SRI's DECIPHER System

Hy Murveit, Michael Cohen, Patti Price, Gay Baldwin,
Mitch Weintraub, and Jared Bernstein
Speech Research Program, SRI International
Menlo Park, CA 94025

Abstract

SRI has developed a speaker-independent continuous speech, large vocabulary speech recognition system, DECIPHER, that provides state-of-the-art performance on the DARPA standard speaker-independent resource management training and testing materials. SRI's approach is to integrate speech and linguistic knowledge into the HMM framework. This paper describes performance improvements arising from detailed phonological modeling and from the incorporation of cross-word coarticulatory constraints.

1 Introduction

The hidden Markov model (HMM) formulation is a powerful statistical framework that is well-suited to the speech recognition problem. Systems based on this formulation have improved dramatically, however, as developers have learned how to modify them appropriately to take into account principles from speech research and from linguistics. Concepts arising from the study of the sound system of a language, i.e., phonology, are language-specific; they are done once for English and little additional labor is required when, for example, the vocabulary or the domain changes. Care should be taken, however, in modeling detailed linguistic structure since the practice can lead to models with many additional parameters to be estimated; unless this problem is addressed directly, performance gains will be compromised.

SRI is not the first group to incorporate speech knowledge and concepts from linguistics in an HMM formulation for speech recognition. In fact, many improvements in HMM-based systems are implicitly related to principles from speech and linguistics, even though this was not their original motivation. We survey some of these modifications below.

Phonetic Units. Not all recognizers are based on phonetic units. A number of HMM-based speech recognition systems have been based on word-level models [1, 12, 9]. The use of phonetic units allows for larger vocabularies by sharing training across subword units that repeat more frequently than words do. Phonetic-based units are now common to many HMM-based recognizers (e.g., [7, 8, 10]).

Triphones. Triphones are phonetic models conditioned on the immediately surrounding phonetic units. BBN [3] was able to show significant performance gain (roughly halving the error rate compared to a similar system without the triphone models), provided the context-dependent models were averaged ("smoothed") with the context-independent ones in order to deal with the large number of poorly trained triphone models. Triphones are used extensively now (e.g., at BBN, CMU, Lincoln Laboratories, SRI). Triphones can model major coarticulatory effects described in the speech research literature, as well as phonological variation conditioned on the immediately surrounding phones. In general, more detailed models will perform better than less detailed models, provided there is sufficient data to estimate the parameters. For 60 phones, there are 60 cubed triphones, which represents a significant increase in parameters to estimate. Triphones would not have shown a performance gain had they been introduced without a mechanism to take this into account, i.e., in this case, smoothing with more general models.

Difference Parameters. The use of additional, independently trainable parameters means that more details can be included in the model without a dramatic increase in the amount of training material. In particular, recognition performance has been significantly increased [8, 10] through the use of codebooks that represent the difference between the current value of a parameter and its value several frames previously. Spectral and energy difference parameters are used in addition to their current values. This captures important dynamic patterns exhibited in speech as well as the standard static information.

In this paper we describe SRI's recent work in incorporating linguistic concepts in the DECIPHER system: improved phonological modeling and modeling cross-word coarticulation.

2 DECIPHER's Basic Design

SRI's DECIPHER speech recognition system uses discrete density 3-state hidden-Markov models to represent phones.. Four discrete densities per state model the variation of vector quantized Mel-cepstra, vector quantized Mel-cepstral time-derivatives, and quantized energy and energy time-derivatives. Word models are constructed from network representations of word pronunciations and from a set of phone models (context-independent, left biphones, right biphones, triphones, and phone-in-word models). The more samples of a word available in the system's training set, the more specific the contexts used for the phone models in the word. The most detailed, primary, models are smoothed by averaging in other models of less specific context with weights estimated automatically using an SRI version of IBM's deleted-interpolation algorithm [6].

3 Database

The speech database used for training and testing SRI's DECIPHER system is described in [11]. This database, intended for the design and evaluation of algorithms for continuous speech recognition, consists of sentences read in a sound-isolated room. The sentences are appropriate to a naval resource management task based on existing interactive database and graphics programs. The database includes 160 male and female talkers with a variety of dialects. The design includes a partition of the database into independent training and testing portions.

The training materials used for the results reported here are the 3950 sentences from 97 training and development talkers that do not overlap the test set reported on. The testing materials used for most of the results reported here are the 150 sentences (1287 words) from the 1987 designated test sets designated by the National Institute of Standards and Technology (NIST, formerly NBS).

The results reported here were obtained with and without the use of a grammar to constrain the recognition search. These conditions are not those that would be used in a real application, but they are simply defined, allow recognition systems to be evaluated over more than one condition of grammatical constraint, and they have been accepted as standards of comparisons. The degree of constraint provided by the grammar is measured by *test set perplexity* [7], or, the geometric mean of the number of words allowed by the grammar at each point in the test set, given the previous words. In the case of no grammatical constraint, any word can follow any other word and the perplexity is equal to the vocabulary size, in this case 1000. The DARPA standard word-pair grammar was created by collecting all two-word sequences allowed in the sentence-patterns used to generate the task sentences (as described in [11]). The perplexity of this grammar as measured on several different 25-sentence test sets from the database is about 60.

4 Phonological Modeling

Pronunciation varies significantly across speakers, as well as in the speech of individuals [5]. However, most current speech recognition systems model words with a single pronunciation or a small number of alternate pronunciations. For systems which use statistical training of models of speech segments, this lack of explicit representation of the range of variation of pronunciation causes different phenomena to be averaged together into the same model, resulting in a less precise model. These less precise models are likely to become more problematic as speech recognition systems move from corpera of read speech to the spontaneous speech that can be expected in real applications, since significantly more pronunciation variability occurs in spontaneous than in read speech (c.f. [2]).

Some previous attempts to explicitly model many pronunciations for each word have led to performance degradation possibly resulting from (1) many additional parameters to be estimated with the same amount of training data, and (2) unlikely pronunciations not previously modeled causing new false alarms.

To deal with the first condition, we have designed a method for developing phonological rule sets based on measures of coverage and overcoverage of a database of pronunciations [4] in order to maximize the coverage of pronunciations observed in a corpus, while minimizing the size of the pronunciation networks.

To address the problem of hypothesizing unlikely pronunciations in inappropriate places, the DECIPHER system incorporates probabilities into our network representation of word pronunciations. The incorporation of pronunciation probabilities has been shown to significantly increase the predictive power of our representation [4].

Current databases for training speech recognition systems have too few occurrences of all but the most frequent words to make accurate estimates of pronunciation probabilities. Therefore, we have developed and implemented an automatic method for tying together frequently occurring sub-word units for training. Knowledge embedded in the rule set can be used to determine equivalence classes of nodes that share similar contextual constraints [4]. Nodes in the same equivalence class share training samples. The probabilities in the pronunciation networks combine word-trained probabilities for frequently occurring words with these node equivalence class trained probabilities.

5 Lexicon Performance

We compared performance of the DECIPHER system for a number of different lexicons, based on the rule set development method described above. A rule set with high coverage of a corpus of pronunciations was developed, and pronunciation probabilities were computed for the resulting pronunciation networks using the node equivalence classes described above. The data used to estimate the pronunciation probabilities was the same data used to train the phonetic models. A series of less bushy networks was derived by eliminating the least probable pronunciations from the networks. We refer to this series as *Rule-Single* (each word has only one pronunciation), *Rule-Sparse* (the mean number of pronunciations per word is 1.3), and *Rule-Full* (the mean number of pronunciations per word is 4.2). Performance was also measured using the lexicon from the BBN BYBLOS system (*BBN*), from the CMU SPHINX system (*CMU*), and the lexicon developed for an early version of the DECIPHER system, prior to the incorporation of multiple pronunciations. This latter lexicon is referred to as *Hand-Single* since it consists of a single pronunciation per word and was specified by hand by an expert linguist.

Table 1 shows the results we have obtained with the SRI DECIPHER system using the various lexicons described above. The recognized word strings were aligned against the correct reference word string and differences tallied using the DARPA-NIST software package. The word correct is $1 - 100\frac{substitutions+deletions+insertions}{ref}$, where *ref* is the number of words in the set of reference words. The DARPA-NIST homophone-equivalency table is used for the no-grammar condition (Perplexity P=1000) and not for the grammar condition (Perplexity P=60). The lexicons labeled *BBN* and *CMU* do not compare the DECIPHER system to the BYBLOS system or to the SPHINX system, rather it compares the CMU, BBN and SRI lexicons all as used in the DECIPHER system without cross-word coarticulatory modeling.

LEXICON	Mean Number of Pronunciations per Word	Px:1000 word correct	Px:60 word correct
BBN-lexicon	1.1	67.0	91.6
CMU-lexicon	1.0	67.5	92.4
Hand-Single	1.0	69.3	90.6
Rule-Single	1.0	72.8	92.6
Rule-Sparse	1.3	74.1	93.7
Rule-Full	4.2	72.9	92.6

Table 1: Phonological Effects in DECIPHER: Word accuracy on the 1987 test set with various lexicons and expansions.

The results of Table 1 show that the lexicon has a significant effect on performance: for perplexity 1000, percent word correct ranges from 67.0% (for the BBN lexicon) to 74.1% (for SRI's Rule-Sparse lexicon); for perplexity 60, the range is from 90.6% (Hand-Single) to 93.7% (Rule-Sparse). Within the set of single (or near single) pronunciation lexicons, the range is nearly as large. Thus, a system that explicitly models only a single pronunciation per word can be improved with careful design of the dictionary of pronunciations. Automatically deriving a dictionary of most common pronunciations (as in the Rule-Single lexicon) was shown to improve performance over a dictionary of pronunciations carefully designed by hand by an expert linguist (Hand-Single).

The improvement from the rule-single to the rule-sparse lexicon suggests that modeling multiple probabilistic pronunciations can improve recognition performance. The degradation in performance from Rule-Sparse to Rule-Full illustrates the importance of keeping pronunciation networks from getting too bushy, while maintaining coverage of likely pronunciations.

6 Cross-word Modeling

The use of triphone modeling and models of whole words has been used extensively (e.g., [3]) to take into account coarticulatory effects. However, extending this notion to operate across word boundaries had not been been done before 1989. Word-boundary contexts have not typically been used because the sizes of the resulting word networks can get very large, and because it requires keeping track of which ending arcs can map to which starting arcs. Since we have already dealt with these issues in our large pronunciation net-

SYSTEM	Context Model?	Px:1000 word correct	Px:60 word correct
DECIPHER	No	74.1	93.7
	Yes	78.3	95.0

Table 2: Word Accuracy Results With and Without Cross-Word Coarticulatory Modeling (for the 1987 test set).

Speakers in Training	Px	%sub	%del	%ins	%error
109	60	5.9	2.5	0.4	**8.8**
72	60	6.4	2.5	0.3	**9.2**
109	1000	21.0	5.2	1.5	**27.7**
72	1000	23.4	6.1	1.8	**31.3**

Table 3: DECIPHER'S Performance using DARPA's 1989 Speaker-Independent Test Set

works, cross-word boundary contexts were a natural extension to the SRI DECIPHER system.

Modeling acoustic variations across words was limited to initial and final phones in words with sufficient training data. To illustrate how the algorithm works, let us consider the initial phone "dh" in the word *the*. In the training database, there are many instances of words ending in "n" before *the*. An additional "dh" arc is added to the pronunciation graph of *the*, though this arc is only allowed to connect to arcs with the "n" phonetic label. The original "dh" arc is prevented from connecting with arcs with the "n" phonetic label. The above algorithm is applied to all words in the vocabulary, provided that 15 occurrences of a (previous/next) phone occurred in the training database.

Table 2 shows that the addition of the cross-word context models improves performance of the DECIPHER system in both the perplexity 60 condition and the perplexity 1000 condition. Also shown in the table, labeled SPHINX, are the best previous results reported on for this database (actually using a little more training data) [8].

7 1989 Test Results

Table 3 shows SRI's official results reported at the 1989 DARPA speech and natural language workshop. These results use the Rule-Sparse pronunciation networks and the across-word-boundary pronunciation constraints. Px stands for perplexity.

If the tradeoff for insertions and deletions is appropriately changed, which was not done for the results in Table 3, performance can be improved slightly: 5.7% substitutions, 1.3% deletions, and 0.8% insertions, for an overall error rate of 7.9%.

Speaker-by-speaker performance varies greatly in the DECIPHER system, and in other systems that were reported in this workshop. For the official DECIPHER performance results for the 1989 workshop for the 109 speaker training set, speaker performance ranged from 17.8% to 37.1% (perplexity=1000), and 3.7% error to 14.3% error (perplexity=60). This variability causes difficulties when trying to compare systems from different sites. For instance, when comparing DECIPHER's results to Carnegie-Mellon's Sphinx system, we find that with perplexity=1000, Sphinx outperformed DECIPHER on six speakers, and DECIPHER outperformed Sphinx on four speakers. With perplexity=1000, Sphinx had better performance on 7 speakers. When comparing DECIPHER to the Lincoln Laboratories system, DECIPHER had fewer errors on 6 of 10 with perplexity=1000, and 7 of 10 with perplexity=60. This analysis shows that the variation across speakers swamps the variation among these three systems, and that the apparent system differences may be due to sampling error. To properly differentiate these systems at a high confidence level would require a test with many more speakers.

8 Discussion

In this section we discuss first the results relating to choice of lexicon and then the results relating to cross-word models.

Though we have shown an important performance gain through improved phonological modeling, we believe that more substantial gains will be shown in the future for the following reasons:

1. The system tests were based on read speech rather than spontaneous speech. The significant increase in phonological reduction and deletion in spontaneous compared to read speech [2] should result in a bigger difference between systems that include techniques for modeling multiple probabilistic pronunciations and those that do not.

2. The rule sets used in the studies described here were developed using a corpus of hand phonetic transcriptions, rather than some form of system output. Different types of variation may be more important to model explicitly in different systems,

and are likely to be different from those captured by hand transcriptions.

3. Larger amounts of training data will allow the design of more detailed models of phonological variation.

Lee has suggested [8] that modeling multiple pronunciations is not worth-while because (1) it makes systems run too slowly, (2) it is impossible to estimate pronunciation probabilities, and (3) it unfairly penalizes words with too many pronunciations. Although, we believe that improvements can certainly still be made in the way we estimate and use our pronunciation probabilities, it is clear from our studies that modeling pronunciation has significant positive impact on recognition performance without an excessive cost in speed. We suggest that the reason for our opposite conclusions lies in the difference between the multiple-pronunciation word-networks that CMU and SRI have tried. As shown in Table 1, SRI's best network models on average about 1.3 pronunciations per word. The network shown as an example in [8] allows thousands of pronunciations, partly because of the excessive detail in the example and partly because of the lack of constraint to correlate the many possibilities.

As for the cross-word coarticulatory modeling we report on here, we believe that the performance improvement can be attributed to the following: (1) for short words (and the most frequent words are short, i.e., one to three phones long), the word boundaries form a significant portion of the context that should not be ignored, and (2) many triphones that otherwise would not be observed can be found across word boundaries, and the additional triphone training can help model the less frequent words.

In sum, the use of speech and linguistic knowledge sources can be used to improve the performance of HMM-based speech recognition systems, provided that care is taken to incorporate these knowledge sources appropriately.

Acknowledgement. This work was principally supported by SRI IR&D and investment funding. We also gratefully acknowledge the National Science Foundation and the Defense Advanced Research Projects Agency for partial support.

References

[1] Bakis, R. (1976) "Continuous Speech Recognition via Centisecond Acoustic States," *J. Acoust. Soc. Am.*, Suppl. 1, Vol. 59, S97.

[2] Bernstein, J., G. Baldwin, M. Cohen, H. Murveit, and M. Weintraub (1986) "Phonological Studies for Speech Recognition", *Proc. DARPA Speech Recognition Workshop*, pp. 41-48.

[3] Chow, Y., R. Schwartz, S. Roucos, O. Kimball, P. Price, F. Kubala, M. Dunham, M. Krasner, and J. Makhoul (1986) "The Role of Word-Dependent Coarticulatory Effects in a Phoneme-Based Speech Recognition System," *Proc. ICASSP*, pp. 1593-1596.

[4] Cohen, M. (1989) *Phonological Structures for Speech Recognition*, U. C. Berkeley Ph.D. thesis.

[5] Cohen, M., J. Bernstein, and H. Murveit (1987) "Pronunciation Variation Within and Across Speakers," 113th Meeting Acoust. Soc. Am., P9.

[6] Jelinek, F. and R. Mercer (1980) "Interpolated Estimation of Markov Source Parameters from Sparse Data," pp. 381-397 in E. S. Gelsema and L. N. Kanal (editors), *Pattern Recognition in Practice*, North-Holland Publishing Company, Amsterdam, the Netherlands.

[7] Kubala, F., Y. Chow, A. Derr, M. Feng, O. Kimball, J. Makhoul, P. Price, J. Rohlicek, S. Roucos, R. Schwartz, and J. Vandegrift (1988) "Continuous Speech Recognition Results of the BYBLOS System on the DARPA 1000-Word Resource Management Database," *Proc. ICASSP*, pp. 291-294.

[8] K. F. Lee (1988) *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: the SPHINX System*, CMU Ph.D. Thesis.

[9] Lippmann, R., E. Martin, and D. Paul (1987) "Multi-Style Training for Robust Isolated-Word Speech Recognition," *Proc. ICASSP*, pp. 705-708.

[10] Murveit, H. and M. Weintraub (1988) "1000-Word Speaker-Independent Continuous-Speech Recognition Using Hidden Markov Models," *Proc. ICASSP*, pp. 115-118.

[11] Price, P., W. Fisher, J. Bernstein, and D. Pallett (1988) "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," *Proc. ICASSP*, pp. 651-654.

[12] Rabiner, L. R., B. H. Juang, S. E. Levinson, and M. M. Sondhi (1985) "Recognition of Isolated Digits using Hidden Markov with Continuous Mixture Densities," *ATT Technical Journal*, 64(6):1211-1233.