TWO-STAGE CONTINUOUS SPEECH RECOGNITION USING FEATURE-BASED MODELS: A PRELIMINARY STUDY

Min Tang, Stephanie Seneff and Victor Zue

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 USA
{mtang, seneff, zue}@sls.lcs.mit.edu

ABSTRACT

In recent research, we have demonstrated that linguistic features can be used to improve speech recognition for an isolated vocabulary recognition task. This paper addresses two important new research problems in our attempts to build a two-stage speech recognition system using linguistic features. First, through a controlled study we show that our knowledge-driven feature sets perform competitively when compared with similar classes discovered by data-driven approaches. Secondly, we show that the cohort idea can be effectively generalized to continuous speech. Improved recognition results are achieved using this two-stage framework on multiple speech recognition experiments, on conversational telephone quality speech.

1. INTRODUCTION

The idea of using sub-word linguistic features such as manner and place of articulation in automatic speech recognition (ASR) dates back to some of the earliest research efforts in this field. For example, four of the five ARPA-funded speech understanding systems developed from 1971 to 1976 made explicit use of acoustic phonetic knowledge regarding manner of articulation in their initial acoustic phonetic recognition modules [1, 2, 3, 4]. Even for the remaining one [5], which undertook a template matching approach, lexical networks were generated using manner-based phonological rules. Detailed protocol analysis of human spectrogram reading experiments performed in the late seventies [6] also revealed the extensive use of place and manner features. In a series of studies that attempted to quantify the constraining power of manner-based broad classes for lexical access [7, 8, 9], these researchers discovered that lexical candidates can be reduced significantly if manner classes of the phonemes can first be established. Based on these findings, they proposed a two-stage speech recognition framework, in which the first stage segmented and classified the signal into manner-based "broad classes". Lexical retrieval based on this broad class representation will result in a small "cohort" of possible word candidates. In the second stage, more detailed acoustic phonetic analysis, coupled possibly with the use of analysis-by-synthesis techniques [10] and higher level knowledge sources (KSs) such as syntactic and semantic constraints, are applied to the small cohort to achieve an efficient and accurate recognition.

In a recent study [11], we explored ways to incorporate the feature-based, multi-stage recognition approach into current probabilistic speech recognition systems. We explicitly described each speech segment in terms of its manner and place of articulation properties and used these linguistically motivated features as the basic units for acoustic modeling in a first-stage recognizer. We used the N-best outputs from the first stage as a logical choice for a cohort. In the second stage, a state-of-the-art phone-based recognizer re-scored the cohort. Through this approach we achieved a 10% relative improvement over the best known result on the PhoneBook isolated word task [12].

This paper addresses two questions that were left unanswered in our earlier study. First, how do our knowledge-based feature sets compare with those obtained from bottom-up, data-driven approaches that also seek to *discover* regularities from the speech signal? Since these features are ultimately used in a probabilistic framework, we are particularly interested in how well these features "fit" the data. Second, can the cohort concept (as well as its N-best incarnation) be generalized to handle continuous speech? With a potentially infinite hypothesis space, only a very small fraction of hypothesized utterances can be covered by the N-best list. Is it possible to create a cohort that is constrained enough so that the second stage can efficiently search, and general enough so that it can recover from errors incurred by pruning?

Recently, other researchers have also begun to model auditory or articulatory features of sub-word units for speech recognition, as in [13, 14, 15, 16, 17]. These researchers usually employ very sophisticated statistical machinery to describe the dynamic interactions among features at a high temporal resolution. Our approach differs from the above approaches in that we explicitly characterize the structural organization of features into sub-word units at a formal level, thus avoiding any significant changes to the underlying recognizer itself.

The remainder of this paper is organized as follows: We first describe the recognizer we used and the two conversational domains used in our experiments. In Section 3, we describe a controlled study in which our knowledge-driven features are compared with a data-driven approach in speech recognition experiments. Section 4 discusses the challenges we face when generalizing the cohort concept to continuous speech. Finally, the feature-based models and two-stage framework are evaluated on several continuous speech recognition experiments and we conclude with remarks on future research topics.

2. BACKGROUND

SUMMIT [18], a segment-based speech recognition system developed at the Spoken Language Systems (SLS) group at MIT, is used throughout our experiments. SUMMIT provides a prob-

[1] This research is supported in part by the NSF under subaward number 1120330-133982 and by NTT

abilistic framework to decode graph-based observations. Our experiments use the landmark-based configuration of SUMMIT. In this configuration, two types of landmarks are modeled: 1) the transitional landmarks, which correspond to the transition *between* two segments; and 2) the internal landmarks, which indicate significant acoustic events *within* a segment. The segments are traditionally phonetic units. Figure 1 illustrates the segment graph and the two different types of landmarks. Acoustic and linguistic knowledge, e.g. phonetics, phonology, language models, etc., are pre-compiled into a single finite state transducer (FST) [19].

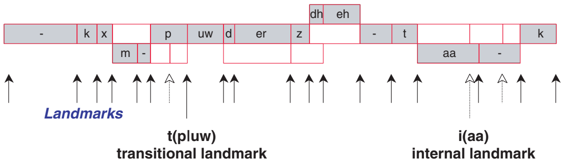

Fig. 1. Graph-based segment representation and two types of landmarks, reproduced from [18].

Our experiments are carried out on two conversational interface domains: Jupiter [20], a weather information domain, and Mercury [21], an air travel information domain. Both systems are continuously operating in our group and provide services via toll-free telephone numbers. Both systems recognize speaker-independent, telephone-quality continuous speech. For the experiments described in this paper, the Jupiter weather domain has a lexicon of 1924 words and a test set of 1888 utterances. The Mercury domain has a lexicon of 1568 words and a test set of 2049 utterances. Each test set contains a "clean" subset which is free of artifacts or out of vocabulary (OOV) words. A fixed training set, which contains over 140k utterances from both domains, is used throughout our experiments. We use as our baseline the same state-of-the-art phone-based landmark models in both domains. In our experiments, the forward search uses a bi-gram and the backward search uses a tri-gram language model.

3. DETERMINATION OF THE BROAD CLASSES

In [11] we described a method to model each speech segment according to its manner- and place-of-articulation features. Speech segments are clustered into broad classes along these two organizational dimensions and acoustic models are subsequently constructed. The question is, when used in a probabilistic recognizer, how well will these knowledge-driven broad-classes fit the data when compared with broad classes derived from a bottom-up, data driven approach? To answer this question, we performed a controlled study to compare these two methods for obtaining the broad classes.

3.1. Manner and Place Features

Manner and place features are explored in [11] because, when they are used as organizational dimensions, sound units belonging to the same "broad classes" demonstrate strong acoustic homogeneity that can be exploited to build more robust acoustic models. These features are also chosen in light of the roles they play in phonological rules [22] as well as in higher level linguistic structure such as syllables [23].

In our approach, phone-sized units are mapped to nine manner and eight place classes as shown in Table 1 and Table 2, respectively. The manner classes are determined based on linguistic theory, but are further refined to maximize acoustic contrast. For example, we distinguish between the closure and release portions of a stop consonant and consider them as two different manner classes, since they are acoustically very different. In another non-standard but effective strategy, vowels are coerced into the traditional place classes of consonants based on their acoustic similarities.

Manner		**Phones**
Schwa	:	ə ɨ ɚ
Vowel	:	ʌ ɛ ɪ ʊ ɑ æ ɔ ɝ iʲ uʷ
Diphthong	:	eʲ ɑʲ ɔʲ ɑʷ oʷ
Semi-Vowel	:	w j ɹ l l̩
Plosive	:	b d g p t k
Closure	:	b̚ d̚ g̚ p̚ t̚ k̚ ʔ ɾ
Fricative	:	f v θ ð s ʃ z ʒ h
Affricate	:	ʧ ʤ
Nasal	:	m n ŋ m̩ n̩

Table 1. Manner assignment to segments used in this research.

Place		**Phones**
Alveolar	:	ɨ ɪ n̩ n s z t t̚ d d̚ ɾ
Dental	:	ð θ
Open	:	ɑ ʌ ə æ h ʔ
Labial	:	ʊ uʷ m m̩ f v w p p̚ b b̚
Lateral	:	ɔ l l̩
Palatal	:	iʲ j ʃ ʒ ʧ ʤ
Retroflex	:	ɝ ɚ ɹ
Velar	:	ɛ ŋ k k̚ g g̚

Table 2. Place assignment for segments used in this research.

Complexity arises when the manner or place property of a segment is dynamic. When the manner of a segment changes in the course of articulation, it often results in a distinct discontinuity in the spectrogram. One way to accommodate such dynamics is through our introduction of additional manner classes, as when we consider a stop to be composed of a closure and a release portion. Phonological rules also help alleviate this problem. For example, the phonological rules in SUMMIT allow affricates to be preceded by a closure as well. When there exist dynamics in place of articulation, the spectral change is often gradual. In this case we allow change of the place of articulation over time for segments, as outlined in Table 3.

Phone		**Place**
ɑʲ	:	Open ⟶ Alveolar
eʲ	:	Velar ⟶ Alveolar
ɑʷ	:	Open ⟶ Labial
oʷ	:	Lateral ⟶ Labial
ɔʲ	;	Lateral ⟶ Alveolar

Table 3. Distinctive left and right place assignments for diphthongs in our scheme.

3.2. Data-Driven Approach

A data-driven approach seeks to discover regularities among speech signals by clustering acoustically similar sound units together. Regularities thus discovered can often have strong correlations to underlying linguistic features. In Table 4 we include a set of ten clusters derived from a data-driven approach as described in [24].

Clustering algorithms will only minimize within-cluster distance while maximizing the between-cluster distance. These derived clusters often organize data around manner or place groups,

Cluster		Phones
1	:	i^{y} ɪ e^{y} ɨ j ü
2	:	o^{w} ɔ ə u^{w} l̩ l w
3	:	ɚ ɝ ɹ
4	:	ʊ ɔ^{j} ʌ ɑ^{w} ɑ ɑ^{y} ɛ æ
5	:	b̚ d̚ g̚ p̚ t̚ k̚ v
6	:	m n ŋ m̩ ŋ̍ n̩
7	:	ʔθ ð f
8	:	t d b p h k g
9	:	z s
10	:	ʃ ʒ ʧ ʤ

Table 4. Clusters derived through a data-driven approach from [24].

but with a heterogeneous decision space sometimes based on manner class membership and at other times on place. In general, vowels and consonants belong to different clusters. There are often two distinct *manner* based clusters for the closure portions and the release portions for stop consonants, which is consistent with the choice we make above. On the other hand, Cluster 1 in Table 4 contains four vowels, a schwa and a glide, which share the "front" property in terms of their *place* of articulation. Thus, the outcome depends solely on which of the competing factors, manner or place, has a greater acoustic manifestation. When we examined the outcome of the landmark measurements clustered using an alternative decision tree algorithm [25], we also observed a mixture of manner- and place-based clusters.

3.3. Speech Recognition Experiments

Although we base our feature sets on linguistic knowledge [26] and acoustic phonetic studies [27, 28], to answer the question of how well they fit the data we compare them with the data-driven clusters in speech recognition experiments.

With our knowledge-based approach, each sub-word unit is decomposed into a manner feature and a place feature. Figure 2 illustrates an effective way to integrate these two feature channels into a single search [11]. We model place of articulation at transitional landmarks and manner of articulation at internal landmarks. A desirable consequence of this mapping is that the resulting model size, i.e. the total number of Gaussian mixture model parameters, is comparable to the size of the three alternatives (see below), which would not be the case if both manner and place were encoded at each boundary.

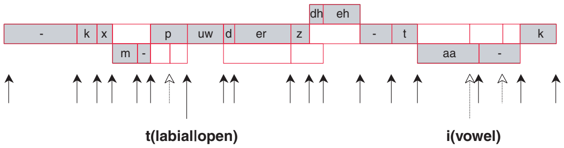

Fig. 2. Modeling place of articulation at transitional landmarks and manner of articulation at internal landmarks, thus integrating information from two feature-channels into one search.

Results are listed in Table 5, where we compare (1) a manner-based system, (2) a place-based system, (3) a cluster-based system and (4) a system integrating both manner and place information as in Figure 2. Our results indicate that cluster-based system performs significantly better than the manner- or place-based system. In part, this may be simply because there were more clusters than in the other models, and thus the cluster-based models can fit the data better. However, with the knowledge-based feature classes and a simple but efficient information fusion scheme (cf. System 4), significantly better performance is achieved than using the data-driven approaches.

System	Model Size	Jupiter	Mercury
(1) Manner	1.78M	30.8	33.1
(2) Place	1.52M	29.5	30.5
(3) Cluster	1.83M	27.9	29.1
(4) Manner + Place	*1.56M*	**25.4**	**25.3**

Table 5. Word error rates (WER) of broad class-based models, in the Jupiter and Mercury domains. Manner- and place-based models perform least satisfactorily because of the incomplete information they provide. However, by integration of both manner and place information, the best performance is achieved.

We conclude that the linguistically motivated feature sets we used in [11] and also shown here perform better than the data-driven clusters, which supposedly best fit the data, with a significant margin in our experiments. The output of a clustering algorithm is sensitive to the acoustic measurement used, while, with knowledge-driven methods, we can explicitly define multiple feature dimensions (in our case, manner and place dimensions) and reuse the training data along these parallel dimensions. The robustness of the overall system can be improved by maximizing the orthogonality of the different feature dimensions, for various front ends and acoustic measurements [29].. For these reasons, we consider that knowledge-driven feature sets are a reasonable starting point in our research.

4. HANDLING CONTINUOUS SPEECH

In [11] we use the N-best list as a logical choice to represent the cohort. The N-best list representation is very concise and can be very efficiently generated [30, 31]. As the *de facto* standard output of a speech recognizer, they are a convenient representation as the protocol between stages.

The challenge lies in the fact that the cohort space needs to be as constrained as possible so that recognition in later stages is efficient, while at the same time it needs to be general enough so that the correct answer is indeed included. With the isolated word task where the hypothesis space is limited (by the lexicon), the N-best representation is adequate. With the infinite hypothesis space of a continuous speech task, the N-best space is too restricted and the correct answer is often inappropriately pruned. Empirically, if we only re-score the N-best paths from the first stage, we observe a significant performance drop.

To generalize to continuous speech, we decided to consider a "cohort" to be the set of words induced from the N-best output of a feature-based first stage. In effect, we restrict the lexicon for the second stage to be only the words that appeared in the N-best list of the first stage. On the isolated word task, this generalization gives us the same cohort as before. With continuous speech, this generalization allows the second stage to hypothesize novel word sequences unseen in the N-best list, thus the capability to recover from mistakes committed in the first stage. Note that the language model for the first and second stages remains the same.

To gain some insights into the effectiveness of this generalization, we computed the word level cohort coverage rate as a function of N, the depth of the N-best list. A word is considered a "hit" if it appears in the N-best list. A word in the transcription but missing from the N-best list will inevitably lead to an error.

Figure 3 shows the word level cohort coverage rate as the N-best depth increases on the Jupiter domain. The lower curve shows that roughly 92% of the reference words are covered with a 50-best list, which is quite encouraging, especially since the OOV rate of this data set accounts for about 2.3% of the words. We also plot, as the upper curve in Figure 3, the coverage rate of words that can be correctly recognized by a state-of-the-art recognizer. With a 50-best list, about 98% of such words are covered. We also notice that both curves level off at a very modest N-best depth.

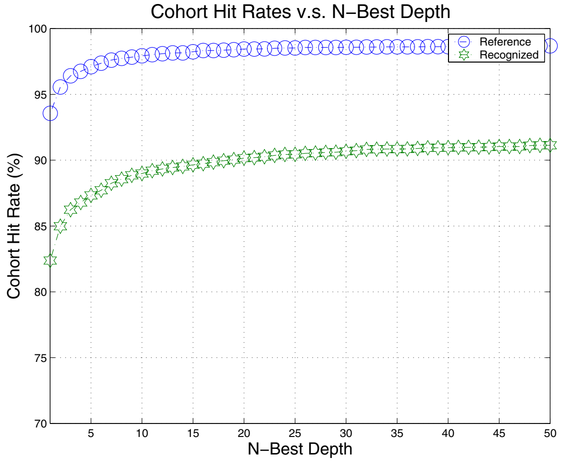

Fig. 3. The cohort hit rates as a function of N-best depth. Above is the hit rate for words that are correctly recognized by a state-of-the-art phone-based recognizer. Below is the hit rate for the correct transcriptions. In both cases, the hit rates level off at a very modest N-best depth. The difference between the two curves is the words which are missing in the cohort but are not correctly recognized by the baseline recognizer either.

Since the cohort space is now the original search space restricted to the N-best lexicon, the reduction in search space can be approximately estimated by the reduction in lexicon size. The upper curve in Figure 4 shows the average N-best lexicon size as N grows. We see that the induced vocabulary is significantly smaller than the original "full" vocabulary. With an N-best depth of 50, the average vocabulary size is only 17.5, less than 1% of the original, 1924-word vocabulary. The lower curve in Figure 4 shows the average number of *correct* words in the induced vocabulary, as the N-best depth grows. With N equal to 50, the N-best vocabulary contains 3.93 correct words on average, which is very close to the average sentence length of 4.28 of this data set, as shown by the solid horizontal line of this figure.

These results suggest that using an N-best list from the feature-based first stage could significantly trim down the search space, as evidenced by the reduction in the size of the N-best lexicon. The amount of "useful" information contained in the N-best list, as shown by Figure 3 and by the lower curve in Figure 4, saturates rapidly as the N-best depth grows, which indicates that only a modest-depth N-best list is necessary for our purposes. Instead of generating deeper N-best lists, some other mechanism is necessary for the system to recover the search space that has been inappropriately pruned.

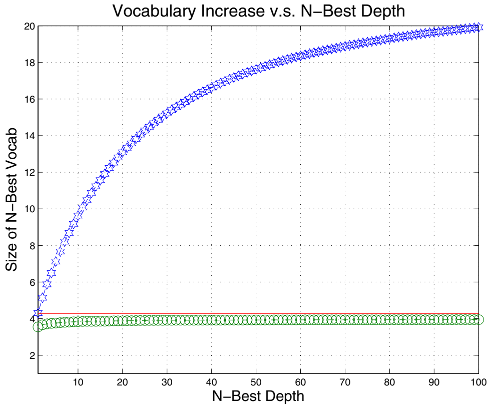

Fig. 4. The size of the vocabulary induced from the first-stage N-best list. With a 50-best list, the average lexicon size is about 17.5, less than 1% of the original, 1924-word vocabulary. The lower curve shows the average number of correct words contained in the N-best lexicon. With N equals 50, the average number of correct words, 3.93, is very close to the average sentence length of 4.28, as shown by the solid horizontal line. Most notably, the lower curve saturates extremely fast, indicating that a shallow N-best list would contain most of the "correct" information.

Analysis of the cohort shows that, in function words such as "I," "you," "yes" and "no," the feature-based models are likely to make errors. Such words are often reduced in their acoustic realization because they contain less information [32]. On the other hand, the feature-based models perform very well on content words, which are information salient and hence acoustically prominent. We considered two ways to help recover the over-pruned search space. One way is to complement the N-best list vocabulary with a small set of most frequently used words to provide the syntactic glue the first-stage recognizer is likely to miss. The other way is to run the first-stage recognizer on a development set to create a complementary vocabulary of words that are missing from the N-best output of the feature-based models. This way we empirically discover from the real data the set of words that the first stage recognizer is likely to miss using the development set.

5. EXPERIMENTAL RESULTS

We report our experiments on the Jupiter weather information domain and the Mercury air travel planning domain. The first-stage recognizer uses the feature-based models as illustrated in Figure 2 on the basis of their compactness and good performance. It uses a "full" FST created with the entire lexicon. A 50-best list is generated for each utterance. Independently, a vocabulary of the 200 most frequent words in each domain, and a vocabulary of the 100 words the first-stage recognizer is most likely to miss[1], are created. The second stage lexicon is the N-best vocabulary augmented with one of these complementary sets. On average this translates to roughly a reduction of 10 in terms of vocabulary size for the second stage recognizer. In the second stage, the "full" FST is pruned to eliminate all arcs outputting words not licensed by the reduced vocabulary. We use as a baseline a state-of-the-art phone-based SUMMIT recognizer in the second stage for both domains. Results are reported on both the clean subset and the full test set.

The speech recognition results are listed in Table 6. When we use the top 200 most frequent words to compensate for the over-pruned search space (System I), the two stage system performs slightly worse than the baseline, for the Jupiter domain. Using the alternative set of words selected from a development set on

[1] determined from an independent development set.

the basis of their absence from the first stage output (System II), the two-stage system outperforms the baseline, for both Jupiter and Mercury. In hindsight, these results are not surprising, since the scheme for System II specifically focuses on words known to present problems in the first-stage recognizer. Figure 5 further illustrates the performance dynamics of System II as the number of compensative words varies. When no compensative words are incorporated, the system performs worse than baseline as the words missing from the first stage can not be recovered. The performance improves as we add more compensative words until it saturates with about 100 words. After this point, the performance slowly decreases and converges to that of the baseline as more compensative words are added.

	Jupiter		Mercury	
	C-Set	F-Set	C-Set	F-Set
Baseline	11.6	18.4	12.7	22.1
Two-stage System I	11.9	18.6	N/A	N/A
Two-stage System II	**11.0**	**17.9**	**12.4**	**21.7**

Table 6. WER's on Jupiter and Mercury. Two-stage System I uses the 200 most frequent words to enhance the second stage vocabulary. Two-stage system II uses a complementary set of words that the first stage tends to make mistakes in. The second two-stage system improves the final performance on both the clean data (C-Set) and the full test data (F-Set), for both domains.

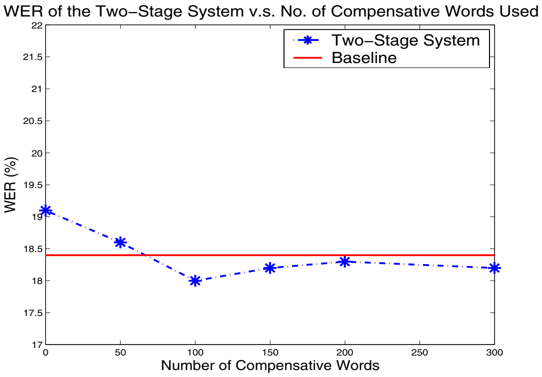

Fig. 5. Performance dynamics of System II as we increase the number of compensative words.

Figure 6 gives an example where the two-stage approach performs better. The utterance is, "update on tornado warnings". The phone-based models recognized it as "date on tornado warnings", as shown in the lower panel. This is a common type of mistake where confusion between utterance-onset noise and the first word arises at the beginning of an utterance, although the first word is prominently articulated. In this particular example, the fact that the /p̚/ (/pcl/ in the figure) in "update" is noisy might also contribute to the error. In the two-stage framework, the first stage rules out the erroneous candidate "date", and enables the second stage to produce the correct result, as show in the upper panel. The feature-based models, probably because they use broad classes and are more robust, are less sensitive to noises and perform well in the presence of reliable acoustic evidence. For this reason, we hypothesize that the feature-based models are able to provide complementary information to a phone-based system. A McNemar test on the full test set of Jupiter shows that the reduction in error is significant at the $p = 0.05$ level.

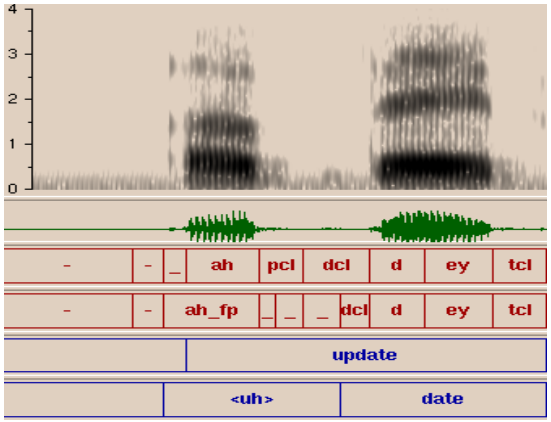

Fig. 6. The output of the two-stage recognizer (upper panel) and that of the phone-based baseline (lower panel) for the utterance ``update on tornado warnings''. The first stage rules out ``date'' and hence enables the second stage to choose the correct answer.

6. SUMMARY AND FUTURE WORK

This paper addresses two important issues in our research to build multi-stage recognizers using linguistic features. Through a controlled study, we show that our knowledge-driven features based on manner and place of articulation perform significantly better than a set of broad classes discovered through data-driven approaches. However, it is important that we integrate information from the parallel manner and place dimensions through proper information fusion techniques.

The N-best list, at the utterance level, is no longer an appropriate representation as a cohort. At the word level, if we consider the vocabulary induced from an N-best list, it is concise and contains most of the useful information. The notion of cohort is generalized to continuous speech as the search space restricted to the N-best lexicon. This cohort is effective when augmented with a small set of complementary words: a second stage recognizer achieves improved recognition accuracy through searching the reduced space.

We have shown a promising approach to improving speech recognition accuracy by using acoustic models based on linguistic units and by taking a two-stage approach. In this preliminary study, experiments are performed off-line. In the near future, we will explore ways to incorporate an efficient implementation of this approach into our live conversational systems.

The performance of the first stage recognizer can probably be further improved by further study of phonological rules based on the linguistic units we used in this research, and by introducing constraints from higher linguistic hierarchies [33]. In its present form, our system uses a reduced vocabulary while maintaining the original language model. It would be interesting to consider modifications to the language model of the second stage on the basis of the first stage cohort, for example, through the simple technique of re-normalizing the probability model once rejected words have been pruned. Alternative techniques similar to boosting may also be effective.

This two-stage framework can potentially be applied to multi-domain speech recognition tasks. In [34] the authors described a mechanism for multi-domain speech recognition by binding together FST's from multiple domains and thus creating a unified search space. The final performance decreases slightly due to the increased complexity during search. Conceivably we can make a domain classification based on the result of the first stage, and perform domain-dependent recognition on the second stage using the two-stage framework we propose in this paper. Whereas in [34] the domains are task-driven, such sub-domains can also be automatically discovered by analyzing the semantic similarity among utterances, and can be coordinated with predictions from the dialogue state. Thus, for example, specific sub-domains can be derived for confirmation/refusal, or for inquiries of dates or cities. When these sub-domains are used in concert with the two-stage framework, we can provide a richer dialog interface to users by supporting seamless domain switching/selection at the second stage.

7. ACKNOWLEDGMENT

Jim Glass gave valuable comments to a draft of this paper. The authors wish to thank Chao Wang, T.J. Hazen and Lee Hetherington for many valuable discussions and for help in setting up some of the experiments.

8. REFERENCES

[1] C. J. Weinstein, S. S. McCandless, L. F. Mondshein, and V. W. Zue, “A system for acoustic-phonetic analysis of continuous speech,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 54–67, Feburary 1975.

[2] J.A. Barnett, M.I. Bernstein, R.A. Gillmann, and I.M. Kameny, “The SDC speech understanding system,” in *Trends in Speech Recognition*, W.A. Lea, Ed., pp. 272–293. Prentice-Hall, Englewood Cliffs, N.J., 1980.

[3] J.J. Wolf and W.A. Woods, “The HWIM speech understanding system,” in *Trends in Speech Recognition*, W.A. Lea, Ed., pp. 316–339. Prentice-Hall, Englewood Cliffs, N.J., 1980.

[4] L.D. Erman and V.R. Lesser, “The Hearsay-II speech understanding system: A tutorial,” in *Trends in Speech Recognition*, W.A. Lea, Ed., pp. 361–381. Prentice-Hall, Englewood Cliffs, N.J., 1980.

[5] B.T. Lowerre and R. Reddy, “The HARPY speech understanding system,” in *Trends in Speech Recognition*, W.A. Lea, Ed., chapter 15. Prentice-Hall, Englewood Cliffs, N.J., 1980.

[6] V. W. Zue and R. A. Cole, “Experiments on spectrogram reading,” in *Proc. IEEE Int. Conf. ASSP*, 1979, pp. 116–119.

[7] V. W. Zue and D. W. Shipman, “Properties of large lexicons: Implications for advanced isolated word recognition systems,” in *Proc. 103rd Meetings of the Acoustic Society of Amrica*, Chicago, IL, April 1982.

[8] D. P. Huttenlocher and V. W. Zue, “A model of lexical access from partial phonetic information,” in *Proc. IEEE Int. Conf. ASSP*, San Diego, CA, March 1984.

[9] V. W. Zue, “The use of speech knowledge in automatic speech recognition,” *Proc. of the IEEE*, vol. 73, no. 11, pp. 1602–1615, November 1985.

[10] K. N. Stevens, “Toward a model for lexical access based on acoustic landmarks and distinctive features,” *Journal of Acoustical Society of America*, vol. 111, no. 4, pp. 1872–1891, April 2002.

[11] M. Tang, S. Seneff, and V. W. Zue, “Modeling linguistic features in speech recognition,” in *Proc. Eurospeech*, 2003.

[12] J. Pitrelli, C. Fong, S. Wong, J. Splitz, and H. Leung, “Phonebook: A phonetically-rich isolated-word telephone-speech database,” in *Proc. IEEE Int. Conf. ASSP*, 1995, pp. 101–104.

[13] K. Kirchhoff, *Robust Speech Recognition Using Articulatory Information*, Ph.D. thesis, Der Technischen Fakultắt der Universitắt Bielefeld, 1999.

[14] M. Richardson, J. Bilmes, and C. Diorio, “Hidden-articulator Markov models: Performance improvements and robustness to noise,” in *Proc. ICSLP*, 2000.

[15] E. Eide, “Distinctive features for use in an automatic speech recognition system,” in *Proc. Eurospeech*, 2001.

[16] J. Sun and L. Deng, “An overlapping-feature-based phonological model incorporating linguistic constraints: Application to speech recognition,” *Journal of Acoustical Society of America*, vol. 111, no. 2, Feburary 2002.

[17] K. Livescue, J. Glass, and J. Bilmes, “Hidden feature models for speech recognition using dynamic bayesian networks,” in *Proc. Eurospeech*, 2003.

[18] J. Glass, “A probabilistic framework for segment-based speech recognition,” *Computer Speech and Language*, pp. 137–152, 2003.

[19] F. Pereira and M. Riley, “Speech recognition by composition of weighted finite automata,” in *Finite-State Language Processing*, Emmanueal Roche and Yves Schabes, Eds., pp. 431–453. MIT Press, Cambridge, MA, 1997.

[20] V. Zue, S. Seneff, J. Glass, J. Polifroni, C. Pao, T. Hazen, and I. L. Hetherington, “JUPITER: A telephone-based conversational interface for weather information,” *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 1, pp. 100–112, 2000.

[21] S. Seneff, R. Lau, J. Glass, and J. Polifroni, “The MERCURY system for flight browsing and pricing,” *MIT Spoken Language System Group Annual Progress Report*, pp. 23–28, 1999.

[22] E. C. Sagey, *The Representation of Features and Relations in Non-Linear Phonolgy*, Ph.D. thesis, Massachusetts Institute of Technology, 1982.

[23] M. A. Randolph, *Syllable-based Constraints on Properites of English Sounds*, Ph.D. thesis, Massachusetts Institute of Technology, 1989.

[24] J. R. Glass, *Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition*, Ph.D. thesis, Massachusetts Institute of Technology, 1988.

[25] C. Wang, S. Cyphers, X. Mou, J. Polifroni, S. Seneff, J. Yi, and V. W. Zue, “Muxing: A telephone-access Mandarin conversational system,” in *Proc. ICSLP*, Beijing, P.R.China, October 2000.

[26] P. Ladefoged, *A Course in Phonetics*, Harcourt Brace Jovanovich, Inc, second edition, 1982.

[27] “Lecture Notes for 6.345: Automatic Speech Recognition, Spoken Language Systems Group, MIT,” .

[28] K. N. Stevens, *Acoustic Phonetics*, MIT Press, 1998.

[29] A. Halberstadt, *Heterogeneous Measurements and Multiple Classifiers for Speech Recognition*, Ph.D. thesis, MIT, Nov. 1998.

[30] F. K. Soong and E.-F. Huang, “A tree-trellis based fast search for finding the n best sentence hypotheses in continuous speech recognition,” in *Proc. ICASSP ’91*, Toronto, Canada, May 1991, pp. 705–708.

[31] I. L. Hetherington, *The Problem of New, Out-of-Vocabulary Words in Spoken Language Systems*, Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, Oct. 1994.

[32] F. Grosjean and J. P. Gee, “Prosodic structure and spoken word recognition,” in *Spoken Word Recognition*, U. H. Frauenfelder and L. K. Tyler, Eds. MIT Press, 1986.

[33] S. Seneff and C. Wang, “Modeling phonological rules through linguistic hierarchies,” in *Proceedings of the ISCA Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language*, Sept. 2002, pp. 71–76.

[34] T.J. Hazen, I. L. Hetherington, and A. Park, “FST-based recognition techniques for multi-lingual and multi-domain spontaneous speech,” in *Proc. Eurospeech*, 2001.