n-gram Frequency Ranking with additional sources of information in a multiple-Gaussian classifier for Language Identification

Ricardo Cordoba, Luis F. D'Haro, Juan M. Lucas, Javier Zugasti

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,juanmak,jzugasti}@die.upm.es

Abstract

We present new results of our n-gram frequency ranking used for language identification. We use a Parallel phone recognizer (as in PPRLM), but instead of the language model, we create a ranking with the most frequent n-grams. Then we compute the distance between the input sentence ranking and each language ranking, based on the difference in relative positions for each n-gram. The objective of this ranking is to model reliably a longer span than PPRLM. This approach outperforms PPRLM (15% relative improvement) due to the inclusion of 4-gram and 5-gram in the classifier. We will also see that the combination of this technique with other sources of information (feature vectors in our classifier) is also advantageous over PPRLM, showing also a detailed analysis of the relevance of these sources and a simple feature selection technique to cope with long feature vectors. The test database has been significantly increased using cross-fold validation, so comparisons are now more reliable.

Index Terms: Language Identification, n-gram frequency ranking, score normalization, feature selection, PPRLM

1. Introduction

The most used technique in Language identification (LID) 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 with a very good performance. 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 [7] they compare the performance of a neural network with a Gaussian classifier as ours. Another recent line of research is the fusion of different sources of information, as in [8] or [9], which we also address.

PPRLM does not model long-span dependencies: with 4-gram language models results are slightly worse, probably due to unreliable estimation. To solve this, we decided to use a ranking of occurrences of each n-gram with higher n-grams [4], in a similar way to [6] where the ranking is applied to written text. Although the information source is very similar to PPRLM (frequency of occurrence of n-grams), results are clearly better.

This paper is a continuation of the work done in [3] with several information sources and [4]. Section 2 describes the system setup and basic techniques. In Sections 3 and 4 the n-gram ranking technique and new information sources are described. In Section 5, results are presented and discussed. Finally, conclusions are presented in Section 6.

2. System description

2.1. Database

We use a continuous speech database (Invoca), which consists of very spontaneous conversations between controllers and pilots. It is a difficult task, noisy and very spontaneous, with one big drawback: 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.

In total, we had some 9 hours of speech for Spanish (4998 sentences) and 7 hours for English (3132 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. To increase the reliability of results we have performed a cross-fold validation, dividing all the material available in 9 subsets. In each pass we dedicated:

- 4 blocks for estimating the acoustic models & the Gaussian distribution for the LMs and the ranking
- 3 blocks for estimating the language models for PPRLM and the n-gram ranking & the Gaussian distribution for the acoustic scores and duration
- 1 block for the test-set and parameter fine-tuning
- 1 block for the validation set

So, results are more reliable because they use 7 times more material and are for a validation set with unseen data. We checked in [2] that to estimate the Gaussian distribution for the LMs we could use the acoustic models training list, as this data does not participate in the LM estimation. The same applies for the distribution estimation of acoustic scores with the LMs training list.

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. 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 α_1, α_2, and α_3 for unigram, bigram and trigram, respectively). All systems using 4-gram LMs provided worse results [2].

2.4. Gaussian classifier for LID

The general PPRLM approach has a bias problem in the log-likelihood score for the languages considered, especially when the phone recognizers have a 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 tend to select that language. To tackle this issue, we proposed in [2] 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 a Gaussian distribution each language. In recognition, the distance between the input vector of LM scores 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.

One nice feature of a Gaussian classifier is that we can increase the number of Gaussians to better model the distribution that represents our classes and have a Multiple-Gaussian classifier. To increase the number of Gaussians we followed the classical HMM modeling approaches (Gaussian splitting and Lloyd reestimation after each splitting).

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

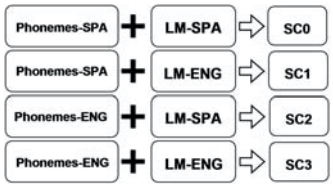

Table 1. Differential score vector

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: instead of multiplying each feature by its weight in the distance measure, it is 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 (which is good for less discriminative features). This smoothing weight is quickly adjusted with good results using the test set.

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

3. n-gram Frequency Ranking

3.1. Description

We use the same input as PPRLM: the sequence of allophones generated by the phone recognizer. As proposed in [6], we use all training data to compute the number of occurrences of each n-gram (n=1 to 5). We sort those counts, and keep only the M most frequent n-grams, which will form the ranking for that input language. It is known ([6]) that the top n-grams are almost always highly correlated to the language. So, we will use this ranking instead of the LM module considered in PPRLM (see Figure 1).

In testing, for each input sentence a ranking is created using the same procedure. Then, the distance between the input sentence ranking and each ranking is computed. The distance measure is the following (we add the difference in the ranking position for all n-grams in the input sentence):

$$d = \frac{1}{L}\sum_{i=1}^{L} abs(pos\ input_i - pos\ global_i) \quad (3)$$

where L is the number of n-grams in the input sentence. If an n-gram does not appear in the ranking (meaning that it has not appeared in training or it is not in the top n-grams selected) it is assigned the worst distance: the ranking size. The language identified by the system will be the one with the lowest distance. For the Gaussian classifier we now have 10 features in our vector (unigram to 5-gram in both languages).

In [4] we obtained the following conclusions for this technique: optimum ranking sizes range in 3000; it is better to have n-gram specific rankings, instead of a global ranking for all n-grams which include too many unigrams and bigrams which are less discriminative; and rankings should be discriminative.

We wanted to give more relevance in the ranking (higher positions) to the items that are actually more specific to the identified language, i.e. n-grams that appear a lot for one language but appear very little, or never, in the competing languages. We propose a variation of tf-idf, which is used for topic classification. Given the following normalized values:

n_1' = occurrences of item *i* in the current language

n_2' = occurrences of item *i* in the competing language (the average to extend the metric to multiple languages)

The best formula with the same philosophy as tf-idf for the final number of occurrences considered for the ranking (which we will call n_1'') is (more details in [4]):

$$n_1'' = n_1' * (n_1' - n_2') / (n_1' + n_2')^2$$

which normalizes the values between 1 and -1: 1 meaning that the n-gram appears in the current language but not in the other competing ones (n_2'=0), so it is especially relevant for that language; -1 meaning just the opposite (n_1'=0), so the n-gram does not appear in the current language.

4. Inclusion of several information sources

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 (phone recognizer score normalized by the number of frames). We have a vector with two features: the acoustic score obtained in the phone recognizers for each language. So, the approach can be easily extended to several languages.

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. To extend this approach to several languages:

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

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 large 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:

$$\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2\ \sigma_2^2} \qquad (4)$$

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. We tested the system using 24, 30, and 35 features, keeping 30 features as the optimum. To get an idea of the information provided by this objective function, in Table 2 we can see the separation which is obtained with PPRLM and n-gram ranking for each n-gram considered applying equation (4). Discrimination for the ranking trigram is very similar to the PPRLM trigram, but now we can use 4-grams and 5-grams. The separation for the sentence acoustic score is 6.84, whereas for the 30 features of the acoustic score for each phoneme it ranges from 3.52 to 0.54.

Table 2. Comparison of feature discrimination

	PPRLM	Ranking
trigram	10.57	10.12
bigram	8.54	7.12
4-gram	-	6.61
5-gram	-	4.25
unigram	3.17	2.19

An alternative to this feature selection algorithm is to apply LDA to reduce the dimensionality, which is oriented to labeled samples, as we have. Unfortunately, results were slightly worse. LDA has one advantage: it projects into a space of dimension "number of classes -1", which is 1 in our case, so the Gaussian distribution is easily estimated. It would probably work better for a multiple class classification. This will be explored as future work.

One reason of the bad results is probably the "missing values" problem: we have an original vector with 68 components corresponding to phonemes, but several of them do not appear in a sentence. The easy solution is to substitute those missing values by their mean taken from the training database, but that implies some loss of information, and the projection of the test vector is worse. So, we still have to tackle this issue.

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 problem is that this duration produced by the recognizer 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.

5. LID results

5.1. Individual features

When mixing several sources of information differences are less evident. So, we will first show in Table 3 the results of each source independently. There are several interesting conclusions:

- The n-gram ranking provides a **15.4%** relative improvement over PPRLM.
- Phoneme acoustic score is 3% better than the Acoustic sentence score.
- Phoneme duration is the worst discriminative, so we still have a normalization problem with the technique.

Table 3. LID results for individual feature vectors

PPRLM	n-gram Ranking	Sentence Acoustic	Phoneme Acoustic	Phoneme Duration
3.69	3.12	8.14	7.90	24.67

5.2. Combination of several features

In Table 4 we can see the results when combining several feature vectors and the relative improvements over the PPRLM and the Ranking base systems from Table 3. We can extract the following comments:

- Rows 1 & 2: "PPRLM + Phoneme Acoustic" is better than "PPRLM + Sentence Acoustic", as the individual results predicted.
- Row 3: The fusion of PPRLM and duration only provides a low improvement, but it could be expected.
- Row 4 & 8: PPRLM / Ranking + both acoustic scores keeps improving the system, so these scores are complementary
- Rows 5-7: The fusion of the Ranking + additional features provides similar improvements to PPRLM, a bit lower probably because they begin from a much better system.
- Row 9: The fusion of PPRLM and Ranking provides a nice improvement. This is even surprising, as they use the same information source, the n-grams.
- Rows 10 & 11: The fusion of PPRLM + Ranking + Acoustic scores provides further improvements, which shows again that they all provide complementary information.

Table 4. LID results for feature vector combinations

Feature vectors	LID	Improv. PPRLM	Improv. Ranking
PPRLM + Sentence Acoustic	3.10	16.0%	-
PPRLM + Phoneme Acoustic	3.08	16.5%	-
PPRLM + Phoneme Duration	3.49	5.4%	-
PPRLM + both Acoustics	**3.00**	18.7%	
Ranking + Sentence Acoustic	2.78	-	10.9%
Ranking + Phoneme Acoustic	2.77	-	11.2%
Ranking + Phoneme Duration	3.07	-	1.6%
Ranking + both Acoustics	**2.63**	-	15.7%
PPRLM + Ranking	2.85	22.8%	8.7%
PPRLM +Ranking+S. Acoustic	2.66	27.9%	14.7%
PPRLM+Ranking+both Acoust.	2.54	31.2%	18.6%
All	**2.52**	31.7%	19.2%

5.3. Longer span of the ranking technique

We also checked the relevance of 4-grams and 5-grams in LID with this technique. In Table 5 we can see that the LID results considering only up to 4-gram or up to trigram are worse than using all n-grams, and the trigram ranking has similar results as PPRLM. So, we are clearly taking advantage of this longer span using this technique.

Table 5. Independent ranking for each n-gram

	Best result
All n-grams	3.12
Up to 4-gram	3.30
Up to trigram	3.59

6. Conclusions

We have demonstrated that the n-gram Frequency Ranking approach can clearly overcome PPRLM thanks to the longer span that can be modeled. Even the combination of this Ranking with more feature vectors keeps improving the results, showing that all the features proposed provide complementary information (phoneme duration being the worse). The acoustic score for each phoneme is a slightly better feature than the sentence acoustic score.

The measure of separation between pdf distributions (Section 4.2) is a good tool to anticipate which features are going to be actually discriminative for the LID task. LDA provides worse results, probably because of the "missing values" problem.

As future work, we will check these results with a bigger and more "standard" database.

7. Acknowledgements

This work has been partially funded by the Spanish Ministry of Education & Science under contracts DPI2007-66846-c02-02 (ROBONAUTA) and TIN2005-08660-C04-04 (EDECAN-UPM) and by UPM-DGUI-CAM under CCG07-UPM/TIC-1823 (ANETO).

8. References

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

[2] Córdoba, R., et al. "Integration of acoustic information and PPRLM scores in a multiple-Gaussian classifier for Language Identification". IEEE Odyssey 2006.

[3] Cordoba, R., D'Haro, L.F., et al. "Language Identification using several sources of information with a multiple-Gaussian classifier". Interspeech 2007, pp. 2137- 2140. Belgium.

[4] Cordoba, R., D'Haro, L.F., et al. "Language Identification based on n-gram Frequency Ranking". Interspeech 2007, pp. 354- 357. Belgium.

[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] Cavnar, W. B. and Trenkle, J. M., "N-Gram-Based Text Categorization", Proc. 3rd Symposium on Document Analysis & Information Retrieval, pp. 161-175, 1994.

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

[8] Gutierrez, J., J.L. Rouas, R. André-Obretch. "Fusing Language Identification Systems using performance confidence indexes". ICASSP 2004, pp. I-385-388.

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