Speech Recognition Using Energy Parameters to Classify Syllables in the Spanish Language

Sergio Suárez Guerra, José Luis Oropeza Rodríguez,
Edgardo M. Felipe Riveron, and Jesús Figueroa Nazuno

Computing Research Center, National Polytechnic Institute,
Juan de Dios Batiz s/n, P.O. 07038, Mexico
{ssuarez, edgardo, jfn}@cic.ipn.mx, j_orope@yahoo.com.mx

Abstract. This paper presents an approach for the automatic speech recognition using syllabic units. Its segmentation is based on using the Short-Term Total Energy Function (STTEF) and the Energy Function of the High Frequency (ERO parameter) higher than 3,5 KHz of the speech signal. Training for the classification of the syllables is based on ten related Spanish language rules for syllable splitting. Recognition is based on a Continuous Density Hidden Markov Models and the bigram model language. The approach was tested using two voice corpus of natural speech, one constructed for researching in our laboratory (experimental) and the other one, the corpus Latino40 commonly used in speech researches. The use of ERO parameter increases speech recognition by 5% when compared with recognition using STTEF in discontinuous speech and improved more than 1.5% in continuous speech with three states. When the number of states is incremented to five, the recognition rate is improved proportionally to 97.5% for the discontinuous speech and to 80.5% for the continuous one.

1 Introduction

Using the syllable as the information unit for automatic segmentation applied to Portuguese improved the error rate in word recognition, as reported by [1]. It provides the framework for incorporating the syllable in Spanish language recognition because both languages, Spanish and Portuguese, have as a common characteristic well structured syllable content [2].

The dynamic nature of the speech signal is generally analyzed by means of characteristic models. Segmentation-based systems offer the potential for integrating the dynamics of speech at the phoneme boundaries. This capability of the phonemes is reflected in the syllables, like it has been demonstrated in [3].

As in many other languages, the syllabic units in Spanish are defined by rules (10 in total), which establish 17 distinct syllabic structures. In this paper the following acronyms are used: Consonant – C, Vocal – V; thus, the syllabic structures are formed as CV, VV, CCVCC, etc.

The use of syllabic units is motivated by:

M. Lazo and A. Sanfeliu (Eds.): CIARP 2005, LNCS 3773, pp. 161 – 170, 2005.

- A more perceptual model and better meaning of the speech signal.
- A better framework when dynamic modeling techniques are incorporated into a speech recognition system [4].
- Advantages of using sub words (i.e. phonemes, syllables, triphones, etc) into speech recognition tasks [5]. Phonemes are linguistically well defined; the number of them is little (27 in the Spanish language) [6]. However, syllables serve as naturally motivated minimal units of prosodic organization and for the manipulation of utterances [7]. Furthermore, the syllable has been defined as "a sequence of speech sounds having a maximum or peak of inherent sonority (that is apart from factors such as stress and voice pitched) between two minima of sonority" [8]. The triphones treat the co-articulation problem to segment words structure as a more useful method not only in Spanish language. The triphones, like the syllables, are going to be nowadays as a good alternative for the speech recognition [5].

The use of syllables has several potential benefits. First, syllabic boundaries are more precisely defined than phonetic segment boundaries in both speech waveforms and in spectrographic displays. Second, the syllable may serve as a natural organizational unit useful for reducing redundant computation and storage [4].

There are not antecedents of speech recognition systems using the syllables rules in the training system for the Spanish language. Table 1 lists the frequencies of occurrence of ten monosyllables used in corpus Latino40 and its percentage in the vocabulary. Table 2 shows the percentage of several syllabic structures in corpus Latino40. Both tables show the behavior of the syllables units for this corpus.

Table 1. Frequency of occurrence of ten monosyllables used in corpus Latino40

Word	Syllable configuration	Number of times	% in the vocabulary
De	Deaf Occlusive + Vocal	1760	11.15
La	Liquid + Vocal	1481	9.38
El	Vocal + Liquid	1396	8.85
En	Vocal + Nasal	1061	6.72
No	Nasal + Vocal	1000	6.33
Se	Fricative + Vocal	915	5.80
Que	Deaf Occlusive + Vocal	891	5.64
A	Vocal	784	4.97
Los	Liquid + Vocal + Fricative	580	3.67
Es	Vocal + Fricative	498	3.15

Table 2. Percentage of several syllabic structures in corpus Latino40

Syllable structure	Vocabulary Rate (%)	Accumulated in the vocabulary (%)
CV	50.72	50.72
CVC	23.67	74.39
V	5.81	80.2
CCV	5.13	85.33
VC	4.81	90.14
CVV	4.57	94.71
CVVC	1.09	95.8

2 Continuous Speech Recognition Using Syllables

In automatic speech recognition research (ASR) the characteristics of each basic phonetic unit in a large extent are modified by co-articulation. As a result, the phonetic features found in articulated continuous speech, and the phonetic features found in isolated speech, have different characteristics. Using the syllables the problem is the same, but in our approach the syllables were directly extracted from the speech waveform, whose grammatical solution were found later using a dedicated expert system. Figure 1 shows the result of the segmentation using STTEF [3].

It can be noted that the energy is more significant when the syllable is present and it is a minimum when it is not. The resulting relative minimum and maximum energy are used as the potential syllabic boundaries. The term syllabic unit is introduced to differentiate between the syllables defined generally on the phonological level and the syllabic segments.

Thus, each syllable can be independently stored in a file. Our database uses 10 phrases with 51 different syllables. For each phrase 20 utterances were used, 50% for training and the remainder for recognition, and there were produced by a single female speaker at a moderate speaking rate.

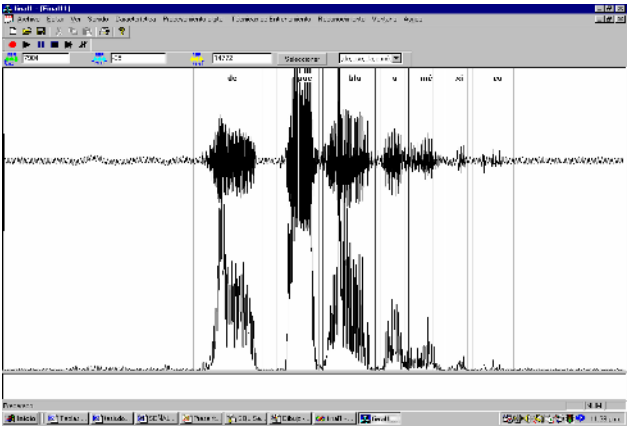

Fig. 1. Syllables speech segmentation labeling

3 Training Speech Model Using Data Segments

The Energy Function of the High Frequency (ERO parameter) is the energy level of the speech signal at high frequencies. The fricative letter, s, is the most significant example. When we use a high-pass filter, we obtain the speech signal above a given cut-off frequency f_c, the RO signal. In our approach, a cut-off frequency f_c = 3500 Hz is used as the threshold frequency for obtaining the RO signal. The speech signal at a lower frequency is attenuated. Afterwards, the energy is calculated from the Equation

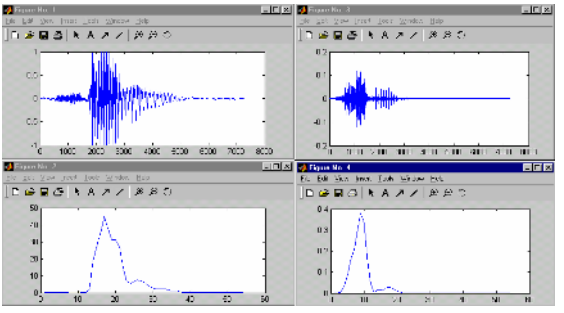

Fig. 2. STTEF (*left*) and ERO (*right*) parameters in the Spanish word 'cero'

(1) for the ERO parameter in each segment of the resultant RO signal. Figure 2 shows graphically the results of this procedure for Short-Term Total Energy Function (STTEF) and ERO parameter in the case of the word 'cero'.

$$ERO = \sum_{i=0}^{N-1} ROi^2 \tag{1}$$

Figure 3 shows the energy distribution for ten different words 'cero' spoken by the same speaker. We found an additional area between the two syllables (ce-ro) using our analysis. In the figure, the dark gray rectangle represents the energy before using the filter, ERO; a medium gray rectangle the energy of the signal after using the filter, STTEF; and a light gray rectangle represents the transition region between both parameters. We call this region the Transition Energy Region -RO.

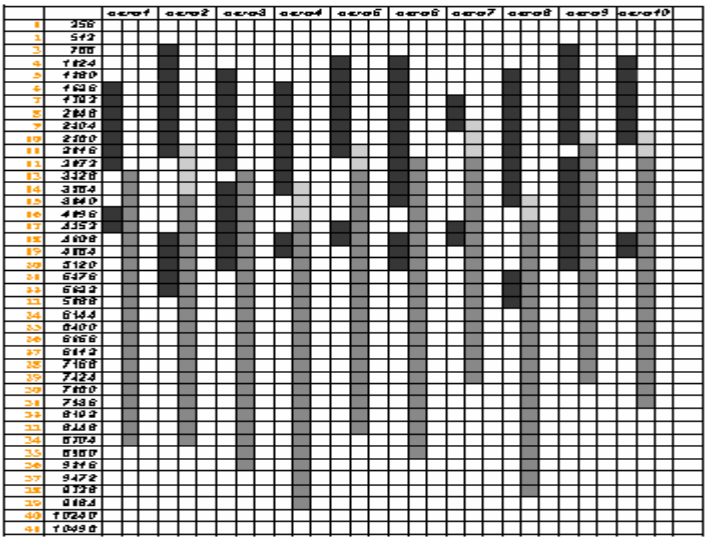

Fig. 3. Energy distribution for ten different words 'cero'

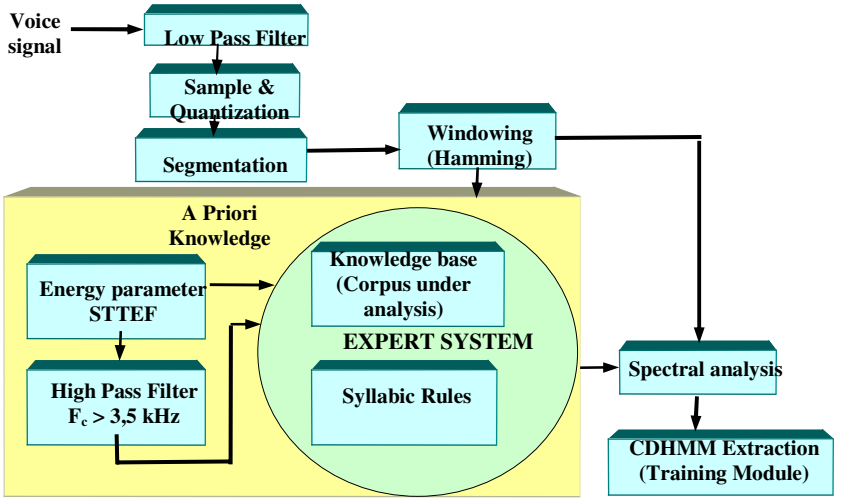

Fig. 4. Functional block diagram for syllable splitting

Figure 4 shows the functional block diagram representing the algorithm used in our approach to extract the signal characteristics.

In the training phase an expert system uses the ten rules for syllable splitting in Spanish. It receives the energy components STTEF and the ERO parameter extracted from the speech signal. Table 3 shows the basic sets in Spanish used by the expert system for the syllable splitting. Table 4 shows the inference rules created in the expert system, associated with the rules for splitting words in syllables.

The rules mentioned above are the postulates used by the recognition system. Syllable splitting is carried out taking into account the spectrogram shape, parameters and the statistics from the expert system. Figure 5 shows graphically the decision trees of the inference rules of the expert system.

After the execution by the expert system and for the voice corpus in process of the entire syllable splitting inference rules, the results are sent to the Training Module as the initial parameters. Then, the necessary models are created for each syllable during the process of recognition.

Table 3. Basic sets in Spanish used during the syllable splitting

CI = {br,bl,cr,cl,dr,fr,fl,gr,gl,kr,ll,pr,pl,tr,rr,ch,tl}	Non-separable Consonant
VD={ai,au,ei,eu,io,ou,ia,ua,ie,ue,oi,uo,ui,iu,ay,ey,oy}	Vocal Diphthong and hiatus
VA={a}	Open Vocal
VS={e,o}	Half-open Vocal
VC={i,u}	Close Vocal
CC={ll.rr,ch}	Compound Consonant
CS={b,c,d,f,g,h,j,k,l,m,n,ñ,p,q,r,s,t,v,w,x,y,z}	Simple Consonant
VT={iai, iei, uai, uei, uau, iau, uay, uey}	Vocal Triphthong and hiatus

Table 4. Inference rules of the expert system

Inference rules		
If CC ^ CC ∈ CI	→	/CC/
If VCV	→	/V/ /CV/
If VCCV	→	/VC/ /CV/
If VCCCV	→	/VCC/ /CV/
If C1C2 ^ C1='h' or C2='h'	→	/C1/ /C2/
If VV ∉ VA, VS	→	/VV/
If VV ∈ VA, VS	→	/V/ /V/
If VCV with C='h'	→	/VCV/
If V1V2 any with accent	→	/V1/ /V2/
If VVV ϵ VT	→	/VVV/

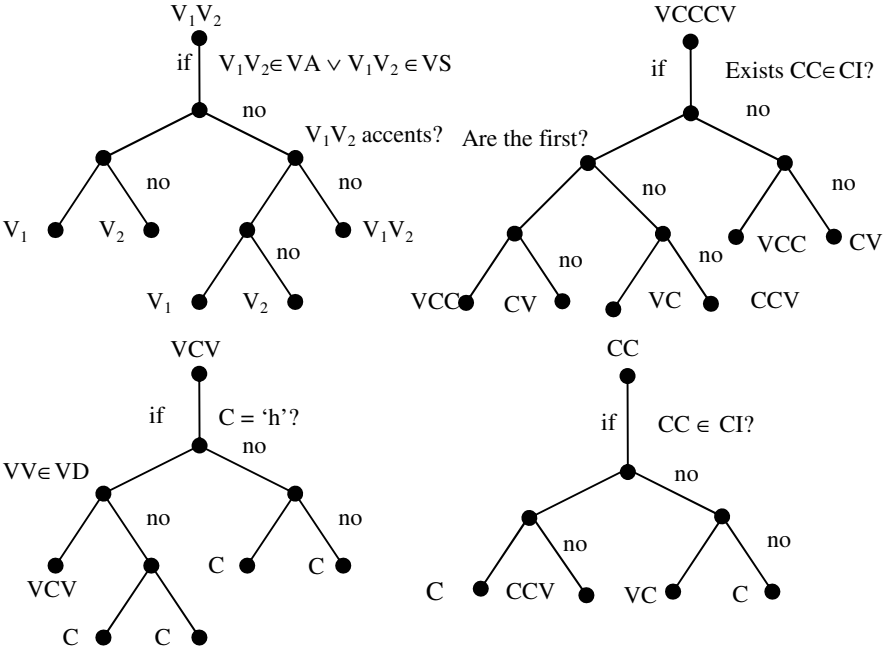

Fig. 5. Decision trees for the inference rules created in the expert system

During the recognition phase, the Recognition Module receives the Cepstral Linear Prediction Coefficients from the signal in processes. They are used to calculate the probabilities of each element in the corpus. The recognized element is that with a higher probability. The final result of this process is the entire speech recognition.

4 Model for Continuous Speech Recognition

In our approach, speech recognition is based on a Hidden Markov Model (HMM) with Continuous Density and the bigram [5] like a language model described by Equation (2).

$$P(W) = P(w_1)\prod_{i=2}^{N} P(w_i \mid w_{i-1}) \tag{2}$$

Where W represents the words in the phrase under analysis w_1 on the corpus; w_i represents a word in the corpus; P (W) is the probability of the language model; $P(w_i)$ is the probability of a given word in the corpus. In automatic speech recognition it is common to use expression (3) to achieve better performance:

$$\overset{*}{W} = \arg\max\left[P(O \mid W)P(W)\right] \tag{3}$$

Here, $\overset{*}{W}$ represents the word string, based on the acoustic observation sequence, so that the decoded string has the maximum a posteriori probability *P* (*O*|*W*), called the acoustic model.

Language models require the estimation of a priori probability *P(W)* of the word sequence $w = w_1 + w_2 + \cdots + w_N$. *P(W)* can be factorized as the following conditional probability:

$$P(W) = P(w_1 + w_2 + \cdots + w_N) = P(w_1)\sum_{i=1}^{N} P(w_i \mid w_{i-1}) \tag{4}$$

The estimation of such a large set of probabilities from a finite set of training data is not feasible.

The bigram model is based on the approximation based on the fact that a word only depends statistically on the temporally previous word. In the bigram model shown by the equation (2), the probability of the word $w^{(m)}$ at the generic time index *i* when the previous word is $w^{(m^1)}$ is given by:

$$\hat{P}(w_i = w^{(m)} \mid w_{i-1} = w^{m^1}) = \frac{N(w_i = w^{(m)} \mid w_{i-1} = w^{(m^1)})}{N(w^{m^1})} \tag{5}$$

where the numerator is the number of occurrences of the sequence $\left\langle w_i = w^{(m)}, w_{i-1} = w^{(m^1)} \right\rangle$ in the training set.

5 Experiments and Results

Taking into account the small redundancy of syllables in the corpus Latino40, we have designed a new experimental corpus with more redundant syllables units, prepared by two women and three men, repeating ten phrases twenty times each to give one thousand phrases in total.

Table 5 shows the syllables and the number of times each one appear in phrases of our experimental corpus.

Table 5. Syllables and the number of each type into our experimental corpus

Syllable	#Items	Syllable	#Items	Syllable	#Items
de	2	es	3	zo	1
Pue	1	pa	2	rios	1
bla	1	cio	1	bio	1
a	5	e	2	lo	1
Me	1	o	1	gi	1
xi	1	ahu	1	cos	1
co	1	ma	2	el	1
cuauh	1	do	1	true	1
te	1	cro	1	que	1
moc	1	cia	1	ri	2
y	1	ta	1	ti	1
cuau	2	en	1	lla	1
tla	2	eu	1	se	2
mo	2	ro	1	ria	1
re	2	pro	1	po	1
los	1	to	1	si	1
ble	1	sis	1	tir	1

Table 6. Percentage of discontinuous recognition

Segmentation	Hidden Markov (%) with 3 states	Models states (%) with 5 states
STTEF	89.5	95.5
STTEF + ERO	95.0	97.5

Table 7. Percentage of continuous recognition

Segmentation	Hidden Markov(%) with 3 states	Models states (%) with 5 states
STTEF	77.5	78.5
STTEF + ERO	79.0	80.5

Three Gaussian mixtures were used for each state in the HMM with three and five states, using twelve Cepstral Linear Prediction Coefficients (CLPCs). Tables 6 and 7 show the results of recognition for the discontinuous and continuous cases, respectively, referred to the experimental corpus. The accentuation of Spanish words was not considered in the analysis.

6 Conclusion

The results shown in this paper demonstrate that we can use the syllables as an alternative to the phonemes in an automatic speech recognition system (ASRS) for the Spanish language. The use of syllables for speech recognition avoids the contextual dependency found when phonemes are used.

In our approach we used a new parameter: the Energy Function of the Cepstral High Frequency parameter, ERO. The incorporation of a training module as an expert system using the STTEF and the ERO parameter, taking into account the ten rules for syllable splitting in Spanish, improved considerably the percent of success in speech recognition. The use of the ERO parameter increased by 5% the speech recognition with respect to the use of STTEF in discontinuous speech and by more than 1.5% in continuous speech with three states. When the number of states was incremented to five, the improvement in the recognition was increased to 97.5% for discontinuous speech and to 80.5% for continuous speech.

CLPCs and CDHMMs were used for training and recognition, respectively.

It was also demonstrated that comparing our results with [9], for English, we obtained a better percent in the number of syllables recognized when our new alternative for modeling the ASRS was used for the Spanish language.

The improvement of the results shows that the use of expert systems or conceptual dependency [10] is relevant in speech recognition of the Spanish language when syllables are used as the basic features for recognition.

References

1. Meneido H., Neto J. Combination of Acoustic Models in Continuous Speech Recognition Hybrid Systems, INESC, Rua Alves Redol, 9, 1000- 029 Lisbon, Portugal. 2000.
2. Meneido, H. Joâo P. Neto, J., and Luis B. Almeida, L., INESC-IST. Syllable Onset Detection Applied to the Portuguese Language. 6th European Conference on Speech Communication and Technology (EUROSPEECH'99) Budapest, Hungary, September 5-9. 1999.
3. Suárez, S., Oropeza, J.L., Suso, K., del Villar, M., Pruebas y validación de un sistema de reconocimiento del habla basado en sílabas con un vocabulario pequeño. Congreso Internacional de Computación CIC2003. México, D.F. 2003.
4. Su-Lin Wu, Michael L. Shire, Steven Greenberg, Nelson Morgan., Integrating Syllable Boundary Information into Speech Recognition. Proc. ICASSP, 1998.
5. Rabiner, L. and Juang, B-H., Fundamentals of Speech Recognition, Prentice Hall
6. Serridge, B., 1998. Análisis del Español Mexicano, para la construcción de un sistema de reconocimiento de dicho lenguaje. Grupo TLATOA, UDLA, Puebla, México. 1993.

7. Fujimura, O., UCI Working Papers in Linguistics, Volume 2, Proceedings of the South Western Optimality Theory Workshop (SWOT II), Syllable Structure Constraints, a C/D Model Perspective. 1996.
8. Wu, S., Incorporating information from syllable-length time scales into automatic speech recognition. PhD Thesis, Berkeley University, California. 1998.
9. Bilmes, J.A. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models, International Computer Science Institute, Berkeley, CA. 1998.
10. Savage Carmona Jesus, A Hybrid System with Symbolic AI and Statistical Methods for Speech Recognition, Doctoral Thesis, University of Washington. 1995.