A Statistical Model of Low-Level Phonological Processes
Author(s): Alan Cole and Michael H. O'Malley
Proceedings of the 2nd Annual Meeting of the Berkeley Linguistics Society (1976), pp. 105-116

A STATISTICAL MODEL OF LOW-LEVEL PHONOLOGICAL PROCESSES[1]

Alan Cole and Michael H. O'Malley
University of California, Berkeley

To correctly recognize differeing phonetic realizations of a word, an automatic speech recognition system must incorporate information about low-level phonological variation. A simple statistical model is proposed to describe this variation, and an analysis technique is developed to estimate the statistical parameters of the model. Preliminary results suggest the usefulness of the model for automatic speech recognition.

Description of the Problem

A great deal of the research in automatic speech recognition has been based on the premise that speech recognition systems will eventually have to incorporate detailed information about the structure of speech and language -- a simple "pattern recognition" analysis of the acoustic signal will never be sufficient. For example, Fry and Denes write in 1956:

> Linguistic knowledge must be added to primary recognition and to be completely successful the machine would have to "know" as much about the language as a human brain does.

The practical effects of this tenet were at first extremely limited, but more recently, a determined effort has been made to solve the many problems of including linguistic knowledge within speech recognition systems.

The particular problem addressed here is how to incorporate a knowledge of low-level phonetic and phonological rules into recognition strategies.

The phonetic realization of a word will in general depend on its context, on speech rate and style, on the speaker's dialect, and on other similar factors. If the differing realizations are all to be correctly recognized as the same word, then some knowledge of phonological variation is clearly necessary.

It should be emphasized that only those low-level phonological processes which describe alternations in pronunciation are considered here; higher level processes (such as the derivation of "sanity" from "sane+ity") are beyond the scope of this paper. Figure 1 shows three examples of the sorts of rules which are of interest.

Rules from the linguistic literature are not necessarily directly applicable to automatic speech recognition. While optional phonological rules generate the possible phonetic realizations of a word, it is also important to know the relative

HOMORGANIC STOP DELETION
(Zue, 1974)

$$\begin{bmatrix} C \\ +\text{stop} \\ \alpha\text{place} \end{bmatrix} \rightarrow \emptyset / \begin{bmatrix} C \\ +\text{nasal} \\ \alpha\text{place} \end{bmatrix} \_\_ (\#)\ C$$

especially if the following C is a nasal, a sibilant, or l.

kindness: [kainnəs]
bends: [bɛnz]

MERGER OF FRONT VOWELS BEFORE NASALS
(Cohen and Mercer, 1975)

(1) ɛ → I/_nasals any: [Ini]
(2) I → ɛ/_nasals him: [hɛm]
(3) I → æ/_ ŋ think: [θæŋk]

(1) is common in many varieties of General American and in the South and Southern Mountain region. (2) and (3) occur sporadically in the South and Southern Mountain region.

---

fθ SIMPLIFICATION
(Hill and O'Malley, 1973)

fθ → θ /_V

fθ → f/_C

fifth avenue: [fIθæ vənu]
fifth street: [fIfstrit]

FIGURE 1. Some examples of low-level phonological variation.

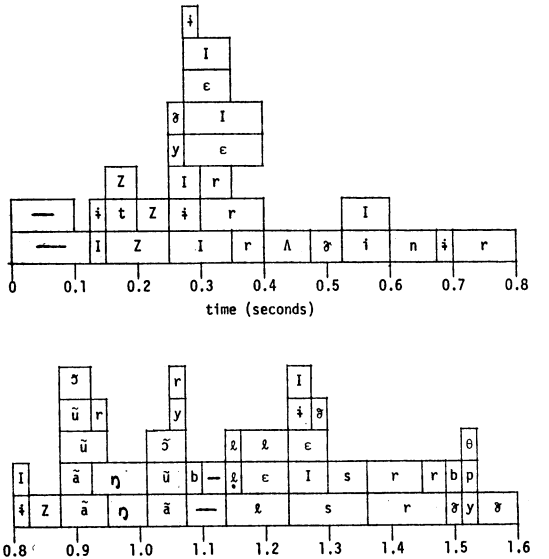

FIGURE 2. A machine transcription[2] of the sentence "Is there any news on the blizzard?"

frequencies with which these different realizations occur -- rule statistics are required.

The treatment of phonological variation in the framework of speech recognition is also complicated by the notorious inaccuracy of machine produced transcriptions. A basic step in most systems is the segmentation and labeling of the acoustic signal to produce what is approximately a phonetic transcription. But errors in both the segmentation and the labeling occur frequently, as illustrated in Figure 2, which shows a typical example of a machine transcription.

The most frequent error is one of substitution. A segment is given the wrong label, and so identified as the wrong phone. Another common error is detection of a segment boundary where none actually exists. This insertion error produces an extra segment in the transcription, but may also be viewed as a decomposition error, since a single real segment is decomposed into two transcribed segments. Conversely, missing a real segment boundary leads to a deletion error in the transcription, or a coalescence error in the sense that two actual segments are merged into a single transcribed segment.

To higher levels in the system, recognition errors and phonological variation look the same, making it difficult to treat either problem in isolation. It may be desirable to describe both sources of variation with a single set of rules. If so, rules from the linguistic literature will be inappropriate.

A Model of Low-level Variation

To deal with the problem of low-level phonological variation in the context of automatic speech recognition, we propose a simple model.

First, each word has one or more base forms, each of which represents a basic pronunciation of the word. More than one base form is allowed per word to handle idiosyncratic variations such as the two forms of "either":

(1) /aɪɚ/

(2) /iɚ/

Since this alternation is not one which would be predicted by low-level rules, both forms are given in the lexicon.

Low-level phonological processes are modeled by a set of generative rules. Each rule maps the base level into the surface level without intermediate stages, and each segment at the surface level is the output of some rule. Each rule must be an instance of one of a small number of rule schemata, defined by the lengths of the input and output strings. Figure 3 shows one possible set of rule schemata. The model does not require that this specific set be used, but it is assumed that each rule falls into one of a limited set of patterns similar to those shown.

FIGURE 3. A set of rule schemata. The symbol "b" represents a base segment, "s" a surface segment.

Substitution	$b \rightarrow s$
Deletion	$b \rightarrow \emptyset$
Insertion	$\emptyset \rightarrow s$
Coalescence	$b_1 b_2 \rightarrow s$
Decomposition	$b \rightarrow s_1 s_2$

Each rule is also characterized by an application probability which is a function of its context.[3] The precise nature of this probability function is unimportant; it might simply be a list of appropriate probabilities for each of the relevant contexts. Another possibility is the use of a variable rule model such as those discussed by Cedergren and Sankoff (1974).

The derivation of an output form from a base form involves the application of several rules. We will assume that the probability of each such derivation is given by some function of the individual rule probabilities. Again, the exact nature of this function is unimportant. One simple possibility is to calculate the probability of a derivation as the product of the probabilities of the individual rules applying in that derivation. This definition corresponds to an assumption of statistical independent of rule applications; different assumptions will lead to different definitions.

As a final point, since there are no intermediate levels between the base and surface levels, there is no need for rule ordering. That is, a rule cannot apply to another rule's output, so the order in which they apply is irrelevant.

Though this model of phonological variation is computationally simple, it may not be immediately applicable in a speech recognition system where transcription errors mask, and are confused with, phonological variation. We may, however, take the previously suggested approach of describing the combined effects of both sources of variation with a single set of rules. In this case, the machine transcription corresponds to the surface or output level of the model.

To show one way in which this model could be used in a speech recognition system, assume that we are given a machine transcription of some utterance which is to be recognized, and that a number of base sequences have been hypothesized, perhaps with the aid of a preliminary examination of the phonetic transcription or with syntactic, semantic, and pragmatic information.

We wish to know which of these base sequences, if any, is the correct one. Since it will normally be impossible to give an absolutely certain answer to this question, we instead simply try to find that hypothesized base sequence which is most likely to have generated the observed transcription.

This process is conceptually straightforward. Find all possible derivations producing the transcription from any one of the

base sequences.[4] Since each such derivation has a probability, we need only find that derivation whose probability, weighted by the a priori probability of the base sequence underlying it, is greatest. That base sequence is then selected as the one most likely to be correct (of the given set of possibilities).

The simplicity of the proposed model of combined phonological variation and recognition error, and in particular its lack of rule ordering, allows finding the most likely derivation without actually finding all possible derivations. Briefly, the mathematical technique of dynamic programming allows us to drop from consideration those partial derivations which cannot possibly turn out to be the most likely.[5]

The model thus appears to be a computationally attractive one for use within automatic speech recognition systems. Two questions remain. First, how do we actually find the rules and their application probabilities? And second, how accurate is the model? The following two sections will suggest answers to these questions.

Estimation of Rule Probabilities

In the past, the determiniation of rule application probabilities has been a laborious task. The data must be examined to find the number of rule applications in a given context compared with the total number of occurrences of that context. Since the context must occur enough times in the data to provide some statistical reliability in the estimates, a large amount of data is normally required. But before the rule counts can be determined, each utterance must be carefully and consistently transcribed. This has been a time consuming and expensive task.

No real solution to the problem exists if only phonological variation is to be studied. But if it is acceptable for the rules to describe both phonological variation and machine recognition error, then machine produced transcriptions may be used instead of hand transcriptions. This allows virtually unlimited amounts of data to be analyzed. The solution is especially useful, of course, in speech recognition, where information about both sources of variation is required.

A second, more theoretical problem is how to tell which rules have actually applied. A complete set of generative rules may well be ambiguous in spite of the rule writer's precautions. That is, more than one derivation may produce a given output from a given base form.

A trivial example may serve to clarify the difficulty. Let us assume that we have both a degemination rule (which deletes one of two adjacent identical consonants) and a dental deletion rule (which deletes /t/ or /d/ between an obstruent and a following consonant). Then, for example, the phonetic form [læ sta^{l}m] for "last time" is produced by either one of the two rules. Which one shall we say has actually applied?

This problem is especially severe when the "rules" describe machine recognition error instead of or in addition to phonological variation. Because of the large number of errors that can, and sometimes will, be made, it is often possible to describe the erroneous transcription of an utterance as the result of any one of several different combinations of specific errors.

We suggest the use of an iterative estimation procedure to answer the question of which rules have applied and to find application probabilities for these rules.

We start with a set of utterances for which the correct base forms are known, and for which transcriptions (hand or machine) are available. We also assume that we have a set of rules whose application probabilities are unknown.

The first step is to make a rough guess at the application probabilities of the rules. These guesses might be totally unmotivated (e.g., every rule applies with probability one-half in all contexts), but should, if possible, be based on previous studies or on an examination of a subset of the data or at least on linguistic intuition.

Using the estimated rule probabilities, it is now possible to determine the most probable derivation out of all derivations which produce the observed transcription from the known base sequence. By assuming that this most probable derivation is in fact the correct one, we now have a probabilistic answer to the fundamental question of which rules have applied.

After repeating this process for each utterance, it is possible to tally the number of times each rule has applied in a given context, and the total number of times that context has occurred in the data. These frequency counts may then be analyzed by, for example, one of the variable rule models proposed by Cedergren and Sankoff (1974) or Sankoff (1975) to obtain new estimates of rule application probabilities.

These new estimates will not be perfect since the decisions about which rules actually applied were based on the original guesses of application probabilities, so that some of these decisions will have been incorrect. The new estimates will be better than the original ones, however, in the sense that they explain a greater proportion of the variation present in the data.

The entire procedure may be performed repeatedly, each time using the most recent set of estimates. This produces successively better sets of estimates which will eventually converge to a final set of values.

One problem of this maximum likelihood method for determining application probabilities of all rules simultaneously is that the final results may depend on the initial estimates. A locally optimum set of probabilities is always found, but finding the global optimum may depend on an auspicious set of initial estimates.

For this reason, it is desirable to make the initial estimates of rule probabilities as accurate as possible. Alternatively, the entire procedure may be repeated with several different sets of initial values in the likelihood that at least one such

set will yield the best possible answer.

The foregoing has assumed that the rules are known and that only their application probabilities as a function of context are unknown. In practice, especially when using machine transcriptions, this will not ordinarily be true. However, this is not a serious problem, for the initial set of rules may include not only all known rules, but all suspected, or even all possible, rules. As the iterative estimation technique is applied, rules which exhibit no descriptive power with respect to the data will be given zero or near zero probabilities, and may be discarded. The remaining rules, which all have a significant probability of applying in at least some contexts, are the rules actually observed in the data.

We have, then, what is in at least a limited sense a discovery procedure for rules. If machine transcriptions are used, then these rules will represent the combined effects of phonological variation and recognition error, while, if hand transcriptions are used, the rules will describe phonological variation together with any possible inconsistencies in the hand transcription process.

Preliminary Results

To test the usefulness of the model proposed here, a small pilot study was performed on a data base consisting of 33 sentences read by a single speaker. A total of 48 different words occurred one or more times in the data. A machine transcription of each utterance was provided by Carnegie-Mellon University's Hearsay II speech understanding project.

The purpose of the study was to determine rule probabilities on the basis of the data, and to evaluate the accuracy of these rules. Because machine transcriptions were used, we expected the rules to describe both phonological variation and machine transcription error.

Not knowing which rules were actually appropriate, we included all possible rules of segment substitution, deletion, and insertion (the first three forms shown in Figure 3). This gave a total of slightly more than 4,000 rules.

Because of the limited size of the data base, it was not possible to determine the effect of context on rule probabilities except in the cases of a few exceptionally frequent rules. Consequently, for the purposes of this study, all rules were assumed to be context-free.

The initial estimates of substitution probabilities were based on a preliminary study of the data; all insertion and deletion probabilities were arbitrarily estimated by two separate constants.

The iterative estimation technique described above was then applied separately to utterances 1-16, utterances 17-33, and to the entire set of 33 utterances, resulting in three sets of final probabilities, based on different portions of the data. Of the original set of more than 4,000 possible rules, only slightly more than 10% ever occurred.

Assuming our model, and given the context-free nature of the rules, the probabilities are the best possible. But a fundamental question still remains: how accurately can the model describe the actual variation in the transcriptions? The answer to this question is not obvious; in fact, it is not even easy to decide what "accuracy" means in this context.

For the purposes of speech recognition, one useful though indirect measure of accuracy is how well an actual recognition system performs when using the model. If the model is hopelessly inadequate, good recognition scores will never be obtained. On the other hand, good recognition scores imply that the model is a useful one.

Consequently, we designed a simple recognition experiment to evaluate the adequacy of the model. For each word in the set of utterances, a single incorrect word was randomly chosen from a list of words often confused (by the machine) with the correct word. Both the correct word and the incorrect word were assumed to be possible, giving, for each sentence, many possible word sequences, depending on which word was chosen at each point.

Using the final set of probabilistic rules, the model was then applied as previously explained to find that word sequence most likely to have generated the observed machine transcription. The percentage of correct words chosen was used as a measure of performance for the model.

Because at any point only one of two words was possible, chance performance was 50%. Using the initial guesses at application probabilities, performance was significantly better than chance, with correct words being selected 80.3% of the time. Using the final calculated values of rule application probabilities, a performance of 93.9% was achieved, which is a statistically highly significant improvement over the initial guesses. Further details are shown in Table 1.

TABLE 1. Word recognition scores with rule probabilities based on different portions of the data.

Score on utterances	Initial estimates	Probabilities based on utterances		
		1-16	17-33	1-33
1-16	78.7%	93.5%	85.2%	93.5%
17-33	81.8%	80.2%	93.4%	94.2%
1-33	80.3%	86.5%	89.5%	93.9%

These results indicate that the model of combined phonological variation and transcription error which we have proposed is accurate enough to be of use in automatic speech recognition, even when the effect of context is ignored. Use of contextual information will further improve this accuracy.

The procedure for estimating rule probabilities also appeared, at least in this instance, to be a robust one. Although seven iterations were required before the probability estimates converged, one or two iterations would have been sufficient to obtain very nearly the same performance. Table 2 shows the recognition performance obtained with the probability estimates resulting from each iteration.

TABLE 2. Improvement in word recognition scores on utterances 1-33 with iteration of the rule application probability estimation procedure.

Iteration	0	1	2	3	4	5	7
Recognition	80.3	93.0	93.4	94.3	93.9	93.9	93.9

To illustrate the sort of rules which might result when a larger data base permits contextual effects to be taken into account, we examined several rules which occurred frequently enough in our data to allow at least a crude estimate of the influence of context.

One such rule was the insertion of an [n] in the transcription. Possible contexts were classified according to the nature of the preceding segment (consonant, vowel, or utterance boundary) and the following segment (stop consonant, non-stop consonant, vowel, or utterance boundary) at both the base and surface levels. The data was analyzed by the variable rule model of Sankoff (1975)[6]. Results are shown in Table 3. In this model, probabilities greater than one-half represent factors favorable to rule application, while factors with probabilities less than one-half tend to block the rule.

TABLE 3. Effect of context on probability of the [n]-insertion rule $\emptyset \rightarrow n$. The symbol "##" signifies an utterance boundary.

Input probability $p_0 = 0.07$

Preceding base segment	##_	V_	C_	
	0.55	0.53	0.42	
Following base segment	_[C +stop]	_V	_[C -stop]	_##
	0.54	0.51	0.49	0.47
Preceding transcribed segment	V_	##_	C_	
	0.65	0.53	0.32	
Following transcribed segment	_##	_[C +stop]	-V	_[C -stop]
	0.87	0.69	0.25	0.16

One of the most favorable contexts for [n]-insertion is seen to be between a vowel and a following stop or utterance boundary (at the transcription level). Since this fails to correspond to any known phonological rule, the [n]-insertion rule clearly represents a recognition error phenomenon.

On the other hand, the /v/-deletion rule shown in Table 4 appears to be similar to a real phonological process previously noted by Shockey (1973). In our case, /v/-deletion is most likely to occur before a (surface level) consonant, and especially likely to occur before a (base level) /m/. The effect of left context could not be estimated for this rule because, except for two cases, the left context was essentially the same for all occurrences of /v/.

TABLE 4. Effect of the following base and surface segments on probability of the /v/-deletion rule $v \rightarrow \emptyset$.

Input probability $p_0 = 0.25$

Following base segment	_m	_C*	_V
	0.75	0.42	0.32
Following transcribed segment	**_C***	**_m**	**_V**
	0.53	0.50	0.46

*The symbol "C" is here used to include all consonants except m.

The final example, a [t]-insertion rule, appears to represent a combination of real phonological tendencies and recognition error. The analysis in Table 5, which includes the effect of a single segment of context at the base level only, indicates that [t]-insertion is most likely between two consonants. In the particular context of a preceding /n/ and a following /s/, this is a special case of the homorganic stop insertion rule (this context occurred twice in the data; both times [t]-insertion took place). However, in many other cases, the rule seems to be an artifact of machine recognition error.

TABLE 5. Effect of the preceding and following base segments on probability of the [t]-insertion rule $\emptyset \rightarrow t$.

Input probability $p_0 = 0.01$

Preceding base segment	C_	V_
	0.68	0.32
Following base segment	**_C**	**_V**
	0.62	0.38

These examples illustrate that when machine transcriptions are used, rules can reflect either true phonological variation or transcription error. But in most cases, the rules will demonstrate a combination of these two factors.

In conclusion, we have described a model of phonological variation and an analysis technique which allows automatic processing of large amounts of data to compute estimates of application probabilities simultaneously for an entire set of rules.

The method appears to be a promising way of including information about phonological variation in speech recognition systems. Furthermore, as machine transcriptions of speech become better, the probabilistic rules will describe true phonological variation more accurately. But even with the current level of machine transcription, we believe the results are both informative and suggestive.

Notes

[1]This research was sponsored by the Advanced Research Projects Agency of the Department of Defense and monitored by the U.S. Army Research Office under grant DAHC04-75-G0088.

[2]This transcription was produced by Carnegie-Mellon University's Hearsay II speech understanding system. Because the system may make several guesses at the identity of a single sound, several possibilities are shown in many locations. Uncertainty about segmentation also produces overlapping phones.

[3]"Context" is interpreted broadly to include both phonological (base level) and phonetic (surface level) context as well as other relevant linguistic and extra-linguistic factors.

[4]If the set of rules is sufficiently "complete", at least one such derivation always exists.

[5]Details of several similar techniques are discussed by Bahl and Jelinek (1975) in connection with a somewhat different model of machine recognition error.

[6]In this model, the probability p of rule application in some context is defined by

$$\frac{p}{1-p} = \prod_{i=0}^{n} \frac{p_i}{1-p_i}$$

where p_0 is an overall input probability, and p_i is a probability associated with that particular factor of the ith (of n) factor groups actually occurring in the context.

References

Bahl, Lalit R. and Frederick Jelinek (1975). "Decoding for Channels with Insertions, Deletions, and Substitutions with Applications to Speech Recognition," IEEE Transactions on Information Theory IT-21, 4 (July 1975), pp. 404-411.

Cedergren, Henrietta J. and David Sankoff (1974). "Variable Rules: Performance as a Statistical Reflection of Competence," Language 50, 2, pp. 333-355.

Cohen, Paul S. and Robert L. Mercer (1975). "The Phonological Component of an Automatic Speech-Recognition System," in Speech Recognition, Invited Papers Presented at the 1974 IEEE Symposium, edited by D. Raj Reddy. New York: Academic Press.

Fry, D.B. and P. Denes (1956). "Experiments in Mechanical Speech Recognition," in Information Theory, edited by C. Cherry. New York: Academic Press.

Hill, Kenneth C. and Michael H. O'Malley (1973). "Fast Speech Rules I." Unpublished SUR Group Note 65 (February 1973). Phonetics Laboratory, The University of Michigan.

Sankoff, David (1975). "VARBRUL Version 2." Unpublished Note (January 1975). Centre de recherches mathématiques, Université de Montréal.

Shockey, Linda R. (1973). "Phonetic and Phonological Properties of Connected Speech," Ph.D. Dissertation reprinted in Working Papers in Linguistics 17 (May, 1974). Department of Linguistics, The Ohio State University.

Zue, Victor W. (1974). "Optional Phonological Rules." Unpublished SUR Group Note 124 (January 1974). MIT Lincoln Laboratory.