# AN UNSUPERVISED APPROACH TO LANGUAGE IDENTIFICATION

*F. Pellegrino, R. André-Obrecht*

IRIT
118, route de Narbonne
F-31062, Toulouse Cedex, France
{pellegri, obrecht}@irit.fr

## ABSTRACT

This paper presents an unsupervised approach to Automatic Language Identification (ALI) based on vowel system modeling. Each language vowel system is modeled by a Gaussian Mixture Model (GMM) trained with automatically detected vowels. Since this detection is unsupervised and language independent, no labeled data are required. GMMs are initialized using an efficient data-driven variant of the LBG algorithm: the LBG-Rissanen algorithm.

With 5 language from the OGI MLTS corpus and in a close set identification task, we reach 79 % of correct identification using only the vowel segments detected in 45 second duration utterances for the male speakers.

## 1. INTRODUCTION

Automatic Language Identification (ALI) is one of the main challenges for the next decade in automatic speech processing. Today, many efforts have been focused on speech technology to provide reliable and efficient Human-Computer Interfaces. With the development of the world communication and of our multi-ethnic societies (European Economic Community…), the demand for multilingual capacities becomes a fact. The language obstacle will remain until ALI systems reach excellent performances and reliability in order not to be the bottleneck of the overall system.

The standard ALI approach is based on phonotactic discrimination via specific statistical language modeling [10]. In most systems, phone recognition is merely considered as a front-end and not exploited for the language likelihood generation. This method yields a sub-optimal use of the phonetic and phonological differences among languages though they carry a substantial part of language identity. Exploiting this information classically involves an HMM modeling that requires a consequential amount of labeled data. We propose an alternative approach that necessitates no labeled data, resulting in an efficient unsupervised modeling. This approach is based on differentiated phonetic modeling: it consists in speech utterance segmentation according to phonetic categories (vowels, fricatives…) and in separated model processing convenient with each category. In the present paper, this method is assessed in the framework of Vowel System (VS) modeling.

The choice of VS modeling is justified from a phonological point of view since languages may be partially classified in an efficient way according to their VS [9]: the 451 languages of the UPSID database share 307 vowel systems, including 271 language-specific ones. Thus, even if phonological vowel system descriptions are not efficient enough to discriminate among all the languages, they provide a relevant information that worth being exploited.

The next section settles the framework of the proposed approach and describes a global segmental ALI system that provides a baseline system for comparison. The VS modeling system is detailed in Section 3. Model topologies are settled by heuristic and entropy based algorithm that is also described. Section 4 deals with the experiments we realize with the OGI multi-lingual telephone speech corpus.

## 2. VOWEL SYSTEM MODELING IN ALI

### 2.1 Description of the segmental reference system

The reference system is similar to the GMM system described in [10]. The main difference is that the cepstral analysis leading to the observations is performed on variable length segments rather than on constant duration frames. This system is noted hereunder Global Segmental Modeling (GSM).

The training procedure consists in the following processing:

- The *a priori* "Forward-Backward Divergence" algorithm [1] provides long steady and shorter transient segments.
- A speech activity detector is applied to discard pauses.
- A segmental cepstral analysis is performed on each segment.
- A GMM per language is computed with the set of language dependent observations.

The same acoustic processing is applied during recognition, and the language is identified *via* a maximum likelihood computation of the utterance according to the language dependent models.

### 2.2 Description of the VS modeling system

In the VS modeling approach (**Figure 1**), language independent vowel detection is performed prior to the cepstral analysis. The detection locates segments that match vowel structure according to an unsupervised language-independent algorithm [7]. For each language, a VS GMM model is trained with the set of detected vowels. During recognition, the utterance likelihood is computed with the detected vowels according to each VS model.

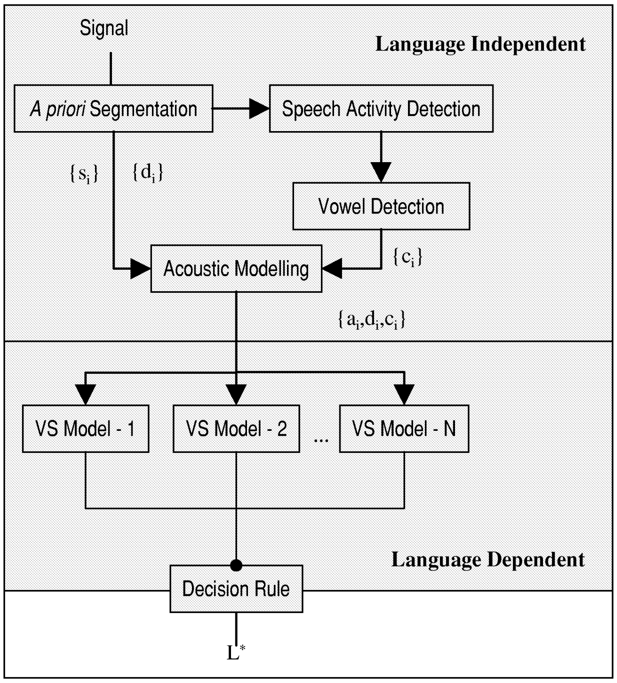

**Figure 1 -** Block diagram of the VS modeling approach. The upper part represents the acoustic preprocessing and the lower part the language dependent Vowel System Modeling.

Let $L = \{L_1, L_2, \ldots L_{NL}\}$ be the $N_L$ languages to identify; the problem is to find the most likely language $L^*$ in the $L$ set.

After the acoustic processing, we obtain for each segment a concatenation of cepstral 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 $o_i$ consists of a parameter vector $y_i$ and a macro-class flag $c_i$, equal to 1 if the segment is detected as a vowel, and equal to 0 otherwise. In order to simplify the formula, we note $o_i=\{y_i,c_i\}$.

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

$$L^* = \underset{1 \le i \le NL}{\operatorname{argmax}} \left[ \Pr(L_i | O) \right] = \underset{1 \le i \le NL}{\operatorname{argmax}} \left[ \Pr(O | L_i) \right] \qquad (1)$$

using Bayes' theorem and assuming that *a priori* language probabilities are identical.

Under the standard GMM assumptions, we assume that each segment is conditionally independent of each other. The VS modeling expression is hence changed to:

$$\Pr(O|L_i) = \prod_{k=1}^{T} \Pr(o_k|L_i) = \prod_{c_k=1} \Pr(y_k|L_i) \qquad (2)$$

since $c_k$ is deterministic and only vowels are taken into account.

According to VS models, the most likely language computed in the log-likelihood space is given by:

$$L^* = \underset{1 \le i \le NL}{\operatorname{argmax}} \left[ \sum_{c_k=1} \log \Pr(y_k|L_i) \right] \qquad (3)$$

# 3. VSM IMPLEMENTATION

## 3.1 Acoustic Processing

Each detected vowel is represented with a set of 8 Mel-Frequency Cepstral Coefficients (MFCCs) and 8 delta-MFCCs. The cepstral analysis is performed using a 256-point Hamming window centered on the detected vowel. This parameter vector may be extended with the duration of the underlying segment and the energy and delta-energy coefficients.

A cepstral subtraction performs both blind deconvolution to remove the channel effect and speaker normalization. For each recording session, the average MFCC vector is computed over all vowels; it is then subtracted from each vowel coefficients. The calculation of the channel effect over the vowel segments rather than over the whole utterance does not show any significant differences but it performs faster.

## 3.2 Vowel System Modeling

Vowel System Models (VSMs) consist in a Gaussian mixture model.

Let $X = \{x_1, x_2, \ldots, x_N\}$ be the training set and $\Pi = \{(\alpha_i, \mu_i, \Sigma_i), 1 \le i \le Q\}$ be the parameter set that defines a mixture of $Q$ $p$-dimensional Gaussians. The model that maximizes the overall likelihood of the data is given by:

$$\Pi^* = \underset{\Pi}{\operatorname{argmax}} \prod_{i=1}^{N} \left\{ \sum_{k=1}^{Q} \frac{\alpha_k}{(2\pi)^{p/2} \sqrt{|\Sigma_k|}} \exp\left[ -\frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right] \right\} \qquad (4)$$

where $\alpha_k$ is the mixing weight of the $k^{\text{th}}$ Gaussian.

The maximum likelihood parameters $\Pi^*$ is performed using the well-known EM algorithm [3]. This algorithm presupposes that the number of components $Q$ and initial values is given for each gaussian pdf. In our system, the LBG and the LBG-Rissanen algorithms fix these parameters.

- **Initializing GMM with the LBG algorithm**

The LBG algorithm [5] elaborates a partition of the observation space by performing an iterated clustering of the learning data into codewords optimized according to the nearest neighbor rule. The splitting procedure may be stopped either when the data distortion variation drops under a given threshold or when a given number of codewords is reached.

- **Initializing GMM with the LBG-Rissanen algorithm**

The LBG-Rissanen algorithm is similar to the LBG algorithm except for the iterated procedure termination. Before splitting, the Rissanen criterion $I(q)$ [8], function of the size $q$ of the current codebook is computed from the expression:

$$I(q) = D_q(X) + 2p.q.\log N \qquad (5)$$

In this expression, $D_q(X)$ denotes the log-distortion of the learning set $X$ according to the current codebook, $p$, the parameter space dimension and N the cardinal of $X$.

Minimizing $I(q)$ results in the optimal codebook size according to the Rissanen information criterion. We use this data driven algorithm to determinate independently the optimal number of gaussian pdf for each language.

- **Identification rules**

During the identification phase, all the vowels detected in the utterance are gathered and parameterized. The likelihood of this set of vowels $Y = \{y_1, y_2, \dots, y_N\}$ according to each VS model $L_i$ is given by:

$$\Pr(Y|L_i) = \sum_{j=1}^{N} \Pr(y_j|L_i) \qquad (6)$$

where $\Pr(y_j|L_i)$ denotes the likelihood of each vowel that is given by:

$$\Pr(y_j|L_i) = \sum_{k=1}^{Q_i} \frac{\alpha_k^i}{(2\pi)^{p/2}\sqrt{|\Sigma_k^i|}} \exp\left[-\frac{1}{2}(y_j-\mu_k^i)^T \Sigma_k^{-1}(y_j-\mu_k^i)\right] \qquad (7)$$

Furthermore, we hypothesize under the *Winner Takes All* (WTA) assumption [6]; the expression (7) is then approximated by:

$$\Pr(y_j|L_i) = \max_{1\le k\le Q_i}\left[\frac{\alpha_k^i}{(2\pi)^{p/2}\sqrt{|\Sigma_k^i|}} \exp\left[-\frac{1}{2}(y_j-\mu_k^i)^T \Sigma_k^{-1}(y_j-\mu_k^i)\right]\right] \qquad (8)$$

# 4. EXPERIMENTS

## 4.1 Corpus description

The VSM approach is tested with the well-known OGI Multilingual Telephone Speech corpus. We limit our experiments to five languages (French, Japanese, Korean, Spanish and Vietnamese) that have been chosen according to their phonological vowel systems. 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.

The data are divided into two corpora, namely the learning and the development sets. Each corpus consists in several utterances (constrained and unconstrained). There is no overlap between the speakers of each corpus. 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 female speakers into account because of the poor number (less than 20 %). The identification tests are made with a subset of the development corpus called 45s since this is the mean duration of the utterances.

## 4.2 Global Segmental Modeling experiments

The reference language identification experiments are performed with several parameter sets. The baseline observation consists of 8 MFCCs. The duration D of the segment may be added. Even if vowels are rather steady sounds, it is well known that a dynamic modeling is more accurate that a pure static one. More complex sets, taking into account the 8 delta MFCCs and the energy coefficients are also examined. The GSM are initialized using the standard LBG algorithm, with a fixed number of codewords Q = 20.

Table 1 – Identification scores with the GSM among 5 languages (45s utterances).

| Parameter Set | | Correct Identification Score |
|---|---|---|
| #1 | 8 MFCCs | 68 % |
| #2 | 8 MFCCs + D | 71 % |
| #3 | 8 MFCCs + 8 DMFCCs + D | 71 % |
| #4 | 8 MFCCs + E +<br>8 DMFCCs + DE + D | 72 % |

Further experiments have been performed in a 3 language identification task (French, Japanese and Spanish) with both male and female speakers in order to compare these results with those given by Zissman with its GMM approach [10]. He reaches 65 % of correct identification (in its system, English replaces French) while with the GSM, we get 68 % of correct identification. Even if the languages are not exactly the same, it shows that segmental GMM is at least as good as a constant duration frame approach.

## 4.3 Vowel System Modeling experiments

Several experiments are reported with the same parameter sets as above, noted #1 to #4.

- **Baseline VSM experiments**

Experiments are performed with several parameter sets and several GMM topologies, either with a number of gaussian pdfs constant among the five languages (ranging from 10 to 30) or with a language specific GMM size determined by LBG-Rissanen (Table 2).

Table 2 – Correct identification scores (%) with the VSM among 5 languages (45s utterances).

| Model<br>Set | LBG-10 | LBG-20 | LBG-30 | LBG-Rissanen |
|---|---|---|---|---|
| #1 | 57 | 57 | 59 | **63** |
| #2 | 63 | 60 | 59 | **67** |
| #3 | **67** | **67** | 64 | 55 |
| #4 | 65 | **69** | 64 | 56 |

Concerning constant size models, the best results (69 %) are reached with the 19 parameter set (8 MFCCs + E + 8 ΔMFCCs + ΔE + D) for 20 gaussian components. Taking the segment duration, the energy parameters and the dynamic cepstral coefficients) into account improves the performance of the static VSM (#1) of about 10 %.

Regarding LBG-Rissanen initializing, it reaches better results than constant size models with the sets #1 and #2 while the performances decrease using the sets #3 and #4. Explanations are given by Table 3. It shows the number of gaussian components computed by LBG-Rissanen algorithm for each language and for each parameter set. It may be deduced that the algorithm behavior is clearly different between low dimension parameter set (#1 and #2) and high dimension sets (#3 and #4). The codebook sizes computed in the first case are about 15-20 components and it is enough to get an efficient modeling, while in high dimension sets, codebook sizes are too small to accurately model each VS. This is probably due to a lack of data that results in a sparse distribution that is not correctly handled with by the LBG-Rissanen algorithm.

Table 3 – LBG-Rissanen codebook size for each language and each parameter set.

| | French | Japanese | Korean | Spanish | Vietnamese |
|---|---|---|---|---|---|
| #1 | 26 | 21 | 19 | 23 | 17 |
| #2 | 20 | 11 | 14 | 17 | 14 |
| #3 | 9 | 5 | 6 | 8 | 4 |
| #4 | 7 | 4 | 5 | 7 | 4 |

- **Improved VSM experiments**

In order to improve the identification robustness, a discriminative pruning procedure is applied during the recognition stage: from each test utterance, the 25 % of the vowel segments that result in the lowest likelihood values according to each VSM are discarded [2]. These more discriminative VSMs result in better identification rates for constant size models: with the #3 parameter set and 20 gaussian pdfs per language, the correct identification rate reaches 77 %.

Finally, a post-processing is applied to take advantage from both the LBG-Rissanen efficiency for low dimension parameter sets and pruning efficiency for high dimension ones. The results provided by the LBG-Rissanen classifier computed with the #2 parameter set (denoted LBG-Rissanen #2) and the LBG classifier estimated with #3 set and including the pruning procedure (denoted LBG-pruning #3) are merged by summing the output likelihood values. Since the two observation streams #2 and #3 are not statically independent, it does not result in joined likelihood values.

Table 4 – Correct identification scores (%) with the improved VSM among 5 languages (45s utterances).

| **Model** | LBG-Rissanen #2 | LBG-pruning #3 | Score merging |
|---|---|---|---|
| **Identification rate** | 67 | 77 | **79** |

- **Discussion**

Including a pruning procedure during recognition improves the performances from 67 % to 77 % with the #3 parameter set (8 MFCCs + 8 ΔMFCCs + D). Even if the #2 set is intrinsically less discriminative, experiments show that data-driven LBG-Rissanen algorithm provides a more efficient codebook than the standard LBG initializing (63 % to 67 %). Moreover, this dimension information is still relevant when used in conjunction with the LBG-pruning #3 model. The resulting correct identification score reaches then 79 % with the 45 second duration utterances from the male speakers.

## 5. CONCLUSION & PERSPECTIVES

This work proves that a significant part of the language characterization is embedded in its vowel system. We show that extracting and modeling this information is possible and efficient. Keeping in mind that vowel segments hardly represent more than 25 % of the overall utterance duration, the differentiated modeling applied to vowel systems is validated. It reaches 79 % of correct identification in a 5 language identification task performed with the male speakers **without requiring any labeled data**.

Similar VSMs may be evaluated with the female speakers and a global system may be designed. At this moment other differentiated models (fricatives, plosives…) are investigated [4]. Taking advantage of several sound-specific models is a quite promising perspective to design totally unsupervised language identification systems.

## 6. REFERENCES

[1] R. André-Obrecht, “A New Statistical Approach for Automatic Speech Segmentation”, *IEEE Trans. on ASSP*, January 88, vol. 36, n° 1, (1988).

[2] L. Besacier and J.F. Bonastre, “Subband approach for automatic speaker recognition: optimal division of the frequency domain”, *in Audio- and Video-based Biometric Person Authentication*, Bigün et al. Eds, Springer LNCS 1206, (1997).

[3] A.P. Dempster, N.M. Laird and D.B. Rubin, “Maximum likelihood from incomplete data via the em algorithm”, J. Royal statist. Soc. Ser. B., 39, (1977).

[4] R. Khoyratty, “ La modélisation acoustique des fricatives en identification automatique des langues ”, Mémoire de DEA, unpublished, (1998).

[5] Y. Linde, A. Buzo and R. M. Gray, “An algorithm for vector quantizer”, *IEEE Trans. On COM.*, January 1980, vol. 28, (1980).

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

[7] F. Pellegrino and R. André-Obrecht, “From Vocalic Detection to Automatic emergence of Vowel Systems”, *Proc. ICASSP ’97*, München, (1997).

[8] J. Rissanen, “A universal prior for integers and estimation by minimum description length”, *The Annals of Statistics*, Vol. 11, No 2, (1983).

[9] J.L. Schwartz, L.J. Boë, N. Vallée and C. Abry, “Major trends in vowel system inventories”, *Journal of Phonetics*, 25, (1997).

[10] 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, (1996).