A PROCEDURE FOR ADAPTIVE CONTROL OF THE INTERACTION BETWEEN ACOUSTIC CLASSIFICATION AND LINGUISTIC DECODING IN AUTOMATIC RECOGNITION OF CONTINUOUS SPEECH

C.C. Tappert and N.R. Dixon
IBM Thomas J. Watson Research Center
Yorktown Heights, New York

Abstract

An adaptive-control procedure is described which is intended to improve both acoustic analysis and linguistic decoding in automatic recognition of continuous speech by bringing into agreement data available at each of these stages. Specifically, hypotheses are formed by the decoder concerning the phonetic transcription derived during acoustic analysis. The procedure then accesses and utilizes relevant acoustic data in an attempt to verify or reject these hypotheses. Depending on the success of such attempts, actions are taken to constrain the decoding in subsequent processing iterations. Preliminary results are presented and discussed.

Introduction

The work to be described here concerns one aspect of a larger effort, the objective of which is to make inroads toward pragmatic solutions for the very difficult problems involved in Automatic Recognition of Continuous Speech (ARCS) (1,2,3,4). The strategy which has been employed is one in which no attempt is made to model human processing in true analog form (5). The model generated consists of hierarchically arranged, essentially independent stages concerned with signal quantization, subword segmentation, acoustic classification and linguistic decoding.

The present topic is an adaptive-control mechanism intended to improve both the acoustic analyzer and linguistic decoder outputs by bringing into agreement data available at each of these stages. In our work the automatic mechanism which performs this verification task is referred to as the linguistic-processor feedback stage (or adaptive controller) of an automatic speech recognition system. Earlier work concerned with this particular utilization of adaptive control has been reported previously (3,4,6,7). The use of adaptive control in the interaction of successive stages in systems of this sort is certainly not new, and has also been tentatively studied in this system relative to the segmentation and classification stages (8).

Since the data presented here represent a larger set of utterances than hitherto processed by the linguistic decoder, the results are more representative than those previously reported. This was made possible primarily by the increased processing speed of an IBM/360, Mod 91 as compared to a Mod 40, on which the system was developed. Also, for purposes of clarity, relevant processing performed on an example utterance will be presented. This should enable readers who are unfamiliar with this system to obtain at least an overview of the ideas behind the main processing stages. The procedures particularly related to this study are described in detail, whereas other procedures are described only briefly. Obviously, length restrictions preclude the possibility of giving a detailed description of all aspects of the system. Further descriptive information on the most recent version of this speech recognition system can be found in Ref. 4.

Procedures

A block diagram of the speech recognition system under consideration here is presented in Fig. 1. The acoustic processor performs signal quantization, subword segmentation and acoustic classification. The input to the acoustic classifier consists of a sequence of digital Spectral Time Samples (STS), one every 10 ms, with time-aligned phonetic-class boundary indicators (Fig. 2). These boundary indicators generally fall within what are usually considered the apparent phonetic states (nodes).

Acoustic Classifier

Classification consists of steady-state and dynamic classification, node consolidation and anchor-point placement. Spectral correlations are performed for each STS (Fig. 2); for any STS they are simply the five highest correlations (in percent) of that STS with each of a set of referent classes. This set roughly corresponds to the broad phonetic classes of General American English. For a boundary STS, steady-state classification consists of the spectral correlations at that point in time. Dynamic classification is concerned with labeling the inter-boundary regions relative to a set of stored classes. The subword segments so defined relate to the transition from phonetic state to neighboring phonetic state, and are referred to as "transemes." The dynamic-classification output is represented as a rank-ordered, five-deep choice table. Node consolidation combines, by rule, steady-state and dynamic classification at the boundaries into a single phonetic transcription. Additional procedures are employed at this point to delete or insert boundaries based on time-domain information. Anchor-point procedures place symbols immediately following the segmentation indicators, marking the certainty with which the boundary is placed and/or the preliminary name is assigned. A bar (!) indicates a definite segmentation point and a plus (+) indicates a definite boundary name.

The output of the classification stage, which serves as input to the linguistic decoding stage, is basically a string of machine-derived phonetic symbols. Using standard phonetic notation (International

Phonetics Association, an essentially broad* transcription of the example utterance was aurally determined to be

[wʌntʰaɪɛd?ɔfsəfɚgɑtðəpʰɑrtʰ?ʌndəðəkʰʌvɚ]

Using our machine-related, alphaphonetic, two-character-per-node method of transcription the corresponding representation would be

WXUHNXTXTQAAIXERDX?XAWFXSXERFXANRXGXAATX

DHUHPXPQAARXTXTQ?XUHNXDXUHDHUHKXKQUHVXER.

The alphaphonetic string actually obtained from the acoustic classifier for a particular utterance can be considered a noisy version of some acceptable phonetic transcription.

Linguistic Decoder

The function of linguistic decoding is that of phoneme-to-grapheme translation, that is, the conversion of this noisy phonetic sequence from a phonetic representation of speaker performance into a standard-orthographical representation of speaker intention. Linguistic decoding utilizes a speech-oriented, graph-search technique based on the Fano algorithm (9), Sentence production is modeled by a 250-word command language** capable of producing approximately 14 million different sentences. The average length of a sentence is eight words. For each word entry the lexicon contains a description of phonetic paths corresponding essentially to the reasonable pronunciations of the word. These pronunciation variations were generated automatically by a set of phonological rules operating on "ideal" or "standard" pronunciations (10). Since the lexicon describes approximately 100 speaker-dependent phonetic paths per word, the tree to be searched contains about ten-to-the-sixteenth-power phonetic paths per sentence, not including paths to account for machine error. The decoder output consists of time-aligned, hypothesized phonetic-symbol and word strings with indications of the types of hypotheses which are necessary to arrive at these strings from the input data.

* Except for stop aspiration we have used broad phonetic transcription in this work. Broad phonetic transcription roughly approximates phonemic transcription. On occasion, however, we have experimented with certain allophonic variations, such as light versus dark [l]. The restriction to essentially broad transcription has been due largely to computer considerations involving acoustic classification.

** The command language is specified by a simple context-free grammar (4). The criteria by which our choice of a command language was made are perhaps worth mentioning. The primary criteria were that i) the grammar must be easily implementable by machine, ii) the statistical properties must be such that phonetic decoding could be controlled and studied independently of syntax and semantics, and iii) it should be of such a complexity that system performance is neither too good nor too poor, so that system improvements can be easily seen.

For the example utterance this output is shown in Fig. 3. These hypotheses are concerned with two types of noise: that considered to originate from the speaker (Dependent, D) and that from the system (Independent, I). Within each type, the noise is further categorized as Substitution (S), Omission (0) or Insertion (I) of phonetic events relative to the hypothesized string. Thus, in the example utterance, the "IS" at the fourth node represents the hypothesis that the input node was a speaker-independent (machine-related) substitution of one sound [p] for another [t], Anchors are utilized to restrict the types of hypotheses allowed at the node level. Thus, I does not permit insertion hypotheses and + does not permit substitution or insertion hypotheses, relative to the hypothesized string.

Adaptive Controller

The aim of the control mechanism is to verify or reject the hypotheses made in arriving at the linguistic-decoder output so as to improve subsequent processing. Feedback is made at the sentence level; the control-mechanism procedures may modify the phonetic and/or anchor information which, in turn, may lead to changes in subsequent decoder output. The procedures make use of information available at the acoustic-classifier level. Specifically, sequential spectral-correlation, ranked dynamic-segment classification and a general "word-fit" criterion are employed. The word-fit criterion concerns the degree to which an hypothesized word requires hypothetical change from the noisy input string.

The results of adaptive control on the example utterance are presented in Fig. 4. The only portion of the example utterance changed by feedback was the prepositional phrase. For the first and second passes the hypothesized word output for that portion was "...between those camps." On the third and fourth passes the output became "...over the cover" and then "...on the cover," respectively, never reaching the intended "...under the cover." It might be noted that of these four passes the alignment of input and hypothesized nodes is considered most accurate (relative to speaker performance) for pass three, due to the proper alignment of "the."

On the three types of hypotheses which may be made at a particular point in the string — substitution, omission or insertion — the action taken by the control mechanism may be positive, negative or neutral. For substitution hypotheses the positive *or* negative action is to change to, or prohibit change to, the hypothesized name. For omission hypotheses, the positive or negative action is to insert, or prohibit insertion of, the hypothesized name. For insertion hypotheses, the positive or negative action is to delete, or prohibit deletion of, the name originally inputted to the linguistic decoder. For all three hypothesis types, neutral action leaves the original input unaltered. The procedure is as follows.

First, the word-fit criterion is applied in terms of information at the word level within the adaptive controller. An hypothesized word is considered in "poor fit" if i) every node in the word required hypothetical change, ii) a gross class substitution was hypothesized (e.g., a vowel hypothesized as a stop) and/or iii) the ratio of independent omission and insertion hypotheses to hypothesized-word length exceeded a threshold. The word-fit criterion is used either to allow or disallow the use of the hypothesized word and corresponding components to direct the attempt to verify an hypothesis; that is, positive action can only be taken if the word-fit criterion is met, whereas negative action can be taken independently.

Second, substitution hypotheses are tested using as confirming or rejecting criteria the classification information at the node (boundary correlation and transeme names) level. Confirmation takes place by changing an input node to the hypothesized node when there is sufficient information to support the hypothesis. Rejection takes the form of i) plus (+) anchor placement if the original name is detected to a high degree in the classification information or ii) if no evidence of the hypothesized name is found, that name is suppressed in subsequent linguistic decoding. Since all actions are cumulative, more than one name could be prohibited as a substitution for a node after several passes. For the example utterance, all rejections were of the suppression variety.

Third, omission hypotheses are tested using as confirmation or rejection criteria the classification information at the inter-node level — that is, correlations between boundaries.* For example, a vowel[i] was hypothesized as having been omitted between the eleventh[f]and the twelfth [s] nodes of the original string. This was an hypothesis of dependent omission (DO). Since the "word-fit" criterion was met, the adaptive controller accessed the information between these boundaries, looking for a sequence of five successive IX[i] names in the correlations within three percent of first-choice correlation (Fig. 2). This would have been sufficient to confirm the hypothesis, inserting into the string for subsequent linguistic decoding. However, since such a search did not produce sufficient evidence of the hypothesized sound, no action was taken. For an independent omission (10), if there is no evidence of the hypothesized sound found in the classification data, a dot anchor {.) is placed after the indicator immediately following the hypothesized omission, disallowing omission hypotheses at this point in the string during subsequent linguistic decoding.

Fourth, insertion hypotheses are tested by searching the correlations between the preceding and following boundaries for confirmatory information, treating the original boundary as a non-boundary state. The general procedure for eliminating nodes hypothesized as insertions is to test the hypothesis that an extra segmentation boundary was adventitiously inserted. The test employed attempts to find a string of names, within a specified percent correlation and corresponding to either the left or right boundary, which will noninterruptedly bridge with the node hypothesized as inserted. If bridgeable, the node hypothesized as inserted is deleted for subsequent processing. On the other hand, if bridging does not occur for an independent insertion, and if strong evidence indicates that the node in question is actually present, a bar (1) anchor is placed so that in-subsequent processing an insertion hypothesis cannot be made.

Plosive-aspiration states (such as TQ) are handled by special procedures in which the goal is to eliminate plosive aspiration as adventitious, relative to the broad-transcription representation in the lexicon used by the linguistic decoder. These states are identified and generally retained in early stages of processing because of the acoustic-phonetic fact that a low-energy fricative following a stop consonant at a word boundary is essentially indistinguishable from plosive aspiration: e.g., "make the"[metkia] versus "make a" [rmik^a]. Syntactic constraints are invoked during linguistic decoding to disambiguate such "minimal-difference" cases.

Several operational criteria for terminating the feedback looping governed by the adaptive controller are feasible. These include the following: i) no change occurs in the hypothesized word string, ii) no change occurs in the hypothesized node string, iii) no increase occurs in similarity measure, iv) no decrease occurs in linguistic-decoder computation time, v) insufficient positive or negative action occurs, vi) a specified number of feedback loops is completed, and vii) some combination of the above. The following criteria were employed for the present system.

1. In order to make another feedback loop, there must be at least one hypothesis rejection, i.e., one or more node suppressions or dot, bar or plus anchors. This was considered reasonable in that supportive action should not cause a change in word output.
2. For computational purposes a specified number of feedback loops cannot be exceeded. A maximum of three was employed here.

Since it was found that simply taking the last output string was not optimal, additional selection criteria were established to select the best output from among the candidates, i.e., the original and feedback passes. It was empirically established on training data that it would be reasonable to take the last complete sentence output, if any complete sentence occurred, and to simply take the last output where only incomplete sentences were obtained. Thus the only outputs selected which were not the most recently obtained

*Medial names in transemes were also used in an earlier set of procedures (3).

were those in which a complete sentence became incomplete through feedback.

Results

Evaluation was performed on a total of 95 sentence-length utterances, 23 test utterances from one speaker under the intra-speaker condition, and 24 test utterances from each of three speakers under the inter-speaker condition. Thus, one of the training speakers was also used for evaluation on test sentences. These performance results are shown in Table 1.

Performance was evaluated in percentages at the following levels, before and after feedback.

1. Sentences correctly decoded relative to all sentences entered. This provides a measure of overall sentence performance with no other factors considered.*

2. Sentences correctly decoded relative to those sentences outputted in complete form. If sentences outputted in incomplete form are considered to be system rejects, and therefore subject to reentry, then this percentage provides a measure of system performance relative to system acceptance. If this percentage were sufficiently high and substantially higher than the first measure, the rejection condition would become operationally useful.

It might be mentioned that it is probably possible to achieve the above by making it more difficult for the system to output complete sentences by tightening system constraints. This would undoubtedly occur only at the expense of an increase in rejection rate, ceteris paribus.

3. Sentences rejected.

4. Correct words in non-rejected sentences. This provides additional analysis at the subsentence level where full sentences are outputted. If enough were known on an operational basis about the relative importance of words according to some linguistic/semantic referent, such an analysis would become increasingly meaningful with regard to information transfer. With this in mind, it is probable that, as such linguistic/semantic referents are brought within operational knowledge, this measure will become increasingly important.

5. Words correct relative to words outputted, whether or not in complete sentences. This corresponds, at the word level, to number 2 above.

6. Words rejected.

7. Words correct relative to all words entered. This corresponds to 1 above.

8. Words correct relative to words outputted from those utterances in which feedback changed the word output. This provides a specific measure of feedback performance.

The speaker involved in both training and test-sentence evaluation will hereafter be referred to as "speaker 1". As might be expected, accuracy within the linguistic processor was somewhat higher for this speaker than that for the other test speakers.

Discussion

Philosophy of Analysis

If the goal of processing is automatic speech understanding, the performance evaluation must be tied to the appropriateness of consequent "actions" taken as a result of semantic factors. However, for the present system, whose goal is automatic conversion of speech signals to printed sentences in standard orthographical form, questions arise in the analysis of performance data concerning the referent to be employed. For linguistic processor output the parameters of concern are accuracy at the word and sentence levels. Two types of errors are considered for each level — errors of omission and errors of commission. The first are typically manifested in terms of word or sentence rejection by the linguistic processor and the second by words outputted in error. This corresponds with human speech-perception behavior insofar as the listener asks the speaker to repeat himself (rejection) or misinterprets what the speaker said (substitution, deletion and/or insertion of words).

If one cares to carry the analogy further it may be reasonably hypothesized that human listeners err in their speech perception behavior, assuming optimal signal conditions, in terms of two major phenomena — those errors of omission which result when the input will not map through the listener's "model" and those errors of commission which result when a wrong path is taken through the model. From a diagnostic point of view, it can be seen that both errors of omission and commission arise from the same two sources — receiver model limitations and input corruption. Further diagnosis of the human situation becomes exceedingly difficult and, in fact, virtually impossible, due to the subtleties of the communication event, listener "set" and point of view of the diagnostician. For example, while some small nuance in the expression of a speaker may be completely misinterpreted by the listener, it is moot whether the miscommunication is due to speaker misevaluation of listener sophistication or to the limited

* In addition, percentage of sentences with only one word in error relative to all sentences is of interest, since there can exist words which are minimally different phonetically, but equally likely at some point in the string. This is one condition under which a very minor error in acoustic analysis can create, without more sophisticated language constraints, an incorrect sentence. The percentage of time that this happens, therefore, can aid in analysis of linguistic processing at the node and word levels.

sophistication of the listener. It must be noted that "miscommunication" involving human processing necessarily involves levels of processing which are much deeper than those operationally involved in automatic processing by machine at the present time. With the exception of some trivial cases, in which lexical items may be viewed independently as operational variables, i.e., simple oppositions such as yes/no and the like, there are presently insufficient operational data extant concerning these deeper levels of natural language as they may be manifested phenomenologically.

With the above in mind, it appears to be reasonable to set as the evaluation referent for system" performance variations at both the phonetic and lexical levels which are not in violation of the model being utilized as our "listener", i.e., the ARCS system, not including statistical phenomena considered to be speaker-independent.

At the node level, if an event in the output string cannot be reasonably accounted for, then it is considered to be an error. In such cases it is possible to determine the source of the error, whether it be in segmentation or classification.

At the word level, if an input string is accounted for and the linguistic-processor output is incorrect relative to the known input, then the word is considered an error arising from the linguistic processor. However, the processor has been designed so as to overcome input contamination resulting from the speaker and/or acoustic analysis. Thus, while components of a word will be considered in error the word itself may not be in error; word errors and phonetic errors may be treated independently.

The most stringent test of system performance is made at the sentence level. In order for a sentence to be correct, its description at all levels of processing must remain within the capability of the system to overcome error. Excessiveness of error at any level of processing will be reflected at the sentence level, since the a priori probability of correct sentence output is virtually zero (1/14 million). Ideally, the system will reject strings in which contamination is excessive. This may be viewed as a reject condition at the sentence level, corresponding to the necessity for the speaker to repeat the sentence. A sentence may be considered in error in varying degrees, depending on the number of words rejected or outputted in error, in reality, a simple counting of correct words in. relationship to total word content of a sentence does not reflect the true degree of error involved in terms of information. It is clear that for purposes of semantics the kind of word outputted in error should be of primary interest; in fact, a one-word error can be more severe than multiple-word errors depending on the semantic weights of the words involved. Since there are no semantic factors incorporated/ as such, in the present system, it is not meaningful to evaluate performance along this dimension.

Diagnosis of System Error

Additional analysis of the performance data beyond pure exposition is extremely difficult at the present time. It should be clear that in the development of a system as complex as this, the optimization and stabilization of performance will progress as a function of depth into the system, i.e., as a function of cumulative complexity. However, these results are certainly encouraging for the present approach.

In analyzing error in the output of the linguistic decoder, it became apparent that a considerable amount of the error term was attributable to interaction between the two major stages of processing. An additional error component was explainable in terms of inadequacies at one or the other stage.

For acoustic analysis the primary problems arose from intrusive phenomena which exceed the present classification nomenclature. Some examples of such are glottal arrests at word boundaries (the S1 at STS 77 and the M at STS 218 in Fig. 2) — normally classified as stops by the system; aspirated releases of initial vowels normally classified as fricatives or stop aspiration; voice breaks and fundamental-frequency and/or integration-interval interaction — normally classified as rapid closures (short, shallow dips in the short-time amplitude measure); and partial dissimilation of vocalics when they occur between fricatives — potentially classifiable as one fricative for fricative-vocalic-fricative. A number of methods for overcoming these errors have been investigated. Some of the techniques were incorporated into segmentation, others into dual classification, and still others into linguistic processing, primarily in terms of lexical representation. In the example utterance, the two fricatives in "officer" were obtained via one of these corrective methods at the segmentation level. It is clear, however, that in many cases there is a need for expanding the nomenclature used in classification so as to allow the identification of some of these events uniquely. Consequently, the decoding at later stages must be able to hypothesize correctly concerning such events, without penalizing as heavily as with "true" errors in the input. An example of this would be the unique classification of glottal arrests and releases.

Some other problems arose from false-positive conditions in the time-domain processing during dual classification. The primary offender was false deletion of nodes needed in order to traverse the correct path in linguistic decoding. In the example utterance d in "under" and the vowel of the second "the" were deleted. These deletions were the major problem preventing correct decoding of this utterance. The opposite condition — false insertion — rarely occurred. It was noted that certain kinds of errors tended to occur in pairs, primarily in the case of adventitious rapid-closure identification. Under this condition a single continuant becomes three events — the

first part of the continuant, the rapid closure, and the final part of the continuant. While methods have been implemented to correct such errors, they continue to represent a point of difficulty for the entire system.

Within linguistic processing the primary decoding problem takes the form of a mismatch between word ranking and phonetic decoding. As the system is presently employed it is possible to rank a word by utilizing a particular word template in which specific errors of omission/insertion are hypothesized and then proceed with decoding at the phonetic level with no knowledge of the template assumptions made. In practice, opposing hypotheses can occur in phonetic decoding relative to word ranking, resulting many times in a wrong path. This problem contributes heavily to two forms of undesirable performance — a high reject rate and low accuracy at the sentence level.

Another problem concerns edge effects. Backtracking is never initiated for the last hypothesized node of the final word of an hypothesized sentence, and a single remaining input node can be left unaccounted for without decreasing the similarity measure {Fig. 4). These techniques were employed to encourage sentence completion within a reasonable processing time and are considered temporary measures during the development of the system. Another technique used to encourage sentence completion was the rather liberal setting of threshold spacing; for the example utterance, a tighter threshold spacing of 2.5 directly yeilded output corresponding to passes 3 and 4.

The present method used for lexical representation also has some shortcomings which reflect themselves in system error. Lexical representation is insufficiently flexible to account for multiple "ideal" transcriptions. It is obvious that for any word there exists more than one transcription which may be considered "ideal," depending on context. Since use is not made of long-term phonetic context, it is not reasonable to allow all "dependent" variations of a word with equal probability, as this would essentially negate the advantages obtained in the present probabilistic processing. For example, provision cannot be made for both [intrists] and [intris] for the word "interests". In fact, the concept of "ideal" is questionable for these purposes. It may be more appropriate to substitute a phonologically-based, expected string for the present "ideal" representation. Conversely, the present lexical representation is sufficiently constrained relative to the method of accounting for dependent omissions in that multiple sequentially occurring omissions are permitted in error. Fortunately, the penalty placed on dependent omissions is sufficient to avoid this problem in the vast majority of cases.

Another problem concerns some limitations of the syntactic model. Most obvious is the way in which the definite article "the" is handled. While this word is treated as a required word in certain paths. in practice it has been found that it is not uncommon for speakers simply to omit the word phonetically and instead rely on prosodic features. Some form of prosodic analysis will be required.

Another component which will undoubtedly be required is some mechanism to account for fusion phenomena at word boundaries such as the phonetic change from[didju]to[d idzu] in the word pair "did you".

Specific versus General Feedback Action

There are two types of verification/rejection handled by the adaptive controller — specific and general: those which are specific concern particular node names (e.g., suppression prohibits the hypothesis of a particular name at a particular point in the string); those which are general concern more than a single node name (e.g., a bar anchor prohibits the hypothesis of insertion for any_ node name at that point in the string). Optimally, a system would contain specific and general actions for each hypothesis type. Because the present system provides no specific action types for certain situations (e.g., where a dot anchor prohibits any omission hypothesis in subsequent processing), it is not difficult to imagine a situation *in* which a general suppression of the omission hypothesis, based on the inability to find a specific name in the data, would lead to difficulty when another event had indeed been missed. At the present status of system development essentially only the general mechanisms have been provided.

It might be noted that the application of specific actions can be considered looser than those which prohibit a whole class of hypotheses. The empirical data demonstrated the need for specific actions regarding hypothesized insertions and omissions in that the main action which led to increased decoding accuracy was the specific action of suppression, accounting for approximately 70 percent of the actions taken.

Comparison with Manual Speech Recognition

The primary source of data presently available which may be used for purposes of comparison is that of Klatt and Stevens (6,7),* in which manual speech recognition was attempted by visual examination of

' Vicens (11) described an automatic system in which multiple-word input was limited by utilizing a 16-word vocabulary and a highly-constrained syntax, capable of generating 192 sentences. In this highly constrained situation, accuracy at the word level was 90 percent when new speakers were used. The system contained no provision for adaptive control of the type under consideration here. Since this system utilized a highly constrained vocabulary and syntax in order to avoid all but the coarsest phonetic distinctions, it would have to be modified significantly in order to achieve useful word recognition accuracy with vocabularies of the size being employed here.

spectrograms. On 19 sentences {158 words) produced by five speakers using entries from a 200-word lexicon with a syntax describing "...questions asked of a computer program whose data base concerned the chemical analysis of moon rocks (12)," they reported overall word-recognition performance of 96 percent correct for each of two experimenters. They also employed a form of verification/rejection akin to that described here.

It is of interest to compare the present performance data at earlier stages of processing with those reported by Klatt and Stevens. In the latter work, 658 phonetic segments were transcribed. At the segmentation level, 10 percent were missed and no mention was made of adventitious segments. At the classification level, 17 percent were incorrectly transcribed and 4O percent were transcribed only partially (in accordance with a phonetic-feature inventory). Thus 33 percent of the segments were completely correct and 73 percent were either partially or completely correct. In the present work, 1800 segments were transcribed from two speakers who were new to the system. At the segmentation level, 2 percent were missed and 4 percent were adventitious. At the classification level, B1 percent of the segments were completely correct, not including those missed and 90 percent were correct at the phonetic class level (the nearest analog to the phonetic-feature measure). Although these measures are not exactly comparable, the automatic acoustic analysis of the present work appears to compare favorably with the reported manual acoustic analysis. Therefore, the difference between word accuracy of these two studies appears to be explainable in terms of the difference between processing by the human system at the word level using all the constraints available to that system and automatic processing with limited syntactic and no semantic constraints.

Klatt and Stevens' "subjective" evaluation of their performance was that it is not encouraging for workers in the field of automatic speech recognition because of the "...seeming complexity of the things we were doing in our heads in order to recognize a feature or word or phrase"(6). The present writers are encouraged, however, by the performance data being reported here, since they were obtained using completely automatic procedures at all levels of processing, and since these data are based on the utterances of several -untrained speakers using a larger vocabulary. In addition, it is felt that adaptive control procedures, such as those herein described, will be significantly helpful in achieving eventual convergence of input and output in ARCS. These results should also be encouraging for other workers who are interested in pragmatic solutions for ARCS problems.

Certainly much remains to be done in this area and the present work can only be veiwed as an initial probe. Feedback of whole sentence-length strings may, in fact, not be the optimal and only level at which feedback should be performed. For example, it is likely that a combination of looping at the word level during initial decoding would be more appropriate when both accuracy and processing time are considered. In such differential looping the confirmation or rejection of hypotheses would probably depend on different levels of analysis, most likely phonetic at the word level and semantic or prosodic at the sentence level.

References

(1) Tappert, C.C.; Dixon, N.R.; Beetle, D.H.; and Chapman, W.D., "A Dynamic-Segment Approach to the Recognition of Continuous Speech: An Exploratory Program," Tech. Rept. No. RADC-TR-68-177, Rome Air Dev. Ctr., Griffiss Air Force Base, N.Y. , 1968.

(2) Tappert, C.C.j Dixon, N.R.; Beetle, D.H.; and Chapman, W,D., "The Use of Dynamic Segments in the Automatic Recognition of Continuous Speech," Tech. Rept. No. RADC-TR-70-22, Rome Air Dev. Ctr., Griffiss Air Force Base, N.Y., 1970.

(3) Tappert, C.C.; Dixon, N.R.; Rabinowitz, A.S.; and Chapman, W.D., "Automatic Recognition of Continuous Speech Utilizing Dynamic Segmentation, Dual Classification, Sequential Decoding and Error Recovery," Tech. Rept. No. RADC-TR-71-146, Rome Air Dev. Ctr., Griffiss Air Force BaBe, N.Y., 1971.

(4) Dixon, N.R. and Tappert, C.C., "Intermediate Performance Evaluation of a Multi-stage System for Automatic Recognition of Continuous Speech," Tech. Rept. No. RADC-TR-72, Rome Air Dev. Ctr., Griffiss Air Force Base, N.Y., 1972.

(5") Dixon, N.R. and Tappert, C.C, "Strategic Compromise and Modeling in Automatic Recognition of Continous Speech: A Hierarchical Approach," J. Cybernetics, Vol. 1, No. 2, 1971.

(6) Klatt, D.H. and Stevens, K.N., "Strategies for Recognition of Spoken Sentences from Visual Examination of Spectrograms," Report No. 2154, Bolt Beranek and Newman Inc., Cambridge, Mass., 1971.

(7) Klatt, D.H. and Stevens, K.N., "Sentence Recognition from Visual Examination of Spectrograms and Machine-Aided Lexical Searching," Proc. 1972 Conf. on Speech Communication and Processing, Newton, Mass., April, 1972.

(8) Tappert, *C.C.,* "A Preliminary Investigation of Adaptive Control in the Interaction Between Segmentation and Classification in Automatic Recognition of Continuous Speech," IEEE Trans. on Systems, Man, and Cybernetics, VoL. SCM-2, Jan., 1972, pp. 66-72.

(9) Tappert, C.C., Dixon, N.R. and Rabinowitz, A.S., "Application of

Sequential Decoding for Converting Phonetic to Graphemic Representation in Automatic Recognition of Continuous Speech," Proc. 1972 Conf. on Speech Communication and Processing, Newton, Mass,/ April, 1972.

(10) Tappert, C.C. and Dixon, N.R., "An Experimental Technique for Establishing Lexical Variants by Rule in Automatic Recognition of Continuous Speech," J. Acoust. Soc. Am., Vol. 53, Jan. 1973", p. 355. (Abstract)

(11) Vicens, P., "Aspects of Speech Recognition by Computer," Report CS-127, Ph.D. Thesis, Computer Science Department, Stanford University, 1969.

(12) Woods, W., "The Lunar Sciences Natural Science Information System," Internal Report under Contract No. NAS9-1115, Bolt Beranek and Newman Inc., Cambridge, Mass., 1971.

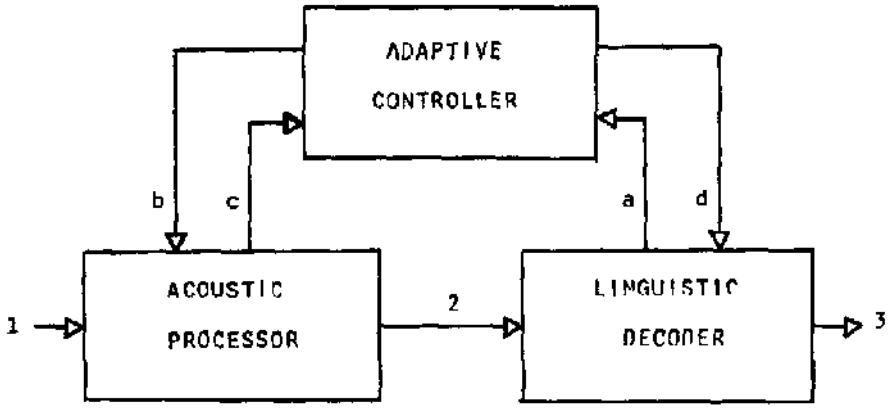

Figure 1. Simplified block diagram of ARCS system with data paths. Main data path conaista of analog speech input (1), node string with corresponding segmentation indicators and anchors (2), and final standard-orthographic output (3). The adaptive-controller paths consist of hypothesized word and node strings (a), adaptive-controller examination of acoustic-processor information (b), confirmation, neutral, or rejection information (c), and modified node string with corresponding segmentation indicators and anchors (d).

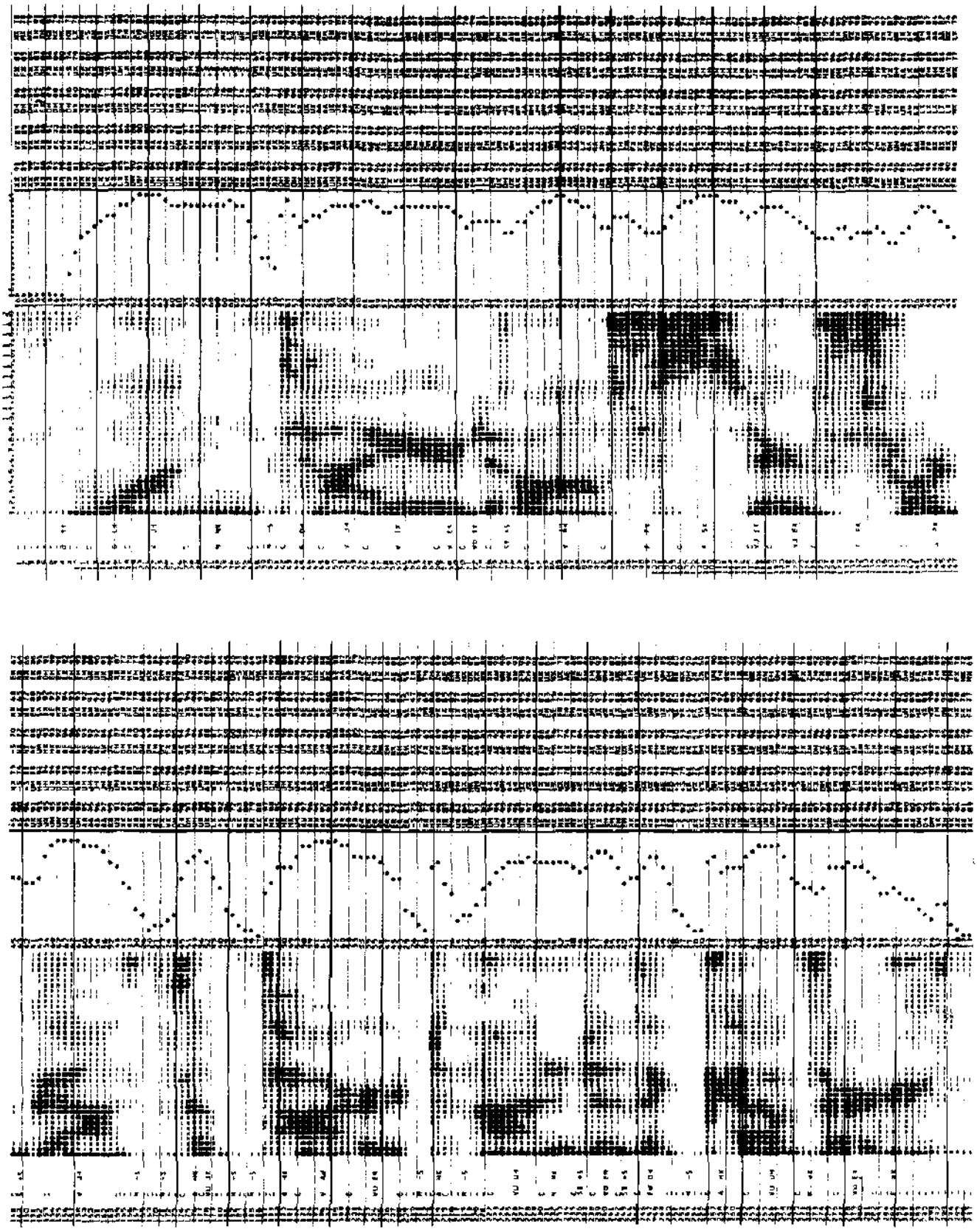

Figure 2. Digital sound spectrogram of example utterance, "One tired officer forgot the part under the cover." From bottom to top are STS numbers, segmentation and auxiliary indicators, steady-state names, amplitude in dB, amplitude plot, 1st-5th node correlations.

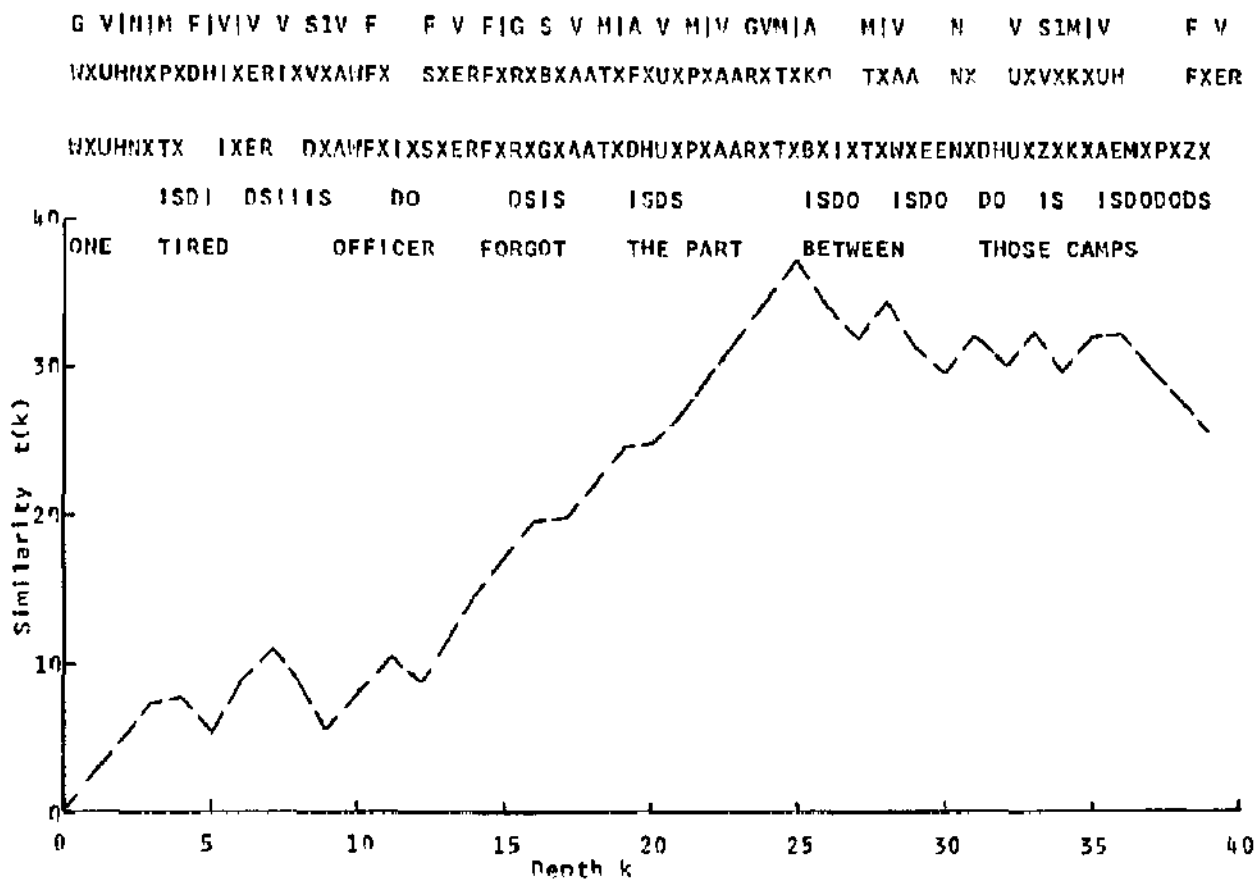

Figure 3. Similarity measure for example utterance before feedback. Decoder threshold spacing was 5. The top two lines represent time-aligned acoustic-processor output, indicators with anchors above the node names. The next three lines represent decoder output, hypothesized nodes, types, and words, respectively.

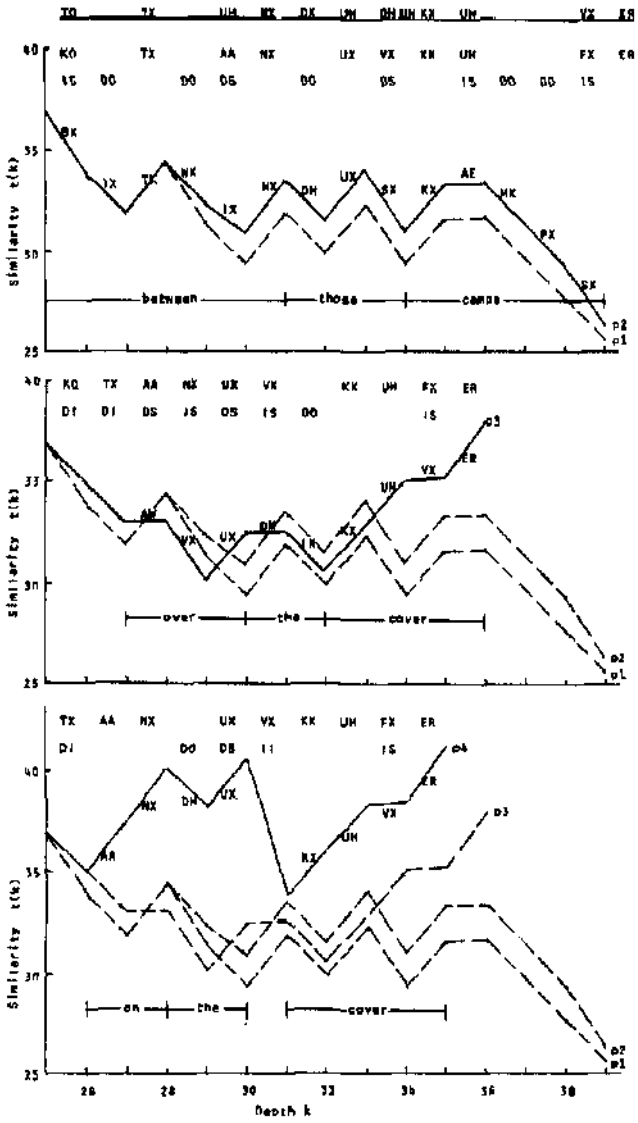

Figure 4. Feedback similarity measures for the prepositional phrase of the example. Above the top solid line is a description of speaker performance (the referent). Above each graph are time-aligned input nodes and hypothesis types; on each graph is the hypothesized node string; below each graph is the hypothesized word String, p1 is the original pass and p2-p4 are the lst-3d feedback passes. Feedback actions were: after p1, AA=WX, VX=ZX, FX=ZX; after p2, AA=IX, VX=SX, FX=SX{ after p3, KQ deleted, NX=VX.

Measure	Before Feedback		After Feedback	
	New Spkrs	Spkr 1	New Spkrs	Spkr 1
Correct Sentences / All Sentences	13	43	19	48
Correct Sentences / Complete Sentences	16	53	22	55
Reject Sentences / All Sentences	22	17	13	13
Correct Words / Words in Sentences Out	77	86	79	86
Correct Words / Words Out	74	83	79	86
Reject Words / All Words	8	8	6	6
Correct Words / All Words	68	77	75	81
Correct Words / Words Out for Sentences Changed by Feedback	68	69	80	82

Table 1. Linguistic-processor feedback performance for new speakers and one training speaker. All entries are in percent.