Reprinted from ICASSP — IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, March 26-29,1985

ISOLATED WORD RECOGNITION USING A SEGMENTAL APPROACH*

Louis C. Sauter

Laboratoires de Marcoussis, Division Informatique
Marcoussis, France

ABSTRACT

Building reference sets for speaker independent isolated word recognition systems using multiple references implies a considerable effort. Whenever the vocabulary is specific to some domain a segment based approach should allow to change the vocabulary without having to record new speakers. We shall examine two possible approaches. The first consists in creating whole word references by concatenating segments. The second uses an algorithm originally designed for connected word recognition. Performances for these two approaches are compared.

INTRODUCTION

Speaker independence in isolated word recognition can be achieved by different methods such as speaker adaptation or multiple references. We shall deal here only with the latter approach. In this case, the training set contains several utterances of each of the words in the vocabulary by enough speakers to take into account inter speaker variations [1,2]. This type of training makes creating a new vocabulary a long process. This is not important for standard vocabularies such as numbers, but is a problem for task specific vocabularies or whenever the vocabulary must change. For these reasons, an approach to speaker independence is being investigated where each word in the vocabulary is represented by a sequence of segments. Each segment must be represented by several templates taking into account inter speaker variations. Given an unknown utterance, the system finds the best sequence of segments corresponding to a word in the vocabulary. This can be done in several ways, with or without previous segmentation, using a top-down or a bottom-up search procedure.

* This work was partly supported by a European Community ESPRIT project.

SEGMENT BASED SPEECH RECOGNITION

Segments for Speech Recognition

One of the major problems in speech recognition is how to deal with coarticulation. The way the context influences different phonemes is not well known and is very difficult to handle automatically. An efficient way of dealing with phonological variability is to use segments containing several phonemes, such as diphones or better, syllables [3-6]. Diphones are difficult to segment reliably and in many cases are not long enough (whenever the influence of the context spans several phonemes such as in [kli] or [kla]). Syllables are longer and contain most of the coarticulation, but the problem for most languages is their number (over ten thousand in French). A good compromise is the demisyllable, that is a segment obtained by cutting a syllable in the middle of the vowel. Many demisyllables are diphones (such as [ka], [at], [ol]) but they can also contain several phonemes (such as in [kla]). Syllables can easily be reconstructed by concatenating two demisyllables. The total number of demisyllables for most languages is a few thousand.

Recognition Algorithms

We will now describe different ways of using segments for recognizing speech.

Whole word reference patterns can be reconstructed by concatenating segment patterns. This has been used for speaker dependent isolated word recognition [6] using demisyllable segments. Results for a large vocabulary were slightly poorer than those obtained with global word recognition. Note that this method requires one dynamic time warp for each entry in the lexicon. We will examine how this method can be used for speaker independent recognition.

A different approach is based on algorithms designed for connected word recognition using whole word templates, yielding a "connected segments" algorithm. The search procedure is guided by the lexicon, in the same way as syntax is used for connected speech. Examples of these types of algorithms can be found in [7-9]. Only Myers and Levinson [10] have examined the problem of speaker independence using multiple templates for each word.

Both of these approaches are promising for small vocabularies. The number of dynamic time warps increases with the size of the vocabulary in both cases. Sub-lexicon prediction could be used to reduce the number of dynamic time warps for larger vocabularies.

In another possible approach the unknown speech is segmented into demisyllables. Each segment is then compared to the reference segments, yielding a lattice of distances, which is then used to identify the word. Knowing the end-points of the segments allows a reduction of the computational complexity. Therefore, this approach would be best when considering large vocabularies. Since most segmentation errors will cause a recognition error, the reliability of the segmentation procedure is of the utmost importance in this approach.

In the following we will examine the first two methods described above : word template construction and connected segments recognition.

TEST CORPUS

An interactive program was implemented in order to build a corpus of segments. For a first series of tests, we have used a corpus of 35 words pronounced by 20 different speakers (10 male and 10 female). Each word was split into a sequence of segments, yielding 87 different segments. The total corpus contains 1740 segments.

We have also picked a test vocabulary of 18 words that can be obtained by concatenating segments belonging to the corpus (Table 1). The vocabulary is difficult : it contains several minimal pairs such as vitre-mitre-titre. Also, the words were chosen so that whenever possible the context for each segment was not the same as in the word from which it had been extracted. The test vocabulary was recorded by 10 speakers (5 male and 5 female), none of which belonged to the original 20 speakers. The end-point detection was done automatically.

CONSTRUCTION OF WHOLE-WORD REFERENCES USING SEGMENTS

In this approach, whole word references are constructed by concatenating demisyllables. Various ways of doing this have been experimented and will now be described.

Building Global Templates

In order to build global templates for a word, it is possible to concatenate segments extracted from different words. When dealing with several speakers, we must choose between using in the same template segments pronounced by only one speaker or mixing segments from different speakers in the same template. In the first case, we can obtain one template for each speaker (in our case, 20 templates for each word). In the second case, a larger number of different templates can be obtained (if n is the number of segments in the word and if m is the number of speakers, there are n to the power m distinct possibilities). We have experimented with both 20 and 100 templates per word.

titre	mitre	vitre
mité	cité	paté
traire	traitre	termite
santé	sente	quantité
fête	fez	cassette
douter	divin	passoire

Table I : Test Vocabulary

In both cases, it is possible to smooth the boundaries between segments.

The available speaker independent isolated word recognition system uses 5 reference templates for each word in the vocabulary. These must be computed from the available templates. The first step in this process is to use a clustering algorithm to find classes of similar templates. The second step is to choose one reference template for each class.

Clustering Algorithm.

The aim of this step is to build clusters of similar templates. We have used two different algorithms :

Dynamic Clusters Algorithm [11]. In this algorithm the clusters are obtained by repeatedly assigning each object to the nearest cluster and then calculating the new centers for the clusters. The center of each cluster can be computed in one of the three manners described below (see "choice of the reference"). Since this algorithm converges to a local optimum, the initial conditions can be changed randomly and the best local optimum chosen.

Minimax Algorithm [2]. Clusters are repeatedly obtained by picking the center (minimax) of the remaining objects and assigning to a cluster all the objects such that the distance to that center is smaller than some threshold.

Choice of the Reference for each Cluster.

We have compared three ways of choosing the reference template for each cluster. In the first, we choose the template for which the cumulative distance to all the members of the cluster is minimum. In the second, we pick the template for which the maximum distance to all the members of the cluster is minimum (minimax). In the third method we compute an average of the members of the cluster using a dynamic programming approach described in [1].

It is also possible to construct whole word templates by concatenating the reference segments obtained as described below.

CONNECTED SEGMENTS APPROACH

Building Reference Segments.

A set of reference segments (5 references for each of the 87 segments) was built in the same manner as described above (dynamic clustering followed by dynamic averaging of each cluster).

Connected Segments Algorithm.

We have used a modified version of a connected words recognition algorithm based on the level building system in [10] in which words are replaced by segments. The syntax which was implemented as a finite state automaton is replaced by a description of the vocabulary. The algorithm also takes into account multiple references for each segment.

Finite State Automaton. Figure 1 represents the finite-state automaton used to define the vocabulary. Each transition in the finite-state automaton represents a segment.It was derived from the list of segments for each word.

Local Dynamic Programming Equation. Because it finds the word with the smallest cumulative difference, the dynamic programming algorithm tends to favor shorter words. Usually, the total difference is normalized. In our case, this is not possible because paths of different lengths can lead to the same cell. We must either use a non-symmetrical local equation or normalize by the length of the path at each step.

Constraints. The connected segments algorithm uses dynamic programming to achieve time warping. It is possible to constrain the path both locally or on a global basis. We have tested type 0 or type 1 local constraints. Global constraints are implemented as upper and lower limits for the path. We have also used segment-based constraints : although the connected segments algorithm does not need any information on segment boundaries, it is possible to constrain the time warp path to pass through locations depending on segment-based information.

Smoothing. Smoothing is not possible, since each segment can be preceded or followed by different segments. One way to achieve a kind of smoothing is to reduce the weight of areas close to segment boundaries.

RESULTS

The following results were obtained using the test corpus described above : the 18 word vocabulary was pronounced by 10 different speakers, yielding a test set of 180 words. No rejection threshold was used.

The speech was low-pass filtered and digitized at 8000 samples/second. 9 Mel Frequency Cepstrum Coefficients (MFCC's) were computed every 16 milliseconds.

Segment-Based Whole Word References

Effect of Mixing Speakers. It seems best not to mix speakers : the number of errors is lower with 20 single-speaker templates (30 errors) than with 20 or 100 mixed-speaker templates (respectively 36 and 39 errors).

Choice of Clustering Algorithm. With 100 templates, the choice of the clustering algorithm does not change the results. With 20 templates, it seems that the dynamic clusters algorithm makes better use of the available information. The best way to pick the center of each cluster is to take the object yielding the smallest cumulative distance to the members of the cluster.

Choice of Reference Template. The best results were obtained with the dynamic average template. The number of errors is then 24 instead of 30 with the template yielding the minimum cumulative distance.

Effect of Smoothing. Smoothing greatly improves the quality of the reference set. The number of errors falls from 24 to 16 (less than 9%). Different ways of smoothing have been tested, and the best results were obtained by replacing 4 vectors on each side of the boundary by a least squares linear approximation.

Use of Reference Segments. Directly using words that have been obtained by concatenating reference segments as defined above yields poor performances.

More research is necessary to improve these results. The recognition rate for the difficult test vocabulary is better than 91%.

Connected Segments Type Recognition

Choice of Local Equation. The best results (20 errors) were obtained by normalizing the cumulative distance at each cell. (Divide by the length of the path before choosing the optimum).

Weighting at Boundaries. This increases the number of errors.

Segment-Based Constraints. Constraining VC type segments to begin on a local maximum of the energy (i.e. at the middle of a vocalic segment), and constraining CV type segments to begin in the region of a local minimum of the energy improves the recognition. The number of errors is then 16.

Local Constraints. Above results use type 0 local constraints. Using type 1 constraints increases the number of errors.

Global Constraints. Above results used global constraints defined by slopes of 2 and 1/2. Better results (15 errors) were obtained by relaxing slightly these constraints (5/3 and 1/3).

The best results obtained by this approach were 15 errors for the 180 word test corpus (the recognition rate is approximately 92%).

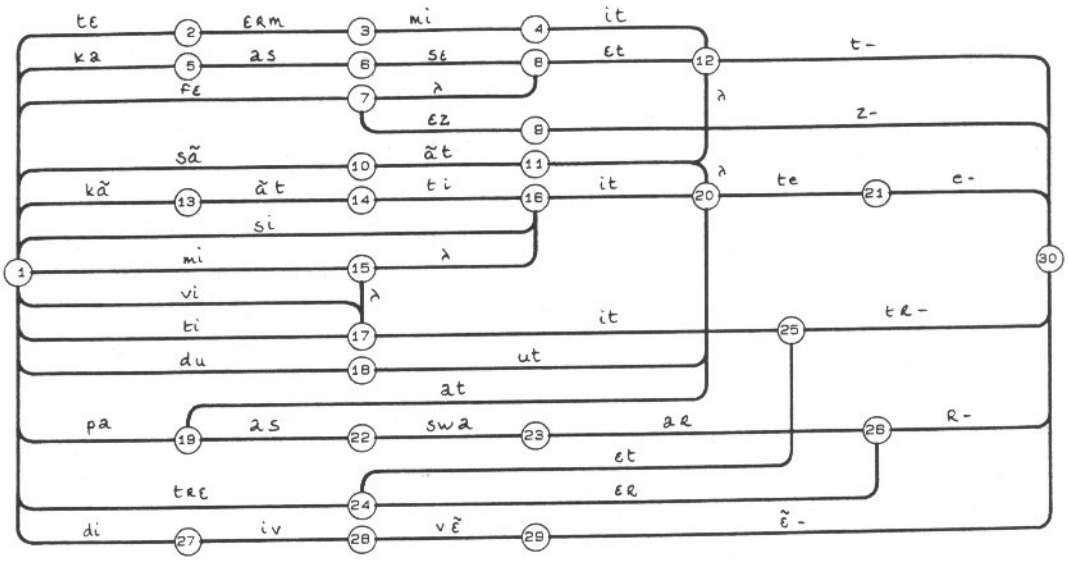

Fig. 1 : Finite-State Automaton

CONCLUSION

The recognition rates for the two approaches discussed above are very similar. The connected segments algorithm has a higher computational complexity than the whole-word reference system, and its recognition time is about twice as long. Since we do not have a whole word-based reference set for this vocabulary, there is no way to compare exactly these results to what would be achieved by the more classical approach. Nevertheless, they are quite promising.

As we have seen above, segment based information improves the quality of the recognition. This advantage would be full exploited in an approach using reliable segmentation.

Current results have all been obtained by systems using Dynamic Time Warping. The segment based approach should also perform well with recognition systems using Hidden Markov Models.

REFERENCES

[1] B. Flocon, N. Briant, "SYRIL: Système temps réel de reconnaissance de mots indépendant du locuteur", 4ème congrès AFCET RFIA, Paris, France, pp. 33-39, Jan. 1984.

[2] L. R. Rabiner and J. G. Wilpon, "Speaker Independent Isolated Word Recognition for a Moderate Size (54 Word) Vocabulary," IEEE Trans. Acoust. Speech and Sign. Proc., Vol. ASSP-27, No. 6, pp. 583-587, Dec. 1979.

[3] O. Fujimura, "Syllable as a Unit of Speech Recognition," IEEE Trans. Acoust. Speech and Sign. Proc., Vol. ASSP-23, No. 1, pp. 79-82, Feb. 1975.

[4] P. Mermelstein, "Automatic segmentation of speech into syllabic units," J. Acoust. Soc. Am., Vol. 58, No. 4, pp. 880-883, Oct. 1975.

[5] Th. Schotola, "On the use of demisyllables in automatic word recognition," Speech Communication, Vol. 3, No. 1, pp. 63-86, April 1984.

[6] A. E. Rosenberg, L. R. Rabiner, J. G. Wilpon and D. Kahn, "Demisyllable-Based Isolated Word Recognition System," IEEE Trans. Acoust. Speech and Sign. Proc., Vol. ASSP-31, No. 3, pp. 713-725, June 1983.

[7] L. Sauter, "RAPACE, un système de reconnaissance analytique de parole continue," 4ème congrès AFCET RFIA, Paris, France, pp. 89-98, Jan. 1984.

[8] H. Ney, "The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition," IEEE Trans. Acoust. Speech and Sign. Proc., Vol. ASSP-32, No. 2, pp. 263-271, April 1984.

[9] J. S. Bridle, M. D. Brown and R. M. Chamberlain, "An Algorithm for Connected Word Recognition", Proc. Int. Conf. Acoust. Speech and Sign. Proc., Paris, France, pp. 899-902, May 1982.

[10] C. S. Myers and S. E. Levinson, "Speaker Independent Connected Word Recognition Using a Syntax-Directed Dynamic Programming Procedure," IEEE Trans. Acoust. Speech and Sign. Proc., Vol. ASSP-30, No. 4, pp. 561-565, Aug. 1982.

[11] E. Diday, "The dynamic cluster algorithm and optimization in non hierarchical clustering", Proc. Fifth IFIP Conf., Rome, Italy, 1973.