Language Identification Techniques based on Full Recognition in an Air Traffic Control Task

Fernando Fernández, Ricardo de Córdoba, Javier Ferreiros, Valentín Sama, Luis F. D'Haro

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

Abstract

Automatic language identification has become an important issue in recent years in speech recognition systems. In this paper, we present the work done in language identification for an air traffic control speech recognizer for continuous speech. The system is able to distinguish between Spanish and English. We present several language identification techniques based on full recognition that improve the baseline results obtained using the most commonly known "PPRLM" technique. We have in our database some task specific critical problems for language identification like non native speakers, extremely spontaneous speech or Spanish-English mix in the same sentence. We confirm that PPRLM is quite sensible to those problems and that a technique based on a Bayesian classifier is the one with the best performance in spite of its higher computational cost.

1. Introduction

Each day, more and more recognition systems are multilingual and need to know in a very short time the language of the user of an automatic system to use the appropriate recognition models specific to that language.

PPRLM [1] is a language identification technique based on phone sequences. It is the most widespread technique and all previous studies show that PPRLM is the technique with the best performance despite of its drawbacks: more processing time and labeled data is needed. This has been the main reason to choose it as our baseline technique in order to evaluate the different alternatives proposed. The main objective of the present work is to optimize, by means of alternative techniques, the language identification rate of our system, improving the baseline results obtained using the PPRLM technique.

There are other popular techniques like a simple GMM classifier. This technique addresses the first differential factor between languages: every language has sounds that are specific to it. Its main advantage is that we do not need labeled data to train the classifier, so it is a very cheap system. Its main drawback is its low performance, worse than PPRLM, due to the fact that it does not deal with any information regarding the sequence of sounds (the second main factor of differentiation between languages.)

There are previous experiences based on the combination of both types of techniques. These have been proposed to try to take the advantages from both techniques: a GMM classifier called "GMM tokenizer" [2] and PPRLM, whose combination improves the overall result.

In summary, there is a general agreement that PPRLM is the best option if you look for performance and have labeled data available to model the phone recognizers. In fact, it has been widely used for speaker recognition and language identification with very good results ([3] and [4]).

So, in this paper, we are going to focus on full recognition based techniques, and we will compare them to PPRLM. Also, we will enumerate a set of task specific critical problems for language identification and will try to evaluate the importance of every information source, acoustic versus linguistic information.

This work has been done under the project INVOCA, for the public company AENA, which manages Spanish airports and air navigations systems [5].

The paper is organized as follows. First, a brief overview of the PPRLM system and the reference results that are going to be compared to those obtained with the proposed techniques. Then, we present the database and the general conditions of the experiments in Section 2. In Section 3 we describe the PPRLM technique and their results, whereas in Section 4 we describe the full recognition techniques that we present in this paper and their results. The conclusions are given in Section 5.

2. System setup

2.1. Database

We use a continuous speech database, which consists of very spontaneous conversations between controllers and pilots. For speech recognition it is a very difficult task. 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. This is a decisive factor in all cases for English identification. We have a second drawback: the controllers use to mix Spanish for greetings and goodbyes even when the rest of the sentence is in English. Also, many company names and airports have the Spanish pronunciation embedded in the English conversation.

Table 1: Database (sentences / hours)

	Spanish	English
HMM training set	**4,026 / 7.1**	**2,200 / 4.7**
LM training set	**500 / 0.9**	**500 / 1.0**
Validation set	**503 / 0.9**	**453 / 0.9**

We have separated each database in three sets:

- A training set, used to generate HMM acoustic models.
- A set dedicated to train the language models.
- A third set dedicated to the validation of all alternatives.

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 (only PPRLM), we have used context-independent continuous HMM models. For Spanish, we have considered 49 different allophones and, for English, 61 different allophones. So, we have tried to cover all possible phonetic variations in both languages, specially including allophones that do not exist in the other language. All models use 10 Gaussians densities per state per stream.

For continuous speech recognizers we have used context-dependent continuous HMM models. For Spanish, we have considered 1506 clustered states, each state with 8 mixture components, and, for English, 901 clustered states, each state with 8 mixture components (as there is less training data, the optimum number of states was lower).

3. "PPRLM" (Parallel Phone Recognition Language Modeling)

3.1. Description

The main objective of this technique is to model the frequency of occurrence of different phone sequences in each language. This system has two stages. In the first stage, a phone recognizer takes the speech utterance and outputs the sequence of phonemes corresponding to it. The sequence of phonemes generated by the phone recognizers is used as input to a language model module. In the second stage, the language model module scores the probability that the sequence of phonemes corresponds to the language.

It can use several phone recognizers modeled for different languages. The advantage is that using many recognizers we can cover most of the phonetic realizations of the languages. Its main drawback is speed: processing time is multiplied by the number of recognizers. Using PPRLM, we can even have phone recognizers modeled for languages different than the languages that have to be identified, but obviously if there is a match between the input language and the language of the models the performance will be better, because you can model explicitly the phonetic variations of each language. In our case, as we want to identify English and Spanish and we have labeled data for both of them, the best option is to use PPRLM with phone recognizers trained for English and Spanish.

In the identification stage a language model module scores the probability that the sequence of phonemes corresponds to the language according to the process illustrated in Figure 1. The overall score is calculated as an average between both scores obtained for the same language according to (1). 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. In our case, we have considered up to trigrams. For a sequence of three consecutive symbols observed in the phone stream, we use the formula (2).

$$SC-CAST=\frac{SC0+SC2}{2}\ ;\ SC-ING=\frac{SC1+SC3}{2} \qquad (1)$$

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

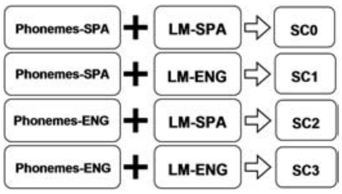

Figure 1: PPRLM Score average.

In this article we are not going to give further details about weight selection. If you want to see more details you could have a look at [2] in order to check several weight selection techniques.

3.2. PPRLM Results

The following results have been obtained for a sentence average duration of 4.6 seconds (~70 phonemes). The Spanish identification error rate is **3.8%** while the English identification error rate is **10.9%**. The overall result is **7.4%**. The English identification performance is very bad because speakers use to mix Spanish and English in their greetings and goodbyes. Another factor that can affect the performance of the system is that the database consists of extremely spontaneous live conversations between controllers and pilots.

4. Full Recognition

4.1. Comparison of Scores in Full Recognition

The low results of the PPRLM technique and the problems mentioned before, like the brief duration of some sentences, were the main reasons that induced us to study alternative language identification techniques.

This technique is based on the comparison of scores between both continuous speech recognizers corresponding to each language. Both recognizers process each sentence. The computational cost of this technique, where the two speech recognizers are working simultaneously, is higher than the PPRLM computational cost (there are two phonetic recognizers plus the continuous speech recognizer for the appropriate language). This is an important drawback especially if we need to identify several languages, but it can be acceptable if just two or three languages have to be identified, as it is our case.

The system will output two different results: the Spanish recognized sentence and the English recognized sentence. The recognition algorithm is "one-pass". In this algorithm we combine acoustic and linguistic models according to (3).

$$p(w_1^N)^\alpha\cdot p(x_1^T \mid w_1^N)IWP^{\ N} \qquad (3)$$

α weights the linguistic information over the acoustic information. These LM weights have been tuned for optimum word accuracy. Those weights, α(Spanish)=9.5 and α(English)=11, are high, so they give the LM an appropriate relevance and it is possible to do language identification from the score difference between both recognizers.

4.1.1. Results

For the validation set, the Spanish and the English identification error rates are **0.4%** and **6.2%** respectively. The overall result is **3.1%**. That gives a relative improvement of **57.6%** over PPRLM.

4.1.2. Error Rate Variation with LM weight

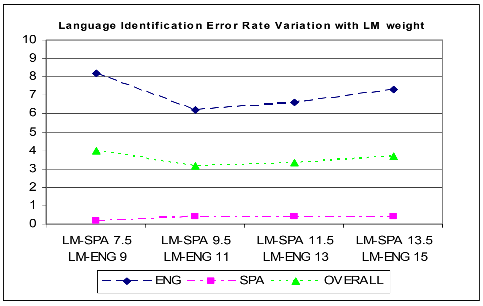

Figure 2: Error Rate Variation with the LM weight.

In Figure 2, the identification error rates for Spanish, English and overall as a function of the LM weight are presented. The best results are those obtained using the weights tuned for optimum word accuracy. This is probably due to the better accuracy of the resulting sentences in a grammatical sense. The experiments show that any different combination of weights has worse performance. Therefore, there is a strong dependency on the linguistic information.

4.1.3. Using Linguistic Information Only

Next, we are going to try to separate the linguistic from the acoustic information. We want to assess the importance of both information sources separately. First, we will evaluate the language identification rate just from the linguistic information source. In order to do that, we will compare the grammar scores obtained applying each LM to its corresponding recognition sentence.

We cannot say that we are strictly using just linguistic information. It is essential to point out that the recognition sentences that have been processed with the LMs have been obtained using the acoustic information encoded by the recognition acoustic models.

The Spanish and the English language identification error rate obtained just from linguistic information are **11.9%** and **4.8%** respectively. The overall result is **8.6%**. This result confirms that, indeed, the combination of both information sources, linguistic and acoustic, gives a better performance than any of them alone. We get a significant improvement considering the acoustic information.

4.1.4. Using Acoustic Information Only

Finally, we will evaluate the language identification rate just from the acoustic information source. In order to do that, we will take the difference between the global score and the grammar score corresponding to the LM.

Once again, we have to emphasize that we are not considering only acoustic information strictly. In the global score estimation from the one-pass algorithm, it is implicit the use of linguistic information from the applied LM. If we wanted to use just acoustic information in a strict way, we will have to assign null weight to the LM during the recognition process.

The language identification error rates for Spanish and English are **1.0%** and **12.6%** respectively. The overall result is **6.5%**. From this result we can extract the same conclusion as in the previous section: the results are worse than those obtained using both linguistic and acoustic information. However, one significant result is that this last technique improves the PPRLM results. This gives us the idea that no optimum PPRLM performance has been reached. The reason is, probably, that the LMs training in PPRLM has been poor as we did not use enough data.

Another very interesting issue is that the result for Spanish is much better than for English, whereas using linguistic information this was just the opposite. The main reason is that all the speakers are Spanish, so the acoustic information tends to predict that the speaker is Spanish.

This also explains in part the worse results for PPLRM for Spanish speakers (from 3.8% to 0.4% error rate): as the acoustic models used in the full recognition are much more detailed they detect more easily that the speaker is in fact Spanish.

4.2. Bayesian Classifier based in Comparison of Scores

As in the previous section, this technique is going to be based on full recognition, but this time we are going to apply a Bayesian classifier for both languages scores.

Unfortunately, in the Bayes classifier, or the minimum error rate classifier, we rarely have complete knowledge of class-conditional pdfs and/or prior probabilities. In this case, the prior probabilities estimation is easy. The class conditional pdf estimation is more complicated and there is always concern to have sufficient training data. We are going to use a parametric method to estimate the class conditional pdf from the training data. We have used both the HMM training set and the LM training set for this. We assume a Gaussian distribution for the pdf.

The classes of our problem are two: Spanish sentences and English sentences. We model each class with a Gaussian pdf of dimension two, (d=2). Thus, every parameter vector consists of a pair of components that are the normalized scores (global score divided by the number of frames) for each full recognition in each language. We have chosen such parameters because of their discriminative behavior between both classes, as it is shown in Figures 3 and 4 for training data.

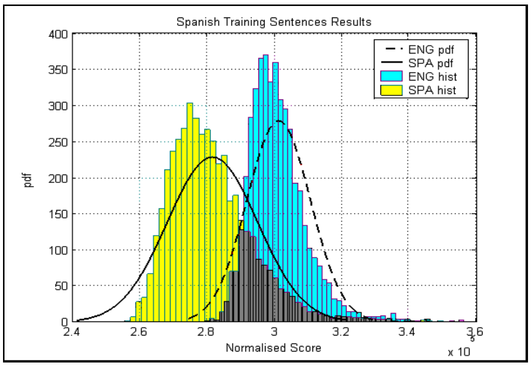

Figure 3: Class-conditional pdfs for Spanish training data

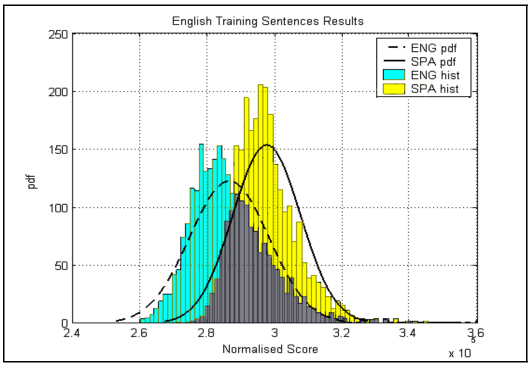

Figure 4: Class-conditional pdfs for English training data

From both figures it can be observed that the parameters considered are more discriminative for the Spanish language, as the common dark area is smaller. This imbalance is due to the fact that the speakers for the English sentences database are non-native.

Once we have estimated the μ mean vectors and the Σ covariance matrices that model both classes, we use (5) as our discriminative function for our classifier. The decision process assigns the class j to the sentence x according to (6).

$$\begin{aligned} d_i(x) &= \log p(x \mid \omega_i)P(\omega_i) = \\ &= -\frac{1}{2}(x-\mu_i)^t\,\Sigma_i^{-1}(x-\mu_i) + \log P(\omega_i) - \frac{1}{2}\log|\Sigma_i| - \frac{d}{2}\log 2\pi \end{aligned} \qquad (5)$$

$$j = \arg_i \max d_i(x) \qquad (6)$$

4.2.1. *Bayesian Classifier Results*

Using the validation test we get a relative improvement of **46.8%** over the scores comparison technique and of **77.4%** over PPRLM. The Spanish and the English language identification error rates are **0.4%** and **3.1%** respectively. The overall result is **1.7%**.

4.2.2. *Bayesian Classifier Without Prior Probabilities*

In this case we are going to assume that both classes have the same prior probability, which will be more realistic. The Spanish and the English language identification error rates are **0.4%** and **2.9%** respectively. The overall result is **1.6%.** If we compare this result to the one obtained taking into account prior probabilities, we get a relative improvement of **6%**.

4.3. Summary of Results

Table 2: Summary of results (error rates)

Tech.	SPA	ENG	ALL	Rel. Improv.
PPRLM	3.8	10.9	7.4	-
Scores	0.4	6.2	3.1	45.2%
Scores + Bayesian	0.4	3.1	1.7	77.4%
Scores + Bayesian (no priors)	0.4	2.9	1.6	78.4%

5. Conclusions and Future Work

It has been identified a set of critical problems for language identification. Most of such problems are task specific: the sentences duration, an extremely spontaneous speech, the speakers' mixing of different languages in the same sentence (like greetings and goodbyes), or non-native speakers.

PPRLM has advantage over the proposed techniques in its lower computational cost. Nevertheless, it is more sensible than the others to the mentioned set of problems.

The technique based on comparison of scores is more robust than PPRLM in that sense. Moreover, it is an extremely easy technique. It also doesn't need neither additional training, as for phonetic recognizers in PPRLM, nor pdfs parametric estimation, as in the Gaussian classifier. On the other hand, its computational cost is higher than PPRLM. Another drawback is that it is an ad-hoc solution which cannot be always applied with success.

We have tried to assess the importance of both types of information, linguistic and acoustic, in terms of language identification. For the task that has been evaluated in this article, it has been detected that the acoustic information is more helpful than the linguistic, mainly due to the lack of data to train the LMs. However, the most significant conclusion is that the joint use of both information sources results in an optimum performance.

The Bayesian Classifier technique is the one with the best performance with an overall relative improvement of 78% over PPRLM.

As future work, we plan to apply the same Bayesian Classifier technique using additional useful sources of information. We want to extend the type of parameters used by the classifier. We want to check the effect of introducing the number of frames divided by the number of recognized words as a measure of differentiation between languages.

We also plan to test this technique with an English continuous speech database with native speakers.

6. References

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

[2] 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.

[3] Córdoba, R., et. al., "PPRLM Optimization for Language Identification in Air Traffic Control Tasks". Eurospeech 2003, pp. 2685-2688.

[4] Jin, Q., Schultz, T., Waibel, A., "Phonetic Speaker Identification", ICSLP 2002, pp. 1345-1348.

[5] INVOCA Project Synopses. Eurocontrol. Analysis of Research & Development in European Programs. Available at http://www.eurocontrol.int/eatmp/arde-parda/servlets/SVLT014?Proj=AEN043.