# Automatic Language Identification : Alternative Approaches to Phonetic Modeling[1]

François PELLEGRINO, Arnaud GALINIER, Régine ANDRE-OBRECHT

Institut de Recherche en Informatique de Toulouse
IRIT - 118, Route de Narbonne
F-31062 Toulouse Cedex - France

(pellegri, galinier, obrecht@irit.fr)

*ABSTRACT*

This paper deals with our research on vowel system modeling in a Language Identification (LId) purpose.

The phonological study of vowel systems shows that a typology of the languages may be issued from the description of their vowel system (vowel number and acoustic space topology). We try to exploit this knowledge, to improve the phone recognizer usually used as front end of the Automatic LId systems.

In a previous work, we have developed a vowel detection algorithm based on a spectral analysis of the acoustic signal and requiring no learning stage. Currently, we propose two alternative approaches to model the phonetic space :

- for the first one, the vocalic space is modeled as a Gaussian mixture and we use the LBG-Rissanen algorithm to estimate it. A language identification is provided through this simple knowledge.
- for the second one, we have used Hidden Markov Models (HMM) to describe the global phonetic space. The decoding process involves an automatic classification into four broad classes and a modeling of each class.

We use the Oregon Graduate Institute Multi-Language Telephone Speech Corpus (OGI_MLTS) to assess our approaches. We limit presently our experiments to five languages (French, Japanese, Korean, Spanish and Vietnamese). As we don't use any phone language modeling stage, the correct identification rate (about 61%) is a promising result and offers many perspectives.

## 1. INTRODUCTION

Many efforts have been focused on the automatic extraction of information from speech signals, to provide Interactive Voice Systems (IVS) : Automatic Speech Recognition systems have reached good performance and Text to Speech Synthesis systems are currently used.

With the development of the world communication, the need for an automatic Language Identification system as an IVS front end is crucial and speech processing in the framework of multilingual applications, becomes a great challenge.

Most LId systems operate in two phases ; after an acoustic processing, a digital-symbolic translation is provided by an acoustic phonetic decoder ; classification methods are based on Gaussian mixtures, classical HMMs, Neural Networks... and the symbolic sequence is evaluated with a stochastic language model [Muthusamy 94]. To train the models, the systems require an orthographic transcription of each utterance and a pronunciation dictionary, or a phonetic transcription. These systems have became more accurate, but the number of identified languages have increased more slowly as they require too much labeled data and time to be extended to a new language. An alternative approach consists in extracting automatically and storing a set of prototypical spectra for each language, but the acoustic phonetic decoder role is generally limited to provide a phonetic transcription of the sounds, the acoustic identification score is rarely taken into account.

Our work focuses on the improvement of the phonetic decoder efficiency by introducing phonological knowledge about each language vowel system, and getting an identification score as early as this stage.

The study of vowel systems is a main topic of phonetics and linguistics [Maddieson 84, Vallée 95, Carré 95]. The phonological studies show that one of the more discriminative features among languages is their vowel systems. Therefore, taking advantage of this knowledge in automatic language identification tasks is a rational idea, but it is an open problem.

Our approach is based on an automatic vowel detection and a stochastic vowel model by language. Both identification scores and input features are provided to a language modeling stage.

The second section provides basic information on vowels and vowel detection [Pellegrino 97]. These specific sound patterns may be characterized by different methods and this topic is also discussed.

The third section deals with the problem of a global vocalic space characterization : the goal is to reduce the set of features necessary to handle all the phonetically different vowels of the language. It results in a vocalic system model for the language. Our approach is based on Vector Quantization (VQ) ; from a statistical point of view, we reduce the vowel space description size by extracting a set of relevant acoustic references. In our

[1] This research is supported by the French "Ministère de la Défense" as part of an agreement with DGA.

multilingual context we compare the classical LBG algorithm [Linde 80] to an improved one where the LBG algorithm is coupled to an information theory criterion , the LBG-Rissanen [Rissanen 83]. The fourth section describes the global phonetic decoder based on an automatic pre classification and an HMM model. The last section describes language identification experiments, performed with a 5-language subset of the OGI Telephone Speech Database [Lander 95].
At this time the vocalic simple approach seems to give better results as its performance reaches 61 % of correct identification for 5 languages.

## 2. FROM VOWELS TO VOCALIC SPACE

Vowels can be characterized from several points of view. They are produced by both a vibration of the vocal cords and an opening of the vocal tract. From the acoustic point of view the vocal tract behaves like a resonator and it results in resonance frequencies in the speech spectrum. These so-called formants can be used as a model for each vowel. Numerous previous works proved that the acoustic space in which vowels are produced is limited [Vallée 95]. Formants are distinctive features and ideally, they adequately describe vowels. However, complex phenomena as coarticulation or vowel reduction [Schwartz 89, Van Bergem 95] outcome a merging of the different vowels in a more complex representation of the vocalic space. Furthermore, male and female speakers develop different strategies of speech production and their vocal tracts present distinctions. It results in an intricate vocalic space [Henton 95]. Even if the formantic description is the best way to manually handle vowels, an automatic system would be limited by the complexity of formant detection and by the necessity to have robust parameters.

The cepstral coefficients offer a solution to these problems, they have proved their efficiency in automatic speech recognition. MFCCs (Mel Frequency Cepstral Coefficients), calculated in a perceptual frequency scale, are undoubtedly relevant in automatic systems. Nevertheless, the automatic vowel detection is a difficult task because of the existence of voiced consonants similar to vowels when they occur in a vocalic context. We develop a multilingual algorithm [Pellegrino 97] that requires no learning phase. This algorithm is able to extract features for each acoustic sample, but the identification of the underlying basic quality requires a more global modeling. To match this purpose in a speaker or language identification task, we propose to use a probabilistic model. Our approach is exposed in the next section.

## 3. VOWEL SYSTEM MODELING

Among the numerous statistical methods available to model a set of data, we explore the efficiency of Vector Quantization. From the raw vocalic data given by the automatic detection stage, we search a multi-dimensional reference map of the patterns. This way, a vocalic system model appears as a Gaussian Mixture Model where each Gaussian density is characterized by the mean and the variance of each acoustic reference. To build this reference map, we propose two alternative, the LBG algorithm and the LBG-Rissanen algorithm.

### 3.1. Baseline Vector Quantization Algorithm

The baseline system we implement is a standard LBG splitting method [Linde 80]. Assuming that each vocalic quality can be modeled by a Gaussian distribution, the algorithm reaches at least local optimality. To adjust the number of codewords, the user fixes a threshold for the relative distortion of the codebook and/or imposes a maximum number of references.
Identifying a vowel system while ignoring the number of vocalic references is similar to build a VQ codebook of unknown size and to estimate this size. We say above that the standard LBG splitting algorithm requires a threshold adjustment. It appears that optimizing the codebook size for each set of data may result in a more adequate model for each vocalic system. Using an information criterion calculated from the data set and the LBG modeling is the proposed method.

### 3.2. LBG-Rissanen Vector Quantization

- Algorithm description

In the LBG-Rissanen algorithm, the standard splitting LBG method is applied to the vowel segment set, and before splitting, we compute the Rissanen criterion I(n) , function of the size n of the current codebook. Minimizing I(n) results in the optimal number of references.

$$I(n) = -Ldg(n) + 2n.p.\log N$$

where : - Ldg is the log likelihood of the vowel set, when classing a codebook as a multigaussian distribution,
- p is the parameter space dimension,
- n is the number of codewords,
- N is the cardinal of the vocalic set.

- Decision Rule

The multi dimensional reference map of vocalic patterns may be considered as a simple codework, and the decision rule is the classical K-mean classification (based on the Euclidean distance in the cepstral space). Otherwise, it may be interpreted as a Gaussian mixture model ; the classification results of the maximization of the likelihood score of each acoustic observation.

- Basic Assessment

In order to estimate the relevance of the LBG Rissanen algorithm, we perform a basic evaluation : the corpus of the experiment is composed of utterances of 10 vowels of french. These data have been recorded by the ICP (Institut de la Communication Parlée) at 16kHz. The data set is mono speaker and it consists of 100 occurrences for each vowel. The set is split in a learning subset (700 occurrences) and a test set (300 occurrences). We extract a MFCC vector for each vowel and the LBG Rissanen is in charge with automatically clustering the data. A classification is performed with the test set, based on euclidean distance and likelihood.

- Results

Figure 2 shows the evolution of the likelihood *Ldg* and the Rissanen criterion *I*, in function of the number of codewords, during the quantization process. Note that the minimum (equal to 10) reached by the criterion corresponds exactly to the number of vowels. We verify that the automatically generated codewords match each expected vowel quality (the correlation rate is 100%). The classification performance with the test subset reaches 100% of correct identification. Even if we experiment the LBG-Rissanen algorithm with a basic task (one speaker, good condition recording, with no coarticulation effect), it confirms results from earlier experiments [André-Obrecht 93]. This algorithm is able to provide information about the optimal number of codewords. The challenge is to use these approaches to find an optimal vocalic space model for each language, for a Language Identification task.

## 4. PARALLEL PHONE RECOGNITION (PPR)

In order to get an identification score from a phone recognizer and to exploit the previous vocalic models, we develop two parallel phone recognition systems. Each one is based on HMMs.

1) PPR System by language :

   During a pre processing stage, we detect automatically four classes of sounds : the vowels, the consonants, the short silences (less than 150ms as bursts) and the long silences (pauses) (Figure 3). A specific HMM is built for each class and each language ; the pdfs are given by the LGB-Rissanen algorithm. A global HMM is obtained for each language (figure 1). During the test, no pre classification is performed, the classical Viterbi algorithm is used to find the best path in each language and to give an identification score : the identified language corresponds to the best score.

2) Parallel Vocalic Recognition (PVR) System by language :

   We use exactly the same models as previously, but the identification score is computed differently : the Viterbi algorithm is used to find the best path for each language model, but the identification score takes only the likelihood of the pre classified vocalic into account. With this kind of language identification, we try to quantify the vowel discriminant power.

## 5. LANGUAGE IDENTIFICATION TASK

The vowel system and the PPR and PVR systems are tested with a subset of the OGI_TS [Lander 95].

- **Corpus description**

We use the Oregon Graduate Institute Multilanguage Telephone Speech Corpus (OGI-TS). We limit presently our experiments to five languages (French, Japanese, Korean, Spanish and Vietnamese).
According to the National Institute of Standards and Technology (NIST), the utterances are divided into three corpora, namely the learning set, the development set and the test set. We use only the learning and the development sets in our experiments. There is no overlap between these subsets (sentences and speakers).
Each set is composed of several utterances (the day of the week, the digits, some unconstrained spontaneous speech...) pronounced by each speaker. There are about 20 speakers per language in the development subset and 50 speakers by language in the training one

- **Acoustic Pre processing**

During the detection and pre classification step, we use the a priori « Forward Backward Divergence » segmentation algorithm [André-Obrecht 88] ; in particular, this algorithm provides stationary segments of variable size. Consequently, when a segment occurs, an analysis window of 20 ms is centered and 8 MFCC vector is computed. Furthermore, to explore an eventual interest of the segment duration in the identification task, we add the duration D of the underlying segment.

- **Vowel System Modeling Experiments**

The data set consists of the vowels automatically detected by the algorithm described in [Pellegrino 97] (figure 4). The task we propose is to identify the language spoken in the test utterances.
We use the vowel collection of the learning set to perform several quantization experiments with the LBG-Rissanen algorithms. It provides two codebooks for each language, function of the acoustic parameters. The codebook sizes are not constrained and the algorithm is in charge with computing the optimal size for each language (see Table 1).

| | FR | JA | KO | SP | VI |
|---|---|---|---|---|---|
| 8MFCC | 35 | 34 | 36 | 43 | 48 |
| 8MFCC+D | 34 | 34 | 35 | 36 | 41 |

Table 1: LBG-Rissanen codebook size

Once the VQ codebooks are generated, a classification is made for each speaker from the development subset of the OGI corpus. After computing a matching score for each language, a decision strategy is applied to identify the spoken language. As said previously, we experiment two matching scores, one is based on the likelihood of each language and the other ones on the Euclidean distance. The difference between the two decision rules are not significant and we give quantitative results when using the Euclidean distance. Table 2 displays the results obtained from the 5 languages and Table 3 from a limited set of 4 languages (French, Korean, Spanish and Vietnamese). The decision is taken on the set of vowels, so we distinguish two cases : the identification is performed on 2 minutes of speech, or only on 45 seconds of spontaneous speech.

| | 2 minutes | 45 seconds |
|---|---|---|
| 8MFFC | 52.5 % | 35% |
| 8MFCC +D | 61.2% | 52.5% |

Table 2: Identification scores for a 5 language identification task

| | 2 minutes | 45 seconds |
|---|---|---|
| 8MFFC | 63 .1 % | 43.1% |
| 8MFCC +D | 73.8% | 61.5% |

Table 3: Identification rates for a 4 language identification task.

The duration improves significantly the performances of our system and the correct identification reaches 61 % for the 5-language task while the identification rate for the 4-language task gets to 74 %. The performance decreases obviously with the length of speech, ie the number of detected vowels : for example, during 45 seconds, the detected vowel number is about 180.

- **PPR and PVR Experiments**

The experimental protocol is the same as previously. The test set is limited to the utterances of 45 seconds and the observation vector is composed of the 8MFCC.

| | PPhoneR | PVocalicR |
|---|---|---|
| 5 language | 37% | 38% |
| 4 language | 43% | 47% |

Table 4 : Identification rates with the Parallel Phone and Vocalic Recognitions.

We observe that the PVR approach gives better rates than the PPR one (table 4). This first result is very promising because the protocol experiment is very simple and may be largely optimized. More it confirms the vocalic discriminant power, and it is comparable to the vocalic model identification. Its advantage is that a phone symbolic translation is given and may be processed by a language model.

## 5. CONCLUSION

This work proves that it is possible to extract vowel information from the acoustic signal and to model the vocalic space in a Language Identification purpose. With only this modeling, a good identification rate is reached. Our present purpose is to optimize our Parallel Vocalic Recognizer by using phonological studies. We will add an adequate language model to take the phonotactic rules into account and a prosodic model to build a performant Language Identification system.

## 6. REFERENCES


[André-Obrecht 88] R. André-Obrecht, “A New Statistical Approach for Automatic Speech Segmentation”, IEEE Trans. on ASSP, Jan. 1988, vol. 36 no 1 pp. 29-40

[André-Obrecht 93] R. André-Obrecht, “Segmentation et Parole ?”, Habilitation à diriger des recherches, Université de Rennes, IRISA, June 1993

[Bergem 95] D. R. Van Bergem, “Experimental Evidence for a Comprehensive Theory of Vowel Reduction”, Eurospeech `95 Madrid, pp. 1319-1322

[Carré 95] R. Carré, M. Mrayati, “Vowel transitions, vowel systems, and the Distinctive Region Model” , in Levels in Speech Communication, Elsevier Science B. V., 1995

[Henton 95] C. Henton, “Cross-Language Variation in the Vowels of Female and Male Speakers”, ICPhS `95, Stockholm, pp. 420-423

[Lander 95] T. L. Lander, R. A. Cole, B. Oshika, M. Noel, “The OGI 22 Language Telephone Speech Corpus”, Eurospeech `95, Madrid, pp. 817-820

[Linde 80] Y. Linde, A. Buzo, R. M. Gray, “An Algorithm for Vector Quantizer Design”, IEEE Trans. on COM. Jan. 1980, vol. 28 pp. 84-95

[Maddieson 84] I. Maddieson, ``Patterns of Sounds", Cambridge University Press, 1984

[Muthusamy 94] Y. K. Muthusamy, E. Barnard, R. A. Cole ``Reviewing Automatic Language Identification" IEEE Signal Processing Magazine 10/94, pp. 33-41

[Pellegrino 97] F. Pellegrino, R. André-Obrecht, “From Vocalic Detection to Automatic Emergence of Vowel System”, ICASSP '97, Munich, to appear

[Rissanen 83] J. Rissanen, “A Universal Prior for Integers and Estimation by Minimum Description Length”, The Annals of Statistics, 1983, Vol. 11, No 2, pp. 416- 431

[Schwartz 89] J. L. Schwartz, L. J. Boë, P. Perrier, B. Guérin, P. Escudier, “Perceptual Contrast and Stability in Vowel Systems: A 3-D Simulation Study”, Eurospeech '89, Paris, pp. 63-66

[Sirigos 96] J. Sirigos, V. Darsinos, N. Fakotakis, G. Kokkinakis, “Vowel-Non Vowel Classification of Speech using an MLP and Rules”, Eusipco '96, Trieste, pp. 1059-1062

[Vallée 95] N. Vallée, L. J. Boë, Y. Payan, ``Vowel Prototypes for UPSID's 33 Phonemes", ICPhS `95,Stockholm

[Yan 95] Y. Yan, E. Barnard, “An Approach to Language Identification with Enhanced Language Model” Eurospeech `95, Madrid, pp. 1351-1354

[Zissman 96] M.A. Zissman, “Comparison of Four Approaches to Automatic Language Identification of Telephone Speech”, IEEE Trans. on SAP, Jan. 1996, Vol. 4, No 1, pp. 31-44


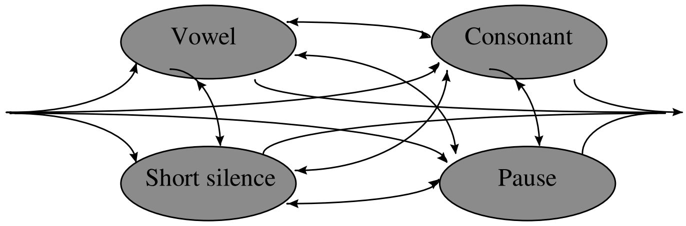


Figure 1: HMM topology in the PPR system

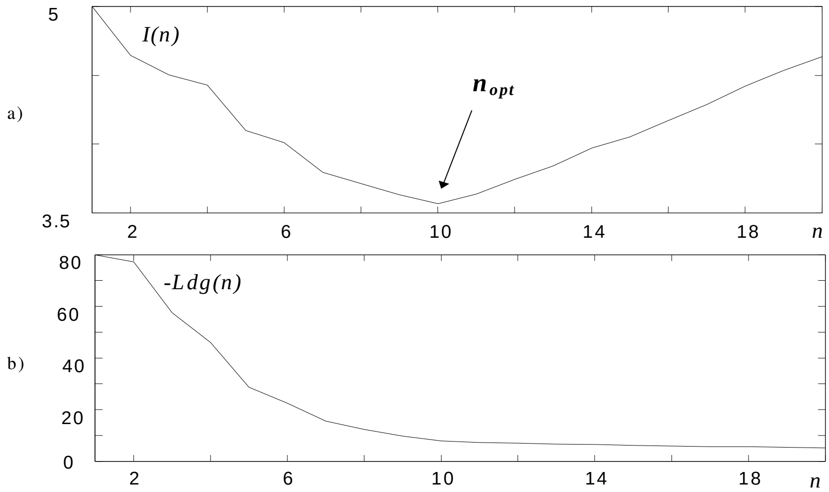


Figure 2: Variations of the vocalic set likelihood Ldg/N and Rissanen criterion I(n)/N

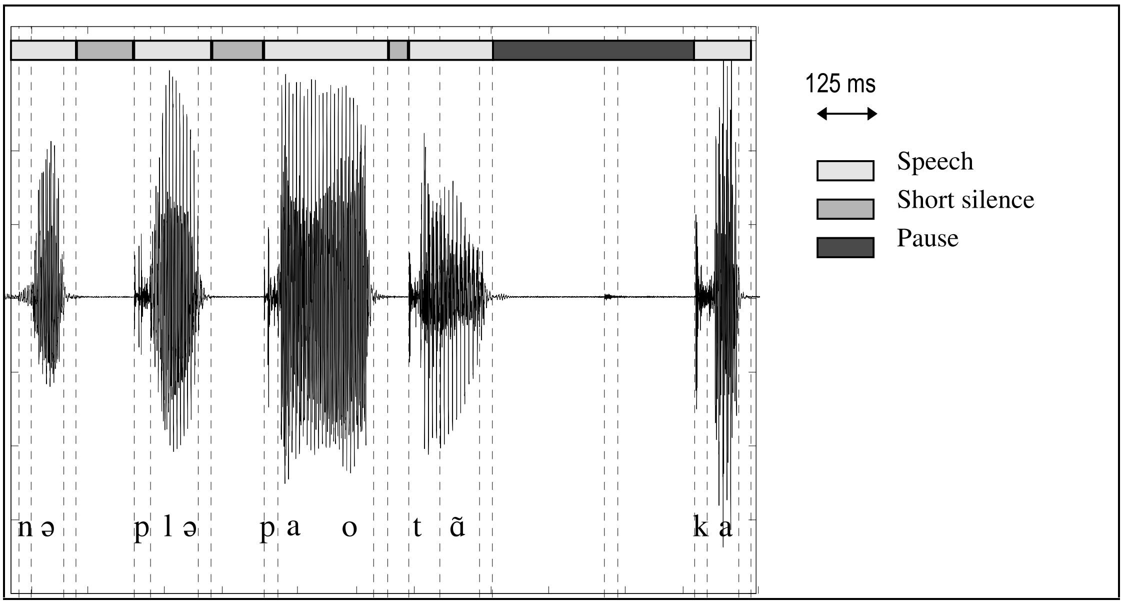


Figure 3 - Segmentation and speech activity on the sentence "...ne pleut pas autant qu'à..."

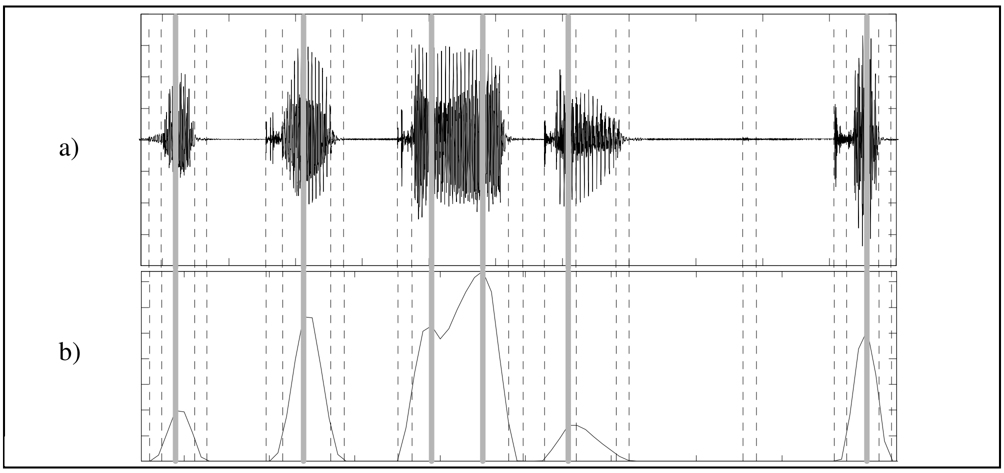


Figure 4 - Segmentation (dashed lines) and vowel detection (grey lines):

a) Acoustic signal

b) Spectral criterion