# On-line Garbage Modeling with Discriminant Analysis for Utterance Verification

*J. Caminero, C. de la Torre, L. Villarrubia, C. Martín and L. Hernández*[1]

Speech Technology Group
Telefónica Investigación y Desarrollo
Emilio Vargas 6, 28043 - Madrid, SPAIN
e-mail: {jcam, celinda, luigi, cma}@craso.tid.es, luish@gaps.ssr.upm.es

## ABSTRACT

Out-of-vocabulary (OOV) utterance detection and rejection are specially important and difficult problems in large-vocabulary and continuous speech recognition. In [1] we proposed an utterance verification procedure based on the use of frame-by-frame best acoustic state scores instead of using explicit garbage models. This procedure is usually referred to as *on-line garbage modeling*.

In this contribution we extend our previous work in two major directions: a) we analyze, through the use of Discriminant Analysis, the possibilities of using L-best local scores and N-best utterance hypotheses scores for utterance verification; b) we present experimental results not only for a spontaneously spoken natural number recognition task, as in [1], but also for a flexible large vocabulary recognition task. All the results, based on a telephone database, show that the proposed on-line garbage modeling procedure outperforms, both in performance and computational cost, to other approaches based on the use of explicit garbage models.

## 1. INTRODUCTION

One of the main problems still demanding new proposals in Automatic Speech Recognition (ASR) is the ability to verify utterance hypotheses provided by the recognizer. Out-of-vocabulary (OOV) utterance detection and rejection is a major problem in ASR [2], where real life applications demand high robustness capabilities.

The discrimination between OOV and vocabulary words is specially difficult in large vocabulary and continuous speech. The high acoustic coverage of the acoustic units in large vocabulary ASR makes very difficult to discriminate between vocabulary and OOV words. On the other hand, for a continuous ASR special grammars should be designed to cope with the possible presence of OOV anywhere into the utterance.

Until now, most of the existing techniques for hypothesis verification have three major drawbacks:

- They are based on the use of garbage models and alternative recognition networks [3] [4] that are difficult to design and train.
- Although some techniques have shown to provide improvements in the overall system performance, their computational cost is high, specially looking for a real-time implementation on a moderate hardware.
- Only acoustic data is used in the verification process, when, for a wide range of systems, valuable information from application-dependent knowledge is available and should be used.

In this paper, we will address the first two previous drawbacks. On-line garbage modeling was originally proposed for keyword-spotting applications [5]. In our work, we will discuss different alternatives on the application of this technique over word and utterance verification in continuous speech and large vocabulary recognition. We will consider the application of on-line garbage modeling in two different recognition tasks: in a natural connected number recognition task and vocabulary dependent and independent isolated word recognition. We will compare results using on-line garbage modeling with other based on the use of explicit garbage models.

The rest of the paper is organized as follows. Utterance verification using on-line garbage modeling is introduced in Section 2. Section 3 is dedicated to introduce Discriminant Analysis for OOV utterance discrimination. Experimental results for both natural numbers and vocabulary dependent or independent isolated recognition based on subword units, are given in Section 4. Finally, some conclusions are provided at the end of the paper, Section 5.

[1] E.T.S.I. Telecomunicación. Universidad Politécnica Madrid.

## 2. ON-LINE GARBAGE MODELING

On-line garbage modeling is a very attractive technique that has been shown good results in keyword-spotting applications (see [5]). This technique does not attempt to explicitly define a garbage model: "instead, this approach computes the local garbage scores directly on-line for each time frame as the average of the N best local scores of the models used to describe the keyword models".

In Figure 1, we represent our grammar, before including on-line rejection techniques. For this "traditional" verification scheme we have been using different explicit garbage models depending on the particular recognition task. That is, for natural number recognition we use a four-state left-to-right HMM, while for vocabulary-independent recognition we use a net of context-independent units in parallel [6]. Following this scheme an utterance is rejected when the garbage model provides a garbage score or garbage likelihood which is better than any of the vocabulary words.

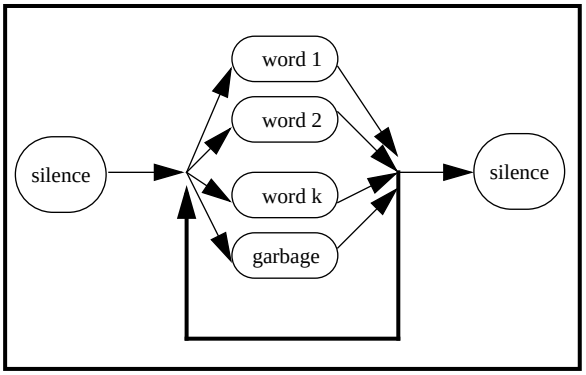


Figure 1: Recognition grammar including garbage model.

When using on-line garbage modeling, we do not use any explicit garbage HMM. In this case the garbage score is obtained from the frame-by-frame scores of the L-best active states in the recognition network only composed of vocabulary words. Thus the garbage score is evaluated during the forward pass of the Viterbi search at very low computational cost. At the end of the utterance, the accumulated garbage score is used to perform the utterance verification. The details of this procedure are described in the next section.

## 3. DISCRIMINANT ANALYSIS

As we have described in the previous Section, the on-line garbage score is obtained using the L-best local scores of the states representing the acoustic units of the recognizer. Therefore the L-best local scores need to be combined in order to provide the final on-line garbage score. From the resulting garbage score, utterance verification can be done using a verification threshold.

However, in this work, trying to extend the discrimination capabilities of the on-line procedure, we considered the possibility of using additional information complementary to the L-best local scores. As complementary information we have included the N-best utterance hypotheses scores provided by the recognizer. This is an information that has been traditionally used for keyword-spotting[7]. So far, in the general case, our OOV utterance verification will be made using a feature vector composed by both the L-best local scores and the N-best utterance hypotheses scores:

$$x = \{d, h\} \ : \text{ feature vector} \begin{cases} d\text{: L-best scores} \\ h\text{: N-best scores} \end{cases}$$

And the utterance verification will be made through the use of a discrimination function, f() and a verification threshold:

$$f(x) \leftrightarrow Threshold \ : \text{ utterance verification}$$

As a first approximation for the discrimination function we will test a linear function obtained through Discriminant Analysis. Linear Discriminant Analysis (LDA) is a well-known technique in statistical pattern classification [8]. The basic idea of LDA is to find a linear transformation of the feature vector x from a D-dimensional space to vector y in a d-dimensional space (d<D) such that the class separability is maximum. In our case, for a two-class classification task (OOV vs. vocabulary utterances), we will use a linear transformation to a 1-dimensional space, using a weight vector w:

$$y = w^t x$$

and the final decision will be made using a decision threshold:

$y > threshold$ --> OOV

$y < threshold$ --> vocabulary word

Several optimization criterion to find w are possible, we will use the traditional one based on the evaluation of the transformation according to the maximization of:

$$tr(T^{-1}B)$$

where tr(X) is the trace of the matrix X and T and B are the within and between covariance matrices respectively.

Using this criterion, it can be seen that the feature vector projection w can be obtained directly as:

$w = T^{-1}(m_{oov} - m_{vw})$, where

$m_{oov}$ = sample mean of the OOV word

$m_{vw}$ = sample mean of the vocabulary word

Several experiments were performed so as to test the

efficiency of the different set of possible features x={d,h}:

*Experiment 1*: x=d; only the L-best local scores are used.

*Experiment 2*: x=h; only the N-best recognition hypotheses are used.

*Experiment 3*: x={d,h}; use the L-best local scores and the N-best recognition hypotheses.

In each case a linear discrimination function was obtained using the Discriminant Analysis previously described. The performance of each case was tested using a database composed of 2948 vocabulary files and 878 OOV files. In all cases we set N=15 and L=20. Results are showed in Figure 2.

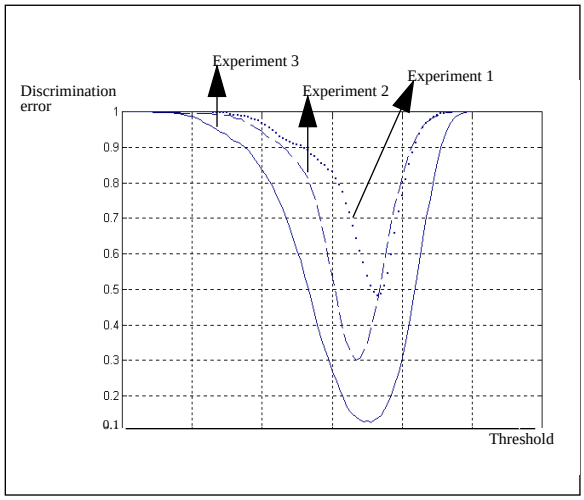


Figure 2: Several experiments using discriminant analysis.

The figure represents the discrimination error (false alarm + rejection) between Vocabulary and OOV pronunciations as a function of the discriminating error. As can be seen the lowest error is achieved by using both the L-local scores and the N-best recognition hypotheses, and the highest error is obtained by only using the L-local scores. Intermediate error values are obtained by using only the N-best recognition hypotheses.

## 4. EXPERIMENTAL RESULTS

In the previous section, we concluded, that the best discrimination is achieved by using both the N-best local scores and the L-best recognition hypotheses. Now we will test our method in two different tasks.

The first one is a natural connected numbers recognition task, which corresponds to the different ways telephone numbers are said in Spanish (not always digit-by-digit), and the second one is an isolated word recognition task of spanish surnames with high acoustic similarity.

In both tasks, the training and the recognition sets were obtained from the VESTEL [9] database, which is a telephone speech corpus collected at Speech Technology Group of Telefónica Investigación y Desarrollo, designed to support research in speaker-independent automatic speech recognition, (ASR) based on word and subword units. The database contains speakers throughout Spain, covering all dialects of Castilian Spanish.

For the spontaneously spoken natural connected numbers recognition task, we use the recordings which correspond to telephone numbers. Each file contains a variable length number string, mostly between 5 to 10 words. We use 4000 files for training and 3000 for testing, both sets balanced respect to the number of pronunciations of all dialectal zones of Castilian Spanish. In order to evaluate the False Acceptance Rate (FAR), we use another set of 900 files, containing out-of-vocabulary pronunciations of non-cooperatives speakers from VESTEL.

In Table 1 we show the results of applying on-line garbage model to this task, together with the previous results obtained by using an explicit garbage model, corresponding to an improved baseline system of the one presented in [10]. We use two different rejection levels, to compare both systems, and a 3-candidates N-best recognition system.

| System | Rejection Level | SER | | SRR | FAR |
|---|---|---|---|---|---|
| | | $1^{st}$can | $3^{rd}$can | | |
| Baseline | not applied | 9.6% | 4.4% | 0% | 100% |
| Explicit garbage model | Medium | 8.6% | 4.1% | 6.8% | 11.8% |
| | High | 8.3% | 4.0% | 9.7% | 9.7% |
| On-line garbage model | Medium | 7.6% | 3.2% | 5.1% | 7.7% |
| | High | 5.5% | 2.1% | 13.1% | 3.9% |

Table 1: Compared performance of explicit and on-line garbage modeling in a natural connected numbers recognition task

We can see that for the same rejection level, similar SRR (Sentence Rejection Rate) and much lower SER (Sentence Error Rate) and FAR (False Acceptance Rate), are obtained by using of on-line rejection techniques.

Now we will test our method in another different task, that is, an isolated word recognition task based on context-dependent units. In these systems, we previously used a garbage model constructed as a combination of context independent subword HMMs trained using only the keyword training set, as we presented in [6].

Table 2 presents the comparative results of applying explicit garbage modeling and on-line rejection

techniques, to a vocabulary-dependent isolated recognition task. We perform recognition tests for two different vocabulary sizes: one with 89-words that include the ten basic digits, the name of the most important cities of Spain and control words, with 2134 keyword files and 712 garbage files, and another with 448-words that are spanish surnames, where 2944 keyword files and 982 garbage files have been used in the recognition test.

| Vocabulary size | Rejection level | System | WER | SRR | FAR |
|---|---|---|---|---|---|
| 89-words | Medium | explicit | 1.8% | 4.4% | 28.8% |
| | | on-line | 1.9% | 4.4% | 16.4% |
| | High | explicit | 1.2% | 11.1% | 11.4% |
| | | on-line | 1.1% | 11.2% | 6.9% |
| 448-words | Medium | explicit | 9.0% | 6.6% | 41.9% |
| | | on-line | 8.9% | 6.6% | 35.2% |
| | High | explicit | 7.9% | 12.5% | 28.2% |
| | | on.line | 7.2% | 12.4% | 24.8% |

Table 2: Compared performance of explicit and on-line garbage modeling in a vocabulary-dependent isolated recognition task.

In Table 3, we present the comparison between explicit garbage techniques and on-line rejection techniques, in a vocabulary-independent isolated recognition task of spanish surnames. We use two vocabulary sizes: one of 955-words, with 1683 keyword files and 561 garbage files, and another of 2000-words, with 3037 keyword files and 1012 garbage files.

| Vocabulary size | Rejection level | System | WER | SRR | FAR |
|---|---|---|---|---|---|
| 955-words | Medium | explicit | 9.9% | 7.7% | 54.9% |
| | | on-line | 9.8% | 7.7% | 55.6% |
| | High | explicit | 5.64% | 25.0% | 22.3% |
| | | on-line | 5.41% | 25.0% | 27.6% |
| 2000-words | Medium | explicit | 12.4% | 13.9% | 44.3% |
| | | on-line | 12.5% | 13.9% | 45.8% |
| | High | explicit | 9.7% | 24% | 28.4% |
| | | on.line | 9.3% | 24% | 31.7% |

Table 3: Compared performance of explicit and on-line garbage modeling in a vocabulary-independent isolated recognition task.

From the results showed in Tables 2 and 3, we conclude that for a fixed rejection level, on-line rejection techniques provide better performance in vocabulary-dependent tasks. In vocabulary-independent tasks, the computational cost reduction is great enough in these complex tasks to allow us to obviate the slight increase in terms of false acceptance rate, and to consider this method more suitable for our real-time recognizer. Moreover it is important to note that a lower WER (Word Error Rate) is achieved in most of the cases.

## 5. CONCLUSIONS

On-line garbage based rejection techniques with discriminant analysis have shown to be an excellent way to reject OOV pronunciations, at low computational cost. We have tested this method in connected and in isolated word recognition tasks, obtaining in both cases an important improvement respect to explicit garbage modeling.

Besides, this method does not require the training of a garbage model, and provides a robust way to measure the quality of the recognized pronunciations.

## References


[1] C. de la Torre, L. Hernández-Gómez, F.J. Caminero-Gil and C. Martín-del Álamo. “On-line Garbage Modeling for Word and Utterance Verification in Natural Numbers Recognition”, ICASSP-96, Atlanta, pp. 845-848.

[2] I.L. Hetherington. “A Characterization of the Problem of New, Out-of-Vocabulary Words in Continuous-Speech Recognition and Understanding”, Ph. D. Thesis, MIT, 1995.

[3] R.C. Rose, B.H. Juang and C.H. Lee. “A Training Procedure for Verifying String Hypothesis in Continuous Speech Recognition”, ICASSP-95 Detroit, pp. 281-284.

[4] M.G. Rahim, C.H. Lee and B.H. Juang. “Robust Utterance Verification for Connected Digits Recognition”, ICASSP-95, Detroit, pp. 285-288.

[5] H. Boulard, B. D’hoore and J.M. Boite. “Optimizing Recognition and Rejection Performance in Wordspotting Systems”, ICASSP-94, Adelaide, pp. 373-376.

[6] J.C. Torrecilla, D. Tapias, J. Caminero and L. Villarrubia. “Rejection Techniques based on Context Independent Subword Units”, EUROSPEECH-95, Madrid, pp. 1633-1636.

[7] M. Weintraub. “LVCSR Log-Likelihood Ratio Scoring for Keyword-Spotting”, ICASSP-95, Detroit, pp. 297-300.

[8] R.O. Duda, P.E. Hart. “Pattern Classification and Scene Analysis”, Wiley, New York, 1973.

[9] D. Tapias, A. Acero, J. Esteve and J.C. Torrecilla. “The VESTEL Telephone Speech Database”, ICSLP-94, Yokohama, pp. 1811-1814.

[10] F.J. Caminero, C. de la Torre, L. Hernández and C. Martín. “New N-Best based Rejection Techniques for Improving a Real-Time Telephonic Connected Word Recognition System”, EUROSPEECH-95, Madrid, pp. 2099-2102.