S2.10

THE DECIPHER SPEECH RECOGNITION SYSTEM

Michael Cohen, Hy Murveit, Jared Bernstein,
Patti Price and Mitch Weintraub

SRI International
Menlo Park, CA 94025

Abstract

DECIPHER is SRI's HMM-based speaker-independent continuous speech recognition system. DECIPHER performs well on the speaker-independent DARPA resource management task, as described in last year's ICASSP Proceedings [10]. To determine whether speaker-specific acoustic and phonological adaptation can further improve performance, the current paper describes DECIPHER's performance on a speaker-dependent task.

1. Introduction

The Speech Research Program at SRI International has designed and implemented several speech recognition systems in the last six years. SRI's current large-vocabulary, continuous-speech system, DECIPHER, is based on a hidden Markov model (HMM) approach and was designed to achieve high word accuracy in a speaker-independent mode. It has been trained and tested on DARPA's Resource Management database [9]. The DECIPHER system was described at last year's ICASSP meeting [10]. That paper presented results showing that speaker-independent recognition performance could be improved by incorporating certain kinds of linguistic knowledge into the Markov model framework, including cross-word coarticulatory modeling and detailed modeling of phonological variation.

This paper presents the results of a series of experiments that tested acoustic and phonological adaptation of the DECIPHER system to the pronunciations of a single speaker in a speaker-*dependent* task.

2. The DECIPHER System

The DECIPHER system uses an HMM framework similar to that used in a number of other systems [2, 7, 8]. The overall structure of such a system is well described in [7]. The overall structure of SRI's DECIPHER system is shown in Figure 1.

DECIPHER's front end samples an analog acoustic signal 16,000 times per second after passing the signal through a 6.4 Khz low-pass, anti-aliasing filter with 0.95 pre-emphasis. Signal analysis starts with a 512-point discrete Fourier transform (DFT) calculated every 10 msec on a 25.6 msec Hamming window. Four discrete acoustic features are calculated every 10 msec. The features are based on a 13-dimensional cepstral transform of the logarithms of the energies in 25 overlapping filters (approximately equally spaced on the mel scale) in the range from 100 Hz to 6400 Hz. An optional noise-robust spectral estimation process is described in [6] in this volume.

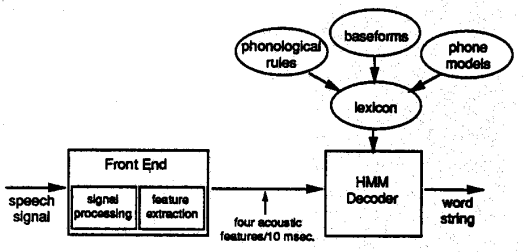

FIGURE 1: The DECIPHER System

The phonetic models in the DECIPHER system are discrete density 3-state hidden-Markov models. There are four discrete densities per state, one for each of the four acoustic features produced by the front end. Word models are directed graphs of phonetic models (combining context-independent and context-dependent phonetic models). The lexical graph for a vocabulary item is generated by the application of a set of phonological rules to a baseform pronunciation (similar to previous efforts at modeling multiple pronunciations [4]). The modeling of multiple pronunciations in the DECIPHER system differs from previous efforts in two important respects:

[1] A new technique for developing phonological rule sets was used, with the goal of maximizing the coverage of the pronunciations found in a corpus of speech while minimizing the size of networks.

CH2847-2/90/0000-0077 $1.00 © 1990 IEEE

[2] A new algorithm was used to estimate the probabilities of alternate pronunciations. The new algorithm defines sub-word units which can share training data based on equivalence classes of nodes.

These two techniques are described in the following two sections.

3. Developing Phonological Rule Sets

Previous efforts to model multiple pronunciations of words have suffered because many new parameters were introduced which had to be estimated with a fixed amount of training data. The approach to rule set development SRI uses has the goals of maximizing the coverage of observed forms in a corpus of speech while minimizing the size of the networks, and therefore minimizing the number of parameters which need to be estimated.

A number of software tools were developed which allow the measurement of the coverage of pronunciations in a corpus as well as overgeneration (generation of pronunciations not used), both for a full rule set and for the individual rules in a rule set. These tools can be used to optimize the definition of the contextual constraints of individual rules, as well as the choice of rules to include in a rule set.

The development of phonological rule sets proceeds as follows:

[0] Start with a lexicon of base forms, a corpus of pronunciations, and (optionally) a phonological rule set (i.e., we can start with an existing rule set and refine it, or start with just baseforms).

[1] Measure coverage of output forms (resulting from the application of current rules, if any, to baseforms) on observed pronunciations. Get diagnostic information on uncovered pronunciations.

[2] Write rules to cover pronunciations.

[3] Measure coverage and overgeneration of individual rules. Analyze and refine contextual specifications of rules based on individual rule diagnostics.

[4] Repeat from step 1 to achieve high coverage rule set.

Using the method outlined above, we have been able to develop a phonological rule set with significantly higher coverage and significantly lower overgeneration than rule sets developed by more traditional methods both at SRI and elsewhere [3].

4. Estimating the Probabilities of Alternative Pronunciations

Previous efforts to model multiple pronunciations of words have suffered because the unlikely pronunciations (not previously modeled) caused false alarms. This was a problem because the systems lacked accurate estimates of the probabilities of the many pronunciations modeled. Achieving accurate estimates is difficult because current databases for training recognition systems have too few occurrences of all but the most frequent words to make accurate estimates.

In order to reliably estimate pronunciation probabilities for words which don't happen frequently enough to provide adequate training data, it is necessary to tie together sub-word units which do happen frequently. Thus, reliable probabilities can be estimated for these sub-word units, which can then be concatenated to form estimates for word pronunciations. Because extended context can play an important role in determining the allophonic form of a segment in a word, we want to tie together the largest units possible that have adequate training data, in order to capture the greatest amount of contextual information. We have developed an approach which attempts to automatically determine the best grouping of sub-word units into node-equivalence classes for common training.

In the DECIPHER system, the training of pronunciation probabilities is incorporated into the training of the HMM models using the forward-backward algorithm. The forward-backward algorithm provides estimates of the number of transitions for each arc at the end of each iteration through the training data. The estimated transitions for arcs which correspond to arcs in pronunciation networks are used to reestimate pronunciation probabilities allowing arcs to share training samples when they occur in the same node-equivalence class, as defined above.

We have shown improvements in speaker-independent performance using the rule set development and node-equivalence class training techniques outlined above [10]. The next section reports the evaluation of these techniques on a speaker-dependent database.

5. Speaker-Dependent Phonology

A set of experiments were performed in which pronunciation models were adapted to individual speakers. Initially, each speaker started with a set of pronunciation networks which resulted from the application of a phonological rule set, developed using the method described above, to a set of baseforms. The mean number of pronunciations represented per word with these networks was approximately 35. These networks were then trained separately for each speaker in the speaker-dependent test set. The training set for each speaker included 600 read sentences (the DARPA speaker-dependent resource management training set). Two iterations of the forward-backward algorithm were run, and the node-equivalence class algorithm referred to above was used to estimate speaker specific pronunciation probabilities for these networks. The networks were then pruned by removing low probability arcs, using an algorithm that includes constraints to prevent the creation of disconnected components of word networks and to avoid the creation of word models which can't connect to other words due to cross-word phonological constraints. In addition, node types with less than a specified minimum

number of training instances were constrained so that only the most likely arc was left after pruning.

These pruned speaker-dependent word networks had an average of approximately four pronunciations per word. An additional two iterations of the forward-backward algorithm were then run in order to train the acoustic HMM models with the pruned speaker-dependent word networks.

Tests were run with the DARPA 1000-word resource-management database using both the DARPA February 89 speaker-dependent test set and the 100-speaker development set. The DARPA perplexity-60 word pair grammar was used. Results are shown in Table I. The single networks were derived by pruning out all but the single, most likely, path in all of the word networks after training pronunciation probabilities using the node-equivalence class training algorithm. The multiple pronunciation networks were pruned, as described above, until there were an average of approximately four pronunciations per word. Table I compares performance using networks with pronunciation probabilities based on a speaker-independent training set and a speaker-dependent training set. (Only the training and pruning of pronunciation networks was varied for these runs - in all cases the acoustic HMM models were trained speaker-dependently.) Percent word correct was measured as

$$1 - \frac{insertions + deletions + substitutions}{total}$$

where total = number of words in the correct sentences

phonological training	networks	dev set	Feb 89 set
SI	single	97.5	97.0
SI	multiple	97.6	97.4
SD	single	97.6	97.4
SD	multiple	97.8	97.7

Table I: Speaker-dependent phonology.

As can be seen in Table I, a reduction in error rate of 12% was achieved for one test set and 23% for another test set going from speaker-independently determined single most-likely pronunciations to speaker-dependently determined multiple pronunciations. It can be seen that part of that gain can be achieved with speaker-specific adaptation of pronunciation networks, and part with the representation of multiple pronunciations. In all four cases shown, going from single pronunciations to multiple pronunciations improved performance, and going from speaker-independent to speaker-dependent phonological training improved performance.

6. Discussion

The results shown here suggest that:

[1] Speaker specific phonological training can improve recognition performance, both for single and multiple pronunciation systems.

[2] Multiple pronunciation models can improve the performance of a speaker-dependent system.

In both cases, the improvements observed were small, but consistent. A larger speaker-specific training set would be likely to improve the results reported here. With a larger training set, bushier word networks could be used while maintaining the accuracy of the estimates of pronunciation probabilities, as well as the estimates of the acoustic parameters of the HMM models.

All the results presented in this paper are based on experiments that both trained and tested the DECIPHER system on carefully collected, read speech. In the future, we intend to evaluate these techniques on goal-directed, spontaneous speech. These techniques are likely to become more important when DECIPHER is used with spontaneous speech where there is significantly increased in phonological reduction and deletion. [1,5].

References

[1] Bernstein, J., Baldwin, G., Cohen, M., Murveit, H. and Weintraub, M., "Phonological Studies for Speech Recognition," *Proceedings: DARPA Speech Recognition Workshop,* pp. 41-48, February, 1986.

[2] Chow, Y.L., Dunham, M.O., Kimball, O., Krasner, M., Kubala, F., Makhoul, J., Roucos, S., Schwartz, R.M., "BYBLOS: The BBN Continuous Speech Recognition System," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. 89-92, April, 1987.

[3] Cohen, M.H., *Phonological Structures for Speech Recognition,* PhD thesis, Computer Science Department, UC Berkeley, April 1989.

[4] Cohen, P.S. and Mercer, P.L., "The Phonological Component of an Automatic Speech Recognition System," in *Speech Recognition,* R. Reddy, ed., Academic Press, New York, p. 275-320.

[5] Dalby, J., *Phonetic Structure of Fast Speech in American English* PhD thesis, Linguistics Department, Indiana University, December, 1984.

[6] Erell, A., and Weintraub, M., "Estimation Using Log-Spectral Distance Criterion for Noise Robust Speech Recognition," this volume.

[7] Lee, K.F., *Large Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System,* PhD thesis, Computer Science Department, Carnegie Mellon University, April 1988.

[8] Paul, D., "Site Report and Benchmark Tests," presented at *DARPA Speech Recognition Workshop,* June, 1988.

[9] Price, P., Fisher, W.M., Bernstein, J. and Pallet, D.S., "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* pp. 651-654, April, 1988.

[10] Weintraub, M., Murveit, H., Cohen, M., Price, P., Bernstein, J., Baldwin, G., and Bell, D., "Linguistic Constraints in Hidden Markov Model Based Speech Recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing,* May, 1989.