# A multiple-Gaussian classifier for Language Identification using acoustic information and PPRLM scores

*R. Córdoba, L.F. D'Haro, R. San-Segundo, J. Macías-Guarasa, F. Fernández, J.C. Plaza*

Speech Technology Group. Dept. of Electronic Engineering. Universidad Politécnica de Madrid
E.T.S.I. Telecomunicación. Ciudad Universitaria s/n, 28040-Madrid, Spain
cordoba@die.upm.es

## Abstract

We present several innovative techniques that can be applied in a PPRLM system for language identification (LID), obtaining a 61.8% relative error reduction from our base system. First, the application of a variable threshold in score computation, dependent on the average scores in the language model, provided a 35% error reduction. A random selection of sentences for the different sets and the use of silence models also improved the system. Then, to improve the classifier, we compared the bias removal technique (up to 19% error reduction) and a Gaussian classifier (up to 37% error reduction). Then, we included the acoustic score in the Gaussian classifier (2% error reduction) and increased the number of Gaussians to have a multiple-Gaussian classifier (14% error reduction). Finally, we included additional acoustic HMMs of the same language with success (18% relative improvement). We will show how all these improvements have been mostly additive.

## 1. Introduction

Automatic language identification (LID) has become an important issue in recent years in speech recognition systems. To do language identification, first we have to identify which factors are more critical to distinguish between languages. We can identify several factors of differentiation: the realization of allophones and sounds and information related to the sequence of allophones, which has demonstrated to be vital: some sequences of allophones do not exist in one language (or occur very little), so the identification of those sequences is crucial for LID. Another possibility is to use prosodic features as the intonation may differ drastically between languages.

The most used technique is the phone-based approach, like Parallel phone recognition followed by language modeling (PPRLM) [1][2], which classifies languages based on the statistical characteristics of the allophone sequences and has a very good performance. Another popular technique is the GMM classifier, which we will not consider here. In [3] the "GMM tokenizer" is described.

Another possibility is to base the identification on the score given by a full continuous speech recognizer. As we demonstrated in [4], the results obtained with this technique are probably the best that can be obtained, as it models both acoustic and phonetic information, together with the sequence of allophones and words, but it has some important disadvantages: a complete speech recognition system has to be trained, a lot of labeled data is needed and it would be difficult to have a real-time system for several languages.

An interesting variant of PPRLM is presented in [5] with several proposals: different ways to combine the allophone sequence information with the acoustic models, use of durations (prosodic information) and a tree-based language model. It is remarkable the integration of several sources of information.

In [6] they use PPR, include bias removal to improve the classification, and include acoustic and allophone sequence information in the classifier, using a Gaussian classifier similar to the one we propose.

This is a continuation of the work done in [2] and [7]. We are going to focus now on improving the classifier, using bias removal and a multiple-Gaussian classifier mixing acoustic and allophone sequence information. This work has been done under project INVOCA, for the public company AENA, which manages Spanish airports and air navigation systems.

## 2. System description

### 2.1. Database

We use a continuous speech database (Invoca database from now on), which consists of very spontaneous conversations between controllers and pilots. For speech recognition it is a very difficult task, noisy and very spontaneous, as in "Lufthansa four two seven nine start up approved clear to Frankfurt standard departure Somosierra one echo three six left squawk one zero two three report parking position".

We have one big drawback with the database: all speakers are native Spanish. So, many of them do not reflect all the phonetic variations in English. Besides, the controllers use to mix Spanish for greetings and goodbyes even when the rest of the sentence is in English. Also, many company/ airport names are pronounced in Spanish inside the English sentence.

For the training set, we had some 8 hours of speech for Spanish and 6 hours for English. For the validation set, we had some 1 hour for both languages and 700 sentences. We have considered sentences with a minimum of 0.5 sec., and a maximum of 10 sec., with an average duration of just 4.5 sec., which is an another important limitation in our system.

### 2.2. Brief description of PPRLM

The main objective of PPRLM (Parallel Phone Recognition Language Modeling) is to model the frequency of occurrence of different allophone sequences in each language. This system has two stages. First, a phone recognizer takes the speech utterance and outputs the sequence of allophones corresponding to it. Then, the sequence of allophones is used as input to a language model (LM) module. In recognition, the LM module scores the

probability that the sequence of allophones corresponds to the language. It can use several phone recognizers modeled for different languages. Interpolated n-gram language models are used to approximate the n-gram distribution as the weighted sum of the probabilities of the n-grams considered (weights $\alpha_1$, $\alpha_2$, and $\alpha_3$ for unigram, bigram and trigram, respectively). In our case, we have considered up to trigrams.

### 2.3. Results presentation

In all our experiments we have obtained the results for all possible combinations of weights $\alpha_1$, $\alpha_2$, and $\alpha_3$, in 0.1 steps. Throughout the paper we will present the results (Sentence error rate) for the average of all weight combinations (Average column in the tables) and for the best result (Minimum column). In general, best (minimum) results occur with the biggest contribution from the trigram score, reflecting that the trigram is the most discriminative feature for language identification. In all tables, we present in parenthesis the relative improvement in relation to the base system considered.

## 3. Initial improvements to the base system

### 3.1. Threshold in score computation

As the size of the database is small, there are quite a big number of trigrams that do not have enough training samples and, so, their estimates are not reliable. We tested several alternatives for LM smoothing, but the results were very similar, showing little improvement. We decided to apply a fixed threshold or additive factor to the score value, in a similar way to the variance flooring applied in HMM estimation: use as the minimum variance a fraction of the average variance. The objective is to give more importance to the allophone sequences that have a high probability in one language and, at the same time, reduce the effect of sequences that have not appeared in training. We considered two alternatives (in the log domain):

1) **n-gram specific fixed additive factor**. We propose the following formula for the score:

$$S(F) = 10\log\left(\prod_{i=0}^{N} P_i(F)\right) = -\sum_{i=0}^{N} 10\cdot\alpha_i \cdot\log\left(P_i(F) + \beta_i\right) \qquad (1)$$

where N is the order of the N-gram, $\alpha_i$ is the weight for the $i^{th}$ n-gram and $P_i(F)$ is its probability. $\beta_i$ is the additive factor. The optimum values were $\beta_{uni}$=0.027, $\beta_{bi}$=0.04 y $\beta_{tri}$=0.08. Obviously, it is not a nice approach as β values are too empiric.

2) **Variable additive factor**. We made the $\beta_i$ dependent on the average scores in the LM ( $\overline{p}_i$ for the $i^{th}$ n-gram). λ is a smoothing factor.

$$S(F) = -\sum_{i=0}^{N} 10\cdot\alpha_i \cdot\log\left(P_i(F) + \frac{\overline{p}_i}{\lambda}\right) \qquad (2)$$

To estimate the optimum λ factor, we observed that very little differences in performance were observed using λ values between 4 and 8, which is a nice feature. In Table 1 we can see the results obtained with both alternatives: the improvement is outstanding, showing the suitability of this approach, especially for the second approach. Even though it is simple, it has been the best improvement in the experiments of this paper.

*Table 1.* Results for different additive factors

| Thresh. technique | Average | Minimum |
|---|---|---|
| None | 8.27 | 6.80 |
| n-gram specific | 7.70 (6.9%) | 5.84 (14.1%) |
| Variable | 6.26 (24.3%) | 4.46 (34.3%) |

### 3.2. Random selection of sentences / silence models

In our database, the same controller uttered a large group of sentences sequentially in the database until there was a shift change. We were afraid that our system was making some kind of speaker modeling instead of language modeling, which was confirmed by the results in Table 2. So, for similar cases we recommend to do a random selection of sentences. Besides, in our original system, we did not consider silence models in the output of the phone recognizer, but it seems that we were not estimating important trigrams which are especially relevant for language identification, e.g. 'ai-t-sil' in the word 'flight', which is extremely rare in Spanish. So, we run an experiment considering the silence models with the results shown in Table 2, with a relevant improvement.

*Table 2.* Results with random selection / silence models

| | Average | Minimum |
|---|---|---|
| Original lists | 6.26 | 4.46 |
| Randomly selected | 5.24 (16.6%) | 4.24 (5.0%) |
| + silence models | 5.03 (4.0%) | 3.92 (7.55%) |

## 4. Bias removal in the classifier

As is described in [6], the general PPRLM approach has a flaw: there is the possibility of having a different bias in the log-likelihood score for the languages considered. This is especially relevant when the phone recognizers have a different number of units (we have 49 phonetic units for Spanish and 61 for English). The language with fewer units will have higher probabilities in the LM score, and so the classifier will tend to select that language. To eliminate this bias, two options are proposed in [6]. We have experimented with the first one: we use the original LM score minus the average of all LM scores in the training database (a language-dependent bias).

**Database for bias estimation**: We can divide the training database in 3 different sets: the first one to train the acoustic models, the second one to train the LMs and the third one to estimate the bias value. This could be the optimal option if the database were large enough, as all estimations are independent. The problem is that, as our database is small, all results worsened due to insufficient training data. So, we had to discard this option. Another option is to estimate the bias value in the original training sets. We considered two possibilities:

1. Estimate the bias with the LMs training list. This is the worst option: as the LMs have been estimated using this list, the bias value estimated is not reliable because it is too optimistic.

2. Estimate the bias with the acoustic models training list. Even though this data does not participate in the LM estimation, it could be a dangerous option. But we observed that the LM score distribution in this set was very similar to the score distribution in the test set. So, we followed this option.

In Table 3 we present the results obtained using bias removal in a system without the threshold described in Section

3.1. We can see an outstanding improvement, showing that this technique is effective when there is an obvious bias in the log-likelihood score as we had presumed.

*Table 3.* Results for bias removal

| | Average | Minimum |
|---|---|---|
| No threshold | 8.27 | 6.80 |
| Bias removal | 6.98 (15.6%) | 5.5 (18.9%) |

But we have to admit that the same technique applied to the best system so far – after the threshold technique – showed no relevant improvement, just 0-1% relative. Most probably, the additive factor compensates the bias effect.

## 5. Gaussian classifier

Another possibility to tackle the issue of different bias in the LM scores is to use a Gaussian classifier instead of the usual decision formula applied in PPRLM. With all the scores provided by every LM in the PPRLM module we prepare a score vector. With all the sentences in the training database we estimate the Gaussian distribution of their respective score vectors for every language. So, we will have a Gaussian distribution for each language in the system. Now, the recognized language is not the one with the largest average score. The distance between the input vector of LM scores and the Gaussian distributions for every language is computed, and the distribution which is closer to the input vector is the one selected as identified language.

**Database for Gaussian estimation**: For the Gaussian classifier, the same considerations as for bias removal can be made. In this case, the problem addressed is even more notorious, as we need more data to estimate a reliable Gaussian distribution than we need to estimate just the bias in the score. So, again we decided to use the acoustic models training list.

**Score vector for the Gaussian classifier:** As we have several scores in the PPRLM system, there are several options for the feature vector of scores:

1. Basic. Use all PPRLM scores as is (M acoustic models x N language models, 2 x 2 in our case). This would be the typical option. The problem is that these scores are quite unstable.

2. Individual scores. We then considered the possibility to model the distribution of each n-gram in the score computation for our feature vector: the score for unigram, bigram and trigram. The drawback now is the increase in dimensionality.

3. Differential scores. Instead of using absolute values, we considered differential scores, which for every sentence are computed as the difference between the score obtained by the LM of the same language of the acoustic models considered (Spa-Spa or Eng-Eng) and the score obtained by the other 'competing' language: SC0 – SC1 and SC3 – SC2 in Figure 1. So, this score can be computed both in training and testing. We also considered the differentiation between individual scores: unigram, bigram and trigram, with 6 features in total.

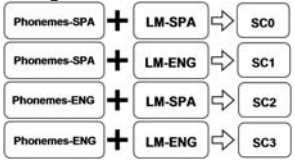


*Figure 1.* PPRLM Score average

We observed that these differential scores are much more homogeneous, being the result that the estimated distributions exhibit a much smaller overlap with the competing language.

In a multiple language system the proposal for the differential score would be:

$SC_{\text{current language}} - \text{Average}\ (SC_{\text{other languages}})$

In Table 4, we can see the results for the 3 techniques in a system without the Threshold described in Section 3.1. As can be seen, the results for the Basic and Individual options are similar and quite bad, probably because two facts: the great variations in score and the insufficient size of the database. Nevertheless, the results for the Differential scores are outstanding, more than 30% relative.

*Table 4.* Results for the Gaussian classifier

| Score vector | Average | Minimum |
|---|---|---|
| No threshold | 8.27 | 6.80 |
| Basic | 11.43 (-38.2%) | 7.7 (-13.8%) |
| Individual | 10.94 (-32.3%) | 7.7 (-13.8%) |
| Differential | 5.82 (29.6%) | 4.3 (36.8%) |

If we apply the technique with the best system so far, the minimum goes to **3.71%** (5.41% improvement), which is a smaller improvement, but are better than for the bias removal technique. Again, the improvement of the threshold technique is not additive with the Gaussian classifier. In any case, these results are a fantastic starting point, as it is easy to include acoustic information using this Gaussian classifier, and use multiple Gaussians.

### 5.1. Inclusion of acoustic information

One drawback in PPRLM modeling is that the basic technique only takes into account information regarding the allophone sequence. We propose the inclusion of acoustic information using our Gaussian classifier, adding new features to our score vector: the acoustic score obtained in the phone recognizers of both languages. We observed that the values of the acoustic score were not homogeneous at all. So, we decided to use again the "differential scores" idea: we used the difference between the score for the Spanish phone recognizer and the score for the English phone recognizer as feature value. So, we just have one feature in the acoustic score vector.

**Database considered**: Obviously, we need to estimate the acoustic score distributions using non-training data. So, the dataset chosen for this task is the LMs training list, because those sentences have not been used to train the phone models. So, we have estimated Gaussian distributions for allophone sequence scores and acoustic scores separately, as they use different lists for the estimation.

The new result using acoustic information was **3.67%** with a 2% relative improvement for the Minimum, but 13.5% for the Average. So, results show that acoustic information complements better the least robust systems.

### 5.2. Multiple-Gaussian classifier

One of the nicest characteristics of a Gaussian classifier is that we can grow up to multiple Gaussians to better model the distribution that represents our classes. We have used different number of Gaussians for allophone sequence score and acoustic

score, as their feature vector dimension is completely different: 6 and 1 features respectively. To increase the number of Gaussians we have followed the classical HMM modeling approaches (Gaussian splitting and Lloyd reestimation after each splitting), so we will not describe them here. In Table 5 we can see a summary of results obtained using different numbers of Gaussians for both scores.

*Table 5.* Multiple-Gaussian classifier

| Number of Gaussians | | Average | Minimum |
|---|---|---|---|
| LM | Acoustic | | |
| 1 | 1 | 4.69 | 3.67 |
| 2 | 1 | 4.35 (7.2%) | 3.52 (4.1%) |
| 2 | 2 | 4.14 (11.7%) | 3.31 (9.8%) |
| 3 | 1 | **4.01 (14.5%)** | 3.24 (11.7%) |
| 3 | 2 | 4.12 (12.1%) | 3.23 (12.0%) |
| 3 | 3 | 4.06 (13.4%) | 3.31 (9.8%) |
| 4 | 2 | 4.20 (10.5%) | **3.16 (13.9%)** |
| 4 | 3 | 4.08 (13.0%) | 3.17 (13.6%) |

We can extract several interesting conclusions:

- The improvements are really remarkable, up to 14% in minimum value and almost 15% in average.
- As we expected, the best system uses more Gaussians for LM score than for acoustic score.
- It is a nice feature that all systems provide better results than the mono-Gaussian system, showing that there is enough training data for the multiple-Gaussian system.
- There are better improvements in the Average value. Again, the more powerful estimation of multiple Gaussians has more relevance in the less robust systems (the ones with bigger weights for unigram and bigram).
- The difference between the Average and Minimum values has reduced drastically, showing these techniques' robustness, so the n-gram weights are less relevant.

### 5.3. Additional acoustic HMMs for the classifier

We considered the inclusion of new HMM models in our system, as it was quite easy with our Gaussian classifier. So far, nobody has reported the use of several models of the same language but different channel conditions in PPRLM. We had two additional acoustic models for Spanish: one based on SpeechDat, telephone noisy speech, and another one with speech recorded in quiet conditions ('Quiet'). So, both of them are quite different to the original Invoca database used so far. We wanted to test if some additional improvements could be obtained using them. These are the conclusions:

- Using them with no adaptation, results do not improve.
- Using them with task adaptation (models adapted using MAP with the Invoca training list) the improvements are remarkable: **2.60%** error rate with an 18% relative improvement.
- Only one of them is needed, the inclusion of both SpeechDat and 'Quiet' did not improve, probably because the increase of dimensionality in the feature vector causes a poor estimation.

So, it is clear that they can provide complementary information to the classifier when task adaptation is used.

### 5.4. Inclusion of four-grams in PPRLM

The inclusion of four-grams did not improve our system. They were poorly estimated in all cases. So, we discourage it unless for a huge database system.

## 6. Conclusions

We have described several improvements in a language identification system using PPRLM scores and acoustic information. The system has improved remarkably, up to **2.60%** error rate with an overall 61.8% relative improvement, especially considering that the average duration of the sentences is just 4.5 seconds. Increasing the sentence minimum duration to 2 seconds instead of 0.5 (5.3 seconds average duration) we obtain a **0.82%** error rate. So, most errors in our system come from extremely short sentences.

The application of the variable additive factor in score computation provided a significant error reduction in all cases. It even compensated the bias mismatch in the LM scores, as the results have shown.

For the classifier, we compared the bias removal technique (up to 19% error reduction) and a Gaussian classifier (up to 37% error reduction), showing that the last one provides better results and has the potential to include additional information. The use of differential scores to estimate the Gaussian distributions is also crucial for the technique.

The inclusion of acoustic score in the Gaussian classifier provided a 2% error reduction and the increase in the number of Gaussians provided an additional 14% error reduction. The inclusion of additional HMMs of the same language but different channel conditions provides a nice improvement if task adaptation is used.

## 7. References

[1] Zissman, M.A., "Comparison of four approaches to automatic language identification of telephone speech," IEEE Trans. Speech & Audio Proc., vol. 4(1), pp. 31-44, 1996.

[2] Córdoba, R., G. Prime, J. Macías-Guarasa, J.M. Montero, J. Ferreiros, J.M. Pardo, "PPRLM Optimization for Language Identification in Air Traffic Control Tasks". Eurospeech 2003, pp. 2685-2688.

[3] Torres-Carrasquillo, P.A., Reynolds, D.A., Deller Jr., J.R., "Language identification using Gaussian mixture model tokenization", IEEE ICASSP 2002, pp. I-757-760.

[4] Fernández, F., R. de Córdoba, J. Ferreiros, V. Sama, L. F. D'Haro, J. Macías-Guarasa. "Language Identification Techniques based on Full Recognition in an Air Traffic Control Task". ICSLP 2004, pp. II-1565-1568.

[5] Navratil, J. 2001. "Spoken Language Recognition – A Step Toward Multilinguality in Speech Processing". IEEE Trans. Speech&Audio Proc., Vol. 9(6), 2001, pp. 678-685.

[6] Ramasubramaniam, V., et al. 2003. "Language Identification using Parallel Phone Recognition". Workshop on Spoken Language Processing, India.

[7] Córdoba, R., R. San-Segundo, J. Macías, Juan M. Montero, R. Barra, L.F. D'Haro, J.C. Plaza, J. Ferreiros. 2006. "Integration of acoustic information and PPRLM scores in a multiple-Gaussian classifier for Language Identification". IEEE Odyssey 2006.