Continuous Speech Recognition using Loudness and RO Parameter Segmentation into Syllables Units

José Luis Oropeza Rodríguez[1]. Sergio Suárez Guerra[1]

[1]Center for Computing Research.
Digital Systems Department
Polytechnic National Institute.
j_orope2002@hotmail.com, ssuarez@cic.ipn.mx

Abstract. In this paper we present a new approach for the automatic speech recognition, where segmentation task was made using the Short-Term Total Energy Function (STEF) and the energy function of the Cepstral High Frequency, more than 4 Khz. (RO parameter) of the speech signal. The recognition was made using the Continuous Density Hidden Markov Models (CDHMM). In resent years, the syllables represent one attractive paradigm units in the speech recognition [5, 9]. Using STEF and RO parameter, the speech signal is segmented at boundaries of syllabic units. We used a small size corpus of natural-speech, construed for research in our laboratory. This experiment involves the partition of the waveform into distinct consecutive non-overlapping syllables and the association of a set of rules for the Spanish language. We extracted 52 different syllables and achieved good accuracy in the recognition. We analyzed the corpus with CDHMM composed of a sequence of three and four states for each syllable. Also, we used three Gaussian Mixtures for each state.

1 Introduction

It is well known that the dynamic speech nature and production systems must be obtained by characteristic model. Segmented-based systems offer the potential to integrate the dynamics of speech, at least the phoneme boundaries. This capacity of the phonemes is reflected in the syllables, like we have demonstrated in our last work [7].

As in many other languages, the syllabic units in the Spanish are defined by rules (11 for this language) and we have 17 syllabic different structures (CV, VV, CCVCC, etc.). The motivation for using syllabic units is:

- Because is a model more perceptual and the speech signal have a meaningful better.
- This method provides a better framework for incorporate dynamic modeling techniques into the speech recognition; and [9].
- In present days, continuous speech recognition systems consider phoneme as a basic unit and use syntactic and semantic rules of the language. The limitation of such implementations is that phonemes are extremely context sensitive because its predecessors and successors potentially affect each unit. Due to this acoustic variability, phoneme is not a good choice for speech models for recognition [6, 9].

Exist a lot of advantages of using sub words (i.e. phonemes, syllables, triphones, etc), into speech recognition task. Phonemes are linguistically well defined; the number of them is little (27 for the Spanish language) [2]. Syllables serve as naturally motivated unit of prosodic organization and manipulation of utterances. 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 stress and voice pitched) between two minima of sonority" [9].

In general we can say that the syllables serve as naturally motivated minimal units of prosodic organization and manipulation of utterances [3, 9].

The syllabic level confers several potential benefits; for one, syllabic boundaries are more precisely defined than phonetic segment boundaries in both the speech waveform and in spectrographic displays. Additionally, the syllable may serve as a natural organizational unit useful for reducing redundant computation and storage in decoding [6].

Before to explain the theory related with the speech recognition, we show the next tables 2 and 3, where we have made the analysis of Latino40 corpus for the Spanish relative to the syllables. This establishes the syllable utility in the speech corpus:

Table 1. Frequency of 10 mono syllables used in Latino40

Word	*Syllable configuration*	*Number of times*	*% Vocabulary*
De	*Oclusiva sorda +vocal*	*1760*	*11.15*
La	*Liquida+vocal*	*1481*	*9.38*
El	*Vocal+liquida*	*1396*	*8.85*
En	*Vocal+nasal*	*1061*	*6.72*
No	*Nasal+vocal*	*1000*	*6.33*
Se	*Fricativa + vocal*	*915*	*5.80*
Que	*Oclusiva sorda + vocal*	*891*	*5.64*
A	*Vocal*	*784*	*4.97*
Los	*Liquida + vocal + fricativa*	*580*	*3.67*
Es	*Vocal + fricativa*	*498*	*3.15*

Table 2. Percentage structural of syllables, in Latino40 corpus

Structure	*% Vocabulary rate*	*% Vocabulary accumulated*
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

Within the research activity in the Automatic Speech Recognition (ASR) area, the phonetic characteristics of each basic phonetic unit are to a large extent modified due to co-articulation. As a result, the phonetic units meet from continuous speech and phonetic units articulate in isolation have different characteristics. Using the syllable the problem is the same, but in this work, we extracted the syllables directly from the waveform of the speech, and with a expert system, we find out the grammatical solu tion for the syllable. The next figure shows the segmentation stage using amplitude of energy for that [7]:

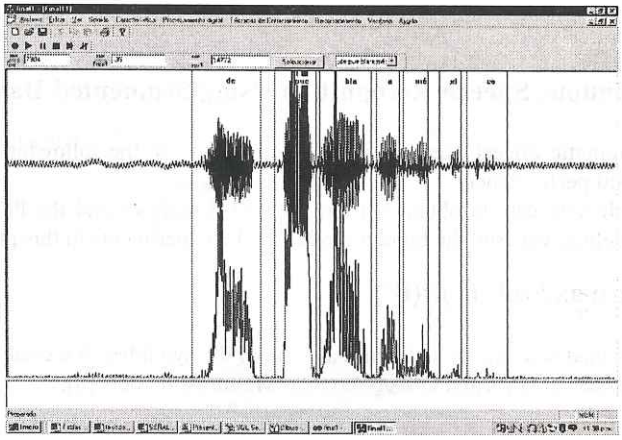

Fig. 1. Labeling speech segmentation, using syllables

As you can see, the energy is more significantly when the syllable is present that was used to have the possibility to extract the syllable units directly from the waveform speech. Such as relative energy maxim are interpreted as potential syllabic boundaries. To differentiate between syllables generally defined on the phonological level and the speech segments that may be located in the signal by phonetic criteria; we introduce the term "syllabic unit".
After that, each of the syllables was stored independently in a file. The database used was created in our laboratory; we used 10 phrases with 52 different syllables. We used 20 utterances for each of the phrases, 50% for training and the rest for recognition. The next figure shows the corpus structural syllable:

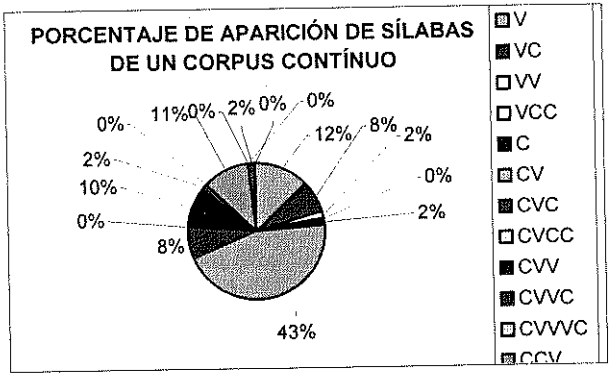

Fig. 2. Percent syllable structural experimental corpus

3 Continuous Speech Recognition Using Segmented Data

In the Automatic Speech Recognition is common to use the following expression to reach a good performance:
The first element can be obtained from the HMM analysis and the P(W) is the language modeling, we used the bigram model for the experiments in this paper.

$$\underline{W} = \arg\max_{W} P(O \mid W)P(W) \tag{1}$$

We obtained separate models for each one of the syllables. We used the next equations for update HMMs and GMs (Gaussian Mixtures) models [8]:

$$\pi_i = \frac{\sum_{e=1}^{E} \gamma_i^e(1)}{E} \tag{2}$$

$$c_{it} = \frac{\sum_{e=1}^{E}\sum_{t=1}^{T_e} \gamma_{it}(t)}{\sum_{e=1}^{E}\sum_{t=1}^{T_e} \gamma_i^e(t)} \tag{3}$$

$$\mu_{it} = \frac{\sum_{e=1}^{E}\sum_{t=1}^{T_e} \gamma_{it}^e(t) o_t^e}{\sum_{e=1}^{E}\sum_{t=1}^{T_e} \gamma_{it}^e(t)} \tag{4}$$

$$\Sigma_{it} = \frac{\sum_{e=1}^{E}\sum_{t=1}^{T_e} \gamma_{it}^e(t)(o_t^e - \mu_{it})(o_t^e - \mu_{it})^T}{\sum_{e=1}^{E}\sum_{t=1}^{T} \gamma_{it}^e(t)} \tag{5}$$

$$a_{ij} = \frac{\sum_{e=1}^{E}\sum_{t=1}^{Te} \zeta_{ij}^e(t)}{\sum_{e=1}^{E}\sum_{t=1}^{Te} \gamma_i^e(t)} \tag{6}$$

EM (Expectation Maximization) updating HMM values, algorithm equations.

Gaussian Mixtures (GMs) used have different representations; the expression to represent a Gaussian is:

$$g(\mu, \Sigma)(x) = \frac{1}{\sqrt{2\pi}^{d} \sqrt{\det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{7}$$

When we used Gaussian Mixtures, the expression is the following:

$$gm(x) = \sum_{k=1}^{K} w_k * g(\mu_k, \Sigma_k)(x) \tag{8}$$

with

$$\sum_{i=1}^{K} w_i = 1 \quad \forall \quad i \in \{1,.....,K\} \ : w_i \geq 0 \tag{9}$$

the following graphic in 3-D showed three mixtures joints:

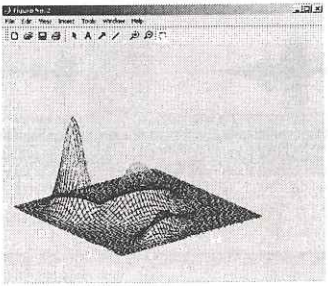

Fig. 3. Gaussian Mixtures used in the experiments

When we used HMM with GMs, we must update every parameters, the next figure shows the activity into states of the HMM that have Gaussian Mixtures:

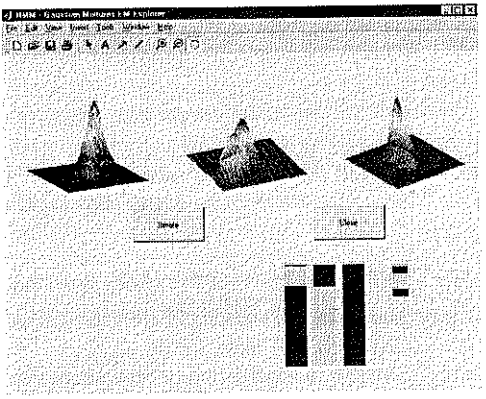

Fig. 4. Mixture Gaussians used for each state

We used the next expression to represent the language model:

$$P(W) = P(w_1)\prod_{i=2}^{N} P(w_i \mid w_{i-1}) \tag{10}$$

The parameter RO is an extraction to find the level of high frequencies in the speech signal. The fricative "s" is the most relative example of that. When we use a high-pass filter, we can obtain the signal above of a threshold. In this work, we use a frequency f_c=4000 Hz, to obtain a level of signal for extract the RO parameter. The speech signal lower at this frequency is reflected. After that we obtained the energy for the resultant signal. The following graphic demonstrate this analysis:

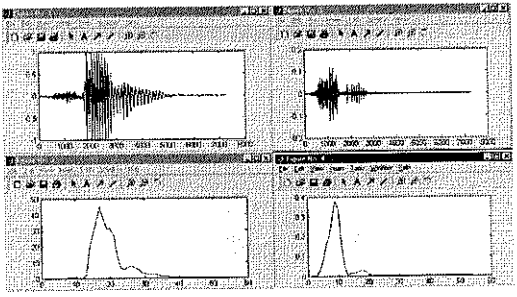

Fig. 5. Parameters STEF and Ro in word 'cero'

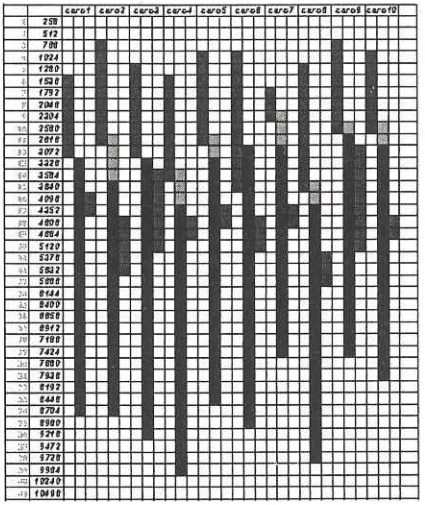

Fig. 6. Distribution energy of STEF and RO parameters

We find an area between the two signals obtained with the last analysis, in the last figure, we can see a brown rectangle that represents the energy of the signal after to use the filter, a blue rectangle before to use the filter and a red rectangle that represents the transition region between two elements, we have called at this region "Transition Region energy-RO parameter".

4 Experiments and Results

We use the following phrases to be recognized for the system:

1 De Puebla a México
2 Cuauhtémoc y Cuautla
3 Cuautla Morelos
4 Espacio aéreo
5 Ahumado
6 Croacia esta en Europa
7 Protozoarios biológicos
8 El trueque marítimo
9 Ella es seria
10 Sería posible desistir

Using a software tool (expert system) programmed in C++, we found the following analysis for this corpus:

Table 3. Syllables into our 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	ló	1
mú	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	rí	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

We used Gaussian Mixtures with three of them for each state; so we used HMM with four and three states, we use 12 MFCCs how observation elements and we probed with five and ten iterations, the programs running on MATLAB and some results obtained are the following:

```
>> HMM1.trans

ans =
   0.7937    0.2063    0.0000         0
        0    0.8118    0.1882    0.0000
        0         0    0.7677    0.2323
        0         0         0    1.0000
>> HMM10.trans

ans =
   0.8332    0.1668    0.0000         0
        0    0.8537    0.1463    0.0000
        0         0    0.8558    0.1442
        0         0         0    1.0000
```

We obtained the following results:

Iterations	Hidden Markov 3	Models states 5
5	82.5%	80.5%
10	85%	82.5%

Table 4. Percentage of recognition

5 Conclusions

This paper showed syllable utility in the Continuous Speech Recognition, the results obtained demonstrate that we can use the syllabic-unit like alternative to the phonemes in a Automatic Speech Recognition System (ASRS) for the Spanish. Using the syllables for speech recognition are (like we saw) very important to avoid the dependency contextual that we can find when used phonemes. The number of syllables could be seeing how a problem, but so much of that can be finding it more than any time.
MFCCs were used in preprocessing; CDHMMs for training and recognition respectively, and we demonstrated that is irrelevant make using phonemes or syllables for these algorithms. So, we used the concatenate property of Hidden Markov models, concatenated HMMs to create larger models that represent the phrase utterance for a speaker. The results demonstrated that we can use this and in comparison with another works [1] for the English, we found out increment the number of syllables and establish a new alternative for modeling the ASRS in the Spanish language, for future works into this language. The use of RO parameter increases the speech recognition in 5% more, that when only use STEF of signal.

References

1. Speech Recognition Using Syllable-Like Units, Zhihong Hu, Johan Schalkwyk, Etienne Barnard, Ronald Cole. Center for Spoken Language Understanfing. Oregon Graduate Institute of Science and Technology. 1998.
2. Análisis del Español Mexicano, para la construcción de un sistema de reconocimiento de dicho lenguaje, Ben Serridge, GRUPO TLATOA, UDLA, Puebla México 1998.
3. Syllable Structure Constraints, A C/D Model Perspective, Osamu Fujimura. 1998.
4. Sobre el uso de la sílaba como unidad de síntesis en el Español. Feal L. Pinto, Informe técnico, Departamento de Informática, Universidad de Valladolid, 2000.
5. Continuos speech recognition using automatically segmented data at syllable units. V. Kamakshi Prasad, T. Nagarajan and Hema A. Murphy. Department of Computer Science and Engineering, Indian Institute of Technology, Madras, Chennai, ICSP, 2002.

6. Integrating Syllable Boundary Information Into Speech Recognition. Su-Lin Wu, Michael.
7. Pruebas y validación de un sistema de reconocimiento del habla basado en sílabas con un vocabulario pequeño. Sergio Suárez Guerra, Karen Suso, Mariana del Villar, Congreso Internacional de Computación CIC2003. México, D.F.
8. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Jeff A. Bilmes. International Computer Science Institute Berkeley CA, 1998.
9. Wu, S., "Incorporating information from syllable-length time scales into automatic speech recognition", Tesis de doctorado, Universidad de California en Berkeley, 1998.