Automatic Language Identification using Wavelets

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
Luis Enrique Erro #1, Tonantzintla, Puebla, 72840, 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. This method applies a wavelet transform to extract the acoustic features of the speech signal. The experimental results on a pairwise discrimination task among nine languages demonstrated that this approach considerably outperforms other previous methods based on the sole use of acoustic features.

Index Terms: spoken language identification, acoustic features, wavelet transform.

1. Introduction

The language identification problem consists on recognizing a language based on a sample of speech from an unknown speaker. There are two main approaches for this task. The most popular one uses the phonotactic content of each language. It is based on the segmentation of the speech signal in phonemes, and on the use of language models –which capture all possible combinations of phonemes from a particular language– to determine the language at issue [1, 2]. The other approach does not take into consideration any 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, 6]. However, this approach requires carrying out a study on the target languages in order to determine all valid phoneme combinations as well as their probabilities of occurrence (i.e., the phonotactics of the language). This study can only be completed for well-systematized languages, which have a fixed writing standard and an ample set of digital documents available. Unfortunately, this is not the case for most marginalized languages, and especially, it is no the case for most of the 62 indigenous languages of Mexico.

In this paper, we describe a method specially suited for the identification of languages that lack of phonothactic information. This 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 monolingual indigenous speakers. However, it is important to mention that due to its generality, this method may be applied to recognize any language, including those clearly systematized.

The proposed method uses the wavelet transform to characterize the speech signal. It is supported on previous applications of wavelets in image, speech recognition and speaker identification [7, 8, 9]. Nevertheless, in our knowledge, this is the first attempt on using wavelets for language identification. In particular, our method characterizes the speech signal by a set of features that captures the variation in the wavelets coefficients.

The rest of the paper is organized as follows. Section 2 describes the speech processing using the wavelet transform. Section 3 shows the experimental results on a pairwise discrimination task among nine languages. It also analyses the pertinence of proposed characterization for language identification. Finally, section 4 depicts our conclusions and future work.

2. Speech processing

The wavelet transform decomposes a signal into successive levels of low-to-high frequency, in what is known as multi-resolution [10]. This characteristic allows wavelets to produce a detailed description of signals and to make a clear distinction between the low and high frequencies. This is especially important for our application, since low frequencies enclose some acoustic phenomena such as rhythm, which is fundamental for language identification. On the other hand, the wavelets do not require dividing the signal sample in small segments in order to obtain its global description. This property differentiates them from other common approaches used in speech recognition.

The method proposed in this paper uses the Daubechies db2 wavelet transform with four coefficients and normalized to [-1, 1]. As we mentioned, the central idea of this method is to support the language identification on the low frequencies of the speech signal. In order to distinguish the low frequencies, we consider the magnitude of the coefficients. It is well-known that large-magnitude coefficients represent low frequencies, and low-magnitude coefficients correspond to high frequencies [11]. Thus, we truncate the wavelet coefficients according to their magnitude. Basically, we maintain the wavelet coefficients larger than some given threshold, and remove (set to zero) the rest of them. Adjusting the threshold value it is possible to vary the fraction of relevant coefficients. In our experiments (refer to section 3), we applied a threshold value that allowed maintaining just 1% of the original set of wavelet coefficients.

Before the construction of classifiers –one for each pair of languages–, it is necessary to apply a procedure for dimensionality reduction. This procedure consists of two main steps. The first one eliminates all coefficients that were truncated out. That is, it eliminates the coefficients set to zero for all instances of both languages. The second step, on the other hand, applies the information gain measure [12] in order to identify the more useful coefficients for discriminating between the languages at hand. For instance, when we constructed the classifier for English/German, we computed the wavelet coefficients for each sample, obtaining 131,072 coefficients. Then, we removed the lower coefficients, maintaining just the 1% (1,310) of the originals. After that, we selected the relevant coefficients for both languages (i.e., those with non-zero values for at least one sample). In particular, for this classifier, we identify 37,875 relevant coefficients. Finally, we applied the information gain measure and reduced the coefficients to only 641. Obviously, this process was done for each pair of languages.

3. Experiments and results

In order to evaluate and compare our proposal with other methods, we decide to carry out some experiments using the standard OGI_TS corpus [13]. Particularly, we considered nine languages from this corpus: English, German, Spanish, Japanese, Chine Mandarin, Korean, Tamil, Vietnamese and Farsi. We excluded the French, since it was recently eliminated from the corpus.

The OGI Multilanguage Telephone Speech Corpus consist 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. In total we used 450 different speakers.

To validate our method we made several comparisons between different pairs of languages. The first experiment considered five languages (English, German, Spanish, Japanese and Mandarin), and the second one nine (including also the Korean, Tamil, Vietnamese and Farsi). We selected these languages because they were formerly used by Cummins et al. [3] and Rouas et al. [4].

Table 1. *Accuracy rates using samples of 10 seconds*

	Ger	Spa	Jap	Man
Eng	**94** (52)	**96** (62)	**94** (57)	**85** (58)
Ger	-	**80** (51)	**84** (58)	**83** (65)
Spa	-	-	**86** (66)	**90** (47)
Jap	-	-	-	**89** (60)

Table 2. *Accuracy rates using samples of 45 seconds*

	Ger	Spa	Jap	Man
Eng	**97** (52)	**97** (62)	**96** (57)	**93** (58)
Ger	-	**93** (51)	**98** (58)	**94** (65)
Spa	-	-	**92** (66)	**91** (47)
Jap	-	-	-	**95** (60)

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 10-fold cross-validation as evaluation scheme.

Tables 1 and 2 show the results from a first experiment using samples of 10 and 45 seconds respectively. These results were achieved using Naïve Bayes. These tables also compare our results with those obtained by Cummins et al. [3] (indicated in parenthesis).

Cummins used the fundamental frequency of the signal (prosody) as the main feature, and a neuronal network (LSTM model) as the classification method.

In all the cases our approach outperforms the results by Cummins, confirming that the wavelet transform enables a good frequency resolution and therefore the extraction of a pertinent set of features. These results also show that the greater the speech samples, the greater the discrimination rates.

Table 3 shows the results corresponding to nine languages and samples of 45 seconds. These results were achieved using Naïve Bayes. From this table, it is clear that our results constantly outperformed those reported by Rouas et al. [4] (indicated in parenthesis), which used the rhythm units of the signal (e.g., the relationship between the vocalic and consonantal intervals) as main features, and the Gausssian Mixture Models (GMM) as classification technique.

Table 4 shows the results obtained when using samples of 10 seconds. It exhibits the behavior of wavelets when using small samples, which are –indeed– commonly used for language identification.

In order to emphasize the accuracy variation caused by the size of the sample, table 5 presents the average accuracy per language. Again, the conclusion is that the greater the speech samples, the greater the discrimination rates.

Table 3. *Discrimination rates using nine languages and samples of 45 seconds.*

	German	Spanish	Mandarin	Vietnamese	Japanese	Korean	Tamil	Farsi
English	**97** (59.5)	**97** (67.7)	**93** (75.0)	**94** (67.7)	**96** (67.6)	**95** (79.4)	**99** (77.4)	**96** (76.3)
German	-	**93** (59.4)	**94** (62.2)	**93** (65.7)	**98** (65.8)	**98** (71.4)	**94** (69.7)	**91** (71.8)
Spanish	-	-	**91** (80.6)	**86** (62.1)	**92** (62.5)	**98** (75.9)	**91** (65.4)	**94** (66.7)
Mandarin	-	-	-	**95** (50.0)	**95** (50.6)	**93** (73.5)	**89** (74.2)	**94** (76.3)
Vietnamese	-	-	-	-	**93** (68.6)	**96** (56.2)	**95** (71.4)	**95** (66.7)
Japanese	-	-	-	-	-	**93** (65.7)	**89** (59.4)	**94** (66.7)
Korean	-	-	-	-	-	-	**95** (62.1)	**91** (75.0)
Tamil	-	-	-	-	-	-	-	**90** (69.7)

Table 4. *Discrimination rates using nine languages and samples of 10 seconds.*

	German	Spanish	Mandarin	Vietnamese	Japanese	Korean	Tamil	Farsi
English	94	96	85	88	94	83	98	83
German	-	80	83	87	84	83	80	82
Spanish	-	-	90	84	86	88	87	79
Mandarin	-	-	-	85	89	83	85	94
Vietnamese	-	-	-	-	85	84	83	86
Japanese	-	-	-	-	-	83	75	89
Korean	-	-	-	-	-	-	86	87
Tamil	-	-	-	-	-	-	-	86

Table 5. *Comparison of accuracies using samples of 45 and 10 seconds*

	English	German	Spanish	Mandarin	Vietnamese	Japanese	Korean	Tamil	Farsi
45 seconds	96	95	93	93	93	94	95	93	94
10 seconds	90	84	86	87	85	86	85	85	86

Finally, 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 comparable results using different classification techniques. Figure 1 shows the average accuracy of each classifier per each language.

The figure 1 indicates that Naïve Bayes, SVM (polynomial kernel) and KNN (k=5) reached the best results. On the contrary, C4.5 achieves 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.

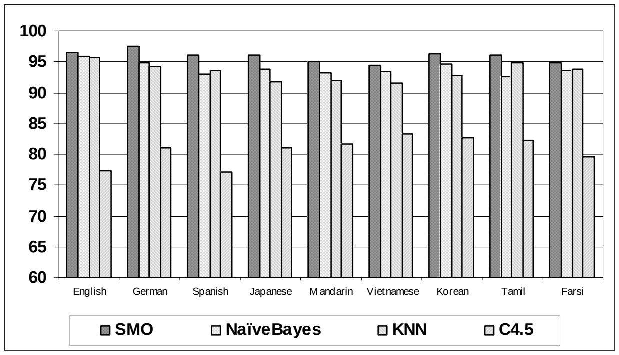

Figure 1. *Average accuracy using different classifiers*

5. Conclusions and future work

The present work demonstrated the benefit of using the wavelet transform to extract the acoustic features from a speech signal for the language identification task. The proposed method allows treating all languages, including marginalized languages, since it does not depend on any kind of linguistic information. The achieved results are encouraging because they clearly outperform the results from other similar methods based on the sole use of acoustic features.

Additionally, our experiments using different classification techniques demonstrated the pertinence of the signal characterization, because they confirmed that the results are a consequence of the characterization and not only an effect of the selected classifier.

Finally, we can assert that proposed method is quite sensible to the sample size, the greater the speech samples, the greater it discrimination rates. This observation indicates that it is necessary to enhance the signal characterization to work with small samples.

As future work, we plan to extend the method in order to work with multi-class classifiers (remember that the reported results correspond to a set of pairwise –binary– classifiers). This modification will allow comparing our approach with other methods, such as that of Sai Jayram, et al. [14] and that of Casseiro, et al. [1]. The first one does not use any phonothactic information and achieves 68% of accuracy for discriminating among 6 languages. On the other hand, the second method considers phonothactic information and achieves accuracies as high as 80% for discriminating among 6 languages.

6. Acknowledgements

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

7. References

[1] D. Casseiro, I. Troncoso, "Language Identification Using Minimum Linguistic Information", in 10th Portuguese on Pattern Recognition (RECPAD'98), Lisbon Portugal, 1998.

[2] O. Andersen, P. Dalsgaard, "Language Identification based on Cross-Language Acoustic models and Optimized Information Combination", in EUROSPEECH-97, Rhodes, Greece, pp. 67-70. 1997.

[3] F. Cummins, F. Gers, and J. Schmidhuber, "Language Identification from Prosody without explicit Features", Proc. EUROSPEECH'99, Budapest, Hungary, 1, pp. 371-374, 1999.

[4] J.-L. Rouas, J. Farinas, F. Pellegrino and R. André-Obrecht "Modeling prosody for language identification on read and spontaneous speech" in Proc. IEEE ICASSP 2003, vol 1, pp. 40-43, 2003.

[5] A. Samouelian, "Automatic Language Identification using Inductive Inference", in 4th International Conference on Spoken Language Processing (ICSLP 96), Philadelphia, USA. 1996.

[6] J. Navrátil, W. Zühlke, "Double-bigram decoding in phonotactic language identification", Proc. ICASSP-97, Munich, Germany, 1997.

[7] M. Gupta and A. Gilbert, "Robust speech recognition using wavelet coefficient features", in IEEE Automatic Speech Recognition and Understanding Workshop, USA, pp. 445-448, 2001.

[8] R. Modic, B. Lindberg, B. Petek. "Comparative wavelet and MFCC speech recognition experiments on the Slovenian and English speechDat2", in ISCA Tutorial and Research Workshop on non-linear speech processing (NOLISP 03), Le Croisic, France, 2003.

[9] Ching-Tang Hsieh, Eugene Lai, You-Chuang Wang: "Robust Speaker Identification System Based on Wavelet Transform and Gaussian Mixture Model". J. Inf. Sci. Eng. 19(2): 267-282 (2003)

[10] I. Daubechies, "Ten lectures on Wavelets", Vol. 61, SIAM Press, Philadelphia, PA USA, 1992.

[11] S. Mallat, "A Wavelet Tour of Signal Processing", Academic Press. USA 1998.

[12] Hall, M. and Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data Engineering. 15(3), November/December 2003.

[13] Y.K. Muthusamy, R. Cole, B. Oshika, "The OGI multi-language telephone speech corpus". International Conference on Spoken Language Processing, volume 2, Alberta Canada, 1992.

[14] Sai Jayram, A, Ramasubramanian, V. & Sreenivas (2002). Automatic Language Identification Using Acoustic Sub-Words Units. In 7th International Conference on Spoken Language Processing (ICSLP 2002).