L. Torres, E. Masgrau, and M.A. Lagunas (eds.)

RECOGNITION OF NUMBERS BY USING DEMISYLLABLES AND HIDDEN MARKOV MODELS

J.B. Mariño, A. Bonafonte, A. Moreno, E. Lleida, C. Nadeu, E. Monte

Department of Signal Theory and Communications
Politechnic University of Catalonia. Spain.

Abstract

A continous speech recognition system (called RAMSES) has been built based on the demisyllable as phonetic unit and tools from connected speech recognition. Speech is parameterized by band-pass lifted LPC-cepstra and demisyllables are represented by hidden Markov models (HMM). In this paper, the application of this system to recognize integer numbers from zero to one thousand is described. The paper contains a general overview of the system, a description of the HMM training procedure and an assessment on the recognition performance in a speaker independent experiment.

1. INTRODUCTION

During the last two years, a continous speech recognition system based on demisyllables and discrete hidden Markov models (HMM) has been built in our laboratory. Demisyllables afford a convenient phonetic coding of Spanish utterances, according to the syllabic character of this language. Hidden Markov models have been shown to be a successful tool for describing in a probabilistic way the acoustic features of speech. Our system has been called RAMSES, Spanish acronym for "automatic recognition by means of semisyllables (demisyllables)". In this paper we provide a general overview of RAMSES and report its application to recognize the Spanish integer numbers from zero to one thousand, in both multispeaker and speaker independent tasks.

The paper is organized in the following way: in Section 2 the block-diagram of RAMSES is described, Section 3 addresses the task oriented aspects, in Section 4 the HMM training procedure is outlined, Section 5 is dedicated to report the recognition experiment results, and finally Section 6 contains the main conclusions.

2. RAMSES' OVERVIEW

Figure 1 shows a general block-diagram of the system architecture. The speech signal is band-pass (100 Hz - 3400 Hz) filtered by an antialiasing filter and sampled at 8 kHz. The utterance is isolated by an end-point detection algorithm and pre-emphasized. A linear prediction (LP) based parameterization follows: the signal is segmented into frames of 30 milliseconds by a Hamming window at a rate of 15 milliseconds, and every frame is characterized by a LP-filter with 8 coefficients. Afterwards, 12 band-pass lifted cepstrum coefficients are computed [1]; the energy of the frame completes the parameterization. Before entering the recognition algorithm, the system evaluates the spectral difference d(t) corresponding to the frame t by using [2]:

$$d(t) = \sum_{k=-2}^{2} k\, s(t+k)$$

where s(t) is the ceptstral vector in frame t. This difference implies a time-average along 90 milliseconds. In a similar way, the energy difference e(t) is calculated. The spectral vector and the spectral and energy differences are vector-quantized separately; in that way, every frame of speech signal is represented by three symbols.

According to the most recent proposals, RAMSES considers energy and time evolution information. However, in our system, the energy is not used directly as a parameter of the signal. This is because the energy depends on the prosody of the sentence and the intensity of the utterance, two very fluctuant features of speech. On the contrary, if the energy is expressed by a logarithmic measure, its difference does not vary with a change in the intensity of the overall sentence, and the variation due to prosodic effects is greatly alleviated.

This work was supported by the PRONTIC grant number 105/88

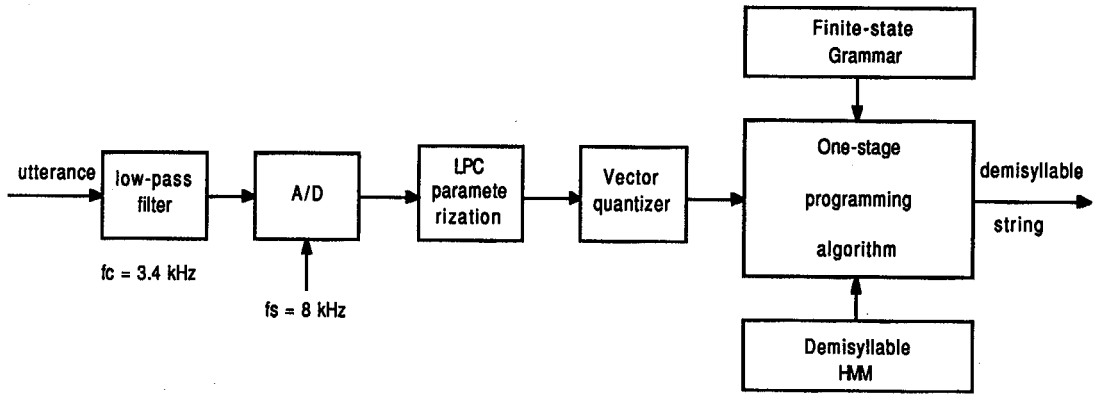

Figure 1.- Recognition System architecture

The recognition algorithm performs an one-stage dynamic programming (described for connected word recognition [3]) driven by a finite state grammar. So, the algorithm computes the string of demisyllable models that provides the most likely path of states throughout the utterance and, at the same time, satisfies the grammar constraints. If necessary, a dictionary provides the semantic meaning of the issued sequence of demisyllables.

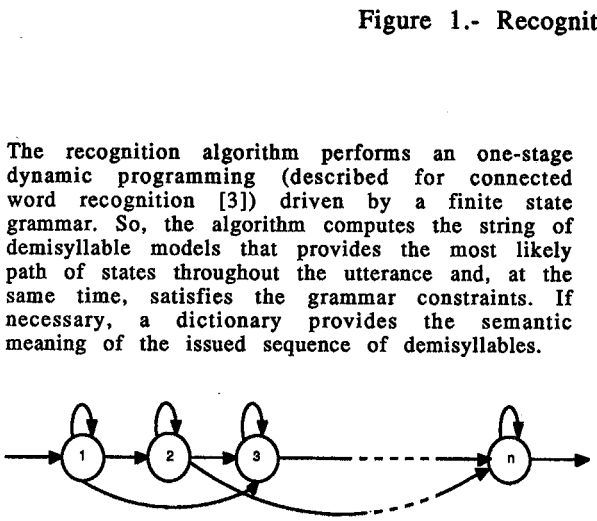

Figure 2.- HMM structure

Figure 2 shows the structure used for the hidden Markov models. It is a typical left-to- right structure, that allows to skip one state when the model makes a transition between states. The emission of symbols is associated to the states, that issue three independent symbols (spectrum, spectrum difference and energy difference) when they are visited. The number n of states is a parameter to be determined. During the recognition task, the transition probability between the final state of a model M_i to the first state of the following model M_{i+1} is determined by the duration probability of the demisyllable modeled by M_i. The length probability of a demisyllable is parameterized by the mean and the variance of a gaussian distribution.

This general architecture can be oriented to a specific application by designing the regular grammar and training the demisyllable models.

3. NUMBER RECOGNITION TASK

The recognition of numbers is an interesting task to try a recognition system. Besides its practical interest, this application exhibits an inherent difficulty due to the subtle acoustic differences that can separate very distinct semantic meanings.

In a previous experiment on number recognition in a speaker dependent environment, the set of necessary demisyllables for this application was stablished. This set, including 67 demisyllables, was designed in order to cope the most usual realizations for numbers issued by Spanish speakers. In that same experiment, the finite state grammar describing the numbers in terms of demisyllable strings was built; the number of states of this grammar was 118. Details on demisyllable definition and grammar inference can be found in [4] and [5].

From ten speakers (6 male and 4 female) a speech data base was acquired in our laboratory. Every speaker uttered one realization of a set of 44 numbers, designed in such a way that: a) every set included at least two samples of each necessary demisyllable in the application, and b) the 44 numbers performed a suitable sampling of the integers from zero to one thousand. The articulation rate of speech spanned from 5 to 7 syllables per second. This data base was segmented by hand into demisyllables and labeled.

4. DISCRETE HIDDEN MARKOV MODEL TRAINING

Demisyllable models were trained following the procedure outlined in Figure 3. Each model was trained independently of the others. Let D_i be the demisyllable which model has to be trained. Every

```
for every demisyllable Di
   - for every speaker
      - collect the samples of Di
      - if the number of samples > 5
         - perform a k-means clustering
         - select 5 representants of Di
      - end if
   - end for
   - train HMM by Baum-Welch algorithm
   - smooth HMM
end for
```

Figure 3.- HMM training algorithm

sample of D_i was collected from the utterances recorded by the first speaker; if the number of samples surpass 5, the 5 most representative samples were selected by a k-means clustering procedure [6]. This strategy aimed to prevent a very dissimilar training for demisyllables with a number of representants very different. Once the samples from every speaker were obtained, the Baum-Welch estimation algorithm was applied. At the same time, the mean and the variance of demisyllable length was computed. Finally, the demisyllable models were smoothed according to the co-occurrence probability method introduced in [7].

Previously to apply this procedure, the values for some important parameters of the models had to be determined, i.e., the size of the three codebooks and the number of states. In order to assess the choice, some training and recognition experiments were accomplished. Specifically, the value for those parameters were fixed and the models were trained with the ten speakers; afterwards, the signals in the data base were recognized. Then the parameter values were modified, and the training and recognition procedures were carried out; and so on. As a result of these trials, we drew the following conclusions:

a) as far as the size of the codebooks is concerned, the most suitable choices are: 64 for the two codebooks dedicated to spectral information and 32 for the codebook devoted to energy differences. Although similar performance can be got with other parameter selections, this option requires the minimum codebook sizes.

b) the recognition performance is noticeably dependent on the number of states of the hidden Markow models. Several criteria to determine the most suitable number of states for every model were tested: equal number of states, number of states according to the number of sounds included in the demisyllable, and number of states as a function of the average length of the demisyllable. This third criterium yielded the best performance for almost every experiment carried out, and when it did not lead to the best choice, it afforded a performance near the optimum. As a consequence, we used this criterion in our finaldesign. In Table 1 the definition of the average length criterion is provided.

average length in frames	number of states
≤ 4	2
5,6	3
7,8	4
9,10	5
>10	6

Table 1.- Criterion to select the number of states of HMM as a function of the average length of demisyllables

5. SPEAKER INDEPENDENT EXPERIMENT

Although we acknowledge that our data base is rather reduced, we were interested in carrying out some experiments that allowed to ascertain the ability of RAMSES to cope with speaker independent tasks. To this aim, we made six different training and recognition trials. In each experiment we trained the system with 8 speakers, and then we recognized the speech signals of the other two; in every case, the outside training speakers were taken with different sex. Table 2 shows the six couples of outside training speakers.

M1 - F1
M2 - F2
M3 - F3
M4 - F4
M5 - F1
M6 - F2

Table 2.- The couples of outside training speakers

Table 3 provides the recognition error percentage achieved for every speaker, when he or she was inside and outside the training set. The results before and after the smoothing of the HMM output probabilities is also shown. We count as one error every number recognized incorrectly, indepently of the number of demisyllables misrecognized. It is worth mentioning that, in most of the cases, the errors affected only one digit in the number (corresponding either the hundreds, or the tenths or the units); for instance, 677 was recognized as 637, or 721 as 621.

	before smoothing		after smoothing	
	trained	not trained	trained	not trained
M1*	0.4	0.7	0.0	0.0
M2	2.3	6.8	2.7	6.8
M3	1.4	4.5	3.2	4.5
M4**	0.0	12.1	2.7	6.9
M5	0.0	27.3	6.4	11.4
M6	0.0	2.3	0.5	0.0
F1	0.0	9.0	0.0	3.4
F2	0.0	3.4	0.0	0.0
F3	0.0	4.5	2.3	2.3
F4	0.0	0.0	0.0	0.0
M	0.6	7.0	1.9	3.8
F	0.0	4.9	0.6	1.5
Total	0.4	6.1	1.5	2.8

The total number of utterances of this speaker is: * 136, ** 58

Table 3. Recognition error percentage

From Table 3 we can observe the following facts:

a) the smoothing afforded a remarkable decreasing of recognition errors outside the traininig set; however, the prize to be paid was an increasing of recognition errors inside the training set.

b) For some speakers (for instance, M4, M5 and F1) the RAMSES recognition ability was very different when the speaker was either inside or outside the training set.

c) The recognition performance fluctuated greatly from some speakers to others. This behaviour is much more evident for the speakers outside the training set.

d) The average performance (2.8% of error percentage) was satisfactory.

6. CONCLUSION

Our interpretation of the enunciated observations is twofold. Firstly, RAMSES seems suitable for recognizing number in continous speech, either in a multispeaker task (every speaker inside the training set) or in a speaker independent application (all of speakers outside the training set); and secondly, the data base required for training this latter case must be increased.

Currently, we are recording a new data base, involving 20 new speakers and utterances with strings of integer numbers from zero to one million.

REFERENCES

[1].- B. H. Juang et al.,"On the use of bandpass liftering in speech recognition", IEEE Trans ASSP-35, pp. 947-954: July, 1987

[2].- S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum", IEEE Trans ASSP-34, pp. 52-59: February, 1986

[3].- H. Ney, "The use of an one-stage dynamic programming algorithm for connected word recognition", IEEE Trans ASSP-32, pp. 263-271: April, 1984

[4].- J. B. Mariño et al., "Finite state grammar inference for connected word recognition", Proc. EUSIPCO'88, pp. 1035-1038: September, 1988

[5].- J. B. Mariño et al., "Recognition of numbers and strings of numbers by using demisyllables: one speaker experiment", Proc. EUROSPEECH'89 vol. 1, pp. 102-105: September, 1989

[6].- J. Wilpon, L. R. Rabiner, "A modified k-means clustering algorithm for use in isolated word recognition", IEEE Trans ASSP-23, pp. 587-594: June, 1985

[7].- K. -F. Lee and H. -W. Hon, "Speaker-independent phone recognition using hidden Markov models", IEEE Trans ASSP-37, pp. 1641-1648: November, 1989