Optimal partition of Markov models and automatic classification of languages 16th ASMDA Conference

Jesús E. García V.A. González-López

June 30, 2015

Introduction

- In this paper we introduce a new methodology for the problem of automatic classification of languages according to rhythmic features, using speech samples.
- The problem is to divide a set of languages in subsets with similar rhythmic properties in an automatic way.
- The set of languages is modeled using a new family of Markov models called Partition Markov Models.
- This family has the characteristic of providing a Markov model with a minimal number of parameters.
- We develop a consistent model selection methodology on which to select a good model is equivalente to find the correct partition for the set of languages.

Dataset

The data set consists of 1648 recorded sentences belonging to eight languages as shown in the following table,

language	number of sentences
Catalan	216
Dutch	228
English	132
French	216
Italian	216
Japanese	212
Polish	216
Spanish	212

The sentences have lengths going from 2 to 3.5 seconds. This data comes from a corpus belonging to the *Laboratoire de Sciences Cognitives et Psycholinguistique (EHESS/CNRS)*.

Dataset

We extract from the speech samples the local energy level on the acoustic signal for two specific frequency bands energy on the signal for frequencies from 80 to 800hz and energy on the signal for frequencies from 1500 to 5000hz. The time is discretized in intervals of 25ms. Fixed the sentence j of language l, at time t, the signal is discretized depending on the sign of the variation of the energies at time t in the following way,

$Z_t^{(l,j)}$	energy in 80 to 800hz	energy in 1500 to 5000hz
0	decreasing	decreasing
1	non decreasing	decreasing
2	decreasing	non decreasing
3	non decreasing	non decreasing

The series $Z_t^{(l,j)}$ have alphabet $A = \{0, 1, 2, 3\}$.

Partition Markov models

Notation

Let (X_t) be a discrete time order M Markov chain

- A the finite alphabet;
- $\mathcal{S} = A^M$ the state space;
- $P(a|s) = \text{Prob}(X_t = a|X_{t-M}^{t-1} = s),\ a \in A,\ s \in \mathcal{S}$ the transition probabilities.

Equivalence relationship on $\mathcal{S}$

Definition

for $s, r \in \mathcal{S}$; $s \sim_p r \iff P(a|s) = P(a|r)\ \forall a \in A$.

- For any $s \in \mathcal{S}$, the equivalence class of s is given by $[s] = \{r \in \mathcal{S} | r \sim_p s\}$.
- The classes defined by $\sim_p$ are the subsets of $\mathcal{S}$ with the same transition probabilities.

Equivalence relationship on $\mathcal{S}$

- The equivalence relation defines a partition $\mathcal{L}$ of $\mathcal{S}$.
- We have $(|A|-1)$ transition probabilities for each "part" (element of $\mathcal{L}$), obtaining a model with $(|A|-1)|\mathcal{L}|$ parameters.
- The elements of $\mathcal{S}$ on the same equivalence class activate the same random mechanism to choose the next element in the Markov chain.

Markov chain with partition $\mathcal{L}$

Definition

Let (X_t) be a discrete time, order M Markov chain on A and let $\mathcal{L} = \{L_1, L_2, \ldots, L_K\}$ be a partition of $\mathcal{S}$. We will say that (X_t) is a Markov chain with partition $\mathcal{L}$ if this partition is the one defined by $\sim_p$.

Model selection problem

Let x_1^n be a sample of the process $\left(X_t\right)$, $s \in \mathcal{S}$, $a \in A$ and $n > M$.

$$N_n(s,a) = \left|\left\{t : M < t \leq n, x_{t-M}^{t-1} = s, x_t = a\right\}\right|, \tag{1}$$

$$N_n(s) = \left|\left\{t : M < t \leq n, x_{t-M}^{t-1} = s\right\}\right|. \tag{2}$$

To simplify the notation we will omit the $_n$ on N_n.

A distance in S

Definition

We define the distance d in S,

$$\begin{aligned} d(s,r) &= \frac{2}{(|A|-1)\ln(n)} \sum_{a\in A} \left\{ N(s,a) \ln\left(\frac{N(s,a)}{N(s)}\right) \right. \\ &+ N(r,a) \ln\left(\frac{N(r,a)}{N(r)}\right) \\ &- \left. (N(s,a)+N(r,a)) \ln\left(\frac{N(s,a)+N(r,a)}{N(s)+N(r)}\right) \right) \end{aligned}$$

for any $s, r \in S$.

A distance in S

Proposition

For any $s, r, t \in S$,

i. $d(r,s) \geq 0$ *with equality if and only if* $\frac{N(s,a)}{N(s)} = \frac{N(r,a)}{N(r)} \quad \forall a \in A$,

ii. $d(r,s) = d(s,r)$,

iii. $d(r,t) \leq d(r,s) + d(s,t)$.

Consistence in the case of a Markov source

Theorem

Let (X_t) be a discrete time, order M Markov chain on a finite alphabet A. Let x_1^n be a sample of the process, then for n large enough and for each $s, r \in S$, $d_n(r,s) < 1$ iff s and r belong to the same class.

Partition of a set of Markov processes

Notation

- $A = \{0, 1, 2, 3\}$
- $M = 4$
- $S = A^4$
- $L = \{1, 2, 3, 4, 5, 6, 7, 8\}$ is the set of languages.

Notation

- $\{z_t^{(l,j)}\},\ t = 1, 2, ..., n_{l,j}$ is the sample j for language l.
- $N_{l,j}(s) = \left|\{t : M < t \leq n, (z^{(l,j)})_{t-M}^{t-1} = s\}\right|$ is the number of occurrences of s in the sample j of language l.
- $N_{l,j}(s,a) = \left|\{t : M < t \leq n, (z^{(l,j)})_{t-M}^{t-1} = s, z_t^{(l,j)} = a\}\right|$ is the number of occurrences of s followed by a in the sample j of language l.

Notation

- $N_l(s) = \sum_j N_{l,j}(s)$ is the number of occurrences of s for language l.
- $N_l(s, a) = \sum_j N_{l,j}(s, a)$ is the number of occurrences of s followed by a for language l.
- For $B \subseteq L$,
  $N_B(s) = \sum_{l \in B} N_l(s)$ and $N_B(s, a) = \sum_{l \in B} N_l(s, a)$.

BIC criterion

- Let $B \subseteq L$ be a set of languages such that meassurements Z for that languages have the same distribution.
- Consider z_B be the set of samples corresponding to the languages in B, and n_B the size of z_B.
- Let $\mathcal{L}$ be a partition of S, then

$$BIC(z_B, \mathcal{L}) = \sum_{a \in A, L \in \mathcal{L}} N_B(L, a) \ln\left(\frac{N_B(L, a)}{N_B(L)}\right) - \frac{(|A| - 1)}{2} |\mathcal{L}| \ln(n_B).$$

BIC criterion

- Let $\mathcal{B}$ be a partition of L.
- For each $B \in \mathcal{B}$, z_B is the set of samples corresponding to the languages in B, and n_B the size of z_B.
- For each $B \in \mathcal{B}$, let $\mathcal{L}_\mathcal{B}$ be a partition of S, then

$$\text{BIC}\Big(z_L, \mathcal{B}, \{\mathcal{L}_B\}_{B\in\mathcal{B}}\Big) = \sum_{B\in\mathcal{B}} BIC(z_B, \mathcal{L}_B).$$

- The BIC estimator of $\mathcal{B}$ is

$$\hat{\mathcal{B}} = arg \max_{\mathcal{B}\in\mathcal{P}} \sum_{B\in\mathcal{B}} BIC(z_B, \hat{\mathcal{L}}_B)$$

where $\hat{\mathcal{L}}_B$ is the partition estimated adjusting a PMM to the dataset z_B.

Consistence in the case of multiple Markov sources

Theorem

Let $(Z_t^1), (Z_t^2), ..., (Z_t^L)$, be a collection of L discrete time, order M Markov chains over the same finite alphabet A. For each $1 \leq l \leq L$ let $(z_t^l)_1^n$ be a size n sample of the process (Z_t^l), then, eventualy almost surely, the BIC estimate $\hat{\mathcal{B}}$ will be the true partition of the set of processes $(Z_t^1), (Z_t^2), ..., (Z_t^L)$.

Application

Consider the following numbering for our eight languages,

Catalan	Dutch	English	Spanish	French	Italian	Japanese	Polish
1	2	3	4	5	6	7	8

Remembering our BIC estimator for the partition of languages,

$$\hat{\mathcal{B}} = arg \max_{\mathcal{B} \in \mathcal{P}} \sum_{B \in \mathcal{B}} BIC(z_B, \hat{\mathcal{L}}_B),$$

where $\mathcal{P}$ is the set of all the partitions of $\{1, 2, 3, 4, 5, 6, 7, 8\}$.
obtaining a BIC value.

Application

The following table shows the 5 partitions of $\{1, 2, 3, 4, 5, 6, 7, 8\}$ with the largest BIC values.

partition ($\mathcal{B}$)	BIC
$\mathbf{\{1, 4\}, \{2, 3, 8\}, \{5, 6, 7\}}$	**-347488.098872141**
$\{1\}, \{2.3.8\}, \{4\}, \{5.6.7\}$	-347509.127775736
$\{1.4\}, \{2.3.5.8\}, \{6.7\}$	-347517.276944946
$\{1.4\}, \{2.8\}, \{3.5.6.7\}$	-347519.546013801
$\{1.4\}, \{2.8\}, \{3\}, \{5.6.7\}$	-347523.086121394

On the first line we can see in bold face the winning partition which correspond to:

{Catalan, Spanish}
{Duth, English, Polish}
{French, Italian, Japanese}.

The only discrepancy with the linguistic conjecture on the winning partition of languages is the placement of Japanese which is the only moraic language in the sample and should be alone. We note that the method proposed in García and González-López[10] is able to capture the singularities of Japanese but it has other weaknesses. This misplacement of Japanese also happened on Cuesta-Albertos *et al.*[2] and García *et al.*[11]-García *et al.*[13]. In contrast, the methodology was particularly efficient in two controversial cases: Polish and Catalan. The first language was included in the stress-timed part while the second language was reported as being similar to Spanish. Despite Polish shows a high syllable complexity, but without the expected vowel reduction for a stress-timed language and Catalan has the same syllabic system as Spanish, although it has some vowel reduction. This suggests that Catalan is not rhythmically different from Spanish.

The authors would like to thank Franck Ramus for providing the 1648 sentences dataset used on this study. We gratefully acknowledge the support for this research provided by (a) USP project: Mathematics, computation, language and the brain (b) Portuguese in time and space: linguistic contact, grammars in competition and parametric change, FAPESP's project, grant 2012/06078-9 and (c) FAPESP Center for Neuromathematics (grant 2013/ 07699-0, S. Paulo Research Foundation).

- D. Abercrombie. *Elements of general phonetics*, Chicago: Aldine (Chapter 5), 1967.
- J. Cuesta-Albertos, R. Fraiman, A. Galves, J. Garcia and M. Svarc. Identifying rhythmic classes of languages using their sonority: a Kolmogorov-Smirnov approach. *Journal of Applied Statistics* 34(6), 749-761, 2007.
- R.M. Dauer. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics* 11, 51-62, 1983.
- S. Frota, C. Galves, M. Vigário, V. Gonzalez-Lopez and B. Abaurre. The phonology of rhythm from Classical to Modern Portuguese. *Journal of Historical Linguistics* 2(2), 173-207, 2012.
- A. Galves, J. Garcia, D. Duarte and C. Galves. Sonority as a basis for rhythmic class discrimination. In *Speech Prosody 2002, International Conference,* 2002.

- A. Galves, C. Galves, J.E. Garcia, N.L. Garcia and F. Leonardi, F. Context tree selection and linguistic rhythm retrieval from written texts. *The Annals of Applied Statistics* 6(1), 186-209, 2012.
- J. Garcia, U. Gut and A. Galves. Vocale-a semi-automatic annotation tool for prosodic research. In *Speech Prosody 2002, International Conference,* 2002.
- J. Garcia and V.A. Gonzalez-Lopez. Minimal markov models. arXiv preprint arXiv:1002.0729, 2010.
- J.E. García and V.A. González-López. Minimal Markov Models. In *Proceedings of the Fourth Workshop on Information Theoretic Methods in Science and Engineering. Helsinki.* v.1. p.25 - 28, 2011.
- J.E. García and V.A. González-López. Modeling of acoustic signal energies with a generalized Frank copula. A linguistic

conjecture is reviewed. *Communications in Statistics-Theory and Methods* 43(10-12), 2034-2044, 2014.


J.E. García, V.A. González-López and M.L.L. Viola. Robust model selection and the statistical classification of languages. In *XI BRAZILIAN MEETING ON BAYESIAN STATISTICS: EBEB 2012,* vol. 1490, no. 1, pp. 160-170. AIP Publishing, 2012.


J.E. García, V.A. González-López and R.B. Nelsen. A new index to measure positive dependence in trivariate distributions. *Journal of Multivariate Analysis* 115, 481-495, 2013.


J.E. García, V.A. González-López and M.L.L. Viola. Robust Model Selection for Stochastic Processes. *Communications in Statistics-Theory and Methods* 43(10-12), 2516-2526, 2014.

 F. Ramus, M. Nespor and J. Mehler. Correlates of linguistic rhythm in the speech signal. *Cognition* 73, 265-292, 1999.