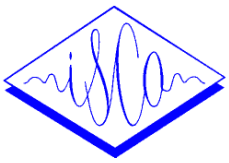

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

# SYLLABIC FILLERS FOR SPANISH HMM KEYWORD SPOTTING

E. Lleida, J.B. Mariño, J. Salavedra, A. Bonafonte

Dept. of Signal Theory and Communications (U.P.C.)
Apdo. 30002, Barcelona 08080, Spain
lleida@tsc.upc.es

## ABSTRACT

In this paper, we describe a preliminary investigation of the use of syllabic fillers to model the out-of-vocabulary utterances or non-keywords in fluent speech for Spanish speech recognition systems. The Spanish language has a syllabic structure which suggest to use syllabic models to deal with the out-of-vocabulary utterances. Furthermore, Spanish demisyllable units has been used successfully as recognition units in the context of continuous speech recognition [7,8]. Thus, our proposal is to use a compact representation of the vocabulary words and the out-of-vocabulary words in terms of demisyllable units and syllabic fillers. The paper evaluates the performances of the syllabic fillers to deal with out-of-vocabulary utterances. As application task, we use the detection of digits in fluent speech.

10.21437/ICSLP.1992-4

## I. INTRODUCTION.

The problem of detecting a given set of words in fluent speech is one of the most interesting topics in speech recognition. Different applications have been proposed in the last few years. One of the most successful approaches is to model the keywords speech and the non-keywords speech or extraneous speech by means of hidden Markov models, driving the recognition process by a null grammar consisting of a parallel network of keywords and non-keywords or Fillers. The output of such systems is a continuous stream of keywords and non-keywords models, given a set of putative hits. The performance of these systems is depending on the robust estimation of the keyword models for high detection probability and the better estimation of the filler models for low false alarm probability. Typically, keywords are represented by words models when the application task has a small number of keywords (task dependent models) or by subwords models when the keyword training is independent from the task (task independent models). By other hand, non-keywords are represented by a great variety of models. In [1], six "function" words and some non-vocabulary words models of equal to syllables are used as filler models. In [2], acoustic word models and subword models as monophone models are used to build the filler models. In [3] alternative models composed of segments of keyword models are used. In [5], the most frequently non-vocabulary words are used as garbage models with a background model. As can be noted, most of this non-vocabulary representations are task dependent (use information of the non-vocabulary words of the application).

Another interesting topics is the detection of new words in continuous speech recognition systems. One approach [4] is to develop an explicit model of new words that will be detected whenever a new word occurs. These models should be general enough to represent any new word. In [4], four types of models were proposed based on the concatenation of context-independent phoneme and diphone (phoneme in the context of the previous phoneme) models. This problem could be also seen as a problem of detecting out-of-vocabulary words or non-keywords.

In both problems, keyword spotting and new word detection, the modeling of out-of-vocabulary words plays an important role. Our approach is based on the use of models related with the linguistic structure of the language. The Spanish language has a syllabic character which suggest to use syllable-based phonetic unit to model both the keyword speech and non-keyword speech. Experiments carried out in the context of continuous speech recognition has shown the high performance of the demisyllable units as phonetic unit for the Spanish language. Furthermore, the inventory of Spanish demisyllables is relatively small: less than 750 units. Thus, demisyllables afford a convenient phonetic coding of Spanish words. In [7] we presented an acoustic processor based on the Demisyllable HMM spotting applied to the recognition of numbers from 0 to 1000. Note that the use of a subword unit as the demisyllable allow us to get task independent models.

Taking the idea of the syllabic structure of the Spanish language, our purpose is to use syllabic fillers to model the non-keyword speech. As the Spanish language has a relative high number of syllables, we classify the sounds in four sets. With these four sets, we can build sixteen different syllabic sets to cover all the legal syllabic structures of the Spanish language. Thus, we model the non-keyword speech as the concatenation of sixteen filler models. With this approach, the syllabic fillers model the general structure of the syllables and the demisyllables model the specific application syllables. Thus, the syllabic models probabilities could be interpreted as the background probability of the utterance and could be used to reject putatives hits of keywords (sequence of demisyllable models).

The paper is organized as follows. Section 2 describes the keyword and filler modeling process. Section 3 describes the training process for the keywords and syllabic fillers models. Section 4 describes the keyword spotting process showing the experimental results, and, finally, conclusions are presented in Section 5.

## II. KEYWORD AND FILLER MODELS.

### 2.1 Keyword models.

Keywords or vocabulary words models should have a high recognition performance to give a high detection probability. In our experience with Spanish continuous speech recognition, the demisyllable units have given excellent results in applications like to recognise the integer numbers from 0 to 1000 [7,8]. Thus, demisyllables afford a convenient phonetic coding of Spanish utterances, which is according to the syllabic character of this language. In order to define the demisyllable set, every possible syllable was divided by the strong vowel into an initial demisyllable and a final demisyllable; accordingly, we distinguished between stressed final demisyllables and unstressed final demisyllables. The main cues of prosodic stress in Spanish are pitch, loudness and syllable length; as pitch and loudness information are not considered in our system, the main difference between stressed and unstressed final demisyllable is the length of their references.

### 2.2. Filler models

By other hand, exploiting the syllabic structure of the Spanish language, we propose to use syllabic fillers to model the

out-of-vocabulary speech. The first problem we have to deal is the definition of syllabic sets. The Spanish language has about several thousand of syllables. Thus, to use syllabic fillers we have to define a small number of syllabic sets and classify all the possible syllables into this sets. So, the first step is to classify the sounds in broad classes. We have defined four classes attending to the similarity of the different Spanish sounds. Dividing the sounds of a language in only four classes could be carried out taking into account similar features as the manner of articulation, the place of articulation and voiced feature. A first classification is defined between vowel-like sounds and consonants. Thus, one broad class for vowel-like sounds and three broad classes for consonant sounds has been defined. Classically, consonants sounds are classified by their manner of articulation in nasals, liquids, stops and fricatives. Liquid and nasal sounds are both voiced sounds and excite the vocal tract solely at the glottis (these sounds are also known as *sonorants*).Thus, liquid and nasal sounds will compose a broad consonant class. The remaining sounds, stops and fricatives, could be used to form the other two broad consonant classes. However, the voiced feature plays an important role in the perceptibility of syllables [9], thus, we have divided the *obstruent* sounds (stops and fricatives) in voiced obstruent and unvoiced obstruent sounds, forming the other two broad consonant classes. In short, all the voiced obstruent consonant sounds are classified in the same broad class named "**s**". The nasals and liquids sounds define the broad class "**n**". Unvoiced obstruent consonants represent the third broad class "**c**" and, finally, all the vowels, glides, diphthong and liquids inside the syllable compose the last broad class "**v**". Table I shows the four classes and their corresponding sounds.

| SETS | SOUNDS |
|---|---|
| **s** | **voiced obstruent consonants** |
| **n** | **nasals, liquids** |
| **c** | **unvoiced obstruent consonants** |
| **v** | **vowels, glides, diphthong, liquids inside the syllable** |

**Table I.** Sound sets.

Once the basic sounds has been classified in four classes, we have to define the syllabic sets. The Spanish language has a simple syllabic structure, if **b** means consonant and **a** means vocal, the 96 % of the syllabic structures of the Spanish are defined by the sequences [9] **ba, bab, a, ab**. The remaining 4 % are more complex structures (**bba, bbab, babb, bbabb**) which can be reduced without significant loss of information to the sequences **ba** and **bab**. With this classification, only sixteen syllabic sets are needed to cover all the possible Spanish syllables (table II). So, we define 16 filler models corresponding to the 16 syllabic sets.

| syllabic structure | syllabic sets |
|---|---|
| **a** | **v** |
| **ab** | **vs, vn, vc** |
| **ba** | **sv, nv, cv** |
| **bab** | **svs,svn,svc,nvs,nvn,nvc,cvs,cvn,cvc** |

**Table II.** Syllabic sets (a, vocal; b, consonant).

This modelization, demisyllables for keywords and syllabic sets for fillers, gives a compact framework where the syllabic sets will match the syllabic structure of the all utterance and the demisyllables models will match with a high probability the syllabic structure of the keywords. Thus, the syllabic filler scoring can be used as a measure of the background probability of the utterances and used to reject putative hits or to define the syllabic structure of a new word.

## III. KEYWORDS AND SYLLABIC FILLER TRAINING.

### 3.1. Keywords training.

The structure used for the demisyllable HMM is the 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 was determined as a function of the average length of the demisyllable, according with table III. Finally, each demisyllable reference is composed by a HMM and the mean and variance of the length of the demisyllable.

| Average length in frames | ≤4 | 5,6 | 7,8 | 9,10 | >10 |
|---|---|---|---|---|---|
| Number of states | 2 | 3 | 4 | 5 | 6 |

**Table III.**Criterion to select the number of states of HMM

The demisyllable HMM´s are trained with an specific database of strings of integers (DB1) consisting in 40 strings of integers uttered by ten speakers (5 male and 5 female), for example, 25011/96, 1019/05/70. This data base was segmented by hand into demisyllables and used for training the HMM of the demisyllable units. The articulation rate of speech spanned from 5 to 7 syllables per second.

Each model was trained independently of the others. Once the samples of every demisyllable were collected from the utterances of DB1, the Baum-Welch estimation algorithm was applied. At the same time, the mean and the variance of the length of the demisyllable was computed. We use three independent codebooks of 64 codewords for the two codebooks dedicated to spectral information and 32 codewords for the codebook devoted to energy differences.

Every frame of speech was vector-quantized with the five nearest codewords, during the training phase; so, for one frame of speech the probabilities of three codewords could be trained. The contribution of a codeword appearance to the probability estimate was weighted inversely with respect to the distance between the frame and the codeword. Thereby, the model estimation and the model smoothing were carried out simultaneously. During the recognition phase, the speech frames were vector-quantized with the nearest codeword only.

### 4.2. Fillers training

The discrete HMM syllabic fillers are trained automatically with a database (DB2) of 120 phonetically balanced sentences uttered by 10 speakers (5 males and 5 females) with a total of 900 sentences (40 minutes of speech). Each syllabic set appears 80 times at least in the training sentences. 20 sentences were hand-segmented in syllabic sets to inicializate the training process. Due to the great variance in the time duration of the syllabic sets, different structures of the HMM has been studied to model the syllabic fillers. Basically, we have tested three different structures of the HMM. The left-to-right structure (**S1**) without skip a state with different number of states for each syllabic set, the one skip state left-to-right structure (**S2**) and the ergodic left-to-rigth structure (**Sn**) where all the transitions between left states to right states are allowed (Fig. 1)

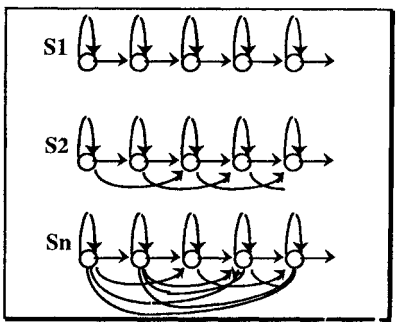


Figure 1. HMM structures for the syllabic fillers

The training process starts with an automatic transcription of the training sentences in syllabic sets. An unsupervised training process performs the training of the sixteen syllabic fillers by iterating 20 times with the viterbi algorithm all the training data base. Due to the great variety of sounds represented by every syllabic filler, HMM fillers inicializated by the 20 hand-segmented sentences in syllabic sets are used to get a good convergence in the training process.

## IV. EXPERIMENTAL RESULTS.

### 4.1. Signal processing.

The speech signal is band-pass (100 Hz - 3400 Hz) filtered by an antialiasing filter and sampled at 8 kHz. A begin/end point detector isolates the utterance to be recogniser. Once the speech signal is pre-emphasizing, a linear prediction based parametrization is used with a Hamming window of 30 ms. every 15 miliseconds. Every frame is characterized by a LP-filter with 8 coefficients. Afterwards, 12 band-pass lifted cepstrum coefficients are computed. Before entering the recognition algorithm, the system evaluates the spectral difference with a time-average of 90 miliseconds. In a similar way, the energy difference is calculated. As we use a discrete hidden Markov modelling, the spectral vector and the spectral and energy differences are vector-quantized separately; in that way, every frame of the speech signal is represented by three symbols of a set of 64 , 64 and 32 codewords respectively.

### 4.2. Keyword spotting process.

The spotting process is a grammar driven continuous word speech recognition system which determines the best sequence of Fillers and Keywords.[2,5,6]. Keywords are represented by demisyllables models and fillers are represented by syllabic sets in our approach. The grammar drives the viterbi search to give as a result a sequence of legal demisyllables (keywords) and syllabic sets. The idea is to extend a classical continuous speech recognition system to deal with the case of unconstrained speech where extraneous speech and out of vocabulary words could appear. The grammar structure will determine the keyword application. Mainly, we use two kind of grammars; a null grammar consisting of a parallel network of keywords and fillers and a limited grammar where we use the a priori knowledge that the sentences has certain structure (i.e. only one or two vocabulary words could appear in any utterance). For a new words detection application, the grammar will allow to recognise syllabic sets where is more likely to appear new words.

The application grammar is obtained by means of a syntactic knowledge inference process which compiles all the legal combinations of demisyllables units to built the keywords or correct sentences with the syllabic sets in a network. Fig. 3 shows an example of the application grammars for the digit spotting (null grammar and limited grammar).

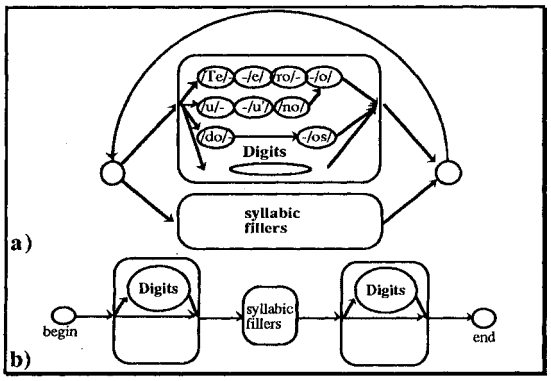


Figure 3. Grammars for detecting digits. a) null grammar b)limited grammar for detecting 0, 1 or 2 digits in a predeterminated position in the sentence.

### 4.3. Syllabic filler experiments.

The first set of experiments were to test the performance of the HMM structure of the syllabic fillers. We have test three structures and different number of states. To delimitate the number of parameters of the system to be tested, we fix the size of the codebooks to the number used in our previous experiments with demisyllables (64 for spectral information and spectral difference and 32 for energy difference). The three structures tested for the syllabic fillers were:

S1) left-to-right without skip any state
S2) left-to-right skiping one state (as the demisyllables models)
Sn) ergodic left-to-right

To decide the number of states of each filler, a first estimation of the duration of each syllabic set were performed with the 20 hand-segmented sentences. Table IV shows the mean length and variance for the sixteen syllabic sets. According with this length and the experience with the demisyllables models, we decide to fix the number of states of each model as shown in table IV. With these number of states, a training process was performed with the structures S1 and S2. For the structure Sn we use a fix number of states for all the models since it allows all kind of transitions and then the training process will adjust the length of the model. To test the performance of the syllabic fillers we make a set of recognition experiments consisting in recognising the syllabic fillers with a null grammar of only syllabic models. We use as test 120 sentences used for training (learning test) with a total of 1288 syllables and a new 41 sentences not used for training (evaluation test) with a total of 598 syllables. Table V shows the performances of the syllabic models in terms of % insertions, % omissions, % errors (# insertions + # omissions + # substitutions) and % recognition (# units - # omissions - # substitutions). Note that S1 models and Sn models with 10 states gives the best results in terms of % errors and % omissions (45,1 % of error and 10,25 % of omissions for S1 and 46,5 % of error and 8,9 % of omissions for Sn with 10 states in the learning test and 56,5 % of error and 13,4 % of omission for S1 and 56,9 % of error and 10,7 % of omission for Sn wit 10 states in the evaluation test). We have made other tests with different number of states for S1, S2 and Sn and different inicialization of the HMM models and we always have found worse results that the presented in table V. We have found a great dependence of the performance with the inicialization, the error rate increase in more than 10 porcentual points if we use an uniform inicialization instead of the hand-segmented sentences.

| syllabic sets | length | variance | # of states |
|---|---|---|---|
| v | 7 | 6 | 5 |
| vc | 9 | 13 | 6 |
| vn | 8 | 18 | 6 |
| vs | 11 | 6 | 8 |
| cv | 11 | 30 | 8 |
| nv | 10 | 25 | 8 |
| sv | 10 | 13 | 7 |
| cvc | 18 | 44 | 10 |
| cvn | 16 | 26 | 10 |
| cvs | 14 | 26 | 10 |
| nvc | 16 | 35 | 10 |
| nvn | 16 | 47 | 10 |
| nvs | 14 | 23 | 10 |
| svc | 14 | 45 | 10 |
| svn | 13 | 31 | 10 |
| svs | 14 | 17 | 10 |

Table IV. Mean length, variance, and number of states for the HMM structures S1 and S2.

| D.B. | Model | %In | %Om | %Error | %Rec |
|---|---|---|---|---|---|
| Learning | S1 | 3,5 | 10,2 | 45,2 | 56,7 |
| Learning | S2 | 10,2 | 13,9 | 56,0 | 48,4 |
| Learning | Sn | 10,4 | 8,9 | 46,5 | 59,1 |
| Test | S1 | 4,8 | 13,4 | 56,4 | 45,6 |
| Test | S2 | 10,2 | 16,7 | 59,6 | 44,5 |
| Test | Sn | 14,7 | 10,7 | 56,9 | 49,3 |

Table V. Results recognising syllabic structures

To evaluate the performance of the syllabic models (structures S1 and Sn with 10 states) versus demisyllable models used to represent the keywords we use a data base consisting in 44 integer numbers from 0 to 1000 uttered by 10 speakers. The recognition task were to recognise the syllabic sets and the demisyllable units of this data base. Table VI shows the average performance over the 10 speakers. Demisyllable models are shorter than syllabic models which explains the higher insertion probability of the demisyllable units.

| Model | % In | % Om | % Error | % Rec |
|---|---|---|---|---|
| S1 | 6,7 | 3,6 | 51,9 | 51,3 |
| Sn | 24,6 | 4,8 | 58,1 | 52,2 |
| Demisyll. | 34,3 | 0,7 | 59,6 | 54,2 |

Table VI. Results with the 0 to 1000 data base.

Unfortunately, at this moment, we don't have operative a data base of sentences with digits or other keywords to test the performance of the fillers in a real application. Nevertheless, we have tested the system in a very difficult task as detect the Spanish digits ( 10 keywords ) uttered in the integer numbers from 0 to 1000. This task has great difficulty because of the similar sounds between digits and non-digits. Table VII shows the recognition results over one speaker using a null grammar of fillers and keywords (sequence of demisyllables) and using a priori knowledge that in the sentence could have one digit at the beginning, one digit at the end or without digits. The results show the keyword probability ($P_k$) and the false alarm probability ($P_f$) computed over the total number of putative words. The analysis of the results reflect some problems in the detection of keywords /uno/ (one) and /dos/ (two) and most of the false alarms with the limited grammar are due to the detection of high confused words as the word /seis/ when the speaker says /sesenta/ (sixty). However, these preliminary results show the capability of the filler models to model the out-of-vocabulary words in a task independent training. With the null grammar we have found an increase in the false alarm rate because of the problems between the words /dos/ and /cientos/ (the syllable /tos/ match the keyword /dos/).

| Grammar | Models | $P_k$(%) | $P_f$(%) |
|---|---|---|---|
| null grammar | S1 | 83,5 | 39,5 |
| null grammar | Sn | 76,9 | 26,6 |
| limited grammar | S1 | 75,6 | 9,5 |
| limited grammar | Sn | 70,6 | 7,5 |

Table VII. Spotting results.

## VI. CONCLUSIONS.

This paper has presented a syllabic approach to the representation of the out-of-vocabulary words in the context of keyword spotting. Keywords or vocabulary words are representated in terms of demisyllables and out-of-vocabulary words are representated in terms of syllabic structures. The work is focused to the definition of syllabic fillers for the Spanish syllabic structure. These filler models provide a task independent models to represent all the Spanish words with a minimum number of models. We have defined 16 syllabic structures with only 4 sound classes (vowel-like sounds, liquids and nasals, voiced consonants and unvoiced consonants). Preliminary results show a recognition rate of more than 50 % recognising the syllabic structure and a detection probability of more than 70 % with a false alarm probability less than 7,5 % detecting digits. Further work will focus in the validation of the syllabic representation over a significative Spanish data base (the ALBAYZIN data base [10] will be available at the beginning of the next year), in the improvement of the training process and in the definition of the sound classes, as well as in the implementation of a spotting strategy according to the syllabic approach of the system.

## VII. REFERENCES.


[1] A.L. Higgins, R.E. Wohlford. "Keyword recognition using template concatenation", ICASSP, pp 1233-1236, 1985.

[2] R.C. Rose, D.B. Paul. "A Hidden Markov Model Based Keyword Recognition System", ICASSP, pp 129-132, 1990.

[3] J.R. Rohlicek, W. Russell, S. Roukos, H. Gish. "Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting", ICASSP-89, pp 627-630, 1989.

[4] A. Asadi, R. Schwartz, J. Makhoul. "Automatic Detection of new Words in a Large Vocabulary Continuous Speech Recognition System", ICASSP, pp 125-128, 1990.

[5] J.G. Wilpon, L.R. Rabiner, C.H. Lee, R. Goldman. "Automatic Recognition of Keywords in Unconstrained Speech Using Hidden Markov Models", Trans. on ASSP, vol 38, no. 11, pp 1870-1878, Nov. 1990.

[6] D. L. Thomson, J. G. Wilpon, R.A. Sukkar, D.P. Prezas. "Automatic Speech Recognition in the Spanish Telephone Network". EUROSPEECH-91, pp 957-960, Genova 1991.

[7] E. Lleida, J.B. Mariño, C. Nadeu, J. Salavedra, " Demisyllable-based HMM spotting for continuous speech recognition",ICASSP, pp. 709-712, 1991.

[8] J.B. Mariño, C. Nadeu, A. Moreno, E. Lleida, E. Monte, A. Bonafonte. "RAMSES: A Spanish Demisyllable Based Continuous Speech Recognition System", in *Speech Recognition and Understanding: recent Advances, Trends and Applications* , NATO ASI Series F, vol 75, pp 113-118, 1991.

[9] E. Martinez Celdrán, *Fonética*, ed. Teide, Barcelona 1984.

[10] F. Casacuberta, , R. García, J. Llisterri, C. Nadeu, J.M. Pardo, A. Rubio. "ALBAYZIN: Spanish Corpora for Speech Research", Workshop Int. on Cooperation and Standardization of Speech, Chiavary (Italy), Sept. 1991.


---


This work was supported by the TIC grant number 92-1026-c02-02