Language Identification using several sources of information with a multiple-Gaussian classifier

R. Cordoba, L.F. D'Haro, F. Fernandez-Martinez, J. M. Montero, R. Barra

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

10.21437/Interspeech.2007-577

Abstract

We present several innovative techniques that can be applied in a PPRLM system for language identification (LID). To normalize the scores, eliminate the bias in the scores and improve the classifier, we compared the bias removal technique (up to 19% relative improvement (RI)) and a Gaussian classifier (up to 37% RI). Then, we include additional sources of information in different feature vectors of the Gaussian classifier: the sentence acoustic score (11% RI), the average acoustic score for each phoneme (11% RI), and the average duration for each phoneme (7.8% RI). The use of a multiple-Gaussian classifier with 4 feature vectors meant an additional 15.1% RI. Using 4 feature vectors instead of just PPRLM provides a 26.1% RI. Finally, we include additional acoustic HMMs of the same language with success (10% relative improvement). We will show how all these improvements have been mostly additive.

Index Terms: Language identification, PPRLM, Gaussian classifier, score normalization, feature selection

1. Introduction

Automatic Language Identification (LID) has become an important issue in recent years. Most dialog systems are multilingual, so the language of the caller has to be identified as soon as possible in order to use the appropriate recognition system specific to that language.

Many techniques have been suggested in recent years for this task. The most widespread technique is the phone-based approach, like Parallel phone recognition followed by language modeling (PPRLM) [1]-[3], 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 [4] they present a GMM classifier called "GMM tokenizer", where the output of the classifier is used as input to a "language model" (LM) module, where the sequence of the different indexes is learnt. The performance of this technique is worse than PPRLM, but its combination with PPRLM improves the overall result.

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. Another approach is to use a lattice instead of the allophone sequence [6] and a neural network at the output of the classifier, instead of doing the average of the scores. This way, there is an improvement in the classifier. In our paper we propose a Gaussian classifier instead.

In [7] 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. In [8] they compare the performance of a neural network with a Gaussian classifier as ours. The neural network provides slightly better results but the use of multiple Gaussians is not mentioned. Another recent line of research is the fusion of different sources of information, as in [9], which we also address.

This paper is a continuation of the work done in [2] and especially [3]. Results for the inclusion of acoustic information are better due to a better construction of the Gaussian classifier and the inclusion of a smoothing factor for the variances. We also present new experiments including more feature vectors in our system for other sources of information. This work has been done under project INVOCA, for the public company AENA, which manages Spanish airports and air navigation systems [17].

The paper is organized as follows. We present the setup, a brief overview of PPRLM and a summary of previous results in Section 2. In Section 3 we describe our Gaussian classifier. In Section 4, we include additional sources of information, namely acoustic and phoneme duration information. Finally, in Section 5 we increase the number of Gaussians in the Gaussian classifier. The conclusions are given in Section 6.

2. System description

2.1. Database

We use a continuous speech database (referred to Invoca database from now on), which consists of very spontaneous conversations between controllers and pilots. It is quite a difficult database, noisy and very spontaneous. 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, and they mix Spanish for greetings and goodbyes even when the rest of the sentence is in English.

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 another important complication for the LID task.

2.2. General conditions of the experiments

The system uses a front-end with PLP coefficients derived from a mel-scale filter bank (MF-PLP), with 13 coefficients including c0 and their first and second-order differentials, giving a total of 39 parameters per frame. For the phone recognizers, we have used context-independent continuous HMM models. For Spanish, we have considered 49 different allophones and, for English, 61 different allophones. All models use 10 Gaussians densities per state per stream.

2.3. 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. 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 α_1, α_2, and α_3 for unigram, bigram and trigram, respectively). 4-gram LMs provide worse results.

2.4. Summary of previous initial improvements

This is a list of the main improvements that we have obtained over the basic PPRLM technique in previous works. More details can be found in [3]:

- Threshold in score: to smooth the LM, we applied an additive factor in the PPRLM formula dependent on the average scores in the LM (34% relative improvement).
- Random selection of sentences: to avoid modeling the speaker instead of the language, we created new lists using a random selection procedure, namely Fisher-Yates, with a 5% relative improvement.
- Bias removal in the classifier: to suppress bias in LM scores, we applied 'bias removal': LM score = original score minus the average of all LM scores in the training database. The improvement can be up to 19%.

3. Gaussian classifier for LID

As is described in [7], 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 true when phone recognizers have different number of units (we have 49 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 be biased to that language. To tackle this issue, we proposed in [3] to use a Gaussian classifier instead of the usual PPRLM decision formula. With all the scores provided by each LM we prepare a score vector. With all the sentences in the training database we estimate the Gaussian distribution of their respective score vectors (one Gaussian / language). The distance between the input vector and the Gaussian distributions for every language is computed, using a diagonal covariance matrix, and the distribution which is closer to the input vector is the one selected as identified language.

To estimate the Gaussian distribution we used the acoustic models training list, as this data does not participate in the LM estimation. We demonstrated in [3] that it was a good option in order to make a better use of the training list, as the LM score distribution in this set was very similar to its distribution in the test set. One important conclusion of that work is that, instead of absolute values, we need to use differential scores: 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(s): SC0 – SC1 and SC3 – SC2 in Figure 1. So, this score can be computed both in training and testing. We applied it to unigram, bigram and trigram separately, with 6 features in total that are listed in Table 1.

Figure 1. PPRLM Scores

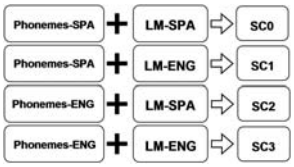

Table 1. Differential score vector

Phonemes-SPA	SCO-SC1 for unigram
	SCO-SC1 for bigram
	SCO-SC1 for trigram
Phonemes-ENG	SC3-SC2 for unigram
	SC3-SC2 for bigram
	SC3-SC2 for trigram

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:

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

One problem that has to be solved is how the weights of the n-grams α_1, α_2, and α_3 from the basic PPRLM equation (1) can be integrated in this approach, as the scores for unigram, bigram, and trigram are independent in our vector.

$$S(w_t, w_{t-1}, w_{t-2}) = \alpha_3 \cdot P(w_t \mid w_{t-1}, w_{t-2}) + \alpha_2 \cdot P(w_{t-1} \mid w_{t-2}) + \alpha_1 \cdot P(w_{t-2}) + \alpha_0 \cdot P_0 \quad (1)$$

We introduce a new contribution not described in [3]: instead of multiplying each feature by its weight in the distance measure, it was much better to divide the variance of the distribution of each score by the corresponding α_i weight (equation (2)). For low α_i, variances increase and so distances are smoothed (good for less discriminative features). This smoothing weight is quickly adjusted with good results.

$$\sigma_i^{final} = \sigma_i^{original} / \alpha_i \quad (2)$$

4. Inclusion of new sources of 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 in two complementary ways: the average acoustic score of the sentence and the average acoustic score for each phoneme. At the same time, phoneme duration generated by the phone recognizer can be very different depending on the input language, so we can take advantage of that too. For these three sources of information we will just add another feature vector in our classifier, as we will see in this section.

4.1. Inclusion of the sentence acoustic score

First, we will consider the global acoustic score of the sentence (normalized by the number of frames, obviously). We have a feature vector with two features: the acoustic score obtained in the phone recognizers for each language. Again, the approach can be easily extended to several languages.

We observed that the acoustic score values were not homogeneous at all, and so, the estimated distributions for competing languages had a big overlap. Then, we decided to use again the "differential scores" idea: we used the difference between the phone recognizer score for Spanish and English as feature value. Again, we observed that the

overlap between the estimated distributions reduced drastically. To extend this approach to several languages:

AcScore $_{\text{current language}}$ – Average (AcScore $_{\text{other languages}}$)

Database considered: Obviously, we need to estimate the acoustic score distributions using non-training data. So, the dataset chosen for this task is the LM training list.

LID results using just this feature is **8.13%** error rate after adjusting the smoothing weight α_i (0.4 as optimum value). In Table 2, we can see the results (classification error rates, with relative improvements in parenthesis) using the Gaussian classifier with two feature vectors, one for the PPRLM scores and the other one for the acoustic scores. As we can see, the improvement is remarkable, so the fusion of both feature vectors provides an additive improvement. This clearly improves the results published in [3], where 3.67 for the minimum was reported. As we have already mentioned, the difference is due to the inclusion of the smoothing weight α_i.

Table 2. PPRLM + sentence acoustic score

	Average	Minimum
PPRLM	5.42	3.74
+ acoustic score	4.40 (18.8%)	3.31 (11.5%)

4.2. Inclusion of the acoustic score for each phoneme

We now considered that the acoustic score for each individual phoneme could also have a strong variation depending on the language. Using our classifier, we modeled the Gaussian distribution for the acoustic score of each phoneme.

For each input sentence we have its corresponding sequence of phonemes using the Spanish and English phone recognizers. We compute the average score for each phoneme appearing in the sentence (averaging the score over all frames belonging to that phoneme) obtaining a feature vector with as many features as the number of phonemes in the system. Obviously, phonemes not appearing in the sentence do not contribute to the final score in the classifier.

Again, the "differential scores" approach is a must, because these scores have a strong variability. To normalize, for every frame: $SC = SC_{Spanish} - SC_{English}$, which is added for all phoneme frames. This approach is clearly better than normalizing using the sentence average score for the "competing" language.

To reduce the size of the feature vector, we grouped some allophonic variations and considered 34 different phonemes for each language. So, we have a vector of 68 features. This vector is obviously too big to have it reliably estimated. In this version of our system we decided to apply a feature selection algorithm to reduce the dimensionality: we keep the *n* features that maximize the following objective function:

$$\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2\ \sigma_2^2} \qquad (2)$$

where μ_1 and μ_2 are the mean values for the feature considering Spanish and English input sentences respectively, and σ_1 and σ_2 are the respective covariances. A high value in this formula means that the feature is very discriminative. There is a very strong correlation among this separation measure and the final results in LID. For a future version we will consider applying LDA to reduce the dimensionality. Using these feature reduction techniques, this approach can be easily applied to multiple language systems.

We tested the system using 24, 30, and 35 features, obtaining the optimum result for 30 features. In LID the best results using just this feature vector and modifying the smoothing weight α_i was **8.78%** error rate.

In Table 3, third row, we can see the results with two feature vectors, one for the PPRLM scores and the other one for acoustic scores for each phoneme, and, in the last row we can see the result using all three vectors together. The nicest conclusion is that all 3 feature vectors improve the system, which demonstrates that the information they provide is complementary. In fact, even though the LID performance of the global acoustic score (8.13) is better than for the acoustic score per phoneme (8.78), the result of the combination of PPRLM and these scores is better for the second case.

Table 3. PPRLM + ac. score per phoneme + both

	Average	Minimum
PPRLM	5.42	3.74
+ ac. score / phoneme	4.30 (20.7%)	3.31 (11.5%)
+ both acoustic scores	4.14 (23.6%)	**3.10 (17.1%)**

4.3. Inclusion of the duration for each phoneme

We considered that phoneme duration could also be different depending on the input language, so we thought that it could be easy to add just another feature vector to our Gaussian classifier. So, we modeled the Gaussian distribution for the average duration of each phoneme in our system. For each input sentence, we computed the average duration for each phoneme and the feature vector had as many features as the number of phonemes.

The bad thing of this feature is that it is quite difficult to normalize. The "differential scores" approach that we should apply here would be to subtract the average duration for the competing language, but, as the phoneme sets are different for each language, this subtraction is not possible. We considered two normalizations: a) Subtract the average phoneme duration of the competing language; b) Subtract the phoneme duration of the competing language for the phoneme which had the largest part in common with the current one, so it will be the most probable "competing" phoneme. B) was a better option.

We reduced the feature vector using the same feature selection technique as in the previous section, keeping this time 22 features as the optimum value. In LID the best results using just this feature vector and modifying the smoothing weight α_i was **22.3%** error rate. It is clearly worse than the results obtained with acoustic scores, showing that we still have a normalization problem with the durations.

Nevertheless, we checked the performance of this feature vector combined with PPRLM (Table 4). The fusion of PPRLM and duration still provided an 8% improvement, but all vectors together provide similar – slightly better – results.

Table 4. PPRLM + duration + acoustic scores

	Average	Minimum
PPRLM	5.42	3.74
+ duration	5.21 (3.9%)	3.45 (7.8%)
+ both acoustic scores	4.15 (23.4%)	**3.08 (17.6%)**

5. 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. Of course, we will need more data to have a reliable estimation of these Gaussians. We will show here that with our data we can estimate reliable multiple-Gaussian distributions using all 4

sources of information. 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.

For the sake of simplicity, to demonstrate the effectiveness of the multiple-Gaussian classifier, we present in Table 5 a summary of results obtained using different numbers of Gaussians for the first two feature vectors (PPRLM and sentence acoustic scores). In parenthesis, we show the improvement relative to the 1-1 Gaussian case.

Table 5. Multiple-Gaussian Classifier 1

Number of Gaussians		Average	Minimum
LM	Acoustic		
1	1	4.40	3.31
2	2	4.09 (7.0%)	3.16 (4.4%)
3	3	4.01 (8.9%)	2.95 (10.9%)
4	4	3.97 (9.7%)	2.95 (10.9%)
5	3	3.92 (11.0%)	**2.88 (13.0%)**

We can extract several interesting conclusions:

- The improvements are really remarkable, up to 13% in minimum value and 11% in average.
- As we expected, the best system uses more Gaussians for LM score than for acoustic score, as the feature vector dimension is 6 for LM and 1 for acoustic.
- The difference between Average and Minimum has reduced drastically, which reduces the importance of the n-gram α_i weights from equation (1) or (2).

The results using all 4 feature vectors are very difficult to present in a paper. So, in Table 6 we decided to present the relative improvement of including additional feature vectors (only for the minimum results) as a function of the number of Gaussians for the first vector (LM), selecting the optimum configuration of Gaussians in all cases. All improvements are relative to just using PPRLM. The last column shows the results using the 4 feature vectors.

Table 6. Improvements with additional vectors

Gaussians in LM	Sentence ac. sc.	Acoustic sc. per phone	Duration per phone	All
1	11.5%	11.5%	7.8%	17.1%
2	17.1%	22.6%	3.7%	24.4%
3	19.0%	19.3%	9.6%	23.0%
4	19.6%	15.8%	4.1%	25.3%
5	21.5%	19.6%	6.0%	26.1%

We can see that in all cases the new sources of information provide remarkable improvements, although they are not completely additive, as could be expected because they are all related to the phone recognizer results. In any case, we can see that their combination always provides better results than the baseline using just PPRLM or PPLRM plus the sentence acoustic scores as in [3]. The best result so far is **2.81%** error rate for the combination of 5-4-5-2 Gaussians, although there a lot of combinations with very similar results.

5.1. 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 used SpeechDat, telephone noisy speech, quite different from the Invoca database. This is a summary of results:

- Using them with no adaptation, results do not improve.
- Using them with MAP task adaptation (with the Invoca training list) the improvements are remarkable: **2.60%** error rate with a 10% relative improvement. (The result is comparable to the 2.88% from Table 5).

6. Conclusions

We have described several improvements in a language identification system using PPRLM scores and acoustic information. The final error rate has improved to 2.60%. The results are outstanding, especially considering that the average duration of the sentences is just 4.5 seconds.

The inclusion of the sentence acoustic score in the Gaussian classifier provided an 11.5% relative improvement (RI), the average acoustic score for each phoneme (11.5% RI), and the average duration for each phoneme (7.8% RI). The increase in the number of Gaussians in our multiple-Gaussian classifier with 4 feature vectors provided an additional 15.1% RI. Using 4 feature vectors instead of just PPRLM provides a 26.1% RI. The inclusion of additional HMMs of the same language but different channel conditions provides a 10% RI if task adaptation is used.

As a future line, we will consider fusion techniques to weigh each feature vector in our system, as in [9].

7. Acknowledgements

This work has been partially funded by the Spanish Ministry of Education & Science under contracts DPI2004-07908-C02-02 (ROBINT) and TIN2005-08660-C04-04 (EDECAN-UPM) and by UPM-CAM under contract CCG06-UPM/CAM-516 (ATINA).

8. References

[1] Zissman, M.A., "Comparison of four approaches to automatic language identification of telephone speech," IEEE Trans. Speech&Audio Proc., v. 4, 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] 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.

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

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

[6] Gauvain, J.L., et al. "Language Recognition using Phone Lattices". ICSLP 2004, pp. I-25-28.

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

[8] Gleason, T.P., M.A. Zissman. "Composite background models and score standardization for Language Identification Systems", ICASSP 2001, pp. 529-532.

[9] Li, J., S. Yaman, et al. "Language Recognition Based on Score Distribution Feature Vectors and Discriminative Classifier Fusion". IEEE Odyssey 2006.