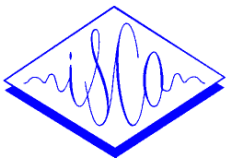

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

A NEW ALGORITHM FOR CONNECTED DIGIT RECOGNITION

S. Cifuentes, J. Colás, M. Savoji, J.M. Pardo

Electronic Engineering Department - Universidad Politécnica Madrid
E.T.S.I. Telecomunicación - Ciudad Universitaria, 28040 Madrid, Spain

10.21437/ICSLP.1992-189

ABSTRACT

In this paper we present an **Unconstrained Level Building (ULB) algorithm** based on the known Level-Building approach [1][11], which allows to deal, in a new particular form, the problem of the interior pauses or silences produced when strings of digits are pronounced in a natural way (strings of digits may contain silences or pauses between digits, due, for example, to the speaker's need of breathing, or to his hesitation in recalling or reading the next digit to pronounce). To evaluate the performance of the new algorithm a speaker dependent connected digits Spanish data base has been used, consisting of 129 digit strings, with lengths ranging from one to five digits, pronounced by a male speaker. The recognition set contains 46 strings, and the training set 64. The 19 remaining strings have been used, as an independent testing set, to control the convergence of the training procedure. Several experiments were carried out to determine the best set of conditions for recognition of digit strings. The improvement obtained on String Accuracy using the ULB algorithm instead of a Level Building with an HMM for the silence is 4.3%, and without any silence modelling is nearly 11%.

INTRODUCTION

One of the most important applications in speech recognition is connected-digit recognition that have significant applications in the area of telecommunications, as well as for recognizing spoken credit-card numbers, stock codes, and so on. For the applications above, speaker-independent systems would be required. However, there are a wide range of applications for speaker-trained connected-digit recognizers, including specialized operator services, insurance-claims entry, quality control and so on.

Because of its vast potential application, a wide variety of approaches to connected-digit recognition has been proposed and evaluated (Sakoe, 1979 [8]; Myers & Rabiner, 1981; Bridle & Brown & Chamberlain, 1982 [9]; Ney, 1984; Bush & Kopec, 1986 [13]). Most of the proposed methods are based on pattern-recognition procedures.

In terms of algorithms for connected-digit recognition, there exist several practical realizations including the two-level dynamic programming approach of Sakoe, the level-building approach of Myers & Rabiner [1][2][3], the one-stage of Bridle & Ney. These algorithms fundamentally are identical, but they differ primarily in efficiency of implementation.

However, the algorithms tested and the tasks evaluated [6][7] typically assume that user input be restricted only to a set of defined vocabulary words [12]. Recently, several trials of speaker-independent isolated and connected-digit recognition technology have been carried out in different places and recognition results showed that customer responses during these trials had the desired vocabulary words (digits), along with extraneous input which ranged from non-speech sounds (background noises, silence, ...) to groups of non-vocabulary words.

Most conventional recognition algorithms have not been designed to handle this type of realistic input. As such, modification of the algorithms has to be made to recognize vocabulary words embedded in speech.

In our opinion, there exist two major ways of dealing with this problem, which are:

1.- Try to model these non-vocabulary words and non-speech sounds using HMM models, as well as the vocabulary words.

2.- Skip over these extraneous and non-modelled sounds or words. The new connected-digit recognition algorithm tested by us, is based on this idea.

Some advantages of this new algorithm are: the possibility to allow non restricted starting points for each connected word, may be extended to other connected word recognition algorithms based on HMM modelling; it is not necessary to model the pauses between digits which is always more difficult and with a higher computational and memory cost (the number of models is less); the necessity of using a grammar (FSN) [4][5] to guide the search process doesn't exist. We think the new idea allows to deal the automatic recognition of vocabulary word sets in unconstrained speech problem without the need of using different garbages or sink models [5] (it would work like a mixed spotting-connected speech recognition algorithm).

DESCRIPTION OF THE ULB ALGORITHM

In this section we are going to describe the modifications included in the standard level-building (LB) algorithm, using the same notation than Rabiner & Levinson in reference [17], to ease the understanding of this changes. Further information can be found in [1][11][3].

The main modification affects the higher level initialization step; it consists of introducing the possibility that for the first state of a level the beginning of the path isn't restricted to the previous frame of the previous level (for the last state of the model). This detail can be seen in the following expressions, in particular in the first one:

$$\delta_t(1) = \max[\hat{P}(l-1,t-x),\ a_{11}^q \times \delta_{t-1}(1)] \times b_1^q(O_t), \quad 1 \le x \le maxskip,\ 2 \le t \le T \tag{1}$$

Equations (2) and (3) create the appropiate initial backpointer array, which records the frame at the previous level in which the previous word ended.

$$\alpha_t(1) = t - x_1, \quad \text{if } \hat{P}(l-1,t-x_1) > \hat{P}(l-1,t-x),\ 1 \le x \le maxskip,\ x \ne x_1, \text{ and } \hat{P}(l-1,t-x_1) > \delta_{t-1}(1) \times a_{11}^q \tag{2}$$

$$\alpha_t(1) = \alpha_{t-1}(1), \quad \text{if } \delta_{t-1}(1) \times a_{11}^q > \hat{P}(l-1,t-x),\ 1 \le x \le maxskip \tag{3}$$

where $\hat{P}$ is the level output best probability, and *maxskip* is a constant which must be estimated. The computational cost is closely related to this constant. Depending on the concrete applications, it might be convenient to penalize the skipping process to avoid too many frames from being ignored.

Besides, it is necessary to add new equations (4) and (5), which must be computed in the higher level inizialization step, which create the appropiate initial skippointer array λ, which records the skip length at the beginning of each higher level, where the previous word ended.

$$\lambda_t(1) = x_1, \quad \text{if } \hat{P}(l-1,t-x_1) > \hat{P}(l-1,t-x)\ 1 \le x \le maxskip,\ x \ne x_1 \text{ and } \hat{P}(l-1,t-x_1) > \delta_{t-1}(1) \times a_{11}^q \tag{4}$$

$$\lambda_t(1) = 1, \ \text{if } \delta_{t-1}(1) \times a_{11}^q > \hat{P}(l-1,t-x),\ 1 \le x \le maxskip \tag{5}$$

During the recursion (step 2 of LB), the skippointer is updated as:

$$\lambda_t(j) = \lambda_{t-1}[argmax_i(\delta_{t-1}(i) \times a_{ij}^q)],\ 1 \le i \le N \tag{6}$$

and at the end of the level, the probability, backpointer and skippointer arrays become:

$$P(l,t,q) = \delta_t(N),\ 1 \le t \le T$$

$$B(l,t,q) = alfa_t(N),\ 1 \le t \le T$$

$$D(l,t,q) = \lambda_t(N),\ 1 \le t \le T$$

Once of the word models have been run at any level, the reduced $\hat{P}$, $\hat{B}$, $\hat{W}$ and $\hat{D}$ arrays are formed using the equations described in [17], and the computation proceeds to the next level.

All this information is used to form the D matrix, which must be reduced as the P and B matrices. This D matrix stores the value of the occurred skips during the levels, and is necessary to normalize the accumulated probability along the path.

The next figure shows an example of the possibilities of the skipping process.

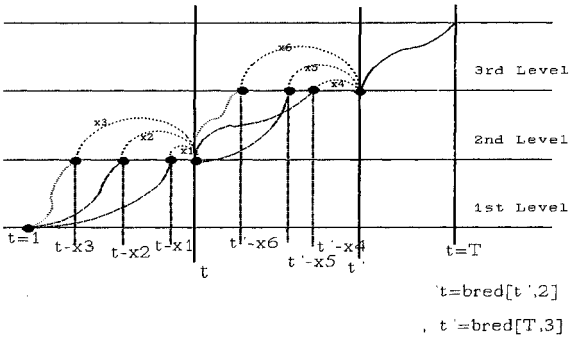

Figure 1. Details of the skipping process in ULB algorithm.

THE CONNECTED-DIGIT RECOGNITION SYSTEM

HMM characterization of digits

The following figure shows the topology of the HMM used to characterize individual digits and the beginning and ending silence models. The models are first order, left-to-right, discrete hidden Markov Models, with five states.

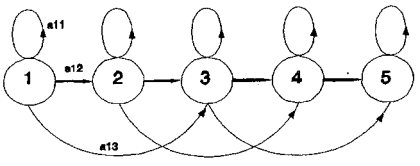

Figure 2. Hidden Markov model structure for digits.

We have also used this kind of model to train the inter-digit pauses, but with two different sizes (three states and five states). So, we could compare the skipping process idea vs. the modelling of these pauses using HMM models and a standard level-building algorithm with a digits-pauses grammar.

Recognition procedure

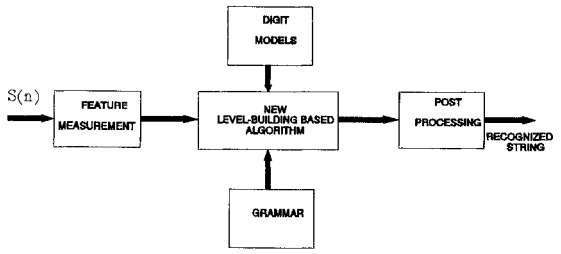

Figure 3. Bolck diagram of HMM/ULB connected digit recognizer.

A block diagram of the overall HMM recognizer is shown in the figure shown above. There are essentially three steps in the recognition algorithm namely:

1. Feature Extraction - The speech signal is bandlimited to 6.4 KHz and sampled at 16 KHz. The frame duration is 16 ms and overlapped Hamming windows are used in this analysis, with 6.25 ms displacement (9.75 ms overlap). Spectral analysis is performed to get 10 MFC coefficients. The frame energy is calculated and integrated with the MFCC to form vectors of 11 parameters that are quantized using a VQ codebook with 64 prototype vectors obtained using the Lloyd algorithm.

2. HMM pattern matching - the sequence of spectral vectors of the unknown speech signal is matched against a set of whole digit patterns (hidden Markov models) using a frame-synchronous ULB algorithm, and a very simple digit grammar that allows background noise to appear at the beginning and the end of each digit-string.

3. Postprocessor (optional) - The output candidate strings may be the subject of further validity tests, e.g., digit and/or state duration penalties are assessed, to eliminate (penalize) strings that are unreasonable with respect to duration constrains, and so on.

Training procedure

In order to build digit models from a training set of labelled connected digit strings, the first step is to optimally segment the connected digits into individual digits. For this task, a segmental k-means training procedure has been shown to be an effective way of converging at the optimum segmentation into digits [14]. A block diagram of the segmentation procedure is given in the following figure.

We assume an initial set of digit models is available. This initial set of models is obtained from an isolated digits database, which has been used to train a HMM for each isolated digit, using a Viterbi algorithm. If no initial models are available, the procedure can be bootstrapped by assuming an uniform segmentation of the connected digit strings into digits.

Given the initial digit model files, and the training files (which consist of connected digit strings of different lengths), an HMM pattern matching procedure is used to optimally segment the training digit strings into individual digit tokens which are stored in digit token files. A digit model building algorithm (i.e., an estimation procedure for determining parameters of the digit HMM's) based on Viterbi algorithm is used to give an updated set of digit models. The above procedure is iterated until the difference in likelihood scores of the digit models, in consecutive iterations, is sufficiently small.

SPEECH DATA BASE

To evaluate the performance of the connected-digit recognizer, in a speaker-trained mode, a small data base was recorded. It consists of one male speaker, who recorded 129 connected-digit strings with lengths ranging from 1 to 5, and 61 isolated digits to train the initial HMM seed-models. The string length was randomized between 1 and 5 digits with, on average, an equal number of occurrences of strings of each length. Similarly, the digits within each string were chosen at random, but, on average, there is an equal number of occurrences of each digit in each string length.

This data base was divided into three different sets, with the following distribution:

Number of	Training set	Testing set	Evaluation set
Strings	64	19	46
Digits	231	63	156

EXPERIMENTAL EVALUATION OF THE ULB RECOGNIZER. COMPARISON WITH OTHER ALTERNATIVES

To evaluate the new proposed algorithm and compare it with other alternative methods, it has been necessary to develop two systems, based on the level-building algorithm, and run the same experiments, in the same conditions, with both of them.

These experiments have been carried out using the speaker-trained database described in a previous section. No further information about the length of the digit strings to be recognized was given to the system.

The results of the experiments to evaluate the new algorithm, and compare it with the other ways of dealing with the pauses between digits problem that we have considered, are presented in the figure below:

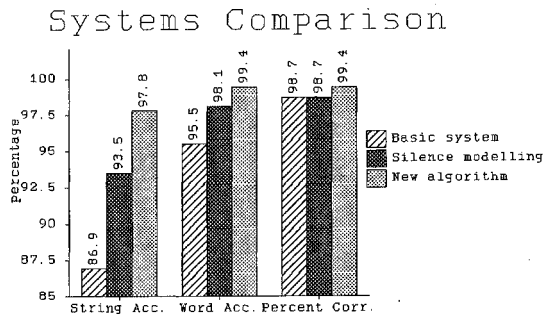

Figure 4. Results of the systems comparison.

Three systems have been compared:

1. Basic system: this system doesn't include any way of treating the inter-digit pauses.

2. Silence modelling system: different HMM models have been used to represent the inter-digit silences. The best results, which can be seen in the figure above, are obtained using a 3-state discrete first order, left to right hidden Markov model. A grammar has to be used with this system, to allow one silence model to be inserted between each two digits.

3. Unconstrained Level Building (ULB) system: this system includes the modified level-building algorithm described before, which uses the skipping procedure to avoid the inter-digit pauses modelling. No grammar is needed with this system.

As can be seen, the improvement obtained on String Accuracy using the ULB algorithm instead of a Level Building with an HMM for the silence is 4.3%, and without any silence modelling is nearly 11%.

SUMMARY

In this paper we have shown that a very high performance connected-digit realistic recognition system can be implemented without the need of modelling (with HMM) extraneous and non-desired sounds.

In spite of the reduced size of the used data base, the results of the experiments using the new idea in order to implement word-spotting systems.

At present, this algorithm is being tested on a speaker independent connected digit Spanish data base.

MAIN REFERENCES

[1] Myers, C.S., Rabiner, L.R., "Connected Digit Recognition Using a Level-Building DTW Algorithm". IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-29, no.3, June 1981.

[2] Rabiner, L.R., Wilpon, J.G., Juang, B.H., "A model-based connected-digit recognition system using either hidden Markov models or templates". Computer Speech and Languages (1986) 1, 167-197.

[3] Rabiner, L.R., Wilpon, J.G., Soong, F.K., "High Performance Connected Digit Recognition Using Hidden Markov Models". IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 8, August 1989.

[4] Lee, C.H., Rabiner, L.R., "A Frame-Synchronous Network Search Algorithm for Connected Word Recognition". IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no.11, November 1989.

[5] Wilpon, J.G., Lee, C.H., Rabiner, L.R., "Application of Hidden Markov Models for Recognition of a Limited Set of Vocabulary Words in Unconstrained Speech", Proceedings of ICASSP 89, Glasglow, Scotland, 1989.

[6] Rabiner, L.R.,Wilpon, J.G., Juang, B.H., "A Performance Evaluation of a Connected Digit Recognizer",IEEE, 1987, pag. 101-104.

[7] Hunt, M.J.,"Figure of Merit for Assessing Connected-Word Recognizers", Speech Communication 9 (1990), pag. 329-336.

[8] Sakoe, H., "Two level DP-Matching - A Dynamic Programming Based Pattern Matching Algorithm for Connected Word Recognition", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-27, No.6,December 1979, pag. 588-595.

[9] Bridle, J.S., "An Algorithm for Connected Word Recognition", Automatic Speech Analysis and Recognition, edit. J. P. Haton, D. Riddle Publishing Co., Holland 1982.

[10] Rabiner, L.R., Wilpon, J.G., Juang, B.H.,"A Segmental k-means Training Procedure for Connected Word Recognition", ATT Technical J.,Vol. 65, iss. 3, May/June 1986.

[11] Rabiner, L.R., Wilpon, J.G., Soong, F.K.,"High Performance Connected Digit Recognition using hidden Markov Models", ICASSP S3.6, 1988.

[12] Ramesh, P., Wilpon, J.G., McGee, M.A., Roe, D.B., Lee, C.H., Rabiner, L.R., "Speaker Independent Recognition of Spontaneously Spoken Connected Digits", ICASSP 1991.

[13] Kopec, G.E, Bush, M.A., "Network-Based Isolated Digit Recognition Using Vector Quantization", IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-33, No. 4, August 1985.

[14] Cifuentes, S., "Connected-Digit Speaker-Dependent Recognition", Internal Report, Electronic Engineering Department, E.T.S.I. Telecomunicación, Universidad Politécnica Madrid, Feb. 1992.

[15] Davis, S. B. and Mermelstein, "Comparison of parametric representations for monosyllabyc word recognition in continuously spoken sentences", P. in IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-28, no. 4, pp 357-365, 1980.

[16] Gray, R.M., "Vector Quantization", IEE ASSP Magazine 1(2):4-29, April, 1984.

[17] Rabiner, L. R., Levinson, S. E. "A Speaker Independent, Sintax-Directed, Connected Word Recognition System based on Hidden Markov Models and Level Building". IEEE Transactions on Acoustics, Speech, and Signal Proccssing, Vol. ASSP-33, no. 3, June 1985.