A Preliminary Study on the Use of Demisyllables in Automatic Speech Recognition

A. E. Rosenberg
L. R. Rabiner
S. E. Levinson
J. G. Wilpon

Bell Laboratories
Murray Hill, New Jersey 07974

Abstract - A speech recognition system is described for recognizing isolated words from reference templates created by concatenating demisyllables from a corpus of about 1000 demisyllables. The composition (in terms of demisyllables) of each reference word is specified in a lexicon with one or more entries for each word of the vocabulary.

Experiments were carried out, using a 100-word vocabulary, to investigate the usefulness of such a representation and the effect on performance of some simple modifications in demisyllable specification and durations of reference patterns. Recognition accuracy of 97.6% was obtained using 132 reference templates for the 100-word vocabulary.

I. Introduction

In this paper we report on preliminary experiments investigating the use of demisyllable (half-syllable speech unit) prototypes for automatic speech recognition. The use of units smaller than words represents a major departure from our previous studies in which whole word prototypes were used exclusively. The change has been made because of the (anticipated) limitations of whole word templates (e.g. storage , access and processing) as vocabulary sizes increase to a range suitable for continuous speech recognition (i.e. 1000-10000 words). Although the change from word templates to demisyllable units 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. The experiments reported here are preliminary, and are intended to provide some insights into the advantages and disadvantages of recognition from an inventory of demisyllable sounds. As such, the vocabulary is limited to 100 words, and test utterances are words spoken in isolation by a single speaker. However the library of demisyllable units containing about 1000 basic units is the one that would be used for unlimited vocabulary sizes.

II. Units for Speech Recognition

Perhaps the most crucial choice for a recognition system is the specification of the recognition unit. For a wide variety of reasons, the demisyllable was chosen as the recognition unit [1-3]. Demisyllables are defined as half-syllable units which occur in strictly constrained initial-final pairs, thereby accounting explicitly for phonological phenomena within the syllable. The use of half-syllable units reduces the inventory size by about a factor of five from that required for whole syllables. In addition, the use of so-called phonetic affixes, typically, the apical consonants /*s*,*z*,*t*,*d*,*θ*/, which may be attached independently to final demisyllables with identical postvocalic voicing conditions, has the potential of reducing inventory size again by about one-half. The result is an inventory of approximately 1000 initial and final demisyllables including five affixes.

The characteristic that distinguishes demisyllables from other half-syllable units is the location of the cut between initial and final demisyllables. The initial demisyllable is generally made quite short, extending just beyond the initial CV transition. In this way any influence of postvocalic consonants (particularly nasals) is largely confined to the final demisyllable, so that initial demisyllables require no special treatment when paired with such final demisyllables. Locating the cut in this way also has the potential of reducing the inventory of initial demisyllables since the same initial demisyllable might be used to precede either stressed or nonstressed final demisyllables or certain diphthongs. For example, the initial demisyllable SHAX might be used equally well in the second syllable of *station*, STEYSHAXN = STEY + EYSH + SHAX + AXN, and as the initial demisyllable in *shove*, SHAHV = SHAX + AHV.

III. Lexicon and demisyllable inventory

To implement the use of demisyllable recognition units we have incorporated two new elements in our recognition system. These are a lexicon and an inventory of demisyllable templates.

The lexicon is a catalogue containing one or more descriptions (in terms of demisyllables) of each word in the vocabulary. Each such description is a sequence of demisyllables corresponding to a standard, pronunciation of a word as an isolated utterance. Examples of lexical entries for the words BACK, MIND, COPPER, FAMILY, and EDUCATION are shown in Figure 1.

WORD	PHONETIC FORM	DEMISYLLABLES
BACK	BAEK	BAE3 + AEK1
MIND	MAYND	MAA3 + AYND1+ D5
COPPER	KAAPER	KAA3 + AA1 + PAX4 + ER2
FAMILY	FAEMIXLIX	FAE3 + AEM1 + MIX4 + IX2 + LIX4 + IX2
EDUCATION	EHJHAXKEYSHAXN	EH3 + EHJH1+ JHAX4 + AX2 + KEH3 + EYSH1 + SHAX4 + AXN2

Fig. 1 Examples of Lexical Entries.

The basic function of the lexicon in our recognition system is to direct the construction of word reference prototypes by concatenating demisyllable templates from the demisyllable inventory according to the specification found in each lexicon entry.

In the preliminary set of experiments reported here, each unknown word is compared exhaustively with every word prototype specified by the lexicon. This word comparison strategy differs from word comparison strategies of previous implementations only by the mediation of the lexicon and demisyllable template inventory.

IV. Basic Speech recognition process

Figure 2 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 [4,5].

As shown in Figure 2, input utterances are compared with whole word reference prototypes consisting of concatenated LPC parameterized demisyllable templates whose composition is specified by the lexicon. A recognition decision is made on the basis of word distance scores.

V. Experimental Evaluation

To evaluate the usefulness of demisyllables for isolated word recognition a series of recognition experiments was run. The vocabulary used for the tests was a 100-word subset of the basic English vocabulary of Ogden [6] The vocabulary consisted of 52 monosyllabic words, 32 two-syllable words, 16-three syllable words, and 2 four-syllable words.

CH1610-5/81/0000-0967 $00.75

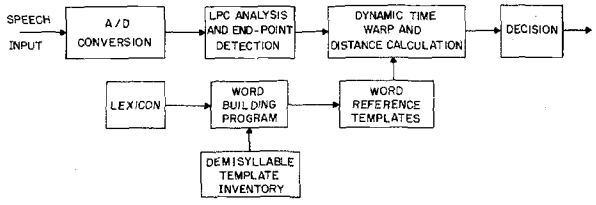

Fig. 2 Block Diagram of Recognizer.

The only talker used in this test (BE) was the one who had originally created the demisyllable inventory for synthesis (albeit several years earlier, and on a different system). Talker BE used a robust isolated word training method [7] to create isolated whole-word reference patterns for the 100 word vocabulary. In addition he spoke the 100-word vocabulary 5 times to provide a test set of 500 tokens. The talker was instructed to articulate naturally and clearly. Pronunciations were prescribed for words in which alternate pronunciations were possible. All recordings were made over a high quality microphone since the original demisyllable inventory was recorded in this manner.

The whole-word templates were used as the reference set for a control experiment together with the 500 test word set. No length normalization of reference or test patterns was used. The resulting recognition accuracy was 99.6%. This result indicates that any recognition errors obtained using demisyllable templates are likely to be due to inadequacies in the templates, not to the confusability among words in the vocabulary.

A series of 5 experiments to test the demisyllable recognizer were performed. Each experiment represented a distinguishable difference in either the template creation procedure, or the lexical description of the vocabulary words, both of which having a direct bearing on recognizer performance. The experiments performed were as follows:

Experiment Number 1 - For each word in the vocabulary, there existed a *single* demisyllable specification in the lexicon (representing a standard dictionary pronunciation). In its crudest form, the lexical entry can be thought of as a variable length vector whose entries are the demisyllables used in the word, and whose length is the number of demisyllables within the word. Thus, if we denote the i^{th} word as W_i, and the j^{th} demisyllable in the inventory as D_j, then the base specification for word W_i is of the form

$$B(W_i) = (q_i(1), q_i(2), \ldots, q_i(L_i)) \qquad (1)$$

where $q_i(k)$ is the inventory index of the k^{th} demisyllable in the word, and L_i is the length of word i (in demisyllables), and the reference pattern $R(W_i)$ is created as

$$R(W_i) = D_{q_i(1)} \oplus D_{q_i(2)} \oplus \cdots \oplus D_{q_i(L_i)} \qquad (2)$$

where $\oplus$ is a concatenation rule.

Each vocabulary word template was created using Eq. (2), and a standard recognition test was performed using the 5-replication 100 word vocabulary. An overall error recognition rate of about 25% was obtained; however the error rate for monosyllabic words (14%) was significantly lower than for 2 syllable words (34%), or 3 syllable words (55%). This was most likely due to large discrepancies in length between reference and test words for polysyllabic words. For example, the length of reference words created from concatenated demisyllable templates increased approximately 50% per syllable while the naturally spoken test words increased approximately 10% per syllable. This led to the next experiment.

Experiment Number 2 - For each word in the vocabulary the overall length of the reference pattern was normalized to either a fixed length (FL) or to the average length (VL) of the 5 tokens of each word as spoken by BE. In addition, at the boundary between demisyllables, the LPC parameter sets were smoothed over a region of Δ frames (Δ=4 was used) using a least-squares quadratic fit for each parameter. The average word error fell to about 16% for both FL and VL normalization. The error rate for monosyllabic words fell to 7.5%, while for 2-syllable words it was 26.5% and for 3-syllable words it was 25%. (The improvements in accuracy for monosyllabic words were primarily due to decreased confusions with polysyllabic words.) The major time alignment problem that remained was that although the overall lengths of the words were correct, the durations of demisyllables, within the words, were often grossly incorrect. This led to the next experiment.

Experiment Number 3 - For each polysyllabic word, the length of the stressed syllable was linearly compressed to q percent of the original length (q=50 was used). This rule was *not* applied when the stressed syllable was the last syllable in the word. The effects of using such a stressed vowel reduction rule are illustrated in Figures 3 and 4 for the words REASON and INDUSTRY. Each of these figures shows plots of the log intensity of the 5 test utterances superimposed (part a), the log intensity of the demisyllable reference pattern with no stressed vowel reduction but with overall length (VL) normalized, (part b), the log intensity plot of the demisyllable reference with 50% length reduction of the stressed vowel (part c), accumulated DTW distance scores for the 5 test utterances compared to the reference of part b (part d), and the accumulated DTW distance scores for the 5 test utterances compared to the reference of part c. Also shown in parts a to c are the segmentation frames (estimated for part a from the DTW warps) between the demisyllables of each word. An examination of the plots of parts a to c shows the vastly improved registration between syllable segments using the vowel length reduced reference, and parts d and e show the lower distance scores that result from the improved templates.

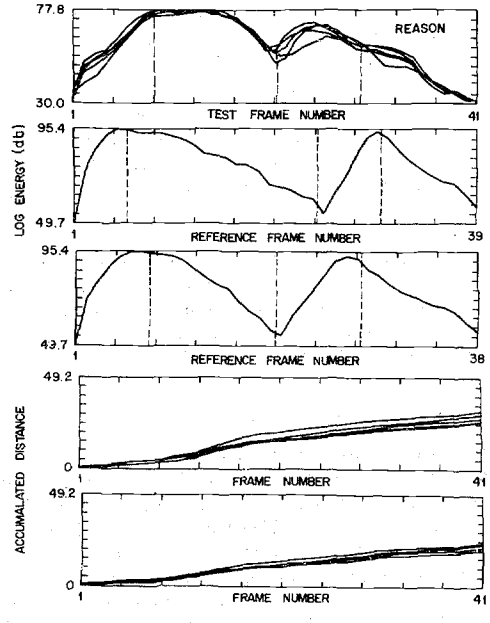

Fig. 3 Illustrative Example of Energy and Distance

Using the vowel reduction rule (with q=50) the overall word error rate fell to about 12%. The improvements in accuracy were 2%, 3.5%, and 7.5% for 1,2, and 3 syllable words respectively. The improvements noted for monosyllabic words were again due to decreased confusability between monosyllabic words and polysyllabic words.

At this point the template creation rules were deemed sufficiently good so that the major emphasis was given to improving the lexicon.

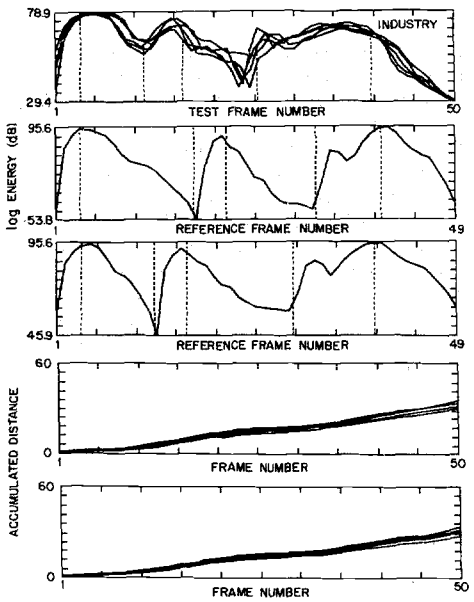

Fig. 4 Illustrative Example of Energy and Distance for word INDUSTRY.

Experiment Number 4 - As noted earlier, initially only a single, standard pronunciation was used for each word in the vocabulary. Informal listening to BE's speech indicated that the vowels on the demisyllable representation did not adequately match the vowels of the naturally spoken word. It was also noted that for some words, especially those with 3 syllables, too much emphasis was given to medial unstressed syllables (e.g. fa-mi-ly instead of fam-li). Thus a number of additions to the lexicon were made in which alternate pronunciations for words were entered. It is pointed out in the discussion section that many of the modifications to the original entries in the lexicon which brought about improved recognitions are not easily correlated with natural changes in pronunciation. A total of 32 such additional entries were made. The word error rate fell to 2.4% with error rates of 0, 6.3%, and 1.6% for 1,2 and 3 syllable words, respectively.

At this point the effects of two of the template creation parameters, namely the smoothing parameter Δ, and the syllable length percentage, q, were investigated. Values of q from 25 to 100, and values of Δ=0 (no smoothing), 2 and 4 were used. Figure 5 shows a plot of the resulting word error rate as a function of q for each value of Δ. It can be seen that for $q > 60$, the error rate increases sharply indicating the importance of some stressed vowel length reduction. For values of q in the range 25 to 50 the error rate did not change significantly. The plots also show a small but consistent reduction in error rate for Δ=4 over Δ=0 or 2 (which have the same error rates).

Experiment Number 5 - Although the results of Experiment Number 4 were considered excellent, it was deemed important to answer the question as to the maximum amount of improvement that could be obtained from improved lexical entries. Thus an automatic procedure was used to obtain the best match to each word of the vocabulary where the best match was defined as the sequence of demisyllables giving the smallest distance. The only constraint was that each odd demisyllable had to be an initial demisyllable, and even demisyllable had to be a final demisyllable. No constraints on the number of demisyllables were used. The procedure used was the recently proposed level building algorithm for connected word recognition [8]. Matches to 1 test version of each word in the vocabulary were obtained. For cases in which the distance of the best automatic match was distinctly smaller than the best match obtained from the current lexical entries (for 55 of the words) additional entries to the lexicon were added. It should be noted that

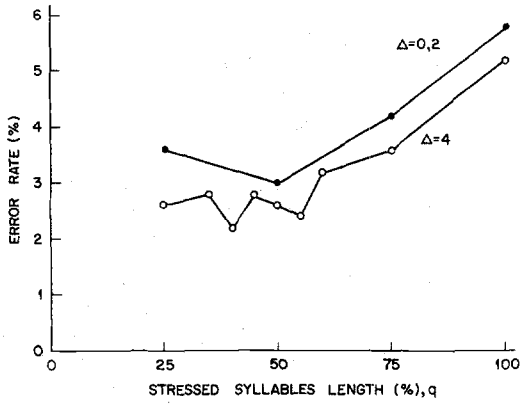

Fig. 5 Recognition Error Rate as a function of stressed syllable length.

the demisyllable computation of the new entries were strictly associated with best matches to words with no additional (phonological or otherwise) significance. This result is due to the relative insensitivity of the distance measure to different consonants.

The overall recognition accuracy using the 187 lexical entries was 99.6% - i.e. the same as that obtained for isolated word (speaker trained) recognition tests.

VI. Discussion

The experimental modifications which bring about improved recognizer performance fall under two general categories, namely duration modification and template specification modification. (A third modification, smoothing the boundaries between demisyllable templates, was found to provide only a small improvement and will not be discussed further). We first consider modifications to the template duration.

6.1 Duration Modifications

An important aspect of our approach is that there is no attempt to explicitly segment syllables but instead, syllable boundaries are by-products of our dynamic programming time alignment process. The time alignment process has the capability of matching templates whose lengths differ by as much as 2 to 1, but performance is improved when the the length ratio is close to 1. For this reason it is important that the durations of reference and test words (both overall and within syllables) be comparable prior to the matching process.

In Experiment 2 overall duration adjustment was handled by normalizing the length of the reference word to either a fixed length (FL), representing a suitable value of duration, or to the average length of the five replications of each test word (VL). It has already been noted that in the absence of this normalization the duration of reference patterns increases about 50% per syllable compared to test words which increase about 10% per syllable. This discrepancy is explained by the fact that reference words are created by concatenating demisyllable templates excised from individual one- or two- syllable source words, whereas in natural speech a word duration constancy phenomenon occurs [9]. Overall duration normalization reduced recognition error rate from 25% to 16%.

The source of the second timing adjustment process, in which the duration of stressed syllables was reduced 50% in polysyllabic words, is the same duration constancy phenomenon noted above. Umeda [10] found that stressed vowel duration decreases approximately 35% from one- to two-syllable words with progressively smaller decreases for additional syllables. In this study, in Experiment 3, a 50% reduction in stressed syllable duration for polysyllabic words was used. The exception to this rule was for final stressed syllables where no adjustment was made. This practice agrees with Umeda's observation [11], that durations of stressed

vowels in the last syllable before a pause do not differ from durations in monosyllabic words. Reducing stressed vowel durations in polysyllabic words in Experiment 3 reduced overall recognition error rate from 16% to 12%.

6.2 Template Specification Modifications

The second category of modifications which brought about improved performance are changes in the demisyllable specifications in the original lexical entries. An obvious requirement for good recognizer performance is that the string of demisyllable units specifying a reference word be a good representation of the actual spoken word. Two conditions are necessary. First, the phonetic transcription specified in each lexical entry must accurately represent the talker's pronunciation of the word. Second, the demisyllable units specified in the lexical entry must accurately reflect the phonetic transcription and, in turn, the talker's pronunciation. In compiling the lexicon a conscientious effort was made to furnish transcriptions in the lexicon which reflected the talker's pronunciation which was standard American English.

In Experiment 4, 32 entries were added to the lexicon in which changes were made to 26 of the original 100 entries. (For some words, more than one additional entry was made). Of these, 18 produced improved recognition scores. Only one of these improved entries could definitely be associated with an error in the phonetic transcription in the lexicon due to pronunciation. This was for the word "cover" in which the original transcription was KAHVER and the additional entry which produced an improved score was KAAVER. The remaining changes reflect either an inadequacy in the demisyllable templates themselves or in the method used to specify a given pronunciation using demisyllable units.

There were three categories of such changes. The first involved reduction of some unstressed syllables. For some unstressed syllables it was found that the general rule which requires that a syllable be specified by an initial followed by a final demisyllable should be violated for improved recognition scores by omitting one or the other or both. In all, 6 of the 18 improved entries fall under this category of reduction of unstressed syllables, in 2 of which medial unstressed syllables were omitted entirely.

In another category of modifications, changes were made in the specification of stressed vowels having no obvious association with pronunciation. The changes were made after listening to the LPC resynthesized versions of reference words and deciding that, for whatever reason, they were not good representations of both the phonetic transcription and the talker's pronunciation, or after examining the kinds of recognition confusions made on the original entries.

The final category of changes, involving 3 of the improved reference words, consisted of adding t or d suffixes to final syllables already including a final t or d. A possible explanation in these instances is that the additional suffix mimics aspiration which may be present in the spoken words.

All the above - mentioned changes reduced word error rate from 12% to 2.4% i.e. approximately as much improvement as for the two timing modifications combined.

The improvements in performance obtained by carrying out these changes in demisyllable specifications raised the question of what is the maximum amount of improvement possible if an exhaustive and unrestricted selection of demisyllable units is allowed to obtain reference words which best match test words. This was the object of Experiment 5. It is not clear whether this exercise has much practical significance and very little meaning can be extracted from the resulting demisyllable transcriptions. The fact that overall word error rate falls from 2.4 to 0.4% indicates that, pushed to its limit, demisyllable specification of word templates can provide as good recognition performance as whole word templates.

VII. Summary

The purpose of this investigation was to determine whether word templates, created from demisyllable units, could adequately represent words for speech recognition applications. Using a set of about 1000 demisyllables word templates were created for a 100 word vocabulary. The composition (in terms of demisyllables) of each word was obtained initially from a standard pronouncing dictionary, but some adjustments to the pronunciations were made. Rules for concatenating demisyllable units were developed so that both word and syllable durations were reasonable. Using the developed set of word construction rules, and a lexicon with 132 entries (1.32 entries/word), a word recognition accuracy of 99.6% was obtained for a single talker. These results indicate that demisyllable reference patterns show promise for use in a continuous speech recognition system.

References

[1] Fujimura, O. and Lovins, J. B., "Syllables as Concatenation Phoentic Units", Chapter in *Syllables and Segments*, A. Bell and J. B. Hooper, eds., North-Holland Publishing Company, 1978.

[2] Fujimura, O., "Syllable as a Unit of Recognition", *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, Vol. ASSP-23, No. 1, pp. 82-87, Feb. 1975.

[3] Mermelstein, P., "A Phonetic-context Controlled Strategy for Segmentation and Phonetic Labelling of Speech", *IEEE Trans. on Acoustics, Speech, and Signal Proc.*., Vol. ASSP-23, No. 1, pp. 79-82, Feb. 1975.

[4] Itakura, F., "Minimum Prediction Residual Applied to Speech Recognition", *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, Vol. ASSP-23, No. 1, pp. 67-72, Feb. 1975.

[5] Rabiner, L. R., Levinson, S. E., Rosenberg, A. E., and Wilpon, J. G., "Speaker-Independent Recognition of Isolated Words Using Clustering Techniques", *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, Vol. ASSP-27, No. 4, pp. 336-349, Aug. 1979.

[6] C. K. Ogden, *Basic English: International Second Language*, Harcourt, Brace and World Inc., 1968.

[7] Rabiner, L. R., and Wilpon, J. G., "A Simplified, Robust Training Procedure For Speaker Trained, Isolated Word Recognition Systems", *J. Acoust. Soc. Am.*, Vol. 68, No. 5, pp. 1271-1276, Nov. 1980.

[8] Myers, C. S. and Rabiner, L. R., "A Novel Dynamic Time Warping Algorithm for Connected Word Recognition", *IEEE Trans. on Acoustics, Speech, and Signal Proc.*, to appear.

[9] Lehiste, I., Suprasegmentals, p. 40, M.I.T. Press, Cambridge, Mass., 1970.

[10] Umeda, N., "Effects of speaking made on temporal factors in speech: vowel durations", *J. Acoust. Soc. Am.*, *56*, 1016-1018, 1974.

[11] Umeda, N., "Vowel duration in American English", *J. Acoust. Soc. Am.*, *58*, 434-445, 1975.