Paper S11.3 to be given at
IEEE ICASSP-80
Denver, April 11, 1980

EXPERIMENTS IN SYLLABLE-BASED RECOGNITION OF CONTINUOUS SPEECH

M.J. Hunt, M. Lennig, P. Mermelstein

Bell-Northern Research
3, Place du Commerce, Nuns' Island, Quebec, Canada H3E 1H6

ABSTRACT

An exploratory implementation of a syllable-based recognizer is described. Continuous speech is first divided into syllabic units, and the units are then matched against syllable templates using a dynamic programming algorithm. A hierarchical transition network is used to limit the syllable search to possible continuations of the current partial sentence hypotheses. Competing hypotheses are pruned by a 'beam search'.

Experiments are reported on automatic recognition of English sentences with a 70-word vocabulary and restricted syntax produced by one male speaker. 85% of the sentences were correctly recognized. Comparable results were obtained for a similar task in French using a female speaker. The method is computationally efficient: real-time performance on dedicated hardware should be obtainable without difficulty. A method of scaling the distance measures used in the syllable matching is described. This scaling takes into account variability in syllable production, both as a function of position within the syllable and as a function of the various spectral parameters being used.

1. INTRODUCTION

At some level in a continuous speech recognition system there is an interface between acoustically derived information about the speech to be recognized and syntactic constraints on the interpretation of that information. For maximum efficiency in the application of the constraints, the interface level should be as low as possible. At the lowest levels, however, the interpretation of acoustic information is highly context dependent. The syllable level is attractive because it is low enough to make the inventory of units to be recognized reasonably small while being high enough to make the phonetic interpretation of each unit reasonably context independent.

In addition, syllable boundaries can be determined independently of the task of identifying the syllable. This property, which does not hold for words or for phonemes, makes for efficient comparison of syllables with stored reference forms and for efficient evaluation of competing sentence hypotheses. The theoretical advantages of the syllable-level representation are discussed by Mermelstein (1) and Fujimura (2). Davis and Mermelstein (3) report results on syllable-based recognition assisted by lexical information but using hand-segmented speech data. Zwicker, Terhardt and Paulus (4) use a similar syllable-based psycho-acoustics-oriented scheme for the recognition of German city names.

The experiments reported pertain to speaker-dependent recognition of pre-specified sentences produced with a limited vocabulary and restricted syntax. One male speaker was used to train and test a 70-word vocabulary English system intended to recognize 'date and time' sentences, e.g., "January fourth at eight forty-five p.m.". A female speaker was used to test and train a comparable French system. The French vocabulary comprised 43 words and an additional 25 word variants corresponding to sequences of two or more words in which syllables had been fused across a word boundary. The number of such word variants of the English system was only four.

A sentence to be recognized is first divided into syllable-sized units. Recognition then proceeds syllable by syllable through the sentence. Each unknown syllable is compared to a set of reference templates using a dynamic programming matching algorithm. The templates are selected on the basis of the preceding syllable matches and the system's knowledge of the syntactic structures allowed. Several alternative sentence interpretations are held in parallel. When the end of the sentence is reached, the syllable sequence which best fits the data is printed out.

2. PREPROCESSING

Speech material is low-pass filtered at 3.7 kHz and digitized at an 8 kHz sampling rate. The power spectrum is computed with a 25.6 ms raised-cosine window with 75% overlap between successive windows, giving a 6.4 ms frame rate.

We believe that the spectral representation should reflect the frequency resolution of the ear and its logarithmic response to intensity. We therefore group the spectrum into twenty channels of equal width on the perceptual mel scale of frequency and take logs of the channel amplitudes. Adjacent channels tend to be correlated, but a compact, uncorrelated representation of the spectral envelope can be obtained by taking the first seven terms of the cosine transform of the channel log amplitudes. We refer to these values as mel-scale cepstrum coefficients. Such coefficients have been shown to compare favorably with other representations of the spectrum for speech recognition task (3).

For the purpose of the syllabifier, a measure of perceptual loudness is computed for each frame. This is the log of the sum of the linear channel amplitudes weighted by the sensitivity of the ear in each frequency region.

3. SYLLABIFICATION

Since a syllabification algorithm very similar to the one we are using in this work has been described in detail elsewhere (5), we shall provide only a brief description here.

Using the loudness contour, continuous regions of speech are first separated from regions of silence. Syllable boundaries are then placed at local minima in the loudness function, subject to the conditions that the dips must be sufficiently deep and the candidate syllables must be sufficiently long and loud and must contain some voiced speech. The first cepstrum coefficient, which is a measure of the difference in low to high frequency energy, is used to make the voiced/voiceless decision.

Some male speakers occasionally produce very low fundamental frequencies associated with creaky voice. The individual glottal pulses impose a modulation on the loudness function. The dips generated in this way tend, however, to be narrower than those occurring at true syllable boundaries, so spurious boundaries can generally be avoided by placing requirements on the width as well as on the depth of acceptable loudness dips.

The 'syllables' produced by the syllabifier do not have to correspond in all cases to what a phonetician would accept as syllables: the word 'twenty', for instance, is often left as one syllable, and this presents no problems to the system. Equally, the syllabification does not have to be perfectly consistent: 'twenty' can appear as one syllable or two. What matters is that any syllabifications occurring in speech to be recognized should have already been encountered in the training material. The syntax is, however, made more complicated when fused pairs of syllables span word boundaries, as when 'twenty-eight' is split into 'twen' and 'ty-eight'.

Minor errors in the location of a syllable boundary can cause all or part of a consonant to migrate to an adjacent syllable. Even when such a migration crosses a word boundary, the phenomenon can be handled effectively by the syntax network described below.

4. SYLLABLE MATCHING

Syllable matching is used in both the recognition task itself and in creating composite reference templates during the training phase. The process is carried out by a variant of the widely used dynamic programming time-warping algorithm. The particular unconstrained symmetric formulation that we use has the property that all paths linking the two ends of the syllables being matched contain the same number of frame comparisons, and that number is equal to the average number of frames in the two syllables. This is useful when generating composite reference templates. Pairs of examples of the same syllable are time aligned and the parameters of corresponding frames are averaged together. The outputs of this combination process are themselves combined in pairs and the process is repeated until a single composite template is produced. The composite has a length equal to the average of the lengths of the component examples. On our data we have found that the use of composite templates produces half as many errors as the use of a single representative example of each syllable type.

The local distance measure used in comparing frames is essentially a squared Euclidean distance summed over the cepstrum coefficients. These distances are then summed along the best time alignment path to get a measure of the similarity of the two syllables. However, some scaling is applied to account for the differences in reliability of information coming from different parts of the syllable and from the various cepstrum coefficients (6). By matching a composite template against the individual examples that went to make it up, we determined that the ends of syllables are more variable than the middles and that the variability for the same speech sound of the individual cepstrum coefficients decreases with increasing coefficient number. These results were found to be consistent across the two male and female speakers examined. When summing local distances along a time-alignment path, then, distance contributions from near the ends of a syllable are scaled down, and in each local distance calculation the contribution of the lower order cepstrum coefficients is reduced. The implementation of these two scaling factors reduced the error rate from 7 incorrectly recognized sentences to 3.

5. REPRESENTATION OF SYNTACTIC AND LEXICAL INFORMATION

Syntactic information is represented in a hierarchical transition network in which the arcs of higher level networks refer to lower level 'subnetworks'. System training consists of providing the system with information on two distinct levels: the higher level syntax and the lexicon. The higher level networks determine which strings of words form grammatical sentences in the task language. Lexical subnetworks recognize individual words by specifying which series of syllables can correspond to them.

The higher level syntax is specified manually by the syntax designer. We have developed a symbolic language for this purpose, in which each line of code specifies one transition network arc. A compiler has been implemented which translates the specified syntax into a numeric code which can be loaded by the Sentence Recognizer. Since the higher level syntax is speaker independent, it need not be modified when the system is trained for a new speaker unless the task language itself is changed.

The lexicon consists of a set of lexical subnetworks, one corresponding to each word in the task language. Each lexical subnetwork is a network similar in form to the networks of the higher-level syntax but instead of specifying possible sequences of words which make up the constituent, the lexical subnetwork specifies possible sequences of syllables

which make up a word. Instead of calling other subnetworks, it employs arcs which make calls to the syllable comparator.

The class of grammatical sentences is defined by the set of paths through the network from the initial state to the final state. Sentence recognition consists of searching this set for the path whose cumulative distance is minimum.

Conditions and actions on arcs are used to handle the problem of syllabification variation in which consonants are transferred across word boundaries. For example, the phrase 'ninth October' is sometimes syllabified as /nain Ɔk to bə/. An action is used to set a status register in any hypothesis recognizing the word 'ninth' to the value /0/. A condition is used to accept /Ɔk/ as the first syllable of 'October' but only if the register has been set to /0/. Similar mechanisms are used to handle backward transference of consonants across word boundaries and to handle liaison in French.

6. TRAINING

During lexical training, a set of syllabified training sentences is transcribed by a human transcriber. Each syllable is played through a loudspeaker and transcribed phonetically. Word boundary locations and standard spellings are also entered by the transcriber. A lexical compiler we have developed is used to convert this data into a set of lexical subnetworks, one for each word in the training set. The lexical subnet corresponding to a word is capable of accepting all the syllabifications of that word which occurred in the training set.

In the manner described in Section 4, a composite template is generated corresponding to each unique phonetic transcription which occurs in the training set. All syllable tokens which are transcribed identically are combined into a single composite template corresponding to their common transcription.

7. SYLLABLE SEQUENCE EVALUATION

As recognition proceeds through the sentence, one syllable at a time, the syntax proposes syllables that can follow the sequence already evaluated. The difference values returned by the syllable comparator and summed over the syllables in a particular sequence are taken to be a measure of the probability that the sequence provides a correct interpretation of the sentence so far. Since information coming later in the sentence can affect the interpretation of earlier parts, a number of sequences are followed in parallel. However, the syntax is constructed in such a way that transitions out of a state do not depend on how that state was reached. Consequently, when two sequences find themselves in the same state at the same time, the Optimality Principle can be invoked to drop the worse fitting sequence.

The number of sequences being considered is also limited by applying a 'beam search' (7): the sequence with the lowest summed distance is determined, and all sequences whose summed distances exceed the lowest value by more than a fixed threshold are dropped. This technique has proved highly effective for our particular task. Since the computation involved in recognizing a sentence is dominated by the syllable comparison process, the number of comparisons made provides a measure of the computational cost of a particular strategy. The strategy which makes the fewest comparisons is that of retaining only the single best syllable sequence at any time. For the English data, this results in sixteen sentences containing syllable recognition errors. The beam search method with a suitably adjusted beam threshold reduces the number of such sentences to three, yet requires only 20% more syllable comparisons. Moreover, the three remaining errors are all single-syllable substitutions (such as 'first' for 'third'), which no alternative syllable sequence evaluation strategy could correct.

As an extra limitation on the number of sequences being held, a check is made as to whether a particular sequence can ultimately lead to a complete sentence given the number of syllables remaining to be recognized. The maximum and minimum number of syllables needed to complete a sentence from a given state can be determined by summing three integers stored in a table. This table is constructed automatically before run time from the syntax network.

8. RESULTS

To test the performance of the system on the English data set, 59 date-and-time sentences were recorded to form the training set and two months later the same 59 sentences were newly recorded by the same speaker to form the test set. The syllable boundary positions found by the syllabifier in six test sentences were incompatible with the syntax. Five of these sentences were rejected by the recognizer because it could not find a substitute sequence that matched the acoustic data well enough. Of the remaining 53 sentences, 50 were correctly recognized when a wide beam was used in the beam search.

The French system was trained on 100 sentences and then tested on 100 new randomly generated sentences. The system correctly recognized 76 of them. Thus, performance on the French sentences in comparable with the 85% obtained in English. Although the English training set was smaller, this is partially offset by the fact that the English test set consisted of new productions of the same sentences, whereas the French test sentences were newly generated using the training vocabulary.

The English and French data present rather different problems to the syllabifier: in English, consonant clusters, particularly stop-fricative combinations, can be troublesome; in French trilled /r/ sounds can generate spurious syllable boundaries, and the common CV and V syllable structures can give rise to syllable fusions. In both languages most syllable recognition errors could be ascribed to small variations in the positions of syllable boundaries.

Running on a PDP 11/45 with the syllable matching carried out on the AP120B fast processor, recognition time per sentence was around 20 sec. About half of that time is taken up in syllable matching.

9. DISCUSSION

The vocabulary used in these experiments is small by comparison with those used by some other research groups and our sentences are relatively short. However, this is balanced to some extent by relatively free syntax, generating over a million sentences, and by the presence of many acceptable sentences differing only in a single syllable, e.g., 'sixteen' and 'fifteen'. Thus, this recognition task is a reasonably exacting test of system performance, perhaps comparable with IBM's New Raleigh language (8).

Representation of both the syntactic and lexical information in transition network of syllables is effective in reducing the number of syllable matches executed. Although the total number of reference syllables is 80, the number of matches per unknown is on the average only 12. By retaining the syntactic and lexical information in a hierarchical structure, we minimize the storage required for specifying the network. Implementation of a syllable-based system in dedicated hardware would allow real-time recognition at a reasonable cost.

The main impediment to achieving higher sentence recognition rates on the current task appears to be the performance of the syllabifier. With the current system organization we have no way of feeding back recognition difficulties to the syllabifier so that alternative segmentations may be proposed. On the other hand, most serious syllabification errors result in sentence rejections rather than misrecognitions, and an increase in task difficulty resulting from an enlarged branching factor would not be expected to increase the number of such errors.

The primary advantage we see for the recognition system described is the conceptual simplicity of the structural and acoustic data representations. Context problems in phone recognition are avoided by representing entire syllables as spectral patterns in frequency and time. Interpretation of these patterns is carried out with the aid of all higher-level information available. The number of active hypotheses is sharply limited by pruning whenever a new syllable is added to the hypothesized sequence.

A significant disadvantage for practical implementation of such a system is the extensive training it requires for each new speaker. A reference form must be generated for each possible syllable or pseudo-syllable. Although storage of such amounts of reference information is reasonable on a speaker-independent basis, use of separate speaker-dependent reference syllabaries would severely restrict the use of such systems.

We do not pretend that the scale of experiments reported here is large enough either to train the system adequately or to evaluate its performance accurately. Nevertheless, our results so far lead us to believe that a syllable-based system is worthy of further consideration.

REFERENCES

(1) P. Mermelstein, "A phonetic-context controlled strategy for segmentation and phonetic labeling of speech", IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-23, pp.79-82, 1975.

(2) O. Fujimura, "Syllable as a unit of speech recognition", IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-23, pp.82-87, 1975.

(3) S. Davis and P. Mermelstein, "Evaluation of acoustic parameters for monosyllabic word recognition", J. Acoust. Soc. Am., Vol. 64, p. S100, 1978.

(4) E. Zwicker, E. Terhardt and E. Paulus, "Automatic speech recognition using psychoacoustic models", J. Acoust. Soc. Am., Vol.65, pp.487-498, 1975.

(5) P. Mermelstein, "Automatic segmentation of speech into syllabic units", J. Acoust. Soc. Am., Vol. 58, pp. 880-883, 1975.

(6) M.J. Hunt, "A statistical approach to metrics for word and syllable recognition", J. Acoust. Soc. Am., Vol. 66, pp.535-536, 1979.

(7) B.T. Lowerre, "The HARPY speech recognition system", Dissertation for Dept. of Computer Science, Carnegie-Mellon University.

(8) F. Jelinek, "Continuous speech recognition by statistical methods", Proc. IEEE, Vol. 64, pp.535-556, 1976.