# APPLYING FEATURE REDUCTION ANALYSIS TO A PPRLM-MULTIPLE GAUSSIAN LANGUAGE IDENTIFICATION SYSTEM

*Juan Manuel Lucas Cuesta, Ricardo de Córdoba Herralde, Luis Fernando D'Haro Enríquez*

Grupo de Tecnología del Habla
Departamento de Ingeniería Electrónica
E.T.S.I. Telecomunicación, Universidad Politécnica de Madrid
Ciudad Universitaria s/n. 28040. Madrid

## ABSTRACT

This paper presents the application of a feature selection technique such as LDA to a language identification (LID) system. The baseline system consists of a PPRLM module followed by a multiple-Gaussian classifier. This classifier makes use of acoustic scores and duration features of each input utterance. We applied a dimension reduction of the feature space in order to achieve a faster and easier-trainable system. We imputed missing values of our vectors before projecting them on the new space. Our experiments show a very low performance reduction due to the dimension reduction approach. Using a single dimension projection the error rates we have obtained are about 8.73% taking into account the 22 most significant features.

## 1. INTRODUCTION

Automatic language identification (LID) has become a cornerstone task in multilingual environments. For an automatic customer care system which could be used for users that speak in different languages, a language-specific speech recognition module has to be used. So, determining the language in what the user speaks is a need in order to adapt further steps of a dialogue system.

The most widespread LID approach consists of using several phoneme recognizers in parallel. At the output of those recognizers, a phoneme language model is applied for each language to be identified. This technique is known as *Parallel Phone Recognition followed by Language Modeling* (PPRLM). Examples of this approach can be seen on [1] or [2].

Each of the phonemes of a given language can be estimated by using Gaussian Mixture Models (GMM, [3]) or Hidden Markov Models (HMM). A GMM-based LID system can be improved with a *clustering* algorithm that groups the feature vectors on an unsupervised approach, according to a distance criterion ([4]).

As an alternative to these probabilistic approaches, [5] or [6] develop neural network-based LID systems that lead to identification rates comparables to the obtained with PPRLM-based systems.

This work continues the presented in [1], [7] and [8], which present a LID system based on PPRLM. The performance of the baseline system is improved with the implementation of a multiple-Gaussian classifier. This subsystem takes its decisions using as input vectors the acoustic score of each phoneme within the input utterance, or the duration of those phonemes.

The number of features of each input vector is high, so the training and the evaluation of the Gaussian models takes an large fraction of processing time. In order to tackle this drawback, a feature selection algorithm such as LDA is proposed.

Since a given phoneme can or cannot appear on an utterance, several features may be missing on an input vector. This fact can cause a reduction of the system performance. To avoid this weakness we have analyzed several missing data imputation algorithms.

The rest of the paper is organized as follows. Section 2 presents a brief description of the dimensionality reduction approach that has been employed. The different imputation methods we have implemented are shown in Section 3. Our baseline LID system is then presented in section 4. Section 5 summarizes the different experiments we have carried out. Finally, Section 6 presents several conclusions of our work.

## 2. FEATURE SELECTION

The Gaussian classifiers we use as a second classification stage make use of 68-dimensional feature vectors. These 68 features are the acoustic scores of the phonemes of each target language (English and Spanish, 34 features each).

We propose a feature selection technique to reduce processing time and resources. Our system can choose the most representative features according to the following criterion:

$$\frac{\mu_1 - \mu_2}{\sigma_1^2\ \sigma_2^2} \tag{1}$$

being $\mu_1$ and $\mu_2$ the arithmetic means of a given feature considering each language, and $\sigma_1^2$ and $\sigma_2^2$, their corresponding variances. Higher values of 1 for a given feature

imply a better separation between the classes.

Despite the goodness of this approach, which can lead to better results using just 22 features instead of the whole feature vector, we want to analyze the behaviour of our system when a more restrictive reduction is applied. The chosen approach is Linear Discriminant Analysis (LDA), which is explained in [9].

We have chosen LDA because it is oriented to labeled samples, that is, the algorithm makes use of the language of each training utterance. Furthermore LDA tries to increase the separability between different class data, so it can be efficiently used for class discrimination.

LDA consists of applying a linear transform over a data set of *d* dimensions which project on a *d'*-dimensional subspace (*d'* < *d*), with *d'* equal to the number of languages minus 1 (in our case, *d'* = 1). The transformation is made in such a way that the between-class variance is maximized, while the within-class variance is minimized. This can lead to an improvement of the separability between the classes.

## 3. MISSING DATA TREATMENT

Our feature vectors consist of the acoustic score for each phoneme that has been observed on each input utterance. This fact implies that a given feature may not appear in a given vector. This could happen if the speaker has not used that phoneme or the system has not recognized it.

This lack of information is a drawback when dimension reduction is applied, because those missing values usually lead to a biased estimation of the optimal transformation vector. So the implementation of an imputation technique ([10]) that can fill those missing features is a must.

We have chosen two different imputation techniques: a substitution based on the arithmetic mean of the non-missing features, and a modification of the imputation approach proposed in [11].

The *mean imputation* procedure consists of evaluating the arithmetic mean of each feature using the non-missing values on the training data. Let $x_1, \cdots, x_n$ be a set of *n* *d*-dimensional feature vectors, which could have several missing values. The arithmetic mean of each feature *j* can be obtained as

$$\overline{X_j} = \frac{1}{n_j} \sum_{i=1}^{n} x_{ij} \tag{2}$$

where $n_j < n$ is the number of vectors for which the feature *j* is not missing.

This method is very simple and is accurate for classification purposes if a supervised training is carried out. However, mean substitution does not take into account the variance of each feature, so it can cause a bias in the estimation.

The imputation procedure proposed in [11] (henceforth referred to as *Bingham imputation*) computes the cross-case mean among the different non-missing features of a given vector *i*,

$$\overline{I_i} = \frac{1}{k} \sum_{j\ non-missing} \left(x_{ij} - \overline{X_j}\right) \tag{3}$$

where *k* is the number of non-missing features in vector *i* and $x_{ij}$ is the value of the non-missing feature *j* of vector *i*.

The imputed value $\tilde{E_{ij}}$ of the missed feature *j* of vector *i* is computed as follows:

$$\tilde{E}_{ij} = \overline{X_j} + \overline{I_i} \tag{4}$$

So, the objective of this imputation is to include an offset in the imputation value that reflects the tendency that acoustic scores exhibit for the non-missing features in the vector.

This imputation method is especially effective when the different features are very correlated, because it weights each feature mean with the rest of the features in the vector. Nevertheless, this method does not take into account the feature variance. To tackle this lack we have modified the former definitions of cross-case mean and final imputation value:

$$\tilde{I}_i = \frac{1}{k} \sum_{jnon-missing} \frac{\left(x_{ij} - \overline{X_j}\right)}{\sigma_j} \tag{5}$$

$$\tilde{E}_{ij} = \overline{X_j} + \sigma_j\ \tilde{I}_i \tag{6}$$

where $\sigma_j$ is the variance of feature *j*.

The inclusion of the variance provides a normalization of this imputed value, so that the values are more stable, avoiding the presence of outliers.

## 4. BASELINE SYSTEM

### 4.1. Database

Our database consists of a set of continuously spoken sentences extracted from conversations between airplane pilots and air traffic controllers. All speakers were native Spanish.

We have used 2929 Spanish sentences and 1053 English sentences. By applying a leave-one-out technique we have used each sentence for both training and evaluating the system, but obviously in separate sets. This way we expand the size of the test set. We have not considered those sentences whose duration is less than 0.5 seconds.

Each phoneme recognizer makes use of context-independent continuous hidden Markov models (HMM). We have considered 49 different phonemes for Spanish and 61 for English. However, we have grouped the less representative phonetic variations and built phoneme vectors of 68 features, 34 for each language.

### 4.2. PPRLM-based LID system

The PPRLM identification system makes use of a phoneme recognizer for each target language. A language model module scores the probability that the sequence of phonemes corresponds to a given language.

We have used smoothed *n*-gram language models to approximate the *n*-gram distribution as the weighted sum of the probabilities of the *n*-grams considered.

We improved the PPRLM approach by taking into account silence models, defining and using a smoothing function in the evaluation of the *n*-gram score, and removing bias in the classifier. The baseline error rate is about 3.7% using only PPRLM.

### 4.3. Gaussian classifier

We have used a second identification system that includes acoustic information of each phoneme. We built a feature vector with the phonemes that the PPRLM system has recognized. We computed an average score for each phoneme appearing in the sentence. Instead of using absolute scores for each phoneme, our previous work ([1]) demonstrated that we can achieve better identification rates by using differential scores obtained by the LM. We then applied equation (1) to get the most representative features. The best results that we have achieved showed an error rate of 7.9% when we use the acoustic score of each phoneme and keep 30 features in the reduced space. If we use the phoneme duration instead, error rate takes a value of 24.7%. This implies that phoneme duration is a much less discriminative feature, at least the way we have implemented it. These results will be our baseline.

## 5. EXPERIMENTS

### 5.1. LDA with mean substitution

The first imputation procedure we have implemented consists of substituting each missing value with the arithmetic mean of the corresponding feature and applying LDA. As the original feature space, we have used the 68-dimensional feature vectors as well as the selection of the most representative features, according to equation (1). The error rates are shown in Table 1 together with the relative improvement over the system without LDA (7.9% error rate) and the average percentage of missing values on the original feature space.

| No of features | Error rate (%) | Improve (%) | Miss feat (%) |
|---|---|---|---|
| 68 | 12.48 | -58.0 | 31.63 |
| 30 | 9.46 | -19.7 | 30.64 |
| 22 | 8.76 | -10.9 | 26.41 |
| 20 | 8.94 | -13.2 | 26.08 |

**Table 1**. *Error rates with LDA and mean substitution.*

The former average is similar for all setups with different number of features (close to 30%). This means that the most discriminant features also present a high number of missing values. So, the imputation of those missing values is still crucial.

We can also see how a pre-selection of the most representative features leads to a more accurate LDA projection. Nevertheless, the use of a low space dimension implies an information loss.

### 5.2. LDA with original Bingham imputation

Our second test makes use of the imputation algorithm presented in [11]. The following table summarizes the results we have obtained as well as the improvement in relation to the previous experiment.

| No of features | Error rate (%) | Improve (%) |
|---|---|---|
| 68 | 11.00 | 11.9 |
| 30 | 9.36 | 1.1 |
| 22 | 8.82 | -0.7 |
| 20 | 9.03 | -1.0 |

**Table 2**. *Error rates with LDA and Bingham substitution.*

If we compare these results with the previous ones we can see that Bingham imputation yields lower error rates when considering 68 and 30 features.

### 5.3. LDA with weighted Bingham imputation

We next weighted the cross-case mean of Bingham imputation by the variance of the corresponding feature, following equation (5). The different error rates for each input feature space are shown in Table 3, together with the relative improvement over the mean substitution-based experiments.

| No of features | Error rate (%) | Improve (%) |
|---|---|---|
| 68 | 12.24 | 1.9 |
| 30 | 9.21 | 2.6 |
| 22 | 8.73 | 0.3 |
| 20 | 9.02 | -0.9 |

**Table 3**. *Error rates with LDA and weighted Bingham imputation.*

This results are slightly better than those obtained with mean substitution, except for the case of 20 features.

All the previous results are shown in Figure 1.

### 5.4. LDA applied to phoneme duration

If we consider the duration of each phoneme and apply both missing value imputation approaches we obtain the following results.

We can obtain a relevant improvement over the original error rate (24.7%). Despite the error rates are clearly higher than those obtained with acoustic scores, when we

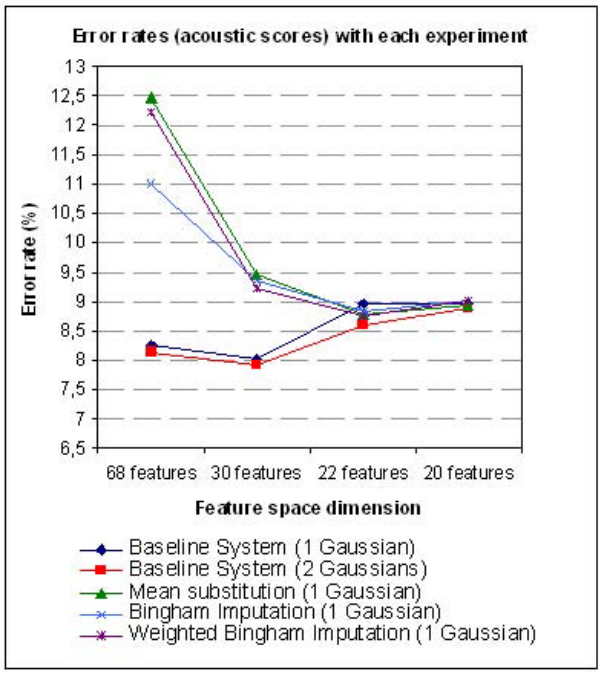


**Figure 1**. *Error rate comparison for acoustic scores.*

| | Mean substitution | | Basic Bingham imputation | |
|---|---|---|---|---|
| **No of param** | **Error rate (%)** | **Relative diff (%)** | **Error rate (%)** | **Relative diff (%)** |
| 68 | 22.77 | 7.70 | 23.90 | 3.12 |
| 30 | 22.79 | 7.62 | 24.31 | 1.46 |

**Table 4**. *Error rates for LDA applied to phoneme duration.*

use both score and duration features we can improve the overall performance (8.6% error rate with 22 features and mean substitution).

## 6. CONCLUSIONS

In this work, we present a feature selection approach that makes use of several missing data imputation techniques in order to complete the input vectors with a low distortion. The increase in error rate due to the dimensionality reduction for the acoustic scores is relatively small, and the identification task becomes easier and faster for a multiple-language task.

The different imputation approaches allows us to accurately predict the values of the most representative features, so the results are very similar to those obtained with the original feature space, but using 1 dimension instead of 22. The best of the applied techniques has been variance-weighted Bingham imputation with 22 original features, with a slight improvement regarding the other techniques. Performance is even similar to the baseline system using 22 features.

## 7. ACKNOWLEDGEMENTS

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

## 8. REFERENCES

[1] R. Córdoba et al., "Integration of acoustic information and PPRLM scores in a multiple-Gaussian classifier for language identification," *IEEE Odyssey*, 2006.

[2] K.C. Sim and H. Li, "Fusion of contrastive acoustic models for parallel phonotactic spoken language identification," *Interspeech*, pp. 170–173, 2007.

[3] Q. Dan, W. Bingxi, and Z. Qiang, "Two discriminative training schemes of GMM for language identification," *IEEE International Conference on Signal Processing (ICSP)*, pp. 630–633, 2004.

[4] B. Yin, E. Ambikairajah, and F. Chen, "Hierarchical language identification based on automatic language clustering," *Interspeech*, pp. 178–181, 2007.

[5] J. Braun and H. Levkowitz, "Automatic language identification with recurrent neural networks," *IEEE World Congress on Computational Intelligence and Neural Networks*, vol. 3, pp. 2184–2189, 1998.

[6] L. Wang, E. Ambikairajah, and E.H.C. Choi, "Multi-layer Kohonen self-organizing feature map for language identification," *Interspeech*, pp. 174–177, 2007.

[7] R. Córdoba et al., "A multiple-Gaussian classifier for language identification using acoustic information and PPRLM scores," *IV Jornadas en Tecnología del Habla*, pp. 45–48, 2006.

[8] R. Córdoba, L.F. D'Haro, F. Fernández-Martínez, J.M. Montero, and R. Barra, "Language identification using several sources of information with a multiple-Gaussian classifier," *Interspeech*, pp. 2137–2140, 2007.

[9] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, Wiley Interscience, second edition, 2001.

[10] E. Acuña and C. Rodríguez, "The treatment of missing values and its effect in the classifier accuracy," *Classification, Clustering and Data Mining Applications*, pp. 639–648, 2004.

[11] C.R. Bingham, M. Stemmler, A.C. Petersen, and J.A. Graber, "Imputing missing data values in repeated measurement within-subject designs," *Methods of Psychological Research*, vol. 3, no. 2, pp. 131–155, 1998.