A Straightforward Method for Automatic Identification of Marginalized Languages

Ana Lilia Reyes-Herrera, Luis Villaseñor-Pineda, and Manuel Montes-y-Gómez

Language Technologies Group, Computer Science Department,
National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico.
{ana_reyes, villasen, mmontesg}@inaoep.mx

Abstract. Spoken language identification consists in recognizing a language based on a sample of speech from an unknown speaker. The traditional approach for this task mainly considers the phonothactic information of languages. However, for marginalized languages –languages with few speakers or oral languages without a fixed writing standard–, this information is practically not at hand and consequently the usual approach is not applicable. In this paper, we present a method that only considers the acoustic features of the speech signal and does not use any kind of linguistic information. The experimental results on a pairwise discrimination task among nine languages demonstrated that our proposal is comparable to other similar methods. Nevertheless, its great advantage is the straightforward characterization of the acoustic signal.

1 Introduction

Automatic language identification consists in recognizing a language based on a sample of speech from an unknown speaker. There are two main approaches for this task. The first approach is based on the use of the phonothactic information of languages. It differentiates languages by the proportion and combination of phonemes in the elocutions. In particular, it considers the segmentation of the speech signal into phonemes and the use of a language model –which capture all possible combinations of phonemes from a particular language– to determine the language at issue [1, 2]. On the other hand, the second approach does not take into consideration the phonothactic information. It identifies languages exclusively using acoustic features from the speech signal such as the prosody [3], the rhythm [4] and some others perceptual features [5].

At present, the best classification results have been achieved by the first approach [1]. However, its application requires carrying out a study on the target languages in order to determine all valid phoneme combinations as well as their probabilities of occurrence. This study can be only completed for well-systematized languages, i.e., it can be done for languages having a fixed writing standard and an ample set of digital documents available. Unfortunately, this is not case for most marginalized languages, and especially, it is not the case for most of the 62 indigenous languages of Mexico.

In this paper, we propose a straightforward method for language identification. This method just considers the acoustic features of the speech signal and does not apply any linguistic information of the languages. In particular, it characterize the speech signals by set of general –language independent– features that capture the variations in the Mel frequency cepstral coefficients and take advantage of the secondary frequencies.

The proposed method will encourage the construction of systems for automatic identification of indigenous languages, which will facilitate the medical and judicial assistance of more than five million of monolingual indigenous speakers.[1]

The rest of the paper is organized as follows. Section 2 describes some previous works on language identification using acoustic features. Section 3 describes a straightforward characterization of the speech signal, which is specially suited for the language identification task. Section 4 shows the experimental results on a pairwise discrimination task among nine languages. Finally, section 4 depicts our conclusions and future work.

2 Related Work

Just a few works has tackled the problem of spoken language identification without using the phonothactic information of languages. These works are founded on the hypothesis that each language has its own rhythm (indeed, linguistics clusters languages in three major rhythmical groups: sylabe-timed, stress-timed, mora-timed). One of the first works in trying to classify languages under this assumption was that of Cummings et al. [3]. In this work, the authors proposed exploiting the variations in the fundamental frequency to perceive the rhythm of the speech. Table 1 shows their experimental results on a pairwise discrimination task among five languages. For the experiments, they implemented a neural net and used the OGI_TS corpus [6]. In particular, they considered 50 different speakers per language for training and 20 for test, and used speech samples of 50 seconds.

Table 1. Accuracy percentages reported by Cummins et al. [3]

	German	Spanish	Japanase	Mandarin
English	52	62	57	58
German	-	51	58	65
Spanish	-	-	66	47
Japanese	-	-	-	60

In other relevant work, Rouas [4] proposed a method for language identification based on the rhythm. It recaptured the linguistic theory of Ramus [7], and tried to characterize the speech rhythm in function of its vocalic and consonantal intervals. According to Ramus, the duration of these intervals determines the rhythm of the languages. Therefore, to obtain the parameters of the rhythm, Rouas segmented the speech signal in intervals formed by vowels and in intervals formed by consonants.

[1] Initially, the idea is assisting a call operator to identify the used language, and therefore to contact an adequate interpreter who provide the required assistance.

In practice, he used the fundamental frequency F0 of each segment to obtain the intonation parameters. He considered four parameters: the stocking, the standard deviation, the F0 skewness and the F0 kurtosis. In order to probe his method Rouas used nine languages of the OGI_TS corpus, and generated a classifier –based on the Gaussian Mixtures Models– for each pair of languages. For the experiments, he considered samples of 45 seconds. Table 2 shows their experimental results.

Table 2. Accuracy percentages reported by Rouas [4]

	German	Spanish	Mandarin	Vietnamese	Japanese	Korean	Tamil	Farsi
English	59.5	67.7	75.0	67.7	67.6	79.4	77.4	76.3
German	-	59.4	62.2	65.7	65.8	71.4	69.7	71.8
Spanish	-	-	80.6	62.1	62.5	75.9	65.4	66.7
Mandarin	-	-	-	50.0	50.6	73.5	74.2	76.3
Vietnamese	-	-	-	-	68.6	56.2	71.4	66.7
Japanese	-	-	-	-	-	65.7	59.4	66.7
Korean	-	-	-	-	-	-	62.1	75.0
Tamil	-	-	-	-	-	-		69.7

Finally, Samouelian [5] proposed an alternative method for the signal characterization. First, he breaks the signal into fixed segments and obtains 12 Mel frequency cepstral coefficients for each segment. Then, he computes the deltas of these coefficients. That is, he calculates the change of each coefficient between to contiguous segments. This way, each signal is represented by a set of deltas. In order to probe the representation, he generated a decision tree (based on the C4.5 algorithm) from a training corpus of 50 speakers of three different languages extracted from the OGI_TS corpus. He obtained 53% of accuracy when using samples of 45 seconds, and 48.6% when using 10 seconds samples.

It is important to mention that the results reported by Samouelain correspond to a multi-class classifier (3-languages: English, German and Japanese), while the other two works report results on a pairwaise (binary) classification task.

The signal characterization proposed in this paper extends Samouelain's ideas. On the one hand, it uses 16 Mel frequency cepstral coefficients instead of just 12. This increment in the number of coefficients allows a better description of the secondary frequencies. On the other hand, it not only considers the changes between contiguous segments, it also includes the deltas among non-contiguous signal segments. The following section details the proposed acoustic characterization.

3 Acoustic Characterization

In this paper, we propose a straightforward characterization of the acoustic signal. This characterization allows differentiating languages by their rhythm, but avoids the demanding representation of the vocalic and consonantal intervals. It is based on two simple ideas.

On the one hand, we represent the acoustic signal by fixed-size segments and characterize each segment using the Mel Frequency Cepstral Coefficients (MFCC). We take this representation from speech recognition. In this task, it is common to use only 12 MFCC since it has been empirically demonstrated that the use of more coefficients does not improve the accuracy [8]. However, for language identification, we suggest to employ additional coefficients in order to obtain more detail on the secondary frequencies. This suggestion is supported in the works by Cummings et al. [3] and Samouelain [5], which indirectly demonstrated that using the fundamental frequency is not sufficient for this task.

On the other hand, we consider that the Mel frequency cepstral coefficients cannot directly capture the rhythm of the speech. Therefore, we propose expressing their information by a set of more general –and time independent– features. In particular, we characterize the signals by their coefficient's variations. That is, we calculate the change of the coefficient's values between two signal segments. Different to Samouelain's proposal, we not only compute the differences between adjacent segments, but also the changes between non-contiguous fragments (two or three positions away from each other). This idea allows our characterization to represent the rhythm, since it presumably captures the changes at the syllabic level.

In order to enrich the acoustic characterization we also compute the averages of the coefficient's variations as well as their maximum and minimum values. In total, we use 192 features to represent each signal sample.

Table 3. The proposed set of features

Description	**Calculation**	**#Features**
Average value of the coefficients	$\tilde{c}_i = \frac{1}{N}\sum_{k=1}^{N} c_{ik}$	16
Maximum value of the coefficients	$\hat{c}_i = \max_{k=1}^{N}\left(c_{ik}\right)$	16
Minimum value of the coefficients	$\breve{c}_i = \min_{k=1}^{N}\left(c_{ik}\right)$	16
Average value of the coefficient's changes	$\tilde{\Delta}_1 c_i = \frac{1}{N-1}\sum_{k=2}^{N} c_{ik} - c_{i(k-1)}$ $\tilde{\Delta}_2 c_i = \frac{1}{N-2}\sum_{k=3}^{N} c_{ik} - c_{i(k-2)}$ $\tilde{\Delta}_3 c_i = \frac{1}{N-3}\sum_{k=4}^{N} c_{ik} - c_{i(k-3)}$	48
Maximum value of the coefficient's changes	$\hat{\Delta}_1 c_i = \max_{k=2}^{N}\left(c_{ik} - c_{i(k-1)}\right)$ $\hat{\Delta}_2 c_i = \max_{k=3}^{N}\left(c_{ik} - c_{i(k-2)}\right)$ $\hat{\Delta}_3 c_i = \max_{k=4}^{N}\left(c_{ik} - c_{i(k-3)}\right)$	48
Minimum value of the coefficient's changes	$\breve{\Delta}_1 c_i = \min_{k=2}^{N}\left(c_{ik} - c_{i(k-1)}\right)$ $\breve{\Delta}_2 c_i = \min_{k=3}^{N}\left(c_{ik} - c_{i(k-2)}\right)$ $\breve{\Delta}_3 c_i = \min_{k=4}^{N}\left(c_{ik} - c_{i(k-3)}\right)$	48

Table 3 describes the used features. It focuses on the description of the features related with each one of the 16 Mel frequency cepstral coefficients. In this table, C_{ik} denotes the coefficient i of the segment k, N indicates the number of considered segments, and Δ_1, Δ_2, and Δ_3 represent the coefficient's changes between fragments separated by one, two and three positions respectively.

4 Experimental Results

The motivation of our work was the identification of marginalized languages, especially, the identification of Mexican indigenous languages. However, in order to evaluate and compare our proposal with other methods we decided to carry out the experiments using the standard OGI_TS corpus [6]. Particularly, we considered nine languages from this corpus: English, German, Spanish, Japanese, Chinese Mandarin, Korean, Tamil, Vietnamese and Farsi. We excluded the French, since it was recently eliminated from the corpus.

The OGI Multilanguage Telephone Speech Corpus consists of recordings of telephone calls (8 KHz), where people spontaneously answer questions such as: describe the way to your work?, describe your house?, how is the weather in your country?, etc. For the experiments, we considered 50 different speakers for each language, and selected samples of 10 and 45 seconds per speaker. We used four different classifiers (KNN, Support Vector Machines, Naïve Bayes and C4.5) in order to be able to validate the proposed signal characterization. In addition, we used the information gain for dimensionality reduction, and the 10-fold cross-validation as evaluation scheme.

Table 4 shows the results corresponding to the samples of 45 seconds. These results were achieved using Naïve Bayes, which was indeed the best classifier in the whole experiments. From this table, it is clear that our results constantly outperformed those reported by Rouas et al. [4] (indicated in parenthesis), even though the proposed characterization method is much simpler than that of them. As explained in section 2, they used the rhythm units of the signal (e.g., the relation-ship between the vocalic and consonantal intervals) as main features, and the Gausssian Mixture Models (GMM) as classification technique.

Table 4. Accuracy percentages using samples of 45 seconds

	German	Spanish	Mandarin	Vietnamese	Japanese	Korean	Tamil	Farsi
English	**77** (59.5)	**88** (67.7)	73 (**75.0**)	**73** (67.7)	**82** (67.6)	79 (**79.4**)	**88** (77.4)	**83** (76.3)
German	-	**50** (59.4)	**75** (62.2)	**58** (65.7)	**62** (65.8)	65 (**71.4**)	**75** (69.7)	64 (**71.8**)
Spanish	-	-	78 (**80.6**)	**77** (62.1)	**72** (62.5)	72 (**75.9**)	**67** (65.4)	63 (**66.7**)
Mandarin	-	-	-	**72** (50.0)	**78** (50.6)	64 (**73.5**)	**79** (74.2)	**75** (76.3)
Vietnamese	-	-	-	-	**72** (68.6)	**71** (56.2)	68(**71.4**)	**79** (66.7)
Japanese	-	-	-	-	-	**65** (65.7)	**70**(59.4)	**76** (66.7)
Korean	-	-	-	-	-	-	**77**(62.1)	63 (75.0)
Tamil	-	-	-	-	-	-	-	**75** (69.7)

We performed the experiments using four different classifiers. In this case, our purpose was to demonstrate the pertinence of the proposed signal characterization. Mainly, we tried to prove that we could obtain similar results using different classification techniques. Figure 1 shows the average accuracy of each classifier per each language. The figure indicates that Naïve Bayes and SVM reached the best results. On the contrary, KNN and C4.5 achieved –in the majority of cases– the lowest results. However, it is noticeable that the four classifiers are relatively consistent. Therefore, we can assert about the pertinence of the characterization. That is, we confirmed that the reached results are a consequence of the characterization and not only a result of the selected classification algorithm.

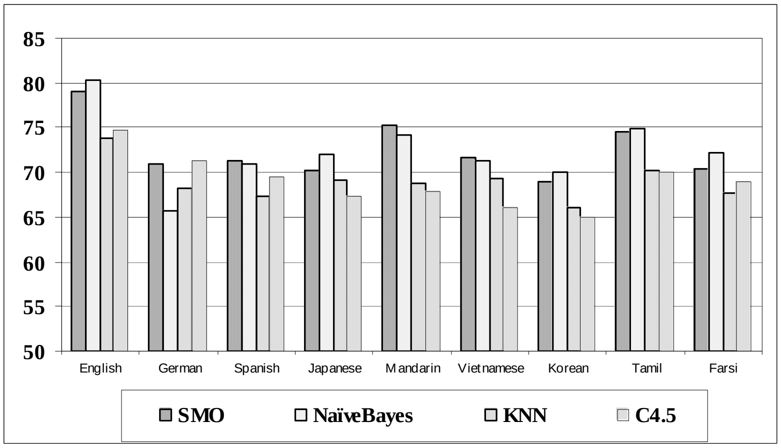

Figure 1. Average accuracy per language using different classifiers

We performed a third experiment using samples of 10 seconds. The objective was to determine the convenience of the proposed characterization when using small samples, which are –indeed– commonly used for language identification. Table 5 shows the results obtained by the Naïve Bayes classifier. Comparing these results with those of table 4 we can observed some variations.

Table 5. Accuracy percentages using samples of 10 seconds

	German	Spanish	Mandarin	Vietnamese	Japanese	Korean	Tamil	Farsi
English	86	87	75	85	87	77	89	84
German	-	75	83	81	68	71	69	77
Spanish	-	-	79	73	69	69	52	61
Mandarin	-	-	-	83	70	61	80	74
Vietnamese	-	-	-	-	68	68	59	64
Japanese	-	-	-	-	-	69	68	61

In order to emphasize these variations, table 6 presents the average accuracy per language. In most cases, the variations were lesser than ± 2%. However, for English and German there is a noticeable difference favoring samples of 10 seconds, while for Tamil the best results were obtained using samples of 45 seconds.

Table 6. Comparison of accuracies using samples of 45 and 10 seconds

	English	German	Spanish	Mandarin	Vietnamese	Japanese	Korean	Tamil	Farsi
45 seconds	80	66	71	72	74	71	70	75	72
10 seconds	84	76	71	70	76	73	69	70	72

Finally, we applied the proposed method for the identification of two indigenous languages of Mexico, namely, the Náhualt and the Zoque [9]. This experiment considered 20 different speakers per language, samples of 10 seconds per speaker, the naïve Bayes classifier, and a 10-cross-fold validation schema. The achieved results were very satisfactory and encouraging (see table 7). However, we believe it is necessary to perform more experiments, with bigger corpora, in order to confirm the pertinence of our method for the treatment of Mexican indigenous languages.

Table 7. Classification between Náhualt and the Zoque

	Náhuatl	Zoque	Accuracy
Náhuatl	16	4	87.5%
Zoque	1	19	

5 Conclusions

In this paper, we presented a straightforward method for spoken language identification task. This method considers an acoustic characterization specially suited for the identification of marginalized languages, where there are not sufficient elements to apply the phonothactic approach.

We evaluated the proposed signal characterization in a pairwise discrimination task among nine languages. The achieved results were comparable to others from similar methods. However, our signal characterization is much simpler. It represents the signal through the changes in the Mel frequency cepstral coefficients and takes advantage of the secondary frequencies.

We also evaluated our signal characterization using four different classification techniques. This evaluation demonstrated the pertinence of the proposed characterization.

Although current results are encouraging, it is still necessary to do more experiments in order to determine with greater precision the scope of the characterization as well as to understand the accuracy variations caused by the sample sizes.

Acknowledgements

This work was done under partial support of CONACYT (scholarship 184660, and project Grants 43990 and U39957-Y) and SNI-Mexico.

References

1. D. Casseiro, and I. Troncoso (1998). *Language Identification Using Minimum Linguistic Information.* 10th Portuguese on Pattern Recognition RECPAD'98. Lisbon, Portugal, 1998.
2. O. Andersen, and P. Dalsgaard (1997). *Language Identification based on Cross-Language Acoustic models and Optimized Information Combination*. EUROSPEECH-97. Rhodes, Greece, 1997.
3. F. Cummins, F. Gers, and J. Schmidhuber (1999). *Language Identification from Prosody without explicit Features*. EUROSPEECH'99. Budapest, Hungary, 1999.
4. J.-L. Rouas, J. Farinas, F. Pellegrino and R. André-Obrecht (2003). *Modeling prosody for language identification on read and spontaneous speech*. IEEE ICASSP-2003, Hong Kong, 2003.
5. A. Samouelian (1996). Automatic *Language Identification using Inductive Inference*. 4th International Conference on Spoken Language Processing ICSLP-96. Philadelphia, USA, 1996.
6. Y. K. Muthusamy, R. Cole, B. Oshika (1992). *The OGI multi-language telephone speech corpus*. International Conference on Spoken Language Processing. Alberta, Canada, 1992.
7. F. Ramus, M. Nespor, J. Mehler. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition* 73(3), pp. 265-293. Elsevier, 1999.
8. X. Huang, A. Acero, H-W. Hon (2001). *Spoken Language Processing. A Guide to Theory, Algorithm ans System Development.* Prentice Hall, 2001.
9. H. Johnson and J. Amith (2005). http://www.ailla.org: *Archive of the Indigenous Languages of Latin America*. Access=public. Texas University. USA.