diphone units such as "st." In all, there are about 140 unique phoneme units. In addition, the syllable-identity symbols are inserted at the end of each syllable, consistent with the phone-phoneme FST. Some examples of lexical entries in ANGIE's format are given in Figure 3. The phone-phoneme FST is then composed with this new baseform file to yield a transducer, $P_A \circ L_A$, mapping phones to words with weights on the arcs. The rest of the recognizer constraints, C and G, are kept the same; 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 sequence "in the." Notice that, for example, "in" can be realized as a syllabic nasal (/en/), and the rules allow for a "stop-like" /dh/ via an optional insertion of a closure interval (/dcl/). The vowels for these function words have several different realizations.

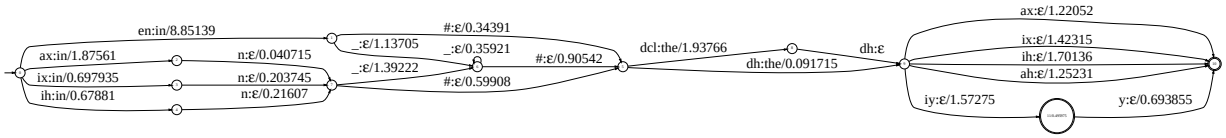

Figure 4: Example of a pronunciation graph created by the system for the word sequence "in the." Each arc is labelled with the input and output symbols, and the corresponding negative log probability.

Set	No. Utts.	Baseline	+ Angie PM
test_all	848	17.3	16.3
test_clean	759	13.9	13.0
test_noisy	89	41.9	39.6

Table 1: Speech recognition performance (in word error rate) for a system which utilized an ANGIE pronunciation model, as contrasted with a baseline system that utilized the same set of phonological rules but lacked probabilities on the arcs. Results are given on the overall set, as well as on the "clean" and "noisy" sub-sets.

Set	No. Parsed Utts.	Baseline	Angie PM
test_all	729	11.9	10.4

Table 2: Speech understanding performance (in concept error rate) using a recognizer which utilized an ANGIE pronunciation model, as contrasted with a system using a baseline recognizer without probabilities on the pronunciation arcs.

4. EVALUATION EXPERIMENTS

4.1. Speech Recognition

To demonstrate the viability of this approach, we trained the system on a corpus consisting of a mixed set of over 80,700 utterances from the Jupiter weather domain [9] and 13,800 utterances from the Mercury flight reservation domain [20], and tested it on an independent set of 848 utterances in the Mercury domain.

The Mercury recognizer has a vocabulary of 1636 unique words (without underbars); however, multiple word units are defined to build class n-gram models. 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, as well as the diphone-to-phone mapping FST. 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 pronunciation probabilities, are tuned on development data, and the final results are reported on unseen test data.

Table 4.1 summarizes the recognition performance of the baseline and the ANGIE systems on the test set. We were able to realize a 5.8% relative reduction in word error rate with the ANGIE pronunciation model, and the improvements are consistent on both clean and noisy utterances.

4.2. 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 about 86% of the utterances (full parse or robust parse). 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 recognizers. The concept error rate was reduced by 12.6% when the ANGIE pronunciation model is used in recognition, which is a substantially greater relative gain than was obtained for speech recognition.

We have two possible explanations for the difference in performance gains for speech recognition as opposed to speech understanding. The first one is that, without probability training, words with many alternative pronunciations obtain an unintended boost because of the multiple ways that they can match against the lexical entries. We have observed that short function words often have much bushier phonetic expansions due to their strong influence from external word context, as well as their tendency to be reduced (see, for example, Figure 4). By supplying probabilities to their alternative arcs, we effectively reduce their relative total word score, leading to a reduction in recognition performance on these function words as contrasted with the content words. However, words like "a" and "the" are typically ignored at the level of concept understanding, and hence their poorer performance is of no consequence to understanding.

Another explanation is that the utterances which fail to parse, on average, perform less well when probability training is included. This could be correlated with their tendency to include words that are rarely used, and hence that might suffer from inadequate observation training.

5. 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. We were able to leverage existing SUMMIT speech recognition tools, including the standard set of phonological rules and the stan-

dard class n-gram language models. An FST mapping subword structure to phonetic realizations with associated probabilities was derived by parsing a large corpus of observed phonetic sequences, and reinserted into the recognizer's full FST, replacing the original component FST, which had no probabilities. In speech understanding experiments, we were able to obtain a 12.6% relative reduction in concept error rate.

An obvious extension of this work is to integrate it with the research described in a companion paper [13]. In their work, Hazen et al. have determined that, in the absence of probability training on the phonological variants, a system with a parsimonious set of phonological rules is superior to a system which has the full set of standard SUMMIT rules, in terms of both memory requirements and recognition accuracy. The parsimonious system was obtained by only retaining rules that involve deletions and/or insertions, thus eliminating schwa reduction, palatalization rules, etc. A retraining of the models is a necessary concurrent step. It would be interesting to see whether a framework utilizing the parsimonious rule set can benefit from our probability training methods to the same extent as was realized in our experiments.

We also plan to apply our approach, which incorporates subword linguistic hierarchy in modeling phonetic variations, to switchboard data. As mentioned in Section 1, a study [16] has demonstrated great potential for recognition improvements with effective pronunciation modeling. In addition, an analysis of phonetically transcribed data [11] showed a need to account for syllable structure and morphology in predicting phonetic variations. It will be interesting to test the effectiveness of our modeling approach in this challenging domain.

6. REFERENCES

1. J. Chang and J. Glass, "Segmentation and Modeling in Segment-based Recognition," *Proc. Eurospeech '97*, Rhodes, Greece, pp. 1199–1202, 1997.
2. M. H. Cohen, *Phonological Structures for Speech Recognition*, Ph.D. Dissertation, U. of California, Berkeley, CA., 1989.
3. G. Chung "Automatically Incorporating Unknown Words in Jupiter, *Proc. ICSLP*, pp. 520-523, Beijing, China, Oct. 2000.
4. G. Chung, "A three-stage solution for Flexible Vocablary Speech Understanding," *Proc. ICSLP 2000*, pp. 266–269, Beijing, China, Oct. 2000.
5. N. Cremelie and J-P Martens, "In Search of Better Pronunciation Models for Speech Recognition," *Speech Communication* 29, pp. 115-136, 1999.
6. V. Y. Gabovich, "A Multi-stage Sound-to-Letter Recognizer," M.Eng Thesis, MIT, May, 2002.
7. J. L. Gauvain, L. F. Lamel, G. Adda, and M. Adda-Decker, "Speaker Independent Continuous Speech Dictation," *Proc. EUROSPEECH '93*, pp.125–128, Berlin, Germany, Sept. 1993.
8. J. Glass, J. Chang, M. McCandless, "A probabilistic framework for feature-based speech recognition," *Proc. ICSLP '96*, Philadelphia, PA, pp. 2277–2280, October, 1996.
9. J. R. Glass and T. J. Hazen, "Telephone-based Conversational Speech Recognition in the Jupiter Domain," *Proc. International Conference on Spoken Language Processing*, pp. 1327–1330, Sydney, Australia, 1998.
10. J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development." *Proc. ICASSP '92*, pp. 517–520, 1992.
11. S. Greenberg, "Speaking in Shorthand – A Syllable-centric Perspective for Understanding Pronunciation Variation," *Speech Communication* 29, pp. 159–176, 1999.
12. A. Halberstadt and J. Glass, "Heterogeneous Measurements and Multiple Classifiers for Speech Recognition," *Proc. ICSLP '98*, pp. 995-998, Sydney, Australia, November, 1998.
13. T. J. Hazen, I. L. Hetherington, H. Shu, and K. Livescu, "Pronunciation Modeling Using a Finite-State Transducer Representation," *These Proceedings*.
14. I. L. Hetherington, "An Efficient Implementation of Phonological Rules using Finite-State Transducers," *Proc. EUROSPEECH '01*, Aalborg, Denmark, 2001.
15. D. Kahn, "Syllable-based Generalizations in English Phonology," Garland Press, New York, 1980.
16. D. McAllaster, L. Gillick, F. Scattone, and M. Newman, "Fabricating Conversational Speech data with Acoustic Models: A Program to Examine Model-data Mismatch," *ICSLP-98*, Vol. 5, pp. 1847-1850, Sydney, Australia, 1998.
17. H. Strik and C. Cucchiarini, "Modeling Pronunciation Variation for ASR: A Survey of the Literature," *Speech Communication* 29, pp. 225-246, 1999.
18. S. Seneff, R. Lau, and H. Meng, "ANGIE: A new Framework for Speech Analysis based on Morpho-phonological Modelling," *Proc. ICSLP '96*, Philadelphia, PA, vol. 1, pp. 110–113, Oct. 1996.
19. S. Seneff, "The Use of Linguistic Hierarchies in Speech Understanding," Keynote Address, *ICSLP '98*, pp. 3321-3330, Sydney, Australia, December, 1998.
20. S. Seneff and J. Polifroni, "Dialogue Management in the MERCURY Flight Reservation System," *Proc. ANLP-NAACL 2000, Satellite Workshop*, pp. 1–6, Seattle, WA, 2000.
21. S. Seneff, C. Chuu, and D. S. Cyphers, "ORION: From On-line Interaction to Off-line Delegation," *Proc. ICSLP '00*, Vol. II, pp. 142–145, Beijing, China, Oct. 2000.
22. H. Shu and L. Hetherington, "EM Training of Finite-State Transducers and its Application to Pronunciation Modelling," to appear in *ICSLP 2002*, Denver, CO, Sep. 2002.
23. M. Weintraub, H. Murveit, M. Cohen, P. Price, J. Bernstein, G. Baldwin, and D. Bell, "Linguistic Constraints in Hidden Markov Model Based Speech Recognition," *Proc. ICASSP '89*, pp. 699–702, Glasgow, Scotland, May, 1989.
24. V. Zue, "The Use of Phonetic Rules in Automatic Speech Recognition," *Speech Communication* 2, pp. 181-186, 1983.
25. V. Zue, J. Glass, D. Goddeau, M. Phillips, and S. Seneff, "The SUMMIT Speech Recognition System: Phonological Modelling and Lexical Access," *Proc. ICASSP '90*, Albuquerque, NM, April, 1990.