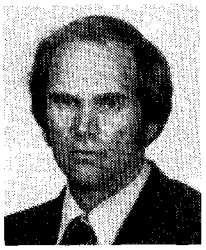


Thomas P. Barnwell, III (M'76) received the S.B., S.M., and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, in 1965, 1967, and 1970, respectively.

From 1965 to 1966 he was a National Science Foundation Fellow. He was a Teaching Assistant during the 1966-1967 school year, and a National Institutes of Health Fellow from 1967 to 1970. Since 1971 he has been at the Georgia Institute of Technology, Atlanta, where he is now a Professor. While at Georgia Tech., he has been involved in the development of the Digital Signal Processing Laboratory, and has introduced several courses in the areas of speech processing and digital systems. His research activities are in the areas of speech processing techniques, digital systems, and digital architecture for signal processing. He has been the principal investigator in numerous research programs in these areas, and is the author of numerous related papers and technical reports.

Dr. Barnwell is a member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.

Demisyllable-Based Isolated Word Recognition System

AARON E. ROSENBERG, MEMBER, IEEE, LAWRENCE R. RABINER, FELLOW, IEEE, JAY G. WILPON, AND DANIEL KAHN, MEMBER, IEEE

Abstract—A speaker-dependent speech recognition system is described for recognizing isolated word utterances using reference templates created by concatenating demisyllable (half-syllable) prototypes. Each word in a vocabulary is specified by one or more entries in a user-supplied lexicon containing a sequence of demisyllables drawn from a corpus of some 1000 units. Experiments were carried out with two talkers using a 1109-word "Basic English" vocabulary to assess the overall effectiveness of demisyllable representations for words. Also, the effects on performance of some simple modifications in demisyllable specifications and adjustments of demisyllable durations were investigated. The recognition error rates obtained for this vocabulary using demisyllable prototypes were 18-33 percent compared with 6-15 percent using whole word prototypes. Although the performance is substantially poorer using demisyllable representations in place of whole words, the approach of using a fixed inventory of smaller-than-word recognition units capable of representing any spoken word in a simple concatenative scheme is clearly an attractive alternative to whole-word prototypes for large-size vocabularies. The approach also has the potential of being effective in representing and recognizing continuous spoken utterances.

I. Introduction

IN THIS PAPER we report on speaker dependent, large vocabulary, isolated word, automatic speech recognition experiments using demisyllable (half-syllable speech unit) prototypes. The use of units smaller than words represents a major departure from our previous studies [1] in which whole word prototypes have been used exclusively. The change has been made because of the (anticipated) limitations of whole word templates as vocabularies increase to a range suitable for continuous speech recognition (i.e., 1000-10 000 words). These limitations involve training, storage, access, and processing of large numbers of whole word templates. Although the change from whole word to demisyllable prototypes is a major one, it will be seen that the impact of the change on our basic system architecture is not so great, and that many of the techniques that have been successfully applied previously are incorporated in the present system.

Our ultimate objective is the automatic recognition of spoken utterances from large (if not essentially unlimited) vocabularies. We are attempting to reach this goal by representing utterances by sequences of demisyllable units taken from a fixed inventory of such units. The purpose of the experiments reported here is to give a partial assessment of that capability, namely the recognition of isolated word utterances drawn from a vocabulary of 1109 words with each vocabulary word represented by one or more sequences of demisyllable units.

This paper is organized as follows. In Section II we motivate our selection of the demisyllable as a recognition unit and establish its definition. In Section III we describe the lexical basis of our system, the methodology for construction of entries in the lexicon in terms of demisyllable units, and the way

Manuscript received April 2, 1982; revised December 31, 1982.
The authors are with Bell Laboratories, Murray Hill, NJ 07974.

0096-3518/83/0600-0713$01.00

in which demisyllable inventories are created. In Section IV the basic speech recognition process is described with particular emphasis on contrasting our template matching technique with segmentation approaches. In Section V we describe the experiments performed to evaluate the use of demisyllables in a large vocabulary, speaker trained, isolated word recognition system and describe and discuss the results obtained. In Section VI conclusions are given.

II. Unit of Representation for Recognition

As stated earlier, our overall objective is the recognition of continuous speech utterances from arbitrarily large vocabularies. Our point of departure is our present word-based recognizer. Our hope is to meet the objective by enhancing the capability of this recognizer with minimal changes to both the signal processing and the pattern recognition algorithms (i.e., the basic system architecture).

In Section IV it is shown that the operation of the recognition system (for isolated words) is based on a template matching procedure with implicit alignment of events in the input speech to events in word templates. Our approach to connected word recognition with this system has been to compare an input sequence of reasonably carefully articulated connected words with concatenated whole word templates [1]. This approach is reasonable if the following conditions are met:

1) phonological variations within words are well represented in the word templates,

2) phonological influences across word boundaries are negligible,

3) the matching process can provide an adequate alignment in time between events in the input sequence and the sequence of templates.

The results obtained to date with matching connected word sequences from whole word templates have led us to believe that the conditions are reasonable [2]. Therefore, we hypothesize that the approach can be extended to recognize continuous utterances from arbitrarily large vocabularies by changing to a recognition unit smaller than a word. This recognition unit should come from a fixed inventory whose aggregate can completely represent all utterances while satisfying the conditions given above. Thus the recognition unit must be capable of being concatenated with other units to represent words, phrases, etc. The unit must have the property that most phonological phenomena are contained within the boundaries, and therefore only minimal care need be exercised at unit boundaries to account for phonological influences across boundaries. Finally, the unit must have the property that events in the unit can be matched and aligned to events in the input speech.

It should be clear that the best choice for such a recognition unit would be the largest speech sound (smaller than a word) with well-defined boundaries and a practical inventory size. The larger the unit, the greater is the amount of phonological phenomena contained *within* the unit and also the greater is the capability of resolving and aligning events in the input speech with events in the sequence of concatenated units. In contrast, a small phone-size unit, representing allophonic variations of phonemes, may be a logical choice for a system based on segmentation and labeling but is clearly a poor choice as a concatenative template because of the well-known boundary phenomenon complexities and complex phonological rules for combining such units.

Given these considerations, a natural choice for a recognition unit to be used in a concatenation-based speech recognition system is the syllable. Syllables are generally considered to contain a vowel nucleus plus optional initial and final consonants or consonant clusters. The syllable thus encompasses both CV (consonant-vowel) and VC (vowel-consonant) transitions, including most coarticulatory and other phonological effects within its boundaries. (Perhaps the best demonstration of the effectiveness of the syllable unit is that quite acceptable synthesis can be achieved by concatenating syllable prototypes using simple concatenation rules [3].) There is also much evidence supporting an important articulatory and perceptual role for the syllable, including syllable-based stress and rhythm phenomena [4]. Syllables have previously been advocated as recognition units [5], and several studies have investigated the possibilities [6]-[8]. In addition, syllables play an intermediate role in other ASR systems [9].

There are some disadvantages to using syllables as recognition units. First of all, there is no general agreement, from a linguistic point of view, as to where syllable boundaries should be located within intervocalic consonants or consonant clusters. This is the so-called syllabification problem. However, in the concatenative template approach to recognition, this problem does not present a significant difficulty since there is no need to explicitly locate syllable boundaries. Any reasonable and consistent assignment of syllable boundaries is sufficient. A more serious difficulty associated with the syllable is the size of the inventory required to represent spoken utterances. It is estimated that spoken English requires some 10 000 syllables [10].

A solution to the inventory size problem can be obtained by constructing syllables from smaller half-syllable units. Each syllable can be considered to be composed of an initial half-syllable containing the initial (optional) consonant cluster and some part of the vowel nucleus plus a final half-syllable containing the remaining portion of the vowel nucleus and the final (optional) consonant cluster. The use of half-syllable units reduces the inventory size required to represent spoken utterances by about a factor of five from that required for whole syllables. However, the use of half-syllables reintroduces some of the boundary complexities we have been seeking to avoid. For example, if the boundary between initial and final half-syllables is made in the center of the vowel nucleus, considerable care must be exercised in matching the vowel on either side of the boundary, especially for diphthongs and vowels followed by liquids and nasals.

This particular difficulty can be essentially eliminated by making use of the concept of the demisyllable as defined by Fujimura and his co-workers originally for use in a high quality concatenative speech synthesis system [11]-[13]. The characteristic that distinguishes demisyllables from other half-syllable units is the location of the boundary between initial and final demisyllables. The initial demisyllable is made quite short, extending just beyond the CV transition. In this way

diphthong transitions and the influence of postvocalic consonants are largely confined to final demisyllables. Locating the boundary in this way is effective in relaxing the matching constraints on either side of the boundary and has the potential of reducing the inventory of initial demisyllables, since the same initial demisyllable might be used to precede a diphthong, an unnasalized vowel, a nasalized vowel, or stressed or unstressed syllables. For example, the initial demisyllable SHAX[1] might be used equally well in the monosyllable *shove*, SHAHV = SHAX + AHV, in the second syllable of *station*, STEYSHAHN = STEH + EYSH + SHAX + AXN, or in *shun*, SHAHN = SHAX + AHN.

Another notion, introduced by Fujimura [14], with a potential for reducing inventory size, is the use of so-called phonetic affixes. Typically these are the apical consonants /*s*, *z*, *t*, *d*, *θ*/ which may be attached independently to final demisyllables with identical postvocalic voicing conditions. The use of affixes reduces the inventory size requirement by an additional factor of 2 so that an inventory of approximately 1000 demisyllables and five affixes is sufficient to provide the needed representations. An example of the use of an affix is constructing the word *tax* by concatenating the units TAE + AEK + S.

There exists another class of units, intermediate between syllables and phone-sized units, which might reasonably be selected as recognition units. There are the units known variously as dyads, diphones, or transemes [15]-[17]. They have in common the concept of incorporating the transition between phonemes and thus represent phoneme pairs. Since much of the coarticulatory phenomena among phonemes is associated with transition regions and since boundaries are more easily defined than for phone-sized units since they occur outside transitions, they represent a more reasonable possibility for concatenative templates than phone-sized units.

We have concluded, however, as it was stated at the beginning of this section, that it is best to select as large a unit as practicable to function as a concatenative recognition template in order to incorporate as many phonological phenomena within the unit as possible while providing stable and well-defined boundaries. For us, the demisyllable seems to represent the best choice according to these criteria.

III. Lexicon and Demisyllable Inventory

In place of the store of parameterized whole word templates of previous versions of our recognizer, there is a lexicon and an inventory of parameterized demisyllable templates. The lexicon contains one or more specifications of each word in the vocabulary in terms of sequences of demisyllable units. For the 1109 word vocabulary used in our tests, a total of 1831 lexical entries (i.e., about 1.6 entries per word) were used. In our experimental evaluation (described in Section V), the input speech consisted solely of isolated words that were compared exhaustively with prototypes corresponding to each whole word specification in the lexicon. Even in future experiments when the input consists of continuous utterances it is expected that the lexicon will continue to play a central role in the recognition process with selective recall of groups of words according to similarities of demisyllable sequences and syntactic constraints.

[1] Phonetic representations in the paper use the ARPABET notation [9].

A. Lexicon Description

Each entry in the lexicon contains three parts, an orthographic transcription of a word, a phonetic transcription, and a demisyllable transcription. The phonetic and demisyllable transcriptions are encoded in ARPABET. There are two types of entries in the lexicon for each word in the vocabulary. There exists at least one entry of the first type, called the primary entry, in which the phonetic transcriptions represent carefully articulated pronunciations of the word obtained by consulting a pronouncing dictionary [18]. There may also exist one or more entries of the second type which represent less carefully articulated versions of primary entries, typically manifested by stress-reduced or deleted syllables. These entries can be generated by rules from primary entries. The rules that are applied are contained in the paper by Oshika *et al.* [19]. The lexicon described in this report was prepared manually, although, more recently, techniques have been implemented for generating secondary entries and demisyllable transcriptions from phonetic transcriptions automatically.

Examples of lexicon entries are shown in Table I. Three levels of syllable stress are specified in the phonetic transcription, primary, secondary and unstressed. The stress marks accompany each vowel, with no mark indicating, by default, primary stress, a "0" indicating unstressed, and "2" indicating secondary stress. The stress marks serve two functions that are described below.

The phonetic transcription is converted to a demisyllable transcription using a set of rules, including rules for syllabification, rules for specifying the vowels in initial and final demisyllables given the stress mark and the syllable vowel, and special rules for handling unstressed syllables and aspirated stops.

1) Rules for Syllabification: The first set of rules that are applied to obtain a demisyllable transcription are syllabification rules. These rules specify the syllable assignment of intervocalic consonants as discussed by Kahn [20]. These are the maximal initial cluster rule and the ambisyllabicity rule, which are applied in order.

a) Maximal Initial Cluster Rule: Assign as many consonants in an intervocalic consonant cluster to the succeeding syllable consistent with the existence of the resulting consonant cluster in word-initial position in English. For example,

attract	AX+TRAEKT
atlas	AET+LAXS.

For atlas, the syllabification AE+TLAXS is not allowed since TL does not exist word initially.

b) Ambisyllabicity Rule: If the preceding syllable ends in a vowel, R, or glide, and the succeeding syllable is unstressed, assign the initial consonant in the consonant cluster to both syllables. For example,

lemon	LEHM+MAXN
distance	DIHS+STAXNS.

TABLE I
EXAMPLES OF ENTRIES IN THE LEXICON. PHONETIC AND DEMISYLLABLE REPRESENTATIONS ARE SPECIFIED IN ARPABET [9]. IN THE PHONETIC TRANSCRIPTIONS, STRESS MARKS MAY FOLLOW EACH VOWEL: 2 = SECONDARY STRESS, 0 = UNSTRESSED; NO MARK INDICATES PRIMARY STRESS (=1).

ORTHOGRAPHIC	PHONETIC	DEMISYLLABLE
brass	BRAES	BRAE+AES
mind	MAYND	MAA+AYND+D
copper	KAAPER0	KAA+AAP+ER
family	FAEMIX0LIX0	FAE+AEM+MIX+LIX+IX
education	EH2JHAX0KEYSHAX0N	EH+EHJH+JHAX +KEH+EYSH+AXN

TABLE II
SPECIFICATION OF INITIAL AND FINAL DEMISYLLABLE VOWELS, V_i AND V_f, GIVEN THE SYLLABLE VOWEL V. THE TOP SECTION OF THE TABLE SHOWS STRESSED VOWELS; THE BOTTOM SECTION SHOWS THE 3 UNSTRESSED VOWELS. WHERE A SECOND CHOICE IS INDICATED FOR V_i, IT IS USED ONLY IF THE DEMISYLLABLE CONTAINING THE FIRST CHOICE VOWEL DOES NOT EXIST IN THE INVENTORY.

V	V_i	V_f
IH	IH	IH
IY	IH	IY
EH	EH	EH
EY	EH	EY
AE	AE	AE
AE+nasal	EH	AE
AA	AA	AA
AY	AA	AY
AW	AA	AW
AW	AE	AW
AH	AH	AH
AO	AO	AO
OW	OW,UH	OW
OY	AO	OY
UH	UH	UH
UW	UW,UH	UW
ER	ER,UH	ER
AX	AX,AH	AX
IX	IX,IH	IX
ER	AX	ER

These rules for syllabification may be violated when a demisyllable transcription cannot be generated because of missing items in the demisyllable inventory.

2) Generating Demisyllables from Syllables: Given a syllable composed of an initial consonant cluster C_i (possibly null), vowel nucleus V, and final consonant cluster C_f (possibly null), the demisyllable representation is given by

$$C_iVC_f \rightarrow C_iV_i + V_fC_f'(+A)$$

where V_i and V_f are functions of V as shown in Table II and A represents one or more possible affixes. In Table II we find that in each case $V_f = V$, since V_f contains most of the vowel portion of the syllable. There are several cases, including all diphthongs, for which $V_i \neq V$, and two cases, for AE followed by a nasal and for AW, where there exist two representations. The second representations were obtained after some experimentation in which it was established that for the two test speakers, in many instances, the second representation provided better recognition scores. In some cases two choices are indicated for V_i. In these cases the second choice is used only if the demisyllable with the first choice vowel does not exist in the inventory.

Affixes are used in two cases. In the first case an affix is appended (almost always at the end of a word) to complete the specification of C_f. That is,

$$C_f \rightarrow C_f' + A.$$

For example,

duct	DAHKT → DAH+AHK+T
length	LEHNXKTH → LEH+EHNXK+TH
round	RAWND → RAA+AWN+D
trousers	TRAWZERZ → TRAA+AWZ+ER+Z.

In the second case, the affixes T and D are used to simulate stop burst releases at the ends of words, as described in Section III-A4).

3) Unstressed Syllable Treatment: There are two sets of unstressed demisyllables in the inventory. One set is associated with the unstressed vowel schwa, /ə/ or AX, as in about, while the second set is associated with the so-called "barred-i", /ɨ/ or IX, as in roses. As will be described in Section III-B, unstressed demisyllables are extracted from unstressed syllables contained in disyllabic source words. Initial unstressed demisyllables are extracted from initial unstressed syllables, while final unstressed demisyllables are extracted from final unstressed syllables. Thus initial and final unstressed demisyllables are extracted from different syllabic environments. Therefore, it does not make the same kind of sense to represent unstressed syllables as a composition of initial and final demisyllables as it does for stressed syllables. For example, to represent the unstressed syllable SAX in the word "solution" we use the initial unstressed demisyllable SAX excised from the source word "supply" but omit the final unstressed demisyllable AX, which is excised from "Cuba," as not necessary and possibly detrimental, since the syllabic environments for the vowel in these demisyllables are distinctly different. In fact, we have found that the recognition results are improved generally when unstressed demisyllables that are not needed to represent consonants are omitted from the representation. The rule that has been applied is to omit unstressed demisyllables, subject to the above constraint, provided that a word beginning with an unstressed syllable is represented by an initial unstressed demisyllable and conversely for final unstressed syllables at the end of a word. The polysyllabic examples in Table I illustrate the application of this rule. For example, for "education" a final demisyllable is not needed following the initial demisyllable JHAX, nor is an initial demisyllable needed before the final demisyllable AXN. In both cases consonants are accounted for in adjacent stressed demisyllables. Note that the final demisyllable in the final syllable for the word "family" is represented by IX. IX generally represents the "barred-i" vowel as in "habit." However, in this case the source for the vowel is the unstressed syllable in "happy."

4) Final Stop Consonant Treatment: When spoken as isolated utterances, words ending in stop consonants are generally accompanied by a strong release. The absence of that release in the recognition prototype can lead to significant deteriora-

TABLE III
EXAMPLES OF SECONDARY ENTRIES IN THE LEXICON. THE FIRST FOUR EXAMPLES SHOW TRANSFORMATIONS BASED ON VOWEL REDUCTIONS OR DELETIONS. THE LAST TWO EXAMPLES SHOW TRANSFORMATIONS BASED ON CONVERSION OF DENTALS TO FLAPS OR GLOTTAL STOPS.

alternate	AOLTER0NAX0T	→ AOLTAX0NAX0T
friday	FRAYDEY2	→ FRAYDIX0
family	FAEMIX0LIX0	→ FAEMLIX0
area	EYRIX0AX0	→ EYRYAX0
automatic	A02TAX0MAETIX0K	→ AO2DXAX0MAEDXIX0K
button	BAHTAX0N	→ BAHQAX0N

tion in recognition performance. Accordingly, the final demisyllable of the final syllable containing the stop should include a strong release following the stop. However, the same stop contained within a word is almost never accompanied by a strong release. Accordingly, all final demisyllables ending in stop consonants have been stripped of strong release in the extraction process from source words. At the end of a word the strong release is simulated by applying the affix D to voiced stops and T to voiceless stops. Both affixes have strong releases. Although logically and preferably, this kind of stop release treatment should have been applied to each kind of stop individually, it was expedient and apparently quite adequate to use the dental stops to simulate all stop releases. These stops already existed as affixes in the inventory. An example of the treatment is the prototype for the word "mind" shown in Table I.

5) Secondary Entries: Secondary entries, representing less carefully articulated pronunciations, are generated from primary entries by applying phonological rules. The transformations, applied to polysyllabic words, are functions of the phonetic transcription including syllable count and syllable stress markings. The rules that have been applied so far include vowel reduction and deletion and transformation of intervocalic T's and D's to flaps, DX, and glottal stops, Q. Some examples of these transformations are shown in Table III. It should be noted that demisyllables with flaps and glottal stops have not been implemented in the inventory. The reason is the following. Currently, all unstressed initial demisyllables are excised from the beginnings of source words, as described in the following section. However, flaps and glottal stops occur intervocalically preceding unstressed syllables. The question arises, which has not been dealt with here, how to excise them reliably from source words. At present, a specification of Q is ignored in a demisyllable transcription and D is substituted for DX.

6) Alternate Entries: There are a few classes of alternate entries that exist for both primary and secondary entries in the lexicon. The distinction between alternate entries and secondary entries is rather arbitrary. At present secondary entries represent additional entries which can be generated by recognized phonological rules associated with rate of articulation. Alternate entries represent all other cases. In most such cases they represent recognized phonological variations that have not been applied systematically or completely, merely pragmatically for the present. For example, it is recognized that in many cases it is not clear whether an unstressed vowel should be represented by AX or IX. We have found it useful to provide alternate entries where every AX in an entry is converted to IX, (but not vice-versa). We have also found it useful, as indicated in Table II, to provide alternate entries for the demisyllable specification of AE+(nasal) and AW. In addition, there are many words in which the vowel might be pronounced either like AA or AO. For example,

wash: WAASH or WAOSH

orange: AORAX0NJH or AARAX0NJH.

Other examples of alternate pronunciations are

length: LEHNXKTH or LEHNTH

exact: IX0KZAEKT or EH2GZAEKT.

It seems likely that these entries are largely associated with dialect characteristics of the talkers. In all likelihood, the lexicon will have to reflect dialect variations, although it is expected that source words from individual talkers will account for a large fraction of such variations.

B. Preparation of Demisyllable Inventories

Following is a description of how the demisyllable template inventory was originally prepared for speech synthesis purposes by Fujimura and his co-workers [11] and then adapted for use in the speech recognition.

Demisyllable prototypes were excised from a corpus of some 1000 recorded source words containing all the required demisyllables as constituents. The source words were recorded in list form by a trained male speaker of General American English. The recordings were made in a sound booth using a high quality microphone. They were then low-pass filtered at 4 kHz and digitized at 10 kHz. Amplitude levels were calibrated and equalized to give roughly uniform subjective loudness across different syllable nuclei. After editing for endpoints, source word utterance files were input to an LPC (linear prediction coefficient) analysis. Using an interactive LPC editing program, providing LPC derived area functions, pitch, amplitude, and a listening facility, demisyllables were excised from the source words. For stressed demisyllables the source words were CVC monosyllables. The consonant complementary to the demisyllable selected for excision was chosen to affect the vowel minimally. Nasals and liquids were never used. Unstressed demisyllables were excised from polysyllabic utterances, so that initial unstressed demisyllables were taken from initial syllables and final unstressed demisyllables were taken from final syllables. The five affixes were excised from source words containing the same environments in which affix templates are used.

The following rule was generally (but not exclusively) used for excision of demisyllables from source words. An initial demisyllable is terminated 60 ms after the start of the CV transition or of onset of voicing. Final demisyllables are defined as the complements of initial demisyllables. Note however that complementary initial and final demisyllables are never taken from the same source words. With excision locations determined in this way, demisyllable prototypes were

extracted in the LPC representation and stored in a demisyllable inventory file.

Since the LPC parameterization employed for recognition differed from that used for synthesis, the entire source corpus was reanalyzed and a new set of LPC demisyllable prototypes was extracted using the location information obtained in the preparation of the synthesis inventory.

Some examples of demisyllables extracted from source words are shown in Fig. 1. These figures show log energy and (LPC-derived) formants plotted as a function of time in frames spaced 15 ms apart for the source words *ked*, *pep*, *fatigue*, *canvas*, and *hips* from which the demisyllables KEH, EHP, FAX, AXS, and the affix S were extracted. The dashed lines show the locations of the excisions that were made based on the rules described above. Note that most of the release following the P in EHP has been stripped from the prototype, as explained in Section III-A4). Note also that the vocalic portions of the unstressed syllables are much shorter than the vocalic portions of the stressed syllables, and consequently, that the initial unstressed demisyllable is a much larger fraction of the syllable than the corresponding stressed initial demisyllable. This represents another reason why unstressed syllables are well represented by single initial or final demisyllables.

Demisyllable inventories for two additional talkers have been obtained using a new bootstrapping technique which extracts demisyllables from the source words of a new talker given the excised demisyllable prototypes of an old talker [21]. The technique makes use of the dynamic programming time alignment facility of the recognition system, to be described in Section IV, in which an optimal time alignment is obtained between an old talker's source word and a new talker's source word. The projection of the demisyllable boundaries for the old talker onto the new talker's time axis through the optimal alignment path provides estimates of the new talker's demisyllable boundaries. The idea of mapping an event contained in a reference (old) or known utterance onto the time axis of a test (new), or unknown utterance is illustrated in Fig. 2 and explained in Section IV. Iterating this technique through the entire corpus and making manual corrections, if needed, provides a complete new set of demisyllable prototypes.

At present, there are 946 prototypes in the demisyllable inventory, consisting of 395 stressed initial units, 445 stressed final units, 60 unstressed initial units, 41 unstressed final units, and the 5 affixes. There are actually 969 entries in the demisyllable catalog, but these include 23 pairs of entries, each of which points to the same prototype. These consist of stressed final demisyllables with rhotacized (/r/ colored) vowels for which the following pairs of vowels are equated: EH and EY, IH and IY, UH and UW, and AO and OW. This inventory is sufficient to represent all the items in our current experimental 1109-word vocabulary with the exception of entries which require flaps or glottal stops, as explained in Section III-A5). In addition, there are some nonessential demisyllables missing from the inventory, for which reasonable substitutions can be made as explained in Section III-A2) and shown in Table II.

C. Storage Requirements

For a practical implementation of a demisyllable-based word recognizer, storage would have to be provided for the LPC representations of the demisyllable inventory and for the dictionary representations of each vocabulary word. The storage for the demisyllable inventory is fixed (independent of the size of the word vocabulary) and requires about 20 000 LPC frames (i.e., an average of about 20 frames per demisyllable), or about 2×10^6 bits of storage. The storage for each dictionary representation is about 12 bytes (96 bits). Hence the storage required for dictionary representations of words is negligible compared to the storage required for the demisyllable patterns for vocabularies on the order of 1000–2000 words.

By way of contrast with a conventional LPC word based recognizer, the storage required for such a system is about 4000 bits for each word pattern. Hence a vocabulary of about 500 words would have the same storage requirements as that of the demisyllable set itself. For the 1109 word vocabulary which we are considering here, the storage requirements of the LPC based word recognizer are more than twice as large as those of the demisyllable-based recognizer.

IV. Speech Recognition System Description

Fig. 3 shows a block diagram of the basic speech recognition process. The front-end processing was first introduced by Itakura and is explained in detail elsewhere [1], [22]. The input signal is digitized at a 6.67 kHz sampling rate (compatible with telephone bandwidth speech), word endpoints are detected using an algorithm based on acoustic energy considerations, and the parameterization, an eighth-order LPC analysis, is carried out over 45 ms frames shifted every 15 ms.

The comparison between a parameterized input utterance and a reference template involves simultaneous time alignment and measurement of pattern similarity. This is accomplished by means of a dynamic programming time warping process. A local distance or measure of dissimilarity is calculated between each input frame and each reference frame within a specified range. Using these local distances the dynamic programming algorithm provides the optimum sequence of reference frames that accumulates the least amount of distance from beginning to end of the utterance. This sequence is a path that specifies a nonlinear time alignment of the reference pattern to the input utterance, and the accumulated distance along this path is an overall measure of dissimilarity between the patterns and is used as a recognition score. An example of such a path is shown in Fig. 2 as the solid line linking the heavy dots. Constraints are imposed to ensure that the path satisfies some natural conditions. These include a specification of the overall endpoints of the path, and a local continuity and monotonicity constraint which specifies that the slope of the path may not be greater than 2 or less than $\frac{1}{2}$ (indicated by the three short path segments leading to a point in the figure). The constraints taken together imply the range shown enclosed by the parallelogram in which paths may be located.

The outcome of a recognition trial is a set of overall distances or recognition scores corresponding to the reference

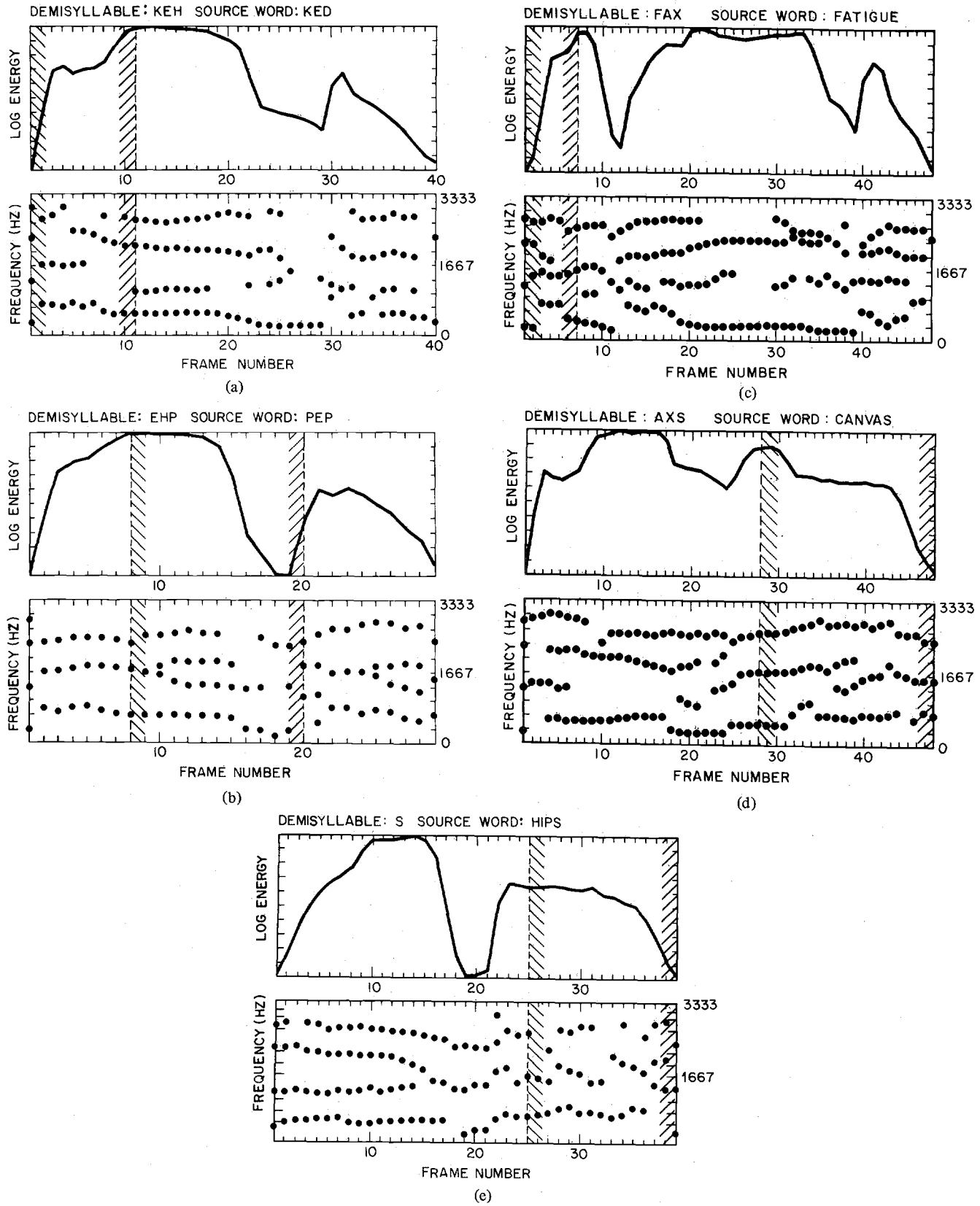


Fig. 1. Five examples of demisyllables extracted from source words. In each part of the figure the top panel shows log energy and the bottom panel (LPC-derived) formants plotted as a function of time in frames. The cross-hatched areas indicate the frames that were excised for each demisyllable from the source words.

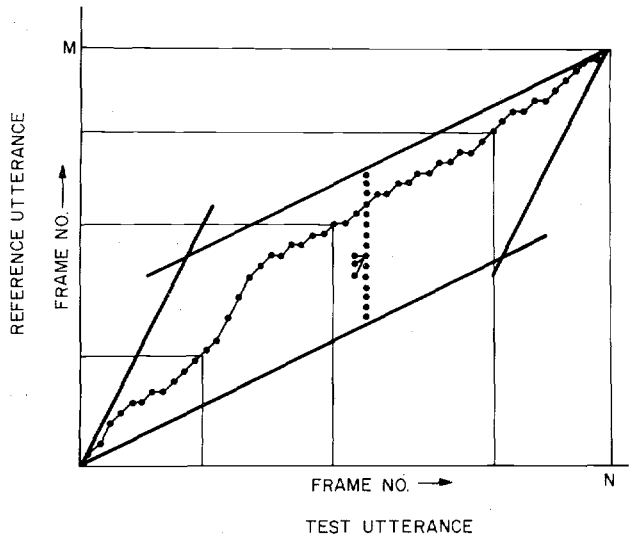


Fig. 2. Example of dynamic programming time alignment of a test utterance to a reference utterance. Each dot represents a reference-test pair of frames for which a local distance is calculated. The solid line connecting dots represents an optimal time alignment path accumulating the least amount of distance from beginning to end of the test utterance. The vertical column of dots represents the distance calculations that are carried out for a particular test frame. Connected to one of these dots are three dots representing three reference frames paired with the preceding test frame. The connections represent the allowable path segments along which distance is allowed to accumulate. The parallelogram encloses the overall range in which distances are calculated and paths are located as implied by the local path constraints and the endpoints. The light horizontal and vertical lines illustrate how events in the reference utterance (for example, word boundaries) can be mapped onto the test utterance via the time alignment path.

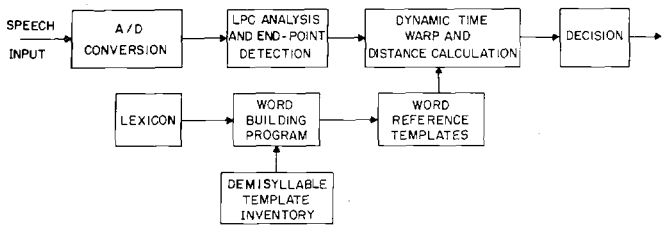


Fig. 3. Block diagram of the demisyllable based word recognizer.

prototypes selected for comparison with the input utterance. A decision rule is applied to these scores to estimate the spoken utterance.

As we have already indicated in Section III, the store of reference templates containing whole-word parameterized prototypes used in previous implementations is replaced by a store of demisyllable prototypes and a lexicon that provides specifications for constructing whole-word prototypes from concatenated demisyllables.

The recognition system that has been described is based on template matching, in contrast to techniques based on explicit segmentation and labelling of speech sounds. This template approach is extended to speech input consisting of connected words by matching strings of concatenated word templates to input utterances. This technique is suggested in Fig. 2 where we suppose that the horizontal lines intersecting the y-axis mark the locations of boundaries between word templates making up a reference string. It is shown how these boundaries can be projected from the time alignment path onto the input axis, providing estimates of boundaries between the words in the input utterance. Such estimates can be used to build up possible word string matches successively as in the level building technique of Myers and Rabiner [23]. Similarly, under the conditions described in Section II, we hope that when the recognition units are demisyllables, this capability can be used successively to build up demisyllable string matches forming words and, in turn, word strings forming phrases and sentences when the input is a continuously spoken utterance.

An important point to note about the dynamic programming time alignment technique is that the greatest range in which paths can be located is obtained when the lengths of the input and reference patterns are equal. The range becomes progressively smaller as the length ratio approaches 2 or $\frac{1}{2}$, thereby reducing the chance of finding the correct alignment between the input and reference. For optimum recognition performance therefore, it is desirable that the duration of input and reference templates be comparable.

The word-building block shown in Fig. 3 prepares reference word prototypes by concatenating demisyllable templates according to the specifications in the lexicon, in the course of which, two kinds of adjustments are carried out.

In the first adjustment, boundaries between demisyllable templates are smoothed by applying a least-squares fit to log area ratios transformed from the LPC parameters across four frames centered at the boundary. In a preliminary set of experiments [24] it was found that 4-frame smoothing produced a small but significant improvement in recognition performance, compared with 2-frame smoothing or no smoothing at all.

The second kind of adjustment, namely adjustment of demisyllable durations, is required because of large discrepancies that often occur between the durations of polysyllabic words and the demisyllable template strings representing them. This condition arises because stressed demisyllables are extracted from monosyllabic words, and syllable durations in monosyllabic words are generally considerably longer than the durations of non-final stressed syllables in polysyllabic words. Although, as mentioned earlier, the dynamic programming time alignment facility can make duration adjustments for overall length ratios between $\frac{1}{2}$ and 2, optimal alignments can be expected only when overall durations are comparable. Therefore, stressed demisyllables are adjusted by applying linear compression in an attempt to make the durations of polysyllabic reference word prototypes comparable to the durations of natural utterances. The compression factor is a variable in the experiments reported in this paper and will be described and discussed in the following sections. In addition, more sophisticated duration adjustments have been investigated and described by Kahn *et al.* [25].

V. Experimental Evaluation

An experimental evaluation was carried out to determine the usefulness of demisyllable representation for isolated word recognition. The vocabulary used for the test was the 1109-

TABLE IV
100-WORD SUBSET OF THE 1109-WORD "BASIC ENGLISH" VOCABULARY

back	page	wood	reason
bite	past	year	respect
bread	play	attempt	science
care	prose	brother	statement
change	range	canvas	sugar
cork	rice	copper	system
crush	rub	cover	transport
death	sand	danger	vessel
dust	self	desire	woman
fact	shake	distance	comfort
fight	sign	event	servant
fold	sleep	feeling	amusement
friend	smell	hearing	attention
gulp	snow	increase	company
group	sort	journey	destruction
heat	start	leather	division
hour	stitch	linen	family
judge	swim	measure	industry
land	test	minute	invention
lift	tin	motion	ornament
loss	trick	number	quality
mass	verse	paper	relation
mind	war	poison	selection
move	wax	power	comparison
night	wind	protest	education

word basic English vocabulary of Ogden [26]. A 100-word subset of the vocabulary, which was used in a series of preliminary experiments [24], is shown in Table IV. This subset is representative of the entire vocabulary, containing approximately the same proportion of one-, two-, three-, four-syllable words. The entire vocabulary consists of 605 monosyllabic words, 317 two-syllable words, 125 three-syllable words, 50 four-syllable words, and 12 five-syllable words. A histogram of the syllable distribution is shown in Fig. 4.

Results are presented for two male talkers, both of whom are experienced participants in speech recognition tests. Each talker provided utterances of both the test vocabulary and demisyllable source words. It required approximately ten 1-h sessions to record the source words for each talker and four 1-h sessions to record the test words. The utterances were obtained with the talkers seated in a sound booth using an ordinary telephone handset. They were transmitted over dialed-up local telephone lines and digitized directly. The talkers were instructed to articulate naturally and clearly. One talker, LR, provided two test sets of the vocabulary, while the second talker, JW, provided just one. JW's test set and LR's first test set were recorded approximately two weeks after the source words. The second test set of LR was recorded approximately 6 months after the first. Demisyllable templates were established for each talker by the bootstrapping technique described in Section III-B with manual corrections applied where required.

Standard recognition experiments were carried out for each talker in which each test word was compared exhaustively with every reference word prototype specified by the lexicon. For each trial, the recognition scores were ordered, and the frequency for which test words were found among the n best scoring reference prototypes was tabulated. Overall performance is generally specified as the n-best-candidate error rate, which is the average frequency over the 1109 words in the vocabulary for which the test word is not found among the n best scoring reference word prototypes. For n equal to 1, this figure represents the conventional recognition error rate.

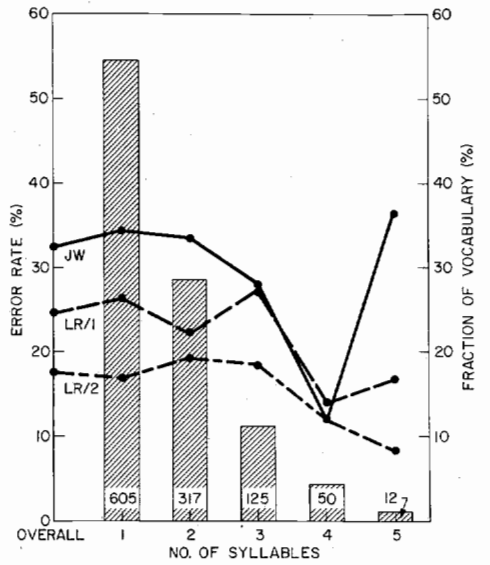


Fig. 4. Histogram of vocabulary words and mean word recognition error rates (best candidates) as a function of number of syllables. The cross-hatched columns represent the histogram with the actual number of words for each number of syllables at the bottom of each column. Error rates for three test sets are shown, one for talker JW, and two for talker LR. Overall mean error rates are shown on the left-hand axis.

Overall recognition results are shown in Fig. 5. For talker JW in Fig. 5(a), two plots are shown, namely, mean recognition error rate as a function of n best candidates for demisyllable prototypes and, for comparison, mean error rate for whole word prototypes, for the same vocabulary using a robust training procedure [27]. For demisyllables, the recognition error rate for n equal to 1 is rather poor at approximately 33 percent but improves considerably with increasing n to approximately 6 percent for n equal to 10. The error rate for whole word prototypes is approximately half the rate for demisyllable prototypes. The results suggest, however, that the difference in error rates decreases considerably with increasing n. It should be noted that the training procedures used for demisyllable and whole word prototypes are not comparable. For demisyllables, the conventional procedure was used, in which a single token is used to establish each prototype. In the robust training procedure used for whole-word prototypes, multiple tokens are obtained for each word until two tokens are similar enough to establish a representative prototype. In this way there is greater assurance of obtaining valid prototypes. For talker LR, two sets of results are shown in Fig. 5(b) for demisyllable prototypes, associated with two test sets. For n equal to 1 error rates of approximately 25 percent and 18 percent are obtained, again, as with talker JW, improving considerably with increasing n to approximately 2 percent for n equal to 10. There is a significant difference in performance between the two test sets for LR. The uniformly poorer performance for TS2 may be accounted for by the fact that TS2 was recorded under slightly different recording conditions and six months after both the recording of the first test set and the source words. For comparison, results are

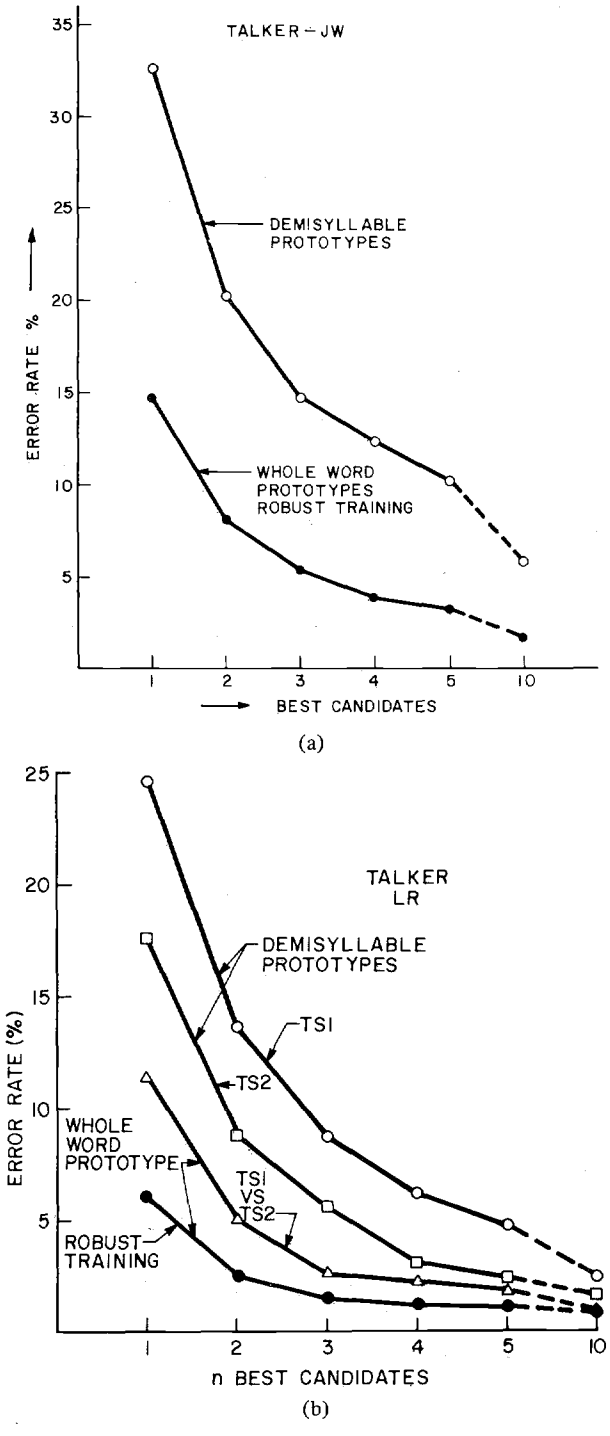


Fig. 5. Mean recognition error rate as a function of number of best candidates n. (a) Results for talker JW. (b) Results for talker LR. Intermediate values of n from 6–9 are omitted. Results with whole-word prototypes are shown along with those for demisyllable prototypes for comparison. Results for two test sets and two kinds of whole-word training procedures are shown.

shown for two whole word prototype experiments. For prototypes established by robust training, the average error rate is 6 percent for n equal to 1 decreasing to less than 1 percent for n equal to 10. In the second set of whole word results, one of the test sets used in the demisyllable evaluation is compared against the other set now serving as whole word prototypes. This is a fairer and more immediate comparison with the demisyllable results. The error rates obtained are intermediate between those obtained for demisyllable prototypes and whole word prototypes using robust training. Again it appears that differences between the whole-word and demisyllable results decrease considerably with increasing n.

The reason for providing recognition error rates as a function of n best candidates is that in almost every implementation of a large vocabulary recognizer higher level constraints are invoked to choose among the best candidates [1]. Performance for such systems is therefore related to the frequency of occurrence of the correct word among the best candidates. Selecting the best candidates can be accomplished by ordering the candidates by their recognition scores and fixing a value of n or by applying a threshold function directly on the recognition scores.

It is useful to examine statistics based directly on the recognition scores since such information can provide more insight into the confusability between the correct word and competing candidates than the candidate order itself. Statistics have been calculated on the difference between the best two candidate recognition scores when the best candidate is correct, Δ_T, when the best candidate is incorrect, Δ_F, and on the difference between the best candidate and the correct candidate when the best candidate is incorrect, Δ_c. Normalizing by subtracting the best candidate recognition score reduces the variance since the best score tends to represent a bias on all the scores associated with a given trial. The medians of these differences are shown plotted in Fig. 6 as a function of the best candidate error rate for JW and the two test sets of LR. (Medians are shown rather than means because of the skewed nature of distance distributions. Each mean is somewhat greater than each median.) It should be noted that the standard deviations of these differences are quite large, only slightly less in each case than the means. It is apparent, however, that on the average, the difference between the top two candidates is 3 times greater when the best candidate is correct, Δ_T, than when it is incorrect, Δ_F. The difference between the best candidate and the correct candidate takes on intermediate values. It is clear that the best candidate, when it is correct, is well separated from other candidates. On the other hand, when the best candidate is incorrect the scores are grouped closer together and the separation of the correct word from the best candidate, Δ_c, is generally smaller than Δ_T. It should therefore be possible to set a recognition score difference threshold that would include the correct word among the candidates with high probability but admit on the average only a few candidates. The data also suggest, albeit tentatively, that Δ_T, the mean difference between the best two candidates when the best candidate is correct, is monotonically related to the recognition error rate, while the other differences remain more or less constant.

There are two experimental variables which are interesting to examine for their effect on error rate. These are the number of syllables and the compression factor used in the adjustment of demisyllable durations discussed in Section IV.

Recognition error rate–best candidate is shown plotted in Fig. 4 as a function of number of syllables along with the histograms showing the distribution of words in the vocabulary

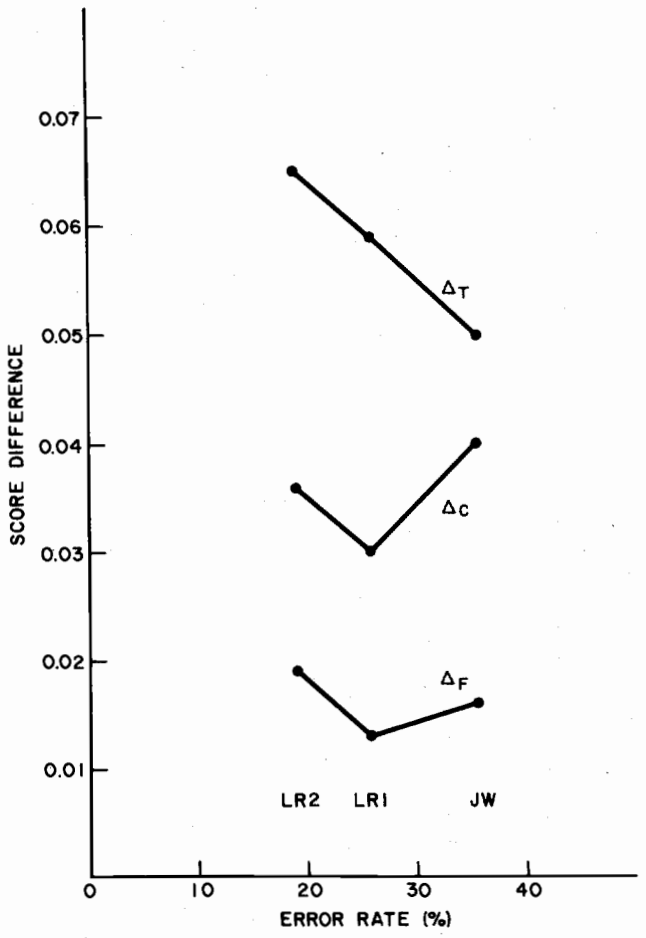


Fig. 6. Three kinds of median difference recognition scores for three test sets plotted as a function of the best candidate error rate for each test set. Δ_T is the difference between the two best candidate recognition scores when the best candidate is correct; Δ_F is the difference between the two best candidates recognition scores when the best candidate is incorrect; Δ_C is the difference between the best candidate and the correct candidate when the best candidate is incorrect.

as a function of number of syllables. Three plots are shown, corresponding to the test set for talker JW and the two tests of LR. Some improvement in performance is suggested with increasing number of syllables, but the effect is not nearly as marked as what has been observed previously for whole word prototypes. For example, in the experiment previously cited for whole word prototypes for this vocabulary [27], talker JW experienced an improvement of 20.8 percent to 6.1 percent error rate from monosyllables to polysyllables and LR 10.0 percent to 2.0 percent. The improvement in recognition performance with increasing number of syllables is generally attributed to the idea that information is fairly independent from syllable to syllable in a given word, making confusion among prototypes less likely. It is not surprising, however, to see this performance advantage decline with the use of concatenated demisyllable prototypes. Although this paper shows that demisyllable sequences provide a reasonable representation for words, there is certainly an element of artificiality in the concatenation of demisyllables. This artificiality with its attendant uncertainties and complexities at boundaries can only increase as more demisyllables are concatenated to represent polysyllabic words thereby impairing performance.

The second experimental variable of interest is the compression factor used to adjust demisyllable durations. In Section IV we noted the occurrence of large discrepancies between the durations of polysyllabic test words and prototypes of concatenated demisyllables representing these words. We also noted that optimum recognition performance is attained only when the durations of prototypes and test words are comparable. We attributed the discrepancies to the fact that stressed demisyllables are extracted from monosyllables whose durations are longer than stressed syllable durations in polysyllabic words. This is the so-called duration constancy phenomenon [28], [29], in which it has been shown, at least for words spoken in isolation or in carrier phrases, that overall duration tends to be maintained as the number of syllables in a word increases by shortening the duration of stressed vowels. Umeda [29] found that stressed vowel duration decreases approximately 35 percent from one- to two-syllable words with progressively smaller decreases for additional syllables.

In the preliminary experiments that were carried out, durations were adjusted by linearly compressing the final demisyllables of non-final stressed syllables by a constant factor of 50 percent. Exempting final syllables from this adjustment agrees with Umeda's observation [30] that the durations of stressed vowels in the last syllable before a pause do not differ from durations in monosyllabic words. The 50 percent compression factor produced satisfactory results in the preliminary experiments and has been used in all the results cited so far.

The effect on error rate of varying the compression factor has been investigated more extensively in this experiment over the entire vocabulary. The results are plotted in Fig. 7 as error rate (best candidate) for talker LR's test set 1. We see that optimum performance is obtained for compression factor approximately equal to 0.45, and that performance is relatively insensitive to the compression factors for values between 0.3 and 0.7.

It is worth noting that, in the preliminary experiments, in addition to the adjustment of stressed syllables, the overall durations of word prototypes were linearly adjusted to the (known) test word durations. This was carried out to insure optimal overall performance for the DTW process given the uncertainty regarding the effectiveness of constant-factor syllable compression in achieving both local and overall alignment. We found (in a separate experiment carried out over the entire vocabulary) that this overall linear prenormalization yielded at most a 1 percent improvement in (best candidate) error rate. We generally obtained satisfactory overall alignments using only the 50 percent compression of stressed syllable durations, as described above.

The question still remains as to whether more sophisticated duration adjustments can effect even better alignments and performance. This question has been investigated by Kahn *et al.* [26] who made adjustments as a function of stress and syllable patterns. The overall conclusion was that sophisticated adjustments did not bring about any significant improvement over the simple constant compression factor adjustment.

We now discuss the effectiveness of adding secondary and

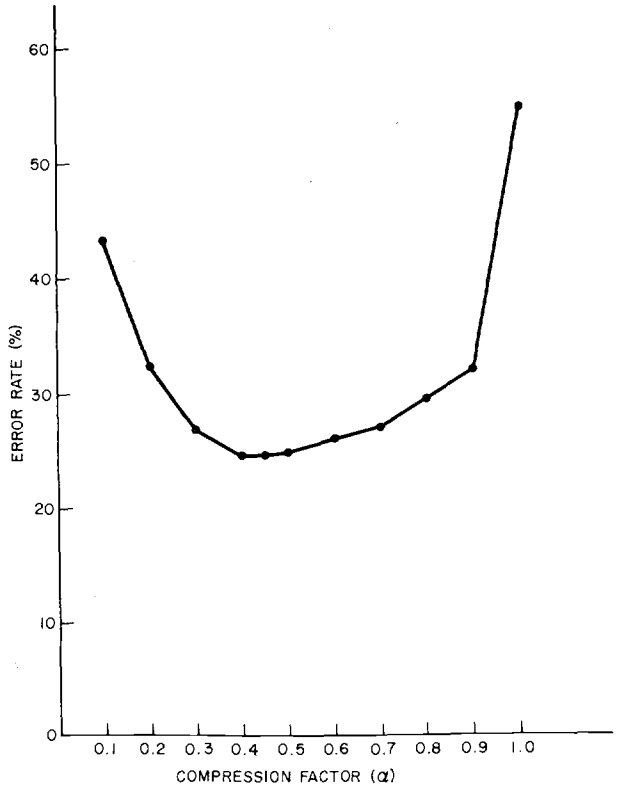


Fig. 7. Mean recognition error rate (best candidate) for LR test set 1 as a function of demisyllable duration compression factor α.

TABLE V

RECOGNITION RESULTS FOR FIVE CATEGORIES OF SECONDARY AND ALTERNATE LEXICON ENTRIES. FOR EACH CATEGORY THE NUMBER OF ENTRIES IS SHOWN, ALONG WITH THE NUMBER OF VOCABULARY WORDS THESE ENTRIES REPRESENT, THE NUMBER OF THESE WORDS FOR EACH OF THE THREE TEST SETS WHOSE BEST SCORES ARE FOUND AMONG THESE ENTRIES, AND THE IMPROVEMENT IN ERROR RATE ATTRIBUTABLE TO THESE ENTRIES FOR TALKER LR'S TEST SET 1. THE FIRST TWO CATEGORIES ARE FOR SECONDARY ENTRIES GENERATED BY PHONOLOGICAL RULES FOR SYLLABLE STRESS REDUCTION. THE LAST THREE CATEGORIES REPRESENT ALTERNATE ENTRIES. THE FIRST OF THESE CONTAINS ENTRIES IN WHICH THE VOWEL, V = AW, IS REPRESENTED BY THE COMBINATION OF INITIAL AND FINAL DEMISYLLABLE VOWELS, V_i = AE AND V_f = AW. (IN THE PRIMARY ENTRIES FOR V = AW, V_i = AA). THE SECOND CATEGORY CONTAINS ENTRIES WHERE THE VOWEL V = AE FOLLOWED BY A NASAL CONSONANT IS REPRESENTED BY V_i = EH (INSTEAD OF V_i = AE IN THE PRIMARY ENTRIES). THE THIRD CATEGORY CONTAINS ENTRIES IN THE VOWEL AA IS SUBSTITUTED FOR AO AND VICE-VERSA.

Partition	number of entries	number of words	number of words with best scores in partition			decrease in error rate (%) for LR1
			JW	LR1	LR2	
reduced/deleted vowels	218	125	55	69	74	9.5
reduced dentals	140	63	44	25	32	6.7
AW → AE+AW	37	30	22	26	26	1.3
AE+nasal→ EH+AE+nasal	22	17	4	11	9	0.9
AO↔AA	20	12	8	9	8	0.9

alternate entries to the lexicon. It should be emphasized again that the designation of many entries as "alternate" is arbitrary. These are entries that were added to the lexicon after it was found that improved recognition scores could be obtained for these words by modifying the phonetic or demisyllable representations. Although these modifications are quite reasonable and recognized, the approach is ad hoc to the extent that they were made with reference to the two talkers in the experiment. However, we avoided any representations which could not reasonably be expected as dialect variations of pronunciation. The examples given in III-A6) are a fair representation of words in this category. There is no reason to believe that the representations are universal or complete. Secondary entries have a different status since they are generated strictly by rule from primary or alternate entries.

Results are shown in Table V for five categories of secondary and alternate entries. The partition associated with the most significant improvement in performance is labeled as "reduced/deleted vowels." This includes all the transformations of polysyllabic words by phonological rules that result in reduced or deleted vowels. There are 218 such entries representing 125 words in the lexicon. The next three columns show the number of words, for each of the three experimental test sets, whose best recognition scores are obtained with entries in the partition. The last column shows the improvement in (best candidate) error rate for test set LR1 attributable to this partition. This figure is obtained by repeating the recognition experiment with all the entries in this partition deleted and noting the increase in error rate. The second partition includes all those entries in which /t/'s and /d/'s are reduced by phonological rules to flaps and glottal stops (in addition to possible reductions or deletions of vowels.) Again, significant improvements are obtained. Recall that in this implementation flaps are represented by D's while for glottal stops, the consonant is simply omitted. It is not surprising that representating glottal stops by omitting the consonant is ineffective. With but one exception (for "mountain"), all the improvements in this partition were obtained for flaps.

The last three partitions shown in the table are for alternate entries. These contain smaller numbers of entries and words than secondary entries, but provide significant improvements within these partitions.

VI. CONCLUSIONS

We conclude that good recognition performance can be obtained for isolated word utterances from a system based on concatenated demisyllable prototypes. The performance is inferior to recognition based on whole word prototypes but the deficiency improves as more candidate prototypes are admitted.

The use of demisyllable representations has significant advantages as vocabulary size increases. The training, processing, and storage of prototypes is fixed at approximately 1000, regardless of the number of words in the vocabulary. A storage with about a quarter million bytes is required for these demisyllable prototypes. However, the demisyllable specification of a word requires some 12 bytes compared to 800 bytes to store the LPC parameterized prototype. Hence, for vocabularies of 500 words, or more, the demisyllable-based system requires less storage than the word based system. Also, the

use of demisyllable units enables us to maintain the basic characteristics of our recognition process, including uniform LPC parameterization, dynamic time warping alignment, and especially concatenative template representation of utterances. The concatenative template approach is relatively simple, in a practical engineering sense compared to the segmentation and labeling approach to recognition. However, important questions remain as to how far this approach can be extended as spoken utterances become less restricted and more natural. A central question is whether demisyllable prototypes extracted from words spoken in isolation can adequately represent all speech sounds in natural discourse. We feel encouraged in this regard since we have shown that simple duration adjustments and straightforward phonological transformations represented in the lexicon provide good recognition performance for polysyllabic words. Some simple experiments with continuous spoken utterances are planned to explore and resolve these questions.

Acknowledgment

The authors wish to express their appreciation to O. Fujimura, C. P. Browman, M. J. Macchi, and E. L. Ohira for their help and cooperation in assembling and transferring the original demisyllable database and to S. E. Levinson as well as the preceding for their valuable comments and suggestions and to K. L. Shipley for programming help.

References

[1] L. R. Rabiner and S. E. Levinson, "Isolated and connected word recognition–Theory and selected applications," *IEEE Trans. Commun.*, vol. COM-29, pp. 621–659, 1981.

[2] C. S. Myers and S. E. Levinson, "Speaker independent connected word recognition using a syntax directed dynamic programming procedure," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 561–565, 1982.

[3] O. Fujimura and J. B. Lovins, "Syllables as concatenative phonetic units," in *Syllables and Segments*, A. Bell and J. B. Hooper, Eds. New York: North-Holland, 1978, pp. 107–120.

[4] I. R. A. MacKay, *Introducing Practical Phonetics*. Boston, MA: Little, Brown, 1978.

[5] O. Fujimura, "Syllable as a unit of speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 82–87, 1975.

[6] P. Mermelstein, "A phonetic-context controlled strategy for segmentation and phonetic labelling of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 79–82, 1975.

[7] M. J. Hunt, M. Lennig, and P. Mermelstein, "Experiments in syllable-based recognition of continuous speech," in *Proc. 1980 IEEE Int. Conf. Acoust., Speech, Signal Processing*, Denver, CO, 1980, pp. 880–883.

[8] G. Ruske and T. Schotola, "An approach to speech recognition using syllable decision units," in *Proc. 1978 IEEE Int. Conf. Acoust., Speech, Signal Processing*, Tulsa, OK, 1978, pp. 722–725.

[9] J. E. Shoup, "Phonological aspects of speech recognition," in *Trends in Speech Recognition*, W. G. Lea, Ed. Englewood Cliffs, NJ: Prentice-Hall, 1980, ch. 6, pp. 125–138.

[10] H. M. Moser, *One-Syllable Words*. Columbus, OH: Merrill, 1969.

[11] O. Fujimura, M. J. Macchi, and J. B. Lovins, "Demisyllables and affixes for speech synthesis," in *Proc. 9th Int. Congr. Acoust.*, 1977, pp. 51–53.

[12] J. B. Lovins, M. J. Macchi, and O. Fujimura, "A demisyllable inventory for speech synthesis," in *Dig. Speech Commun. Papers, 97th Meeting Acoust. Soc. Amer.*, J. J. Wolf and D. H. Klatt, Eds., 1979, pp. 523–526.

[13] C. P. Browman, "Rules for demisyllable synthesis using LINGUA: A language interpreter," in *Proc. 1980 IEEE Int. Conf. Acoust., Speech, Signal Processing*, Denver, CO, 1980, pp. 561–564.

[14] O. Fujimura, "Syllables as concatenated demisyllables and affixes" (abstract), *J. Acoust. Soc. Amer.*, vol. 59, suppl. 1, p. S55, 1976.

[15] G. E. Peterson, W. S.-Y. Wang, and E. Sivertsen, "Segmentation techniques in speech synthesis," *J. Acoust. Soc. Amer.*, vol. 30, pp. 739–742, 1953.

[16] N. R. Dixon and H. D. Maxey, "Terminal analog synthesis of speech using the diphone method of segment assembly," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 40–50, 1968.

[17] N. R. Dixon and H. E. Silverman, "The 1976 modular acoustic processor (MAP)," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 376–378, 1977.

[18] W. Morris, Ed., *The American Heritage Dictionary of the English Language*. Boston, MA: Houghton-Mifflin, 1978.

[19] B. T. Oshika, V. W. Zue, R. V. Weeks, H. Neu, and J. Aurbach, "The role of phonological rules in speech understanding research," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 104–112, 1975.

[20] D. Kahn, *Syllable-Based Generalizations in English Phonology*. New York: Garland, 1980.

[21] L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, and T. M. Zampini, "A bootstrapping training technique for obtaining demisyllable reference patterns," *J. Acoust. Soc. Amer.*, vol. 71, pp. 1588–1595, 1982.

[22] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, pp. 67–72, 1975.

[23] C. S. Myers and L. R. Rabiner, "Connected digit recognition using a level building DTW algorithm," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 351–363, 1981.

[24] A. E. Rosenberg, L. R. Rabiner, S. E. Levinson, and J. G. Wilpon, "A preliminary study on the use of demisyllables in automatic speech recognition," in *Proc. 1981 IEEE Int. Conf. Acoust., Speech, Signal Processing*, Atlanta, GA, 1981, pp. 967–970.

[25] D. Kahn, L. R. Rabiner, and A. E. Rosenberg, "On duration rules in a demisyllable-based isolated word recognition system" (abstract), *J. Acoust. Soc. Amer.*, suppl. 1, vol. 70, p. 560, 1981; *J. Acoust. Soc. Amer.*, to be published.

[26] C. K. Ogden, *Basic English: International Second Language*. New York: Harcourt Brace, and World, 1968.

[27] L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, and W. J. Keilin, "Isolated word recognition for large vocabularies," *Bell Syst. Tech. J.*, vol. 61, pp. 2989–3005, 1982.

[28] I. Lehiste, *Suprasegmentals*. Cambridge, MA: M.I.T. Press, 1970, p. 40.

[29] N. Umeda, "Effects of speaking mode on temporal factors in speech: Vowel durations," *J. Acoust. Soc. Amer.*, vol. 56, pp. 1016–1018, 1974.

[30] —, "Vowel duration in American English," *J. Acoust. Soc. Amer.*, vol. 58, pp. 434–445, 1975.

Aaron E. Rosenberg (S'57–M'63) received the S.B. and S.M. degrees in electrical engineering from Massachusetts Institute of Technology, Cambridge, in 1960 and the Ph.D. degree in electrical engineering from the University of Pennsylvania, Philadelphia, in 1964. He has been a member of the Technical Staff in the Acoustical Research Department, Bell Laboratories, Murray Hill, NJ, since 1964, where his research interests have included auditory psychophysics, speech perception, and currently, speech and speaker recognition.

Dr. Rosenberg is a member of Sigma Xi and a Fellow of the Acoustical Society of America. He is currently a member of the ASSP Society Administrative Committee, the ASSP Conference Board, and Associate Editor for Speech Communication for this TRANSACTIONS.

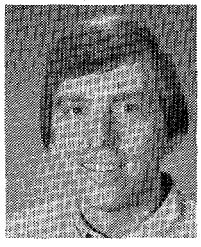
Lawrence R. Rabiner (S'62-M'67-SM'75-F'75) was born in Brooklyn, NY, on September 28, 1943. He received the S.B. and S.M. degrees simultaneously in 1964, and the Ph.D. degree in electrical engineering in 1967, all from the Massachusetts Institute of Technology, Cambridge.

From 1962 to 1964 he participated in the cooperative plan in electrical engineering at Bell Laboratories, Whippany and Murray Hill, NJ. He worked on digital circuitry, military communications problems, and problems in binaural hearing. Presently he is engaged in research on speech recognition and digital signal processing techniques at Bell Laboratories, Murray Hill, NJ. He is coauthor of the books *Theory and Application of Digital Signal Processing* (Englewood Cliffs, NJ: Prentice-Hall, 1975), *Digital Processing of Speech Signals* (Englewood Cliffs, NJ: Prentice-Hall, 1978), and *Multirate Digital Signal Processing* (Englewood Cliffs, NJ: Prentice-Hall, 1983).

Dr. Rabiner is a member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and a Fellow of the Acoustical Society of America. He is a former President of the IEEE S-ASSP AdCom, and is currently a member of the S-ASSP Technical Committee on Digital Signal Processing.

Jay G. Wilpon was born in Newark, NJ, on February 28, 1955. He received the B.S. degree in mathematics and the A.B. degree in economics (cum laude) from Lafayette College, Easton, PA, in 1977 and the M.S. degree in electrical engineering/computer science from Stevens Institute of Technology, Hoboken, NJ, in 1982.

Since June 1977, he has been with the Acoustics Research Department, Bell Laboratories, Murray Hill, NJ, where he is a member of the Technical Staff. He has been engaged in speech communications research and is presently concentrating on problems of speech recognition.



Daniel Kahn (M'81) received the B.S. degree in physics in 1968 and the Ph.D. degree in linguistics from the Massachusetts Institute of Technology, Cambridge, MA, in 1976.

From 1969 to 1971 he was a secondary school teacher of mathematics. In 1976 he worked as a Research Linguist at the University of California, Los Angeles, and from 1977 to 1979 was an Assistant Professor of Linguistics at the University of Massachusetts, Amherst. Since 1977 he has worked on problems in speech synthesis, speech recognition, and theoretical linguistics in the Acoustics Research Department and the Linguistics and Speech Analysis Department, Bell Laboratories, Murray Hill, NJ.

Dr. Kahn is a member of Phi Beta Kappa and the Acoustical Society of America.

Correspondence

Digital Integrators Using Optimal FIR Compensators

A. H. M. ABED, P. J. BLOOM, AND G. D. CAIN

Abstract—**An important class of digital integration algorithms is examined. The complete integration filter is partitioned into a feedback filter combined with a finite duration impulse response (FIR) part which may be thought of as a cascaded compensator subfilter. New integrator designs are obtained using linear programming to achieve minimax optimization of the linear-phase FIR compensator, constrained to have no dc error. Comparisons with other integration algorithms are presented. A tabulation of some optimal integrator filter coefficients is included.**

Manuscript received June 9, 1980; revised December 15, 1981. This work was supported by the Science and Engineering Research Council and the Medical Research Council.

A. H. M. Abed is with the Faculty of Electronic Engineering, Menoufia University, Menouf, Egypt.

P. J. Bloom and G. D. Cain are with the Division of Engineering, Polytechnic of Central London, London, England.

INTRODUCTION

Numerical integration has long been of interest in numerical analysis and is now a topic of considerable importance in digital simulation and digital signal processing, e.g., [1]-[10]. Unlike the related problem of digital differentiation [1], the inherent infinite duration impulse response (IIR) nature of integration precludes a purely FIR approximation if good dc behavior is to prevail. We restrict our attention to an attractive class of integration formulations described by the difference equation

$$y(kT) = y[(k - r)\,T] + T \sum_{n=0}^{N-1} b_n x[(k - n)\,T]. \tag{1}$$

Parameters in (1) are the feedback delay r, the filter order N, and the coefficients $\{b_n\}$ of the feedforward section. In the z domain a transfer function of

$$\frac{T \sum_{n=0}^{N-1} b_n z^{-n}}{(1 - z^{-r})}$$

0096-3518/83/0600-0726$01.00