OUT-OF-VOCABULARY WORD MODELLING AND REJECTION FOR KEYWORD SPOTTING

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

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

ABSTRACT

In this paper, we deal with the problem of non-keyword modelling and rejection in a Hidden Markov Model (HMM) based Spanish keyword spotting. When talking about the performance of a keyword spotting system in terms of false alarm rejection, the non-keyword modelling and the rejection techniques are two relevant topics. With regard to the non-keyword modelling, our approach is to define a set of task independent filler models which can be used in any application. In this paper we investigate the performance of a set of filler definition in the problem of detecting digits embedded in utterances. Particularly, we are working with three filler definitions: phonetic fillers, syllabic fillers and word-based fillers. For false alarm rejection, we handle the problem as a post processor of the HMM word spotting recogniser. We design a specific classifier based on a Neural Network and linear discriminant functions to classify a keyword hypothesis in keyword/non-keyword.

Keywords: Keyword spotting, hidden Markov models, filler models, false alarm rejection, linear discriminant functions, Neural Network.

1. INTRODUCTION

Day by day is increasing the number of speech recognition applications where word spotting technology is putting at work. Two main applications of this technology are the detection of a set of keywords in conversational speech monitoring [1] (i.e. surveillance applications) and telecommunications services [2] (i.e. audiotex or automatic operator services).

Many of the methods used for wordspotting are based on modelling the keyword speech and the non-keyword speech or extraneous speech by means of HMM, driving the recognition process by a null grammar consisting of a parallel network of keywords and non-keywords or fillers. The output of such system is a continuous stream of keywords and non-keywords, given a set of putative hits.

The proposed keyword spotting system uses a word-based HMM to model the keywords and three different sets of filler models to represent the out-of-vocabulary words. We have defined a set of phonetic fillers, syllabic fillers and word-based fillers.

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

We also propose a modification in the keyword spotting process, dividing the process in two steps. The first, called *main* Viterbi, is a null grammar Viterbi search with the keywords and background models, given the keyword with the highest probability frame by frame. The second step, called *background* Viterbi, is a null grammar Viterbi search with the filler models over the segments where the *main* Viterbi makes a keyword hypothesis. After this keyword hypothesis step, a rejection step, based on a Neural Network, classify the hypothesis in valid keywords or false alarms. We take this keyword hypothesis strategy to decrease the probability of lossing keywords that could produced in the case of giving a stream of keyword and non-keyword hypothesis.

We present our keyword spotting approach in Section 2, the filler modelling definitions are discussed in Section 3 and the false alarm rejection approach is definded in Section 4. The experimental results are reported in Section 5, and some conclusions in Section 6.

2. KEYWORD SPOTTING APPROACH

The keyword recognition system is based on three basic blocks, a signal analysis step, a HMM-based keyword hypothesis step and a keyword hypothesis rejection step.

The speech signal is pre-emphasizing, and a linear prediction based parametrization is used with a Hamming window of 30 ms. every 15 miliseconds. Every frame is characterised 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.

The HMM-based keyword hypothesis step makes use of a time synchronous search with the Viterbi algorithm. The task of this step is to give a probability of that certain keyword has been uttered and determinate its position in the utterance. To carry out this task, the spotting algorithm works with two Viterbi algorithms. The *main* Viterbi uses as references the keywords models. With these references, the algorithm gives for each input frame the keyword with the highest probability and its duration. By other hand, a second or *background* Viterbi, using as reference the filler models gives for each keyword hypothesis an estimation of the

background probability over the duration of the hypothesis. Figure 1 shows a block diagram of the keyword spotting system.

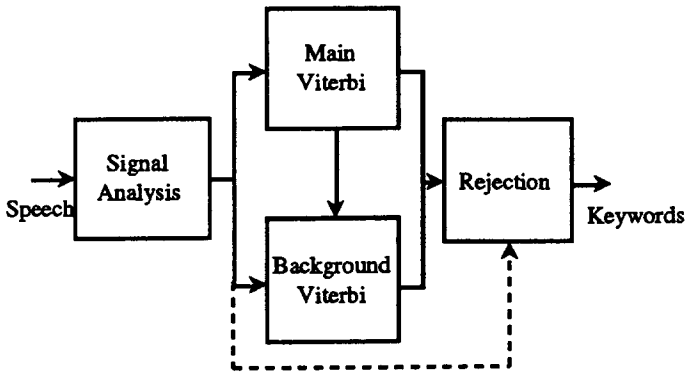


Figure 1. Keyword spotting system

Main Viterbi :

To get the highest probability model in the current frame, the algorithm uses the cumulative probability of each model from the beginning of the speech fragment. The highest probability path in the current frame gives the highest probability model. As all the paths come from the same time instant, all the paths have the same length and thus it is not necessary to normalise the probabilities to carry out the comparison among path probabilities. The score associated to the keyword hypothesis is the probability accumulated from the instant when the path reached the first state of its model. A first rejection is done in this search by removing those hypothesis which their probability is less than a soft threshold, also if a keyword hypothesis appears consecutively less than 3 times is rejected. Once a keyword hypothesis has passed these constraints, we select the duration with the highest probability and look for overlaps with other hypothesis. If there is an overlap greater than a threshold, the hypothesis with the highest probability remains. Finally, the main Viterbi gives as information, the score of the hypothesis (keyword probability), the duration and a new parameter called persistence which is the number of times that the certain keyword appears as the highest probability hypothesis consecutively.

Background Viterbi :

The objective of this new Viterbi is to compute for each keyword hypothesis the probability of the filler models in the same speech segment that the *main* Viterbi gives an hypothesis. This Viterbi search is activated using as reference the filler models over the speech segment which gives the highest probability of the keyword hypothesis. The purpose of this Viterbi is to give a measure of the probability that this hypothesis was a non-keyword.

Thus, the HMM-based keyword hypothesis step gives a set of keyword hypothesis with four informations or rejection parameters, the keyword probability, the duration, the persistence and the non-keyword probability. For the rejection purposes, we use the rate between the keyword and non-keyword probabilities instead of the non-keyword hypothesis. As the probabilities are in logarithms, the rate keyword/non-keyword is defined as the difference between the keyword probability and the background probability. Thus, a positive rate indicates that the hypothesis is more likely to be a keyword and a negative rate indicates a more likely false alarm.

Finally, the rejection step makes use of the information given by the HMM-based keyword hypothesis step to decide if the hypothesis is valid or not. This step will be studied in detail in Section 4.

3. FILLER MODELLING

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-keywods are represented by a great variety of models. Our approach is to define a set of task independent filler models which can be used in any application. Thus, we work with three filler definitions: phonetic fillers, syllabic fillers and word-based fillers.

A first approach for task independent filler definition is to use phonetic models. In Spanish, more than 99% of the allophonic sounds can be grouped in 31 phonetic units as defined in the ALBAYZIN Spanish data base [3]. Thus, the set of phonetic fillers will be composed by the following 31 phonetic units (SAM notation): **p, b, t, d, k, g, m, n, J, N, tS, B, f, T, D, s, z, Z, x, G, l, L, rr, r, i, j, e, a, o, u, w.**

By other hand, exploiting the syllabic structure of the Spanish language, we propose to use syllabic fillers to model the out-of-vocabulary speech [4]. As the Spanish language has about several thousand of syllables, we have defined a small number of syllabic sets by classifying the sounds in broad classes. We have defined four classes attending to the similarity of the different Spanish sounds. 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.

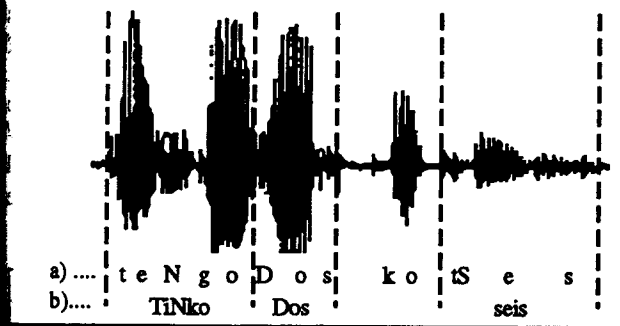
SETS	SOUNDS
s	b,d,g,B,D,z,Z,G
n	m,n,J,N,l,L,r
c	p,t,k,tS,f,T,s,x
v	i,j,e,a,o,u,w,(r,l inside the syllable)

Table I. Sound 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 **a, ab, ba, bab**. 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.

The third out-of-vocabulary word modelling is based on the definition of word-based fillers attending to the number of syllables of the words. Thus, we have defined three word-based filler for monosyllabic words, bi-syllabic words and words with more than three syllables.

In this way, our three filler definition sets could be arranged from specific models as the phonetic fillers to general models as the word-based fillers. More likely, the background probability estimation with the phonetic filler will be best, in the sense of highest probability, than with syllabic fillers or word-based fillers. It isn't clear which is better because a better estimation of the background probability could improve the false alarm rejection but also could decrease the detection probability and a worst estimation of the background probability could improve the detection probability but also increasing the number of false alarms. Thus, syllabic fillers seems to be a compromise between both high detection probability and low false alarm rate. Figure 2 shows the speech signal of the sentence /tengo dos coches/ (I have two cars) and the keyword hypothesis of the *main* Viterbi with the probabilities (log).



Keyword hypothesis	cinco (5)	dos (2)	seis (6)
Keyword Probab.	-4.0122	-3.5696	-4.1633
phonetic rate	-0.0864	0.4606	-0.4633
syllabic rate	0.0893	0.5122	-0.2461
word-based rate	0.2650	0.9500	-0.0135

Figure 3. Speech signal for /tengo dos coches/ with the hypothesis keywords and probability rate (difference between the keyword and the background probabilities) for the three filler sets [in the signal picture, a) phonetic transcription of the sentence, b) keyword hypothesis from the *main* Viterbi].

Filler model training

Our approach is to model keywords and non-keywords models with discrete hidden Markov models (DHMM). To train filler models, a database (DB1) of 120 phonetically balanced sentences uttered by 10 speakers (5 males and 5 females) with a total of 900 sentences (40 minutes of speech) was used.

We use a 3 state model for the phonetic fillers, a 10 state ergodic left-to-rigth model for the syllabic fillers and word-based models. An automatic procedure based on the transcription of the training sentences in phonetic sets, syllabic sets and word-based sets and the Baum-Welch algorithm was used to train the HMM of each filler set.

Keyword model training

The keyword models are a 10 state left-to-rigth without skipping states HMM. To train the models, we use a data base (DB2) of telefone numbers (sets of 6 connected digits) uttered by 26 speakers, (16 males and 10 females). An automatic procedure based on the Baum-Welch algorithm was used to train the keywords

4. FALSE ALARM REJECTION

The false alarm rejection step processes the keyword hypothesis generated by the *main* Viterbi by using information of both Viterbi process and takes a decision over the existence or not of the keyword in the utterance.

There are different methods for false-alarm rejection. The simplest one is to put some thresholds over the information produced by the keyword hypothesis process. In our case, we work with four basic informations that we call rejection parameters. The keyword probability (P_k), the keyword time duration estimation (D), the persistence (P) or number of consecutive times that the same keyword hypothesis is the most likely and the probability rate (R) or diference between the keyword and background probabilities. Thus the simplest method will be to put a threshold for each rejection parameter. However, there is some correlation between them and thus, it will be more efficient to increase the complexity of the process but taking into account the four parameters jointly. One approach could be to build a linear discriminant classifier to clasiffy between keyword and false alarm. In this case, it is a two classes problem where the linear discriminant function is the eigenvector of the non-zero eigenvalue of the matrix $W^{-1}B$, where W^{-1} is the sum of the intra-class scatter matrix of the two classes (keyword or false alarm) and B is the between-class scatter matrix. The eigenvalue gives information about the discriminant properties of the transformation. With this approach, we define one linear discriminat function for each keyword.

Another approach to the rejection problem is to use a Neural Network as classifier. In this work, we have tested a Neural Network with one hidden layer and one output (activated = valid keyword, desactivated = keyword rejected). The basic input layer has 5 inputs (# keyword, P_k, D, P, R). Thus, the basic NN has the following structure: 5 input units, 3 hidden units and 1 output unit. The NN is trained with the backpropagation algorithm .

To give more discriminant information to the Neural Network, the basic input vector could be increased with information of the energy of the signal and the more relevant cepstrum coefficients. To introduce this information to the neural network we have basicaly two ways, reducing the acoustic information to a fixed length (i.e. using trace segmentation technique) or using a feedback neural network from the hidden layer, presenting to the neural network the basic input vector and the acoustic information frame by frame. In this way, this rejection procedure can be seen as a second word recogniser.

5. EXPERIMENTAL RESULTS

To evaluate the performances of the keyword spotting system, we have applied the system to the problem of detecting the Spanish digits (/uno/ /dos/ /tres/ /cuatro/ /cinco/ /seis/ /siete/ /ocho/ /nueve/ /cero/) in uncostrained speech. This vocabulary is very difficult in the sense of the great number of false alarms produced by the monosyllabic digits as /dos/ and /tres/. The test data base (DB3) consist on 8 speakers which have uttered 20 sentences with digits (16 minutes of speech) in a computer room. The total number of digits (keywords) in the data base is 480. Figure 4 shows the detection probability and the number of false alarms obtained from the output of the *main* Viterbi step. The average

detection rate is 90,4 % with 947 false alarms. This high number of false alarms is normal because the *main* Viterbi always gives a keyword hypothesis.

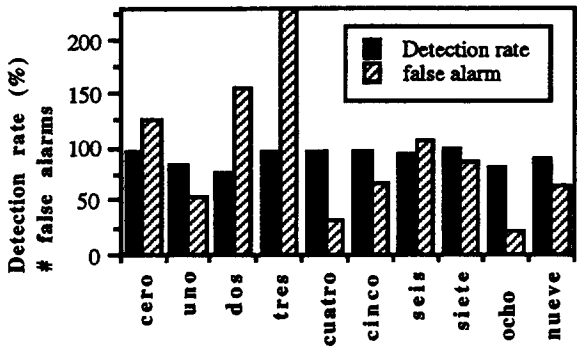


Figure 4. Keyword detection rate and number of false alarm per keyword.

To study the discriminant properties of P_k, D, P and R for the three filler sets we use the Fisher Ratio for each keyword and parameter defined as

$$F=(\mu_k-\mu_{fa})^2/(\sigma_k^2+\sigma_{fa}^2)$$

where μ_k and μ_{fa} are, respectively, the mean of the parameter when there is a valid hypothesis and when there is a false alarm. σ^2_k and σ^2_{fa} are the variance respectively. Figure 5 shows the F ratios for the R parameter for each one of the filler sets. We found that the syllabic filler R ratio is the greatest of the three for all the keywords which means that has better discriminant properties than the others. Figure 6 shows the values of the non-zero eigenvalue of the matrix $W^{-1}B$ for each keyword and using the four rejection parameters (P_k, P, D, R). Again, the set of parameters with the R parameter computed with the syllabic fillers gives the highest scores of discrimination. These ratios shown that the keywords /dos/ and /cinco/ have the worst discrimination rates meaning that it is difficult with these rejection parameters take the decision between keyword and false alarm. New parameters, as acoustic cues, may be introduced in the rejection step to increase the discrimination properties. For instance, many false alarms of the keyword /dos/ are because many Spanish words end with the syllable /dos/, but a difference with the digit /dos/ is the stress. So, incorporing information about entonation will improve the performance of the rejection procedure.

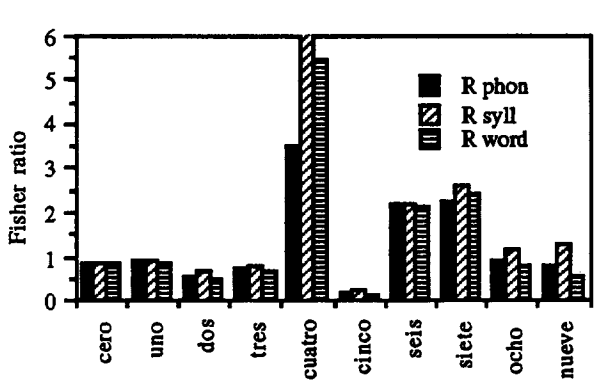


Figure 5. F ratios of the R parameter for each keyword

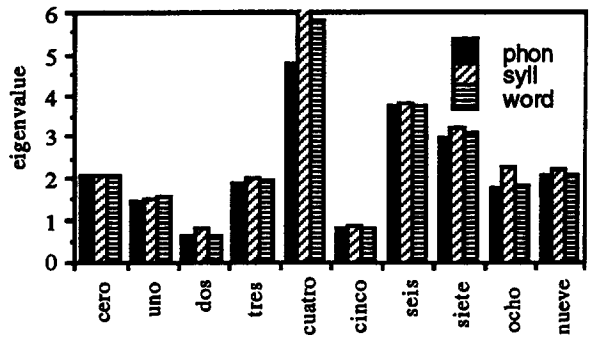


Figure 6. Eigenvalue of the $W^{-1}B$ matrix for each keyword.

Table II shows the results obtained by using as rejection procedure a NN with 5 inputs and linear discriminant functions (LDA). Again the results obtained with the syllabic fillers give a better performance than the other filler sets tested in this work. These results shown also the difficulty of the task. A study of the permformance, keyword by keyword, shows problems on the validation of the hypothesis of the keywords /uno/ (P_d =76,3 %, fa rejection= 76,6 %) , /dos/ (P_d=75% and fa rejection =67 %) and /cinco/ (P_d= 72,7% and fa rejection= 77,9 %) which are the keywords with the smallest eigenvalues (figure 6).

	Phonetic filler		syllabic filler		word-based	
	%P_d	%fa	%P_d	%fa	%P_d	%fa
NN	73,2	87,5	82,1	86,0	76,1	85,3
LDA	81,2	83,8	83,6	83,8	80,8	83,8

Table II. Working points using NN and LDA for the rejection step (%P_d is the detection rate, %fa is the false alarm rejection rate).

6. CONCLUSIONS

In this paper, a keyword spotting system has been presented. To model the out-of-vocabulary words we have tested the performances of three filler definition: phonetic fillers, syllabic fillers and word-based filler. The results shows a better performance of the syllabic fillers over the other two filler sets. The rejection step is based on the use of four parameters obtained from the spotting hypothesis step and a classifier based on a NN. We have also compared the performance with a classical linear discriminant analysis. A future work in this step is to include in the NN acoustic cues as the entonation or spectral information (i.e. a reduced set of cepstral parameters used by the recogniser).

REFERENCES

[1] R.C. Rose, "Discriminant Word-Spotting techniques for rejecting non-vocabulary utterances in uncosntrained speech", IEEE Proc. ICASSP 92, pp. 93-96

[2] M. W. Feng, B. Mazor, "Continuous Word Spotting for applications in telecommunications", Proc ICSLP-92, pp. 21-24

[3] A. Moreno et al, "ALBAYZIN speech database: design of the phonetic corpus", elsewhere of these proceedings.

[4] E. Lleida et al, "Syllabic fillers for Spanish HMM keyword spotting", Proc ICSLP-92, pp. 5-8