# REJECTION TECHNIQUES FOR DIGIT RECOGNITION IN TELECOMMUNICATION APPLICATIONS

*Luis Villarrubia and Alejandro Acero*

Speech Technology Group
Telefonica I+D
Emilio Vargas 6, 28043 Madrid, Spain

## ABSTRACT

In this paper we describe a technique for non-keyword rejection and we will evaluate in the context of an audiotex service using the ten Spanish digits. The baseline keyword recognition system is a speaker-independent continuous density Hidden Markov Model recognizer. We propose the use of an affine transformation to the log-probability of the *garbage* model, an HMM model trained to account for both non-keyword speech and non-stationary telephone noises. The parameters of the transformation for the case of isolated keywords are chosen to minimize a cost function that weighs the keyword error rate, keyword rejection rate and false acceptance rate according to the *a priori* probabilities of keyword/non-keyword and the requirements of the specific application. This technique was also extended to embedded keywords (word-spotting). Use of this rejection technique on the audiotex application reduced the total cost function up to 20% for isolated-word case and 12% for the word-spotting case.

## 1. INTRODUCTION

Evaluations of Voice Response Systems (VRS) over the telephone network show that the performance of speech recognition systems obtained in the laboratory degrades when they are used in telecommunication services. In most of these services, the user is expected to say a keyword in isolation, but often the keyword is surrounded by non-keywords or not present at all. In [1], Wilpon describes a system that uses word-spotting techniques to overcome this problem for a 3-keyword recognizer used in Spain's Intelligent Network. In his application only 1% of the utterances contained no keywords, for which the false alarm rate reported was 20%. We found that this rate, negligible in his case, can be quite high for other applications.

Telefonica I+D has developed an audiotex system that can be accessed through the telephone network through digit recognition. Human factors studies conducted on a field trial [2] suggested the use of an echo-canceller to allow the user to freely jump from menu to menu without having to wait for possibly long messages to finish. The addition of barge-in greatly improved the service for experienced users, allowing them to shorten the transaction time. However, it caused a great deal of recognition misfires, specially for novice users, due to the fact that the recognizer is now exposed all the time to non-keywords and other noises that were ignored before such as breath noises, throat clearing and handset clicks.

While most previous work on word-spotting has been evaluated by keyword detection rate and false alarm rate ([3] for example) for surveillance applications, other authors ([1] for example) found more appropriate the use of other scoring protocols for telecommunication applications. We will propose a new scoring protocol valid for any kind of application.

## 2. EXPERIMENTAL CONDITIONS

For the evaluation of the new rejection and scoring method, we used the Hidden Markov Model Toolkit (HTK) [4] on automatically endpointed speech. To perform this study, we have collected three different databases over the Spanish switched telephone network. All databases were recorded from different adult speakers, both male and female, on long-distance telephone lines.

The system was trained with a database (DB-T) containing 7650 utterances of the Spanish digits. Testing was carried out on another database (DB-R) containing 1708 utterances of the Spanish digits in isolation and 190 embedded digits collected from a field trial of a real VRS. Recognition experiments were ran on DB-R to measure the keyword and word-spotting error rate and the keyword and word-spotting rejection rate.

Finally, another database (DB-G) was collected with 908 utterances containing non-keyword speech of varying duration, and the type of noises found in the field trial (coughs, breath noises, sneezes, laughter, mouth noises, etc). This database was split in half (DB-G1 and DB-G2), with each half containing different speakers and different non-keyword speech and noises. DB-G1 was used to train a garbage model and DB-G2 was used to measure the system's false alarm rate.



0-7803-0946-4/93 $3.00 

The baseline results, when no rejection techniques are used, yield a keyword error rate of 3% and, of course, a keyword rejection rate of 0% and a false alarm rate of 100%.

## 3. PREVIOUS WORK

The use of a garbage model running in parallel with the keyword models to detect the keywords in unconstrained speech has been found to be quite effective [5]. Following this approach, we trained a garbage model with database DB-G1 to provide an explicit representation of out-of-vocabulary speech. The recognizer's output is the word with the highest log-likelihood. Rejection occurs when the top ranked log-likelihood corresponds to the garbage model. This method was tested with the database described before, and as can be seen in Figure 1, the keyword error rate is 2.6%, the keyword rejection rate 2.3%, and the false alarm rate 6.9% (PGM method).

Postprocessing of recognition scores is another technique found in the literature. Moreno [6] found the difference in log-likelihood between the two highest ranked keywords useful to reduce the keyword error rate. The hypothesis is rejected when this difference is lower than a threshold. Our implementation of Moreno's method for this database yielded a reduction in keyword error rate from 3.0% to 0.9% while rejecting 5% of the keywords, which was consistent with the results he reported. Unfortunately, this technique was not derived to minimize the false alarm rate, which was found to be 67% for this point in the ROC, that was unacceptable for our applications.

Chigier [7] uses a similar postprocessing approach, but incorporating a garbage model. He proposes the use of Gaussian classifiers on feature vectors composed of the recognition scores of the keywords and the garbage model, differences of scores, normalized scores and duration.

Recently, word-spotting/rejection algorithms have been improved by the use of discriminant training methods. In all these techniques, either Hidden Markov Models or Neural Networks are trained to maximize the discrimination between keywords and non-keyword speech. For example, Rose [3] adjusts Hidden Markov parameters with MMI and corrective training criteria to maximize the difference between the log-probability of the keyword and the log-probability of non-keyword speech.

## 4. NEW EVALUATION CRITERIA

As Wilpon [1] remarks, there is considerable debate on how to evaluate word-spotting systems. He suggests that the evaluation criterion should be application dependent. The false alarm rate is measured by many researchers in false alarms per keyword per hour. This is a good figure of merit for surveillance applications, but it not be the most appropriate for telecommunication applications such as automating operator services or menu-driven audiotex systems. In the latter applications, both substitution errors and false alarms lead to incorrect actions. False alarms caused by user-generated noises such as breath noises were found to be even more disturbing to the user than substitution errors.

To evaluate isolated word-recognizers, the basic figure of merit is unquestionably the error rate. When rejection capabilities are added, we propose as a figure of merit a cost function that weighs the keyword error rate ($E_k$), keyword rejection rate ($R_k$) and false alarm rate ($F_a$) with parameters $L_e$, $L_r$ and $L_f$ respectively:

$$C = L_e E_k + L_r R_k + L_f F_a \quad (1)$$

with

$$L_e + L_r + L_f = 1 \quad (2)$$

so that the cost function $C$ can be interpreted as a weighted error rate.

Parameter $L_e$ should take into account the contribution of the *a priori* probability of an utterance being an isolated keyword ($P_k$), estimated from field trials of the application, and the penalty associated to keyword errors ($C_{ke}$), given by the application designer. Similarly, $L_r$ and $L_f$ could be computed using penalties associated to keyword rejections ($C_{kr}$) and false alarms ($C_{fa}$). This can be computed by the following expressions:

$$L_e = h P_k C_{ke}\ ; \quad L_r = h P_k C_{kr}\ ; \quad L_f = h (1 - P_k) C_{fa} \quad (3)$$

with $h$ being the normalizing constant so that (2) holds.

We have to note that the *a priori* probability $P_k$ depends highly on the dialog design and whether barge-in is included or not. For example, when the echo canceller was incorporated to the audiotex system, the percentage of non-keywords increased substantially, since the recognizer is active during the whole transaction.

A similar analysis can be performed for recognizers with word-spotting capabilities. In this case, in addition to isolated keywords and non-keywords, we will also consider embedded keywords. In the audiotex application we split the embedded keywords into digits preceded by "*el*", with *a priori* probability $P_e$, and the case of the keyword being surrounded by other non-keywords with *a priori* probability $P_b$. Likewise, new costs $C_{ee}$, $C_{er}$, $C_{be}$ and $C_{br}$ should also be defined respectively, and equations (1) and (2) updated accordingly.

## 5. NEW REJECTION METHOD

We propose the use of an affine transformation to the

log-likelihood of keyword and garbage models:

$$f_i(s_i) = \alpha_i s_i + \beta_i \quad (4)$$

where $s_i$ is the log-likelihood of HMM model $i$, and $\alpha_i$ and $\beta_i$ are the parameters of the affine transformation.

To simplify the approach, we applied the transformation only to the garbage model. The rationale behind this is that the probability of non-keywords is underestimated by the maximum likelihood model and it needs to be compensated. Since log-probabilities are negative quantities, this could be done by choosing $0<\alpha<1$ for $\beta=0$ or $\beta>0$ for $\alpha=1$.

It is important to note that for real time implementation, the $\beta$ parameter can be implemented by simply modifying the transition probabilities to the garbage model. The $\alpha$ parameter can also be taken care of off-line by altering the standard deviation and transition probabilities of the garbage model. Any recognizer can then be used in a transparent fashion with no increase in computation.

## 6. EXPERIMENTS

In this section we describe the experiments conducted on the database of Section 2, with the cost function of Section 4 and the rejection scheme proposed in Section 5. We will analyze separately the case of the isolated word recognizer and the word-spotting recognizer.

### 6.1 Isolated word recognizer

*A priori* probability $P_k$=0.8 was determined from field trials of Telefonica's audiotex application with barge-in. Three different cases of individual costs were studied:

| | $C_{ke}$ | $L_e$ | $C_{kr}$ | $L_r$ | $C_{fa}$ | $L_{fa}$ |
|---|---|---|---|---|---|---|
| **case 1** | 0.4 | 0.3 | 0.4 | 0.3 | 1.6 | 0.3 |
| **case 2** | 0.5 | 0.4 | 0.5 | 0.4 | 1.0 | 0.2 |
| **case 3** | 0.6 | 0.5 | 0.4 | 0.3 | 1.0 | 0.2 |

where we observe that $C_{kr} <= C_{ke} <= C_{fa}$ in all three cases.

Figure 1 shows the keyword error rate, keyword rejection rate, false alarm rate and total cost for the three different cost functions and the two algorithms: parallel garbage model (PGM) and the new linear transformation method (LT). It can be seen that for the first cost function (case 1) LT shows a decrease in total cost of 20%, whereas for the other two cost functions the gain is somewhat smaller.

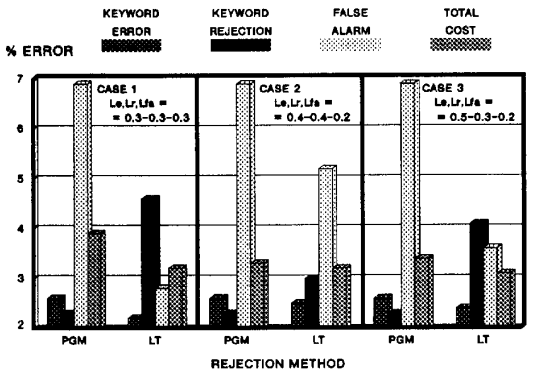


**Figure 1.** Comparison of parallel garbage model (PGM) and linear transformation method (LT). Keyword error rate, keyword rejection rate, false alarm rate and total cost respectively in percentage points are shown for three different cost functions (case 1, case 2 and case 3).

In the LT method, both $\alpha$ and $\beta$ were chosen to minimize the cost function $C$ on the training data. The $\beta$ parameter can be interpreted as the threshold used by many authors to control the point in the ROC. Having both $\alpha$ and $\beta$ allows us more flexibility in choosing the operating point by appropriate selection of the cost function, although having one of them fixed and varying the other yielded essentially the same minimum cost. The range of values studied for $\alpha$ was from 0.9 to 1 and for $\beta$ from -40 to 40. To check the robustness of these estimates, we calculated the optimal $\alpha$ and $\beta$ for the testing database, and they turned out to yield the same performance than those obtained from the training database.

### 6.2 Word-spotting recognizer

In this case, the recognizer has a different grammar that allows the keyword to be preceded and succeeded by non-keywords. The individual costs used are:

$$C_{ke}= 0.7 \quad C_{ee}= 0.8 \quad C_{be}=0.8 \quad C_{fa}=1 \quad (5)$$
$$C_{kr}= 0.3 \quad C_{er}= 0.2 \quad C_{br}=0.2$$

Two cases were studied that differed in the *a priori* probability of utterances containing only keywords ($P_k$), "el" followed by keyword ($P_e$), embedded keyword ($P_b$) and non-keyword ($P_g$):

| | $P_k$ | $P_e$ | $P_b$ | $P_g$ |
|---|---|---|---|---|
| **case 4** | 0.88 | 0.07 | 0.02 | 0.03 |
| **case 5** | 0.73 | 0.06 | 0.02 | 0.19 |

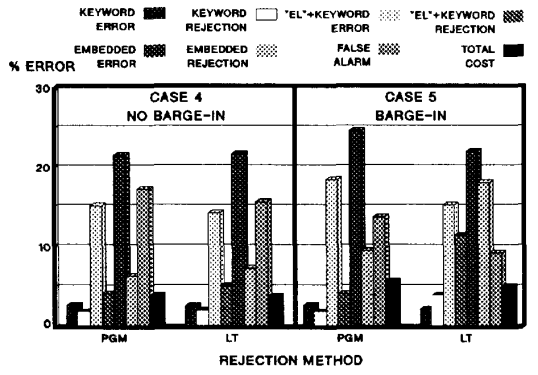


**Figure 2.** Comparison of parallel garbage model (PGM) and linear transformation method (LT) on the word-spotting case for two cost functions (case 4 and case 5).

In this study we set $\alpha$ to 1 and only used the ß parameter for the affine transformation. We had to introduce a new parameter $W$, language weight, to control insertions and deletions in the word-spotting grammar. The rationale for this is that the acoustic model probability is underestimated due to the fallacy of the Markov and independence assumptions[8]. In order to make the acoustic probability and the language probability comparable, a language weight is introduced by raising the word transition probability to a power $W$. In the LT method, ß and $W$ were obtained to minimize the total cost. Figure 2 shows the results obtained for the two cases described above. I can be seen that there is a 12% gain for the case of barge-in.

## 7. CONCLUSIONS

In this paper we described a new criterion for evaluating isolated-word speech recognizers with rejection capabilities on Telecommunications applications. This new cost function weighs the keyword error rate, keyword rejection rate and false alarm rate of the recognizer by individual costs which depend on the *a priori* probability of keyword/non-keyword and the penalties set by the application designer to each type of error. The cost function can easily be extended to the case of word-spotting.

We also described a technique that combines garbage models and score post-processing to allow minimization of the predefined cost function. We observed a reduction of up to 20% in the cost function by using the proposed affine transformation on the log-likelihood of the garbage model for the isolated-word case and a reduction of up to 12% in the word-spotting case. Training the parameters for the affine transformation is a simple task, and this technique is amenable for real-time implementation.

## ACKNOWLEDGMENTS

We owe gratitude to Daniel Tapias ,Carlos Crespo and Celinda de la Torre for their help with the HMM system. We are also grateful to Juan Siles, Emilio Gonzalez, Antonio Golderos and the rest of the speech group for their help and continuing support.

## REFERENCES

[1] J. G. Wilpon, L. G. Miller and P. Modi. "Improvements and Application for Key Word Recognition using Hidden Markov Modeling Techniques". *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*. Toronto, Canada, May 1991, pp. 309-312.

[2] M. Poza, C. De la Torre, D. Tapias and L. Villarrubia. "An Approach to Automatic Recognition of Keywords in Unconstrained Speech Using Parametric Models". *Proc. Eurospeech Conference*. Genova, Italy, Sep 1991, pp 471-474.

[3] R. C. Rose. "Discriminant Word-Spotting Techniques for Rejecting Non-Vocabulary Utterances in Unconstrained Speech". *Proc IEEE Int. Conf. Acoustics, Speech and Signal Processing*. San Francisco, CA, March 1992, pp 93-96.

[4] S.J. Young. "HTK: Hidden Markov Model Toolkit Reference Manual - Version 1.3". Speech Group, Cambridge University Engineering Dept, 1992.

[5] J. G. Wilpon, L. R. Rabiner, C. H. Lee and E. R. Goldman. "Automatic Recognition of Keywords in Unconstrained Speech using Hidden Markov Models". *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*. Vol. 38, No. 11, Nov. 1990, pp. 1870-1878.

[6] P. Moreno, D. Roe, P. Ramesh. "Rejection Techniques in Continuous Speech Recognition using Hidden Markov Models". *Proc. Signal Processing Conference*. Barcelona, Spain, Sep 1990, pp. 1383-1386.

[7] B. Chigier. "Rejection and Keyword Spotting Algorithms for a Directory Assistance City Name Recognition Application". *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*. San Francisco, CA, March 1992, pp 105-108.

[8] K. F. Lee. "Automatic Speech Recognition. The Development of the SPHINX System". Kluwer Academic Publishers.Boston,MA,a989.