AUTOMATIC LANGUAGE IDENTIFICATION:

AN ALTERNATIVE APPROACH TO PHONETIC MODELLING

PAPER SUBMITTED TO SIGNAL PROCESSING

NOVEMBER, 3rd 1997

Paper Title:

Automatic Language Identification: An Alternative Approach to Phonetic Modelling

Authors and affiliations:

François PELLEGRINO	Institut de Recherche en Informatique de Toulouse
	Toulouse, France
Régine ANDRE OBRECHT	CNRS - IRIT
	Toulouse, France

Please send correspondence to:

François PELLEGRINO

IRIT

118, route de Narbonne

F-31062 Toulouse Cedex, France

Phone: +33 5.61.55.60.55

Fax: +33 5.61.55.62.58

e-mail: pellegri@irit.fr

List of unusual symbols:

ð : English consonant (in "The" for example)

Number of pages: 27

Number of tables: 3

Number of figures: 6

Keywords: Language identification; Vowel detection; Vowel system modelling.

AUTOMATIC LANGUAGE IDENTIFICATION:

AN ALTERNATIVE APPROACH TO PHONETIC MODELLING

François PELLEGRINO Régine ANDRE-OBRECHT

Abstract

This paper deals with our research on vowel system modelling in an Automatic Language Identification (ALI) purpose. The study of vowel systems shows that they carry an important part of the language characteristics and taking advantage of this knowledge is promising. We propose an alternative modelling to the standard acoustic-phonetic decoding currently used as front-end in the ALI systems: we model each language vocalic system as a Gaussian mixture that is estimated out from automatically detected vowels.

We use the OGI MLTS (Multi-Lingual Telephone Speech) to assess this approach, and in a 5 language close set identification task, we reach 57.3 % of correct identification with 45 second duration utterances. Taking into account that we use only the vowel information (less than 2.5 seconds per utterance) and no language modelling, these results are very promising and offer many perspectives.

Résumé

Cet article est consacré à la modélisation des systèmes vocaliques et à leur utilisation dans le cadre de l'identification automatique des langues. L'étude des systèmes vocaliques montre qu'ils sont porteurs d'une part importante de l'information caractéristique de chaque langue, et l'exploitation de telles connaissances dans un système automatique est évidemment une perspective intéressante. Dans cette optique, nous proposons de modifier le décodage acoustico-phonétique classiquement utilisé dans les systèmes actuels : un modèle de système vocalique est estimé pour chaque langue étudiée à partir des voyelles détectées automatiquement dans le signal.

Les expériences sont menées à partir des enregistrements du corpus OGI MLTS (Multi Lingual Telephone Speech). Dans une tâche d'identification de 5 langues (sans rejet), nous obtenons 57,3 % d'identification correcte avec des phrases d'une durée de 45 secondes. Ces résultats sont obtenus en utilisant uniquement les voyelles extraites du signal, soit en moyenne moins de 2,5 secondes de parole par phrase ; de plus nous n'utilisons aucun modèle de langage. L'intégration de cette modélisation dans un système d'identification des langues plus complexe est une des nombreuse perspectives.

1 Introduction

Automatic Language Identification (ALI) is one of the main challenge for the XXI[st] century in automatic speech processing. Applications of modern communication technologies are already a reality, and the growing demand for integrated services will confirm this tendency. Today, many efforts have been focused on speech technology to provide reliable and efficient Human-Computer Interfaces (HCIs), especially Interactive Voice Systems and Text to Speech Synthesis Systems. With the development of the world communication and of our multi-ethnic societies (European Economic Community...), customers want to pass through the language obstacle and the demand for multilingual capacities becomes a fact. ALI is the original process to respond to this requirement. The language obstacle will remain until ALI systems reach excellent performances and reliability in order not to be the bottleneck of the entire HCI system. Another situation that requires an efficient ALI system is described in [Muthusamy 94]. When someone calls the 911 emergency telephone number in the United States, a human operator is in charge with identifying the language spoken by the caller and with routing it to the right interpreter. Even if human is one of the most efficient language identifier, the tension and the responsibility can alter the operator capacities. For example, a call from Y. K. Muthusamy results in a 3 minute delay to identify that he was speaking in Tamil. If an ALI system is able to identify the right language or at least to assist the operator in his decision – providing a short list of potential languages for example – a great step in safety will be performed.

The first way to reach multi-linguality in HCI is to design intrinsically multi-lingual systems, i.e. systems able to handle with several languages [Lamel 96]. Such systems are quite difficult to develop, and the extended capacities (more than one single language) are often achieved to the disadvantage of the efficiency in each language.

The other way to get round multi-linguality is to design a specific system that identifies the spoken language and to use it as a front-end for several language-specific HCIs. This way, the efficiency of HCIs is optimal and the use of a specific language identification system allows the designers to independently optimise each stage of the overall system.

This paper deals with the design of ALI systems, and more specifically with their phonetic decoding stage. We investigate the opportunity to integrate some phonological knowledge – like Vowel System information – to the classical phonetic approach[1]. The 2[nd] section provides a framework of the Language Identification, describing what kind of information can be exploited and obviously what features are actually taken into account in *automatic* identification systems. We will also discuss in this section the advantages and lacks of such systems, and our vocalic approach is introduced.

The theoretic framework as well as the speech processing algorithms we use are described in Section 3. The implemented platform is detailed in Section 4 while Section 5 deals with the experiments we realise in a real language identification task using the OGI multi-lingual telephone speech corpus that is briefly described.

2 A Short Review of Automatic Language Identification

From the "Cours de linguistique générale" written by Saussure in 1916 to the most recent researches dealing with language structure [Carré 95, Vallée 94], many linguists have studied similarities and differences among languages. Due to the intrinsic structure of spoken language (speech production and perception of course, but also cognitive structure of the language) many levels can be taken into account to identify and characterise a language. In a recent book [Ruhlen 97] the author presents his theory about a common origin to languages. This publication is the core of a discussion in the linguist community, and it shows clearly that it remains interesting investigations to pursue.

Although linguists study language characterisation for almost a century, Automatic Language Identification is a recent research theme since the first researches arose in the seventies at the demand of the US Air Force [Leonard 78]. During a score years designed systems remained confidential and a wide range of approaches have been investigated. The beginning of the nineties marks the dawn of a new area for ALI: Computational capacities strongly increase and offer new possibilities; Automatic Speech Recognition (ASR) becomes reliant and the demand for real-life systems grows consequently. It results in a renewed interest for connected topics, like automatic speaker identification and automatic language identification. The reader can refer to [Muthusamy 94] for a more exhaustive summary of the past and present researches in Automatic Language Identification.

2.1 Language discriminating features

A wide range of distinctive features are available to characterise a language. They are present in several sources that can be clustered in two categories depending on whether they are low-level or not:

2.1.1 Low level characteristics

We gather in this class the features that can be directly extracted from the acoustic signal. We distinguish the following levels:

- Phonetic Level

  Even if the human speech production is a phenomenon shared by the whole mankind, the diversity of the resulting sounds is quite substantial. The UPSID database, that consists in a 451 language inventory chosen to be representative of the about 5000 languages spoken in the world [Maddieson 86], enumerates 920 different sounds (including 177 vowels, 153 occlusives ...). For example, the phone 'ð' is present in the English phonetic system and not in the French one. Inventorying the sounds present in an utterance is a way to discriminate among candidates languages.

- Phonotactic Level

  When the phones of a given language have been identified, the way they combine each others is greatly discriminating. A given sequence of phones can be statistically significant in one language and totally forbidden in another one. For example, cohorts of 3 consonants are quite unusual in French and their occurrences are strictly ruled while it may be common in other languages. The statistical analysis of these phone sequences is undoubtedly relevant and a lot of systems take advantage from it [Kwan 95, Jardino 96].

- Prosodic Level

  The study of the fundamental frequency (pitch) for different languages shows that each language develops its own patterns, in term of phone duration and intonation. In [Hutchins 94], a classification according to unit of tone pattern is given, from syllable (tonal languages – e.g. Mandarin Chinese) to word (word stressed languages – e.g. English) and phrase (focus accent languages – e.g. Spanish). Numerous studies focus on these differences and their discriminating power [Ivonen 95, Kruckenberg 95, Lehiste 95].

2.1.2 Linguistic discrimination

If a speaker starts telling the sequence *'yes, of course'*, the spoken language is likely English whereas if he tells *'oui, bien sûr'*, it is reasonable to guess that it is French. It means that each language uses its own lexicon; these linguistic differences are linked with the cultural specificity of each people, and language classification can be made using the underlying **morphologic** families (e.g. the Latin languages...). Each language has also its own syntax: sentence patterns are different. It is however interesting to keep in mind that this classification is far from phonological classification.

2.2 Language discriminating strategies

The ultimate goal of an ALI system is to provide a decision (what language is spoken) from the speech utterance. This process roughly requires that the acoustic signal be decoded into a phonetic symbol sequence, and that this string be identified as fitting a language model. If we go further in the description, we will consider that a language identification system consists in 4 parts:

- The first stage, that can be called "Acoustic Modelling", provides features (Cepstral coefficients, F_0, ...) that characterise the pronounced speech signal. This part is language independent.
- The second stage – the "Acoustic-Phonetic Decoding" stage – is in charge with computing discrete phonetic symbol sequences from the features provided by the previous stage. If this stage is language-independent, it results in one single phonetic sequence, else it results in several sequences, up to one decoded sequence per language. The Acoustic-Phonetic Decoders have to be trained with a corpus before the identification procedure can occur.
- The third stage, known as the "Language Modelling" phase, provides likelihood scores of the phonetic symbol sequences for each language to be identify, and it is obviously a language-specific task. Usually speaking, these statistical models are n-grams. Each Language Model is trained with a specific set of phonetic symbols, and if the previous stage provides M phonetic sequences for N languages to identify, the system can provide up to $N \times M$ likelihoods, corresponding to each language score decoded in each phonetic system. This topology is not the only

one, and various connections between the Acoustic-Phonetic Decoders and the Language Models are used [Hazen 94, Hieronymous 97, Zissman 96].

- The fourth stage merges the likelihoods given by the Language Models and eventual additional modules (prosody…) and takes the final decision. Common techniques are based on Bayesian decision or neural network modelling.

As in other speech processing topics, the challenge is to integrate the knowledge gathered by experts in automatic systems. The first features that have been investigated were **phonotactic rules**. While the system developed in Texas Instrument was based on the number of occurrences in each language of automatically selected acoustic patterns, A. S. House and E. P. Neuberg [House 77] proposed what can be considered as a precursor system: they model the different sequences of broad phonetic categories using Hidden Markov Models (HMMs). This architecture is widely used nowadays as we will see later.

Beside that, several low-level features have been studied, from raw spectral characteristics [Cimarusti 82] to pitch and energy parameters [Foil 86]. It cannot be denied that an efficient **acoustic-phonetic decoding** is essential to reach good performances. Existing systems differ from each others in this phonetic decoding strategies. HMMs provide a convenient way to reach a phonetic symbol sequence from the acoustic features and they are most widely used, even if other systems based on different statistical modelling or neural networks [Muthusamy 94] reach also good identification results.

However, the best systems have taken advantage from the phonotactic rules. The most recent experiments confirm [Hazen 94, Hieronymous 97, Jardino 96, Zissman 96] that a language modelling based on n-gram statistics is utterly efficient to capture discriminative structural information from the spoken language. For this reason, it is now well accepted that the language modelling component is the core of ALI systems, and the main research efforts have focused on it. Furthermore, several approaches are applied for the general architecture of the phonetic decoders. The lowest computation cost technique performs only one acoustic-phonetic decoding using a common set of phonetic symbols, resulting in one single symbolic sequence. T. J. Hazen reaches very competitive results with this strategy. On the contrary, the most complex systems achieve one acoustic-phonetic decoding per language to identify, using a language-specific set of phonetic symbols [Lamel 94]. Halfway solutions have been studied by M. A. Zissman: he implements an ALI application using 3 language-specific decoders (English, Japanese and Spanish) to recognise three *other* languages (namely French, Farsi and Tamil). These experiments show first, that increasing the numbers of decoders improves the performances, and secondly, that language specific acoustic-phonetic decoders can be different from the languages to identify.

The most recent achieve very good results: their overall correct identification score reaches about 90 % for 45 second utterances in a 11 language close set identification task. The data are usually taken from the OGI multilingual telephone speech corpus, that is described in section 5. These very good performances are obtained with quite complex systems that include several improvements to the acoustic modelling (gender identification, duration and prosodic models…), and at this moment, other approaches result in lower performances.

Nevertheless, these systems present some disadvantages and non optimality that may be investigated with profit.

Even if the simplest systems require only acoustic data without any additional information, the most common ones can't work without a phonetic transcription (i.e. the sequence of pronounced phonetic symbols) of the training corpus. Generating such transcriptions require an human expert phonetic knowledge for each language, and it is an expensive task in term of time and service.

In addition, HMMs (or related models) require a large amount of data to be efficient as it is demonstrated in the ASR systems. It means that training the Acoustic-Phonetic Decoder for a new language requires an impressive collect of labelled data pronounced by as many speakers as possible. This limitation subsists while the computation time cost - that was until recently a very restrictive element– decreases according to the computer capacities continuous growth.

However, these restrictions are common with most speech processing systems, and for numerous applications, they are not so disastrous. By contrast, and from a phonetic point of view, we present hereafter remarks that lead us to propose an alternative approach to the standard phonetic modelling.

2.3 Motivations for an extended phonetic modelling

It appears that the Acoustic-Phonetic decoding phase is usually not exploited to provide a discriminating information. In his system, T. J. Hazen uses the SUMMIT English decoder only to project the acoustic features to a discrete symbol space. Zissman uses several decoders, but his experiments show that the improvement may merely result from the better overall phonetic handling. Several systems use a broad set of phonetic units gathered among the languages to identify. The main purpose of these Acoustic-Phonetic decoding is to provide the closest phonetic transcription of an utterance. Poor language-specific discriminating information is retrieved and the identification score is entirely given by the language modelling stage. Slightly different topologies optimise globally the likelihood of the phonetic decoding and of the n-gram patterns. In such systems, the phonetic information is taken into account in this embedded scoring, but it seems that an important part of the underlying information is not exploited.

These observations lead us to propose an alternative language identification strategy. The main idea is to take advantage from both phonotactic models and phonologic system models.

Numerous linguist works [Maddieson 86, Lindblom 89] assess that languages can be characterised to a large extent by their phonologic system description. What we call phonologic system is the inventory of the phones pronounced by native speakers for a given language. These systems can be split in a *vowel system*, and a *consonant system*.

Vowel systems have been widely studied by linguists and typologies exist. Such classifications are possible because vowels share an homogenous acoustic structure [Stevens 85, Lindblom 89]: it enables to describe all vowels, and by extend all vowel systems, in a common space, usually a space derived from acoustic analysis.

On the contrary, phonological research on consonant systems are less exhaustive, even if numerous contributions dealing with consonants sub categories are available [Funatsu 95]. The main reason is that consonants are described by default as non-vowel sounds. A wide

diversity of consonants can be produced resulting in very different acoustic patterns (fricatives, plosives, clicks...), and no clear common description exists, except from the articulatory point of view.

In order to integrate phonologic system description in an automatic system, we concentrate on vowel systems, taking advantage from their common space descriptions. This choice is also directed by some observations with the UPSID database [Maddieson 86, Vallée 94]. In this 451 language inventory, 177 vowels, derived from 36 basic vocalic qualities occur at least in one system. The 451 languages share 307 vowel systems, including 271 language-specific systems. Thus, even if phonological vowel system descriptions are not efficient enough to discriminate among all the languages, they provide a relevant information. We guess that this phonological characterisation, exploited with adequate phonotactic models may improve ALI efficiency.

3 The Vowel System Modelling

3.1 Overview of the ALI system

Several influences must be taken into account when designing an ALI system. The exploitation of as many discriminative features as possible is hardly in agreement with reasonable computational requirements and expert knowledge. The system described in this paper makes use of automatically extracted features and knowledge, in order to be generalised to other languages even if no labelled data is available.

The segmental character of speech signal has been assessed by many works. For this reason, our system applies a segment based processing. These segments are obtained applying the "Forward Backward Divergence" algorithm [Andre-Obrecht 88]. A vowel detection algorithm [Pellegrino 97] providing a basic Consonant/Vowel label to each segment follows. A cepstral analysis gives for each vocalic segment a Mel Frequencies Cepstral Coefficient (MFCC) vector.

The acoustic phonetic decoding is performed by parallel language-specific Vowel System Models (VSM) based on HMM approach. The outputs consist of the likelihood of the detected vowels according to each VSM and a Consonant/Vowel sequence.

This sequence can be processed by language-specific N-grams models, in order to catch the phonotactical information. This treatment is not described in this paper, because we focus on the discriminative power of the phonological features. The system that has been actually implemented is displayed on Figure 1 without language modelling. In this system, the decision is directly applied using the likelihoods provided by the VSMs.

Fig. 1

3.2 Statistical framework

The Vowel System Model (VSM) is a simplified Hidden Markov Model where each state corresponds to a specific sound. After the acoustic processing, each segment is labelled consonant or vowel. To focus on the discriminative power of the vocalic system, the state set of the HMM consists of only one state for all consonants, while the other states describe the various vocalic qualities.

Let $L = \{L_1, L_2, \ldots, L_{NL}\}$ be the N_L languages to identify. The challenge is to determinate which language is spoken when an unknown speaker pronounces an utterance and the problem is to find the most likely language L^* in the L set.

Let $\Omega_i = \{C, V_1^i, V_2^i, \ldots, V_n^i\}$ be the state set of the VSM relative to language number i where C is the consonant state.

After the acoustic processing, we obtain for each segment a concatenation of heterogeneous features. Let T be the number of segments in the spoken utterance. $O = \{o_1, o_2, \ldots, o_T\}$ is a sequence of observation vectors. Each vector o_t consists of a spectral feature vector a_t, the duration of the segment d_t, and a macro-class flag c_t, equal to 1 if the segment is recognised as a vowel, and equal to 0 otherwise. In order to simplify the formula, we note $y_t = \{a_t, d_t\}$ and $o_t = \{y_t, c_t\}$.

Given the observations O, the most likely language L^* is defined by the following equation:

$$L^* = \arg\max_{1 \le i \le NL} \left[\max_{\Phi \in \Omega_i^T} \Pr(\Phi_i, L_i | O) \right] \qquad (1)$$

where $\Phi_i = \{\varphi_1^i, \varphi_2^i, \ldots, \varphi_T^i\}$ is a state sequence of the i^{th} language VSM.

Using Bayes' theorem, this expression changes to:

$$L^* = \arg\max_{1 \le i \le NL} \left[\frac{\max_{\Phi \in \Omega_i^T} \left(\Pr(O|\Phi_i, L_i) \cdot \Pr(\Phi_i, L_i) \right)}{\Pr(O)} \right] (2)$$

$$L^* = \arg\max_{1 \le i \le NL} \left[\max_{\Phi \in \Omega_i^T} \left(\Pr(O|\Phi_i, L_i) \cdot \Pr(\Phi_i | L_i) \right) \Pr(L_i) \right] (3)$$

In (3), we can consider that $\Pr(O|\Phi_i, L_i)$ is the phonetic likelihood term and that $\Pr(\Phi_i|L_i)$ is the phonotactic likelihood expression.

Under the standard HMM assumptions, we assume that each segment is conditionally independent of other segments. This assumption – that is common in speech processing – is improper, but it greatly simplifies the statistical expression, and in the case of variable duration homogeneous segments, we guess that this statistical independence hypothesis is less incorrect than in a constant duration frames modelling. The phonetic modelling expression is hence changed to:

$$\Pr(O|\Phi_i, L_i) = \prod_{t=1}^{T} \Pr(o_t | \varphi_t^i, L_i) \qquad (4)$$

For each segment k, the *a priori* vowel detection provides a flag c_k equal to zero for consonants and equal to one for vowels. It is then possible to split (4) and to get a consonant and a vowel part in the phonetic modelling expression:

$$\Pr(O|\Phi_i, L_i) = \prod_{c_t=0} \Pr(o_t | \varphi_t^i, L_i) \prod_{c_t=1} \Pr(o_t | \varphi_t^i, L_i) \qquad (5)$$

In our approach, no language-specific consonant model is computed. It means that for c_k equal to zero, the likelihood $\Pr(y_t|\varphi_i^c, L_n)$ is language independent. (5) simplifies to:

$$\Pr(O|\Phi_n, L_n) = \prod_{c_k=0} \Pr(y_t|\varphi_i^c = C) \prod_{c_k=1} \Pr(y_t|\varphi_i^c, L_n) \qquad (6)$$

Given, (3) and (6), the overall language likelihood yields:

$$L = \arg\max_{L_n \in \Omega_L} \left[\max_{\Phi_n \in \Omega_\Phi^c} \Pr(\Phi_n, L_n|O) \right] = \arg\max_{L_n \in \Omega_L} \left[\max_{\Phi_n \in \Omega_\Phi^c} \left(\prod_{c_k=1} \Pr(y_t|\varphi_i^c, L_n) . \Pr(\Phi_n|L_n) \right) \Pr(L_n) \right] \qquad (7)$$

In our experiments, the *a priori* probability of occurrence of each language is assumed to be equal; it leads finally to:

$$L = \arg\max_{L_n \in \Omega_L} \left[\max_{\Phi_n \in \Omega_\Phi^c} \left(\prod_{c_k=1} \Pr(y_t|\varphi_i^c, L_n) \Pr(\Phi_n|L_n) \right) \right] \qquad (8)$$

According to the remark section 3.1, we implement a version of the algorithm to study the discriminative power of vowel system modelling without other features like phonotactic rules. Under this assumption, the HMM becomes ergodic $\Pr(\Phi|\text{Li})$ is independent of the language L_i. The more likely language is given by:

$$L = \arg\max_{L_n \in \Omega_L} \left[\max_{\Phi_n \in \Omega_\Phi^c} \left(\prod_{c_k=1} \Pr(y_t|\varphi_i^c, L_n) \right) \right] \qquad (9)$$

4 Implementation

In our experimental platform (Figure 2) the Acoustic modelling is language independent, while the Phonetic modelling consists in language specific Vowel System Models.

Fig. 2

4.1 Acoustic Modelling

The purpose of this phase is to provide a sequence of vocalic and consonantal segments. According to our model (equation 9), consonants are ignored, and parameters are computed only for vowel segments. Three pre-processing are performed before the parameter estimation. The "Forward-Backward Divergence" algorithm provides a relevant segmentation more adapted than the classical constant duration frame technique. A Speech Activity Detector flags each segment not to process long silences, and a vowel detector is in charge with locating vocalic segments in the utterance. Finally, a cepstral analysis is performed to generate the inputs of the VSMs.

4.1.1 A priori Segmentation

The segmentation results from a statistical study of the acoustic signal. A divergence criterion is computed at each moment n between two AR models θ_n and θ_I estimated on two different windows. θ_n is estimated on a growing window $[0,n]$ while θ_I is estimated on a short L duration sliding one $[n-L,n]$. The divergence criterion is computed from the cross entropy between the distributions of θ_n and θ_I.

Assuming that these distributions are gaussian, it can be shown that the cumulative divergence W_n has a zero conditional drift under the hypothesis $\theta = \theta_n$, and a conditional drift equal to the opposite of the conditional Kullback's divergence under the hypothesis $\theta = \theta_I$ (Figure 3a). The algorithm is improved applying the Page-Hinkley rule in order to detect inversion of trend rather than a simple negative trend (Figure 3b).

Fig. 3

The actually implemented algorithm is modified by the use of a backward detection that corrects the possible misses of the forward computation. Interested readers are referred to [André-Obrecht 88] for a detailed study of the Forward Backward Divergence Algorithm.

4.1.2 Speech Activity Detection

Our LID system has been tested with data from the OGI MLTS corpus. These data are recorded in conditions close from real life, and the utterances often present long silences with no voice activity. We decide to locate these areas that carry no phonetic or phonotactic significance.

The algorithm performs a basic acoustic statistical analysis and a threshold T_a is computed from the standard deviation of each segment: Let z be the acoustic signal and $S = \{s_1, s_2, \ldots, s_T\}$ the segments computed by the Forward-Backward Divergence algorithm, T_a is given by:

$$T_a = \alpha . \min_{s}(\sigma_s(z)) \qquad (10)$$

where $\sigma_s(z)$ is the standard deviation of the signal z in the i^{th} segment and α is a scale factor. In our experiments, α equals 2.5.

Segments presenting a standard deviation greater than T_a are flagged as speech, otherwise, they are considered as silence. In order to distinguish no activity segments (i.e. pauses) from short silences embedded in speech production (typically silences preceding bursts in plosive sounds or short hesitations), the duration of the segment is used: silence segments shorter than 150 ms are considered as significant silences while longer segments are flagged as no-activity periods. If neighbouring segments are flagged as silence, the total duration is taken into account.

The subsequent processing ignores the no-activity segments.

4.1.3 Vowel Detection

The vowel detection algorithm is based on spectral analysis. Due to the production structure of vowels, the vocal (and nasal in some cases) tract behaves as a resonator; it results in a formant-antiformant structure for the spectrum of a vowel. Additionally, antiformants are usually less remarkable than formants.

We compute a criterion called *Rec* (Reduced energy cumulative) for each frame of the signal. This criterion is derived from 24 energy coefficients achieved through a Mel-scale filter bank spectral analysis. The algorithm is frame based in order to obtain a more accurate vowel locating than with a segmental approach.

Let t be the number of the current frame and $E_i(t)$ be the energy of the t^{th} frame in the i^{th} Mel filter. Let $\overline{E}(t)$ be the related mean of filter energies and $E(t)$ the total energy of the t^{th}

frame. In order to take only voiced sounds into account, let $E_{LF}(t)$ be the energy in the low frequencies (100 Hz – 1000 Hz) Mel filters. The criterion is given by:

$$Rec(t) = \frac{E_{LF}(t)}{E(t)} \sum_{i=1}^{n} \delta_i \left(E_i(t) - E(t) \right)^2 \quad (11)$$

where δ_i is equal to one for filters from 300 Hz to 3200 Hz, and it is equal to zero for all other frequencies.

Fig. 4

The *Rec* criterion is a kind of similarity measure between the spectral structure of the frame and a theoretical formantic structure. Vowels are characterised by the highest values and we consider that the peaks of *Rec(t)* are located in vocalic sounds (Figure 4). Additionally, we do not take into consideration detections occuring in segment shorter than 15 ms.

The main advantage of this algorithm is that it requires no labelled data and no supervised learning. Thus, it is language independent and fully operating whatever language is studied.

4.1.4 Cepstral Analysis

Each detected vowel is represented with a set of 8 Mel-Frequency Cepstral Coefficients (MFCCs). The cepstral analysis is performed using a 256-point Hamming window centred on the *Rec* peak detection, via a Fast Fourier Transform. This parameter vector is extended with the duration of the underlying segment.

A cepstral subtraction is performed to operate a blind deconvolution to remove the channel effect. For each recording session, the average MFCCs vector is computed over all vowels; it is then subtracted from each vowel coefficients. This method is slightly different from the classical cepstral subtraction since the average values are not computed over the whole utterance or the silence frames, but only over vocalic frames. Recent studies tend to show that the resulting channel estimation is not deteriorated and even that it may be more efficient [Puel 97].

Figure 5 summaries the acoustic pre-processing with an example from the OGI database. The results of voice activity and vowel detection are represented and the automatic segmentation is also displayed. We can see that the vowel detection may improve the segmentation and add missed transitions, for example between the two co-articulated vowels 'a' and 'o'.

Fig. 5

4.2 Phonetic Decoding

Vowel System Models (VSMs) consist in a simplified HMM with an ergodic topology and one gaussian pdf (μ, Σ) for each vocalic state.

Let $X = \{x_1, x_2, \ldots, x_N\}$ be the training set and $\Pi = \{(\alpha_1, \mu_1, \Sigma_1), (\alpha_2, \mu_2, \Sigma_2), \ldots, (\alpha_n, \mu_n, \Sigma_n)\}$ be the parameter set that defines a mixture of n p-dimensional Gaussians. The model that maximises the overall likelihood of the data is given by:

$$\Pi^* = \arg\max_{\Pi} \prod_{i=1}^{N} \left| \sum_{k=1}^{n} \frac{\alpha_k}{(2\pi)^{p/2} \sqrt{|\Sigma_k|}} \exp\left[-\frac{1}{2} (x_i - \mu_k)^t \Sigma_k^{-1} (x_i - \mu_k) \right] \right| \quad (12)$$

where α_k is the mixing weight of the k^{th} Gaussian.

Steven Nowlan studies relations between Gaussian mixture models and VQ in detail in his PhD Thesis [Nowlan 91]. He shows that under the assumptions:

- The summation can be approximated by the maximum term in the summation ("Winner-take-all" assumption),
- Gaussians weighting coefficients are equal,
- Gaussians are spherically symmetric (i.e. $\Sigma_i = \sigma^2 I_p$ whatever k),

the estimation of the maximum likelihood parameters reduces to the least squares equation. In the case of the Euclidean distance, this estimation is given by the expression:

$$\Pi^* = \arg\min_{C} \frac{1}{N} \sqrt{\sum_{i=1}^{N} \left[\min_{1 \le k \le n} \|x_i - \mu_k\|^2 \right]} = \arg\min_{C} \sum_{i=1}^{N} \left[\min_{1 \le k \le n} \|x_i - \mu_k\|^2 \right] \quad (13)$$

where $\Pi^* = \{\mu_1, \mu_2, \ldots, \mu_n\}$ is the corresponding codebook.

This way, a Vector Quantization (VQ) algorithm computes a multi-dimensional reference map of the vocalic patterns. From a theoretical point of view, it is incorrect to say that we model the phonological vowel system of the language: In fact, we model its spoken vocalic system: the vowels are in context and strongly co-articulated, diphthongs may also be detected and modelled, and the reference patterns result from all these factors. In our application, we use the LBG-splitting [Linde 80] algorithm to estimate the codebook that we identify as the VSM.

During the identification phase, all the vowels detected in the utterance are gathered and parameterised by the acoustic modelling. The distortion between this set of vowels $Y = \{y_1, y_2, \ldots, y_N\}$ and each of the N_L Vowel System Models is computed. Given that the codebook Π^* is now specific to language L_n, the identified language L^* is given by:

$$L^* = \arg\min_{1 \le n \le N_L} \left[\sum_{i=1}^{N} \min_{1 \le k \le n} \left(\|y_i - \mu_k^*\|^2 \right) \right] \quad (14)$$

5 Experiments

The ALI system is tested with the OGI MLTS corpus [Lander 95]. We limit presently our experiments to five languages (French, Japanese, Korean, Spanish and Vietnamese). These languages have been chosen because of their phonological vowel systems (see Figure 6). Spanish and Japanese vowel systems are rather elementary (5 vowels) and quasi-identical while Korean and French systems are more complex, with several vowels with the same quality. Vietnamese system is of average complexity.

According to the American National Institute of Standards and Technology (NIST), the data is divided into three corpora, namely the learning set, the development set and the test set. Each corpus consists in several utterances (the days of the week, the digits and longer unconstrained speech) pronounced by each speaker. There is no overlap between the speakers of each corpus.

Fig. 6

There are about 20 speakers per language in the development subset and 50 speakers per language in the learning one. In our experiments, we don't take into account female speakers because of the poor number (less than 20 %).

A part of these data is labelled according to broad phonetic categories. This labelling is automatically generated and we use it to test vowel detection. Insertion and omission rates are calculated according to OGI vowel labelling (Table 1).

Table 1

The language identification experiments are performed using two vowel parameter description: the first one is the basic 8 MFCCs description $\{a_i\}$ while the second one is improved by the segment vowel duration $\{a_i, d_i\}$. Each VSM is trained with all the vowels detected for all speakers of the training set of each language (Table 2) and the language identification results are given for the development set. With the LBG-splitting algorithm, we constrain the codebook size to 20 and 70 clusters, in order to test 2 codebooks (respectively called LBG-20 and LBG-70) per language.

Table 2

The language identification decision is taken on the set of vowels detected in the test utterance. Table 3 provides the results according to the duration of the test utterance: in a first experiment, we use only the vowels detected in the 45 second duration unconstrained utterance for each speaker, and in a second test, we gather all vowels for a given speaker (about 2 minute duration) and we use them.

Table 3

Vowel duration improves significantly the performances of the vowel system identification and the correct identification reaches 57.3 % with the 45 second utterances. However, the performance decreases obviously with the length of speech since the more the vowel set is large, the more the estimation is correct. Another effect of the decrease of the vowel number is that the LBG-20 codebook reaches better results than the LBG-70 one (with 45 second duration utterances). It suggests that it needs more data to correctly estimate the likelihood of the vowels from the test utterance in the 70 codeword model than in the 20 codeword model. At last, it is also important to note that the vocalic sound duration represents only 2.15 s in the 45 second utterance (this value is the mean value upon the development set). It means that we get 57.3 % of correct identification using less than 5 % of the data!

6 Conclusion

This work proves that a significant part of the language characterisation is embedded in its vowel system. We show that it is possible to extract this information and to model it for a language identification task. The identification rates that we reach with only the vowel system identification is very good, and it is essential to take into account that more than 95 percent of the data is not exploited.

These results offer many perspectives:

- using a strict Gaussian mixture modelling instead of a constrained one may improve vowel system modelling. The weights represent the occurrences of each vocalic sound in the language and the parameter estimation is performed using the EM algorithm.

- taking advantage from the state sequence through a phonotactic modelling based on a N-gram approach should improve the ALI system.

- another promising approach is to develop a specific model for other sounds categories like fricatives or plosives for example. This way, it will be possible to use together several sound-specific systems that may be a powerful alternative to standard all-sound phonetic models.

7 References

[André-Obrecht 88] R. André-Obrecht, "A New Statistical Approach for Automatic Speech Segmentation", *IEEE Trans. on ASSP*, January 88, vol. 36, n° 1, pp. 29-40

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

[Cimarusti 82] D. Cimarusti and R. B. Ives, "Development of an Automatic Identification System of Spoken Languages: Phase 1", *Proc. ICASSP'82*, Paris

[Foil 86] J. T. Foil, "Language Identification using Noisy Speech", *Proc. ICASSP'86*, Tokyo

[Funatsu 95] S. Funatsu, "Cross Language Study of Perception of Dental Fricatives in Japanese and Russian", *Proc. ICPhS' 95*, Stockholm, pp. 124-127

[Hazen 94] T. J. Hazen and V. W. Zue, "Recent Improvements in an Approach to Segment-based Automatic Language Identification", *Proc. ICSLP'94*, Yokohama, pp. 1883-1886

[Hieronymous 97] J. Hieronymus and S. Kadambe, "Robust spoken language identification using large vocabulary speech recognition", *Proc. ICASSP'97*, Munich, vol. 2, pp. 1111-1114

[House 77] A. S. House and E.P. Neuberg, "Toward Automatic Identification of the Language of an Utterance. I. Preliminary methodological considerations", *J. of Acoustical Society of America* 62(3), 1977, pp. 708-713

[Hutchins 94] S. E. Hutchins and A. Thyme-Gobbel, "Experiments using Prosody for Language Identification", *Proc. Speech Research Symposium XIV, 1994*, Baltimore

[Iivonen 95] A. Iivonen et al., "Comparison of Prosodic Characteristics in English, Finnish and German Radio and TV Newscasts", *Proc. ICPhS' 95*, Stockholm, pp. 382-385

[Jardino 96] M. Jardino, "Multilingual stochastic N-Gram Class Language Models", *Proc. ICASSP'96*, Atlanta

[Kruckenberg 95] A. Kruckenberg and G. Fant, "Notes on Syllable Duration in French and Swedish", *Proc. ICPhS' 95*, Stockholm, pp. 158-161

[Kwan 95] H. Kwan and K. Hirose, "Recognized Phoneme-based N-Gram Modeling in Automatic Language Identification", *Proc. Eurospeech'95*, Madrid, pp. 1367-1370

[Lamel 94] L. F. Lamel and J. L. Gauvain, " Language Identification using Phone-based Acoustic Likelihoods ", ICASSP'94, Adelaide, pp. I.293-I.296

[Lamel 96] L. F. Lamel and al., "Spoken Language Processing in a Multilingual Context", ICSLP'96, Philadelphia, pp. 2203-2206

[Lander 95] T. L. Lander et al., "The OGI 22 language telephone speech corpus", *Proc. Eurospeech '95*, Madrid, pp. 817-820

[Lehiste 95] I. Lehiste, "Cross Linguistic Comparison of Durational Patterns in Finnish and Finland-Swedish", *Proc. ICPhS' 95*, Stockholm, pp. 632-635

[Leonard 78] R. G. Leonard and G. R. Doddington, "Automatic Language Discrimination", *Technical Report RADC-TR-78-5*, Air Force Rome Air Development Center, Jan. 78

[Lindblom 89] B. Lindblom and Q. Engstrand "In What Sense is Speech Quanta?", *Journal of Phonetics* 17, 89, pp 107-121.

[Linde 80] Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer", *IEEE Trans. On COM.*, January 1980, vol. 28 pp. 84-95

[Maddieson 86] I. Maddieson, *Patterns of Sounds*, 2nd Edition, Cambridge University Press, Cambridge 1984

[Mokbel 95] C. Mokbel, D. Jouvet and J. Monné "Blind Equalization using Adaptative Filtring for improving Speech Recognition over Telephone", *Trans. Eurospeech 95*, Madrid, pp 1987-1990

[Muthusamy 94] Y. K. Muthusamy, "Segmental Approach to Automatic Language Identification", *Ph. D Thesis*, Oregon Graduate Institute of Science & Technology, 1993

[Nowlan 91] S. Nowlan, *Soft Competitive Adaptation: Neural Network Learning Algorithm based on fitting Statistical Mixtures*, PhD Thesis, School of Computer Science, Carnegie Mellon Univ. 1991

[Pellegrino 97] F. Pellegrino and R. Andre-Obrecht, "Vocalic system modeling: a VQ approach", *Proc DSP '97*, Santorini

[Pied 97] J.-B. Pied, *Reconnaissance automatique de parole téléphonique : adaptation au GSM*, Thèse de 3ème cycle, Univ. Paul Sabatier, 1997, Toulouse

[Rubden 97] M. Rubden, "L'origine des langues", collection "Débats", Ed. Belin, 1997

[Schwartz 89] J.-L. Schwartz et al., "Perceptual Contrast and Stability in Vowel Systems: A 3-D Simulation Study", *Proc. Eurospeech '89*, Paris, pp. 63-66

[Stevens 85] K. N. Stevens "Spectral Prominences and Phonetic Distinctions in Language", *Speech Communication* 4, 137-144

[Vallée 94] N. Vallée, *Systèmes vocaliques : de la typologie à la prédiction*, Thèse de 3ème cycle, Univ Stendhal, 1994, Grenoble

[Yan 95] Y. Yan and E. Barnard, "An Approach to Language Identification with Enhanced Language Model" *Proc. Eurospeech '95*, Madrid, pp. 1351-1354

[Zissman 96] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech", *Proc. IEEE Trans. On SAP*, January 1996, vol. 4, no. 1, pp. 31-44

Footnotes

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

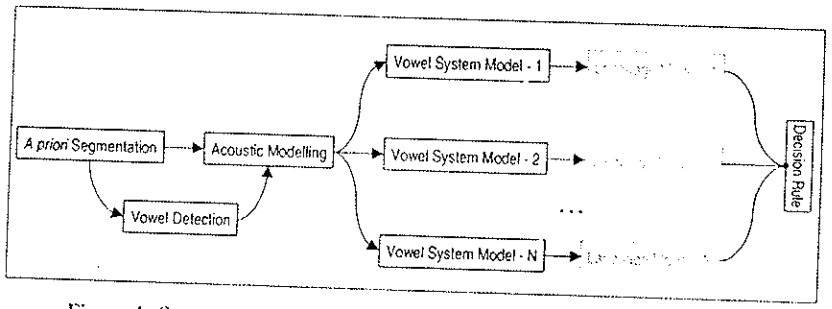

Figure 1: Synopsis of the VSM-based Automatic Language Identification System for N languages.

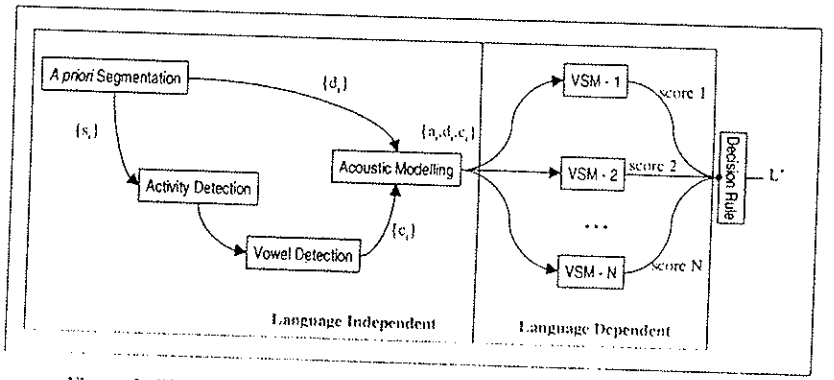

Figure 2: Block diagram of the implemented System

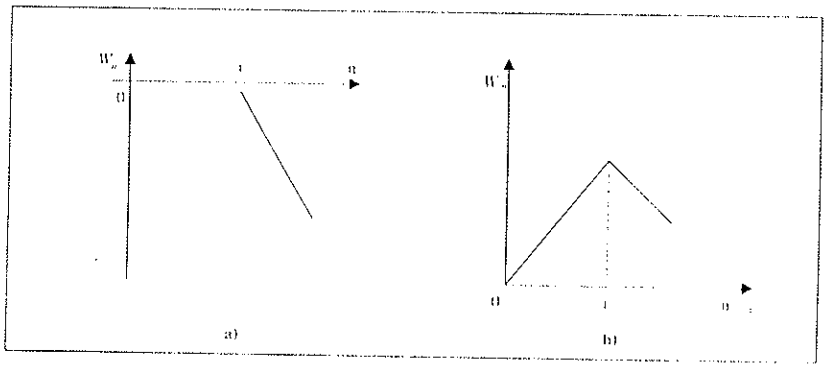

Figure 3: Variations of the cumulative sum test:

a) the statistics W_n;

b) the statistics coupled to Hinkley's stopping rule $\tilde{W}_n$.

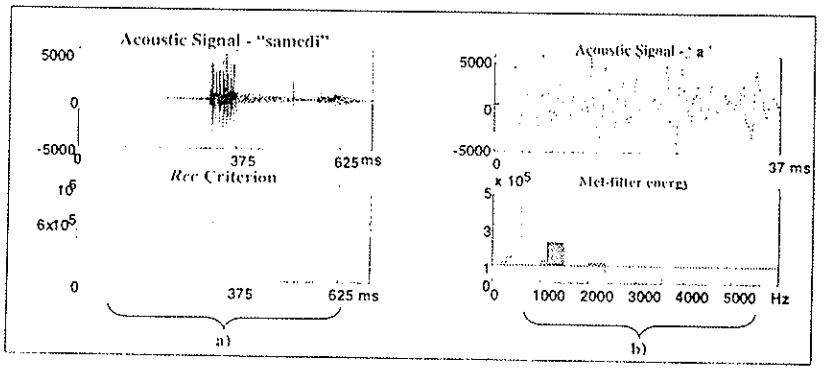

Figure 4: Example of a vowel detection:

a) Acoustic Signal (French word "samedi") and *Rec* criterion;
b) Mel-scale filter analysis for the vowel 'a'

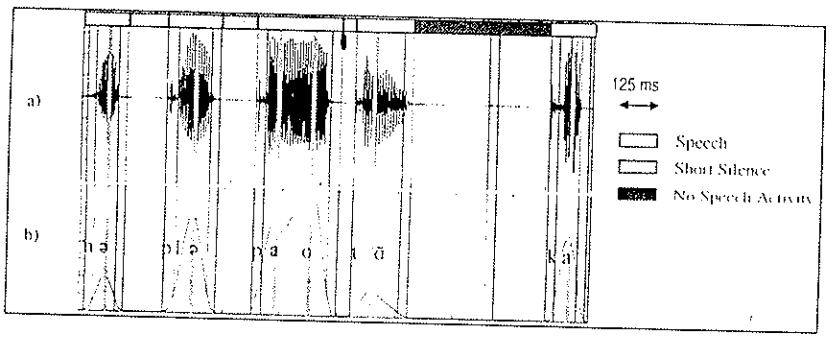

Figure 5: Example of Acoustic pre-processing – the French sentence is "... ne peut pas autant qu'à..."; the vertical black lines are the segment boundaries; the vertical grey lines are the detected vowels.

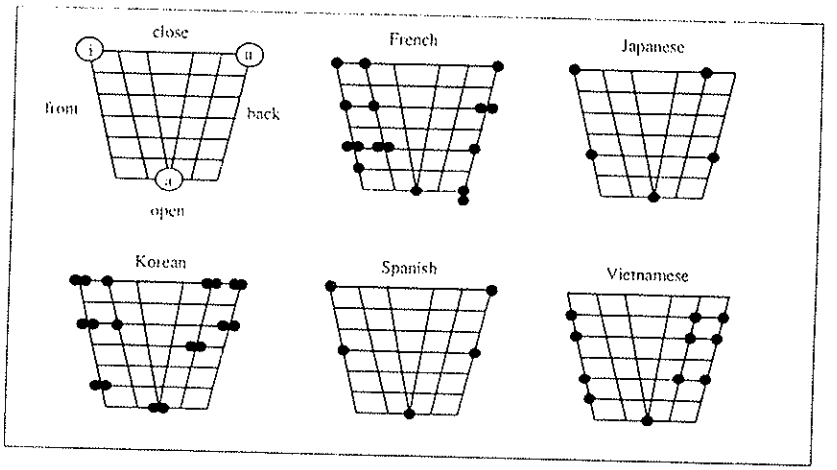

Figure 6: UPSID Phonological vowel systems – The draw in the upper left corner displays how vowels are represented [Vallée 94]

	French	Japanese	Korean	Spanish	Vietnamese
Deletion rate	4.34 %	5.66 %	6.93 %	6.58 %	4.09 %
Insertion rate	10.76 %	8.30 %	15.28 %	9.84 %	16.84 %

Table 1: Vowel detection results with the labelled OGI subset

	French	Japanese	Korean	Spanish	Vietnamese
8 MFCCs	18045	16108	14283	18583	13287

Table 2: Number of detected vowels in the OGI learning set

	LBG-20		LBG-70	
	45 s	2 min	45 s	2 min
8 MFCCs	50.7 %	56.2 %	52.0 %	66.2 %
8 MFCCs + Duration	57.3 %	73.7 %	56.0 %	76.2 %

Table 3: Correct Identification scores for a 5 language ALI task