INVESTIGATING AUTOMATIC LANGUAGE DISCRIMINATION VIA VOWEL SYSTEM AND CONSONANTAL SYSTEM MODELING

Nathalie Parlangeau-Vallès*, François Pellegrino*+ and Régine André-Obrecht*
** IRIT, Toulouse, France - + DDL, Lyon, France*

ABSTRACT

This paper presents an approach to Automatic Language Identification (ALI) based on a differentiated modeling of vowel and consonantal systems. The objective is to consider phonetic and phonological features that are not taken into account in the standard phonotactical approach. For each language, two Gaussian Mixture Models (GMM) are trained respectively with automatically detected vowel and non-vowel segments. Since this vocalic detection is unsupervised and language independent, no labeled data are required. GMMs are initialized using a data-driven variant of the LBG vector quantization algorithm: the LBG-Rissanen algorithm. Experiments show that this algorithm behaves efficiently to take the vowel system structure into account.
With 5 languages from the OGI MLTS corpus and in a close set identification task, we reach 85 % of correct identification for the 45 second duration utterances, considering the male speakers.

1. INTRODUCTION

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 meaningful. This language obstacle will remain until Automatic Language Identification (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 [11]. In most systems, phone recognition is merely considered as a front-end and not exploited for the language likelihood generation.
This method is efficient since good performances are reached in 11 language identification task. However, a quite long utterance (45 seconds) is still necessary to identify languages with a good probability (about 90 % in [11]). Moreover, rather few improvements have been performed since a couple of years. In fact, the phonotactical approach may reach its limit and other methods should be investigated.
The phonotactical method yields a sub-optimal use of the phonetic and phonological differences among languages though they carry a substantial part of language identity. Generally speaking, phonetic modeling is very resource consuming (in term of time and hand-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, voiceless fricatives…) and in separated model processing convenient with each category. At this moment, vowel system modeling has been widely investigated, and a similar consonantal system modeling is proposed as well.
The framework of the proposed approach is settled in the next section and the differentiated modeling system is also described. The system implementation is detailed in Section 3. Section 4 deals with the experiments realized with the OGI multi-lingual telephone speech corpus.

2. DIFFERENTIATED MODELING IN ALI

2.1. Objectives

The main goal of the differentiated modeling is to take phonetic and phonological features into consideration. We develop this topology in order to catch structural features about phonological systems and we first focus on Vowel Systems (VS). This choice is driven by phonological and acoustic considerations.
From a phonological point of view, languages may be partially classified in an efficient way according to their VS [10, 8]: the 451 languages of the UPSID database [10] 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 ‘s worth being exploited.
From an acoustic point of view, it is quite obvious that considering sounds that share the same acoustic structure in an homogeneous model may be more efficient than merging heterogeneous sounds in an unique model : taking both voiceless fricatives and vowels in a single model may result in a less accurate discrimination among fricatives and among vowels that processing separated (or differentiated) models. Moreover, differentiated modeling may enable to take sound specific rules into consideration (vowel space boundaries, etc.).

2.2. Synopsis of the system

The system consists of two Gaussian Mixture Models (GMM) that independently process vowel and consonant segments.
The training procedure (see **Figure 1**) 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 language independent vowel detection locates vowel nuclei [6].
- A segmental cepstral analysis is performed on each segment.
- Two GMMs per language (on for the Vowel System and the other for the Consonant System) is computed with the set of language dependent observations.

Note that, unlike most of acoustico-phonetic decoders, the cepstral analysis is performed on variable length segments rather

than on constant duration frames ; a temporal information, the segment duration is added to the observation vector ; recognition experiments have previously proved its interest [9].
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.

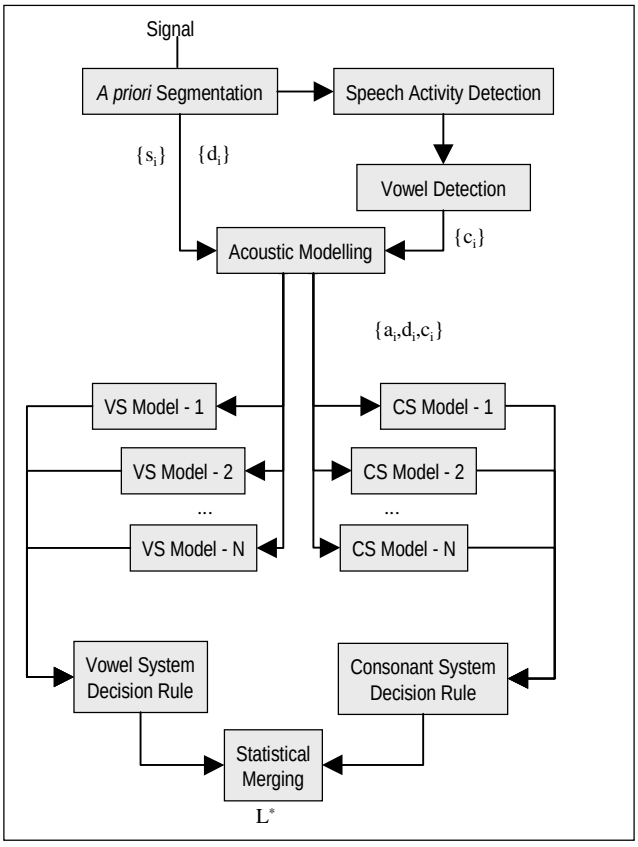

Figure 1 - Block diagram of the Differentiated Modeling approach. The upper part represents the acoustic preprocessing and the lower part the language dependent Vowel and Consonantal System Modeling.

2.3. Statistical framework

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 Differentiated Modeling (DM) is defined by the following equation:

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

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 DM expression is hence changed to:

$$\Pr(O|L_i) = \prod_{k=1}^{T}\Pr(o_k|L_i) = \prod_{c_k=0}\Pr_{CS}(y_k|L_i).\prod_{c_k=1}\Pr_{VS}(y_k|L_i) \quad (2)$$

since c_k is deterministic and considering that $\Pr_{CS}(.|L_i)$ (resp. $\Pr_{VS}(.|L_i)$) denoted the likelihood according to the consonant model (resp. vowel model) in language number i.
According to DM models, the most likely language computed in the log-likelihood space is given by:

$$L^* = \underset{1\le i\le N_L}{\operatorname{argmax}}\left[\sum_{c_k=0}\log\Pr_{CS}(y_k|L_i) + \sum_{c_k=1}\log\Pr_{VS}(y_k|L_i)\right] \qquad (3)$$

3. IMPLEMENTATION

3.1. Acoustic Processing

Each segment 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 either on the detected vowel or on the middle of the consonantal segment. This parameter vector is extended with the duration of the underlying segment providing a 17 coefficient vector.
A cepstral subtraction performs both blind removal of the channel effect and speaker normalization. For each recording session, the average MFCC vector is computed; it is then subtracted from each coefficients.

3.2. Vowel and Consonantal System modeling

Vowel System Models (VSM) and Consonant System Models (CSM) both consist in Gaussian Mixture Models.
Let be $X = \{x_1, x_2, \ldots, x_N\}$ the training set and $\Pi = \{(\alpha_i, \mu_i, \Sigma_i), 1 \le i \le Q\}$ the parameter set that defines a mixture of Q p-dimensional Gaussian laws. The model that maximizes the overall likelihood of the data is given by:

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

where α_k is the mixing weight of the k^{th} Gaussian distribution.
The maximum likelihood parameters estimation is performed using the well-known EM algorithm [3]. This algorithm presupposes that initial values are given for each gaussian pdf and that the number of components Q is also known. In our system, these parameters are fixed using the LBG or the LBG-Rissanen algorithms (see below).

- **Initializing GMM with the LBG algorithm**

The LBG algorithm [4] 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

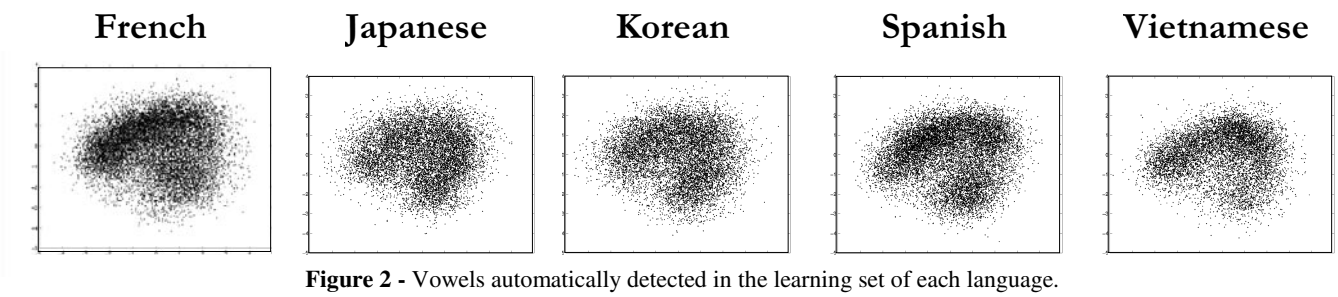

Figure 2 - Vowels automatically detected in the learning set of each language.
Segments are displayed in a common space resulting from Principal Component Analysis performed on 8 MFCC vector processing.

except for the iterated procedure termination. Before splitting, the criterion *J(q)* (derived from the Rissanen criterion [7]), function of the size *q* of the current codebook is computed from the expression:

$$J(q) = D_q(X) + 2p.q.\log(\log N) \quad (5)$$

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 *J(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 pdfs for each language.

- **Identification rules**

During the identification phase, all the vowels (resp. non vowels) detected in the utterance are gathered and parameterized. It results in two sets of vowel and consonantal segments. The likelihood of each set $Y = \{y_1, y_2, \ldots, y_N\}$ according to each DM model L_i is given by:

$$\Pr_{XS}(Y|L_i) = \sum_{j=1}^{N} \Pr_{XS}(y_j|L_i) \quad (6)$$

where $\Pr_{XS}$ is either $\Pr_{CS}$ or $\Pr_{VS}$, according to the set considered. The likelihood of each segment is subsequently given by:

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

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

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

4. EXPERIMENTS

4.1. Corpus Description

The DM 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 [10]. 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. Furthermore, vowel duration is distinctive in Korean. 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. Vowel detection

A part of the OGI MLTS corpus is provided with a broad class labeling (vowels, fricatives…). According to it, the mean rate of correctly detected vowels reaches 93,5 %, with an insertion rate of 10 %. For each language, Figure 2 reports the detected vowels in a common space derived from MFCC analysis.

4.3. Vowel System Modeling (VSM)

Figure 3 displays the results reached unsing only the VSM: only vowel segments are considered. It means that less than 15 second is taken into account for each 45 second utterance.

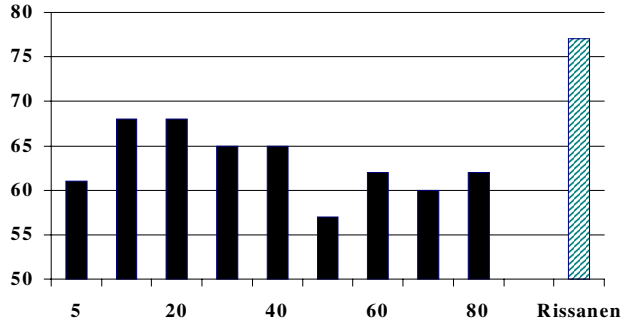

Figure 3 – Correct identification rates using only vowel segments and Vowel System models. Plain bars correspond with models initialized with the LBG algorithm (codebook size is also given). Dashed bar corresponds with LBG-Rissanen initializing.

With constant size models among the 5 languages, the best result

is 67 % of correct identification (with 20 Gaussian components by model). Using the LBG-Rissanen algorithm to estimate the optimal language specific codebook size (given in Table 1) is much efficient since the identification rate is 77 %.
It shows that VS modeling is relevant and that the LBG-Rissanen approach is able to determinate the convenient topology of the model in a language specific way.

	French	Japanese	Korean	Spanish	Vietnamese
Vowel System	29	24	23	22	21
Consonant System	22	23	24	26	27

Table 1 – Codebook size given by the LBG-Rissanen algorithm.

4.4. Consonant System Modeling (CSM)

The same kind of experiments have been conduced using only non vowel segments and CSM. Results (Figure 4) show that the best identification rate is similar to which obtained with VSM: the best topology is given by the LBG algorithm and 30 Gaussian components by model (76 % of correct identification).

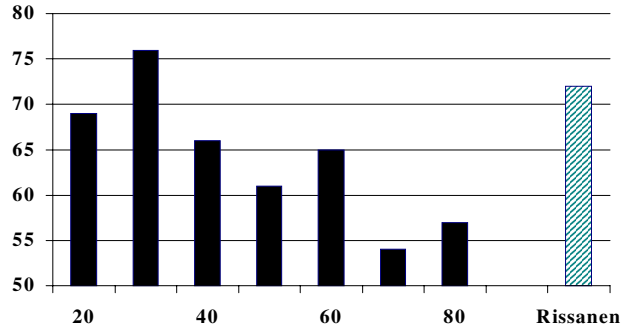

Figure 4 – Correct identification rates using only non vowel segments and Consonant System models. Plain bars correspond with models initialized with the LBG algorithm (codebook size is also given). Dashed bar corresponds with LBG-Rissanen initializing.

The LBG-Rissanen provides less discriminative models than those of constant size. An explanation lays in the data to model. Consonant segments are acoustically heterogeneous; it means that the consonant parameter space is much more complex than the vowel space and the LBG-Rissanen is unable to deal with it.

4.5. Merging VSM and CSM

A pruning post-processing (that discard the less likely segments [2] of each utterance) has been applied to VSM and CSM prior to the statistical merging (equation 3). The results (Table 2) reach 85 % of correct identification using both vowel and non vowel segments.

Model	VSM (Rissanen)	CSM (30 codewords)	VSM + CSM (DM)
Identification rate	78 %	78 %	85 %

Table 2 – Results of the Differentiated Modeling.

5. CONCLUSION & PERSPECTIVES

This work proves that the significant part of the language characterization that is embedded in its vowel system may be used in ALI. The automatic detection of vowel segments is a relevant way to take phonetic and phonological features into account **without requiring any labeled data**. Moreover, experiments show that the LBG-Rissanen algorithm is efficient to model the vowel system structure.
Though the overall identification rate reaches 85 %, it seems that improvements may be done, especially for the consonantal system modeling. Splitting consonant segments according to natural acoustic classes in order to model several homogeneous systems rather than one heterogeneous one is a quite promising perspective.

ACKNOWLEDGMENTS

This work was supported by the French "*Ministère de la Défense*" as part of an agreement with DGA (*Délégation Générale de l'Armement*).

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

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

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

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

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

[9] N. Suaudeau, R. André-Obrecht, "An efficient combination of acoustic and suprasegmental informations in a speech recognition system", *ICASSP 94*, Adelaïde, april 1994.

[10] N. Vallée, *Systèmes vocaliques : de la typologie aux prédictions*, Thèse de 3ème cycle, Univ. Stendhal, Grenoble, (1994)

[11] 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).