Automatic Rhythm Modeling for Language Identification

Jérôme Farinas[1] & François Pellegrino[2]

[1]Institut de Recherche en Informatique de Toulouse
UMR 5505 CNRS – Université Paul Sabatier, France
Jerome.Farinas@irit.fr

[2]Laboratoire Dynamique Du Langage
UMR 5596 CNRS – Université Lumière Lyon 2, France
Francois.Pellegrino@univ-lyon2.fr

Abstract

This paper deals with an approach to Automatic Language Identification based on rhythmic modeling. Beside phonetics and phonotactics, rhythm is actually one of the most promising features to be considered for language identification, but significant problems are unresolved for its modeling. In this paper, an algorithm of rhythm extraction is described. Experiments are performed on read speech for 5 European languages. They show that salient features may be automatically extracted and efficiently modeled from the raw signal: a Gaussian mixture modeling of the extracted features results in a 81 % percent of correct language identification for the 5 languages, using 20 s duration utterances.

1. Introduction

Automatic Language Identification emerged during the last ten years. The major stakes may be divided in two areas: Multilingual Man-Computer Interfaces (Interactive Information Terminal, Speech dictation, etc.) and Computer-Assisted Communication (Emergency Service, etc.).

At present, the standard approach considers a phonetic modeling system as a front-end, and the resulting sequences of phonetic units are decoded according to language-specific statistical grammars [1]. Even if this approach reaches the best results, only marginal improvements have been performed since ’96, and it seems crucial not to underestimate the relevancy of alternative features present in the signal.

Among the different levels of language description, prosodic features carry a substantial part of the language identity (Section 2). However, due to the numerous problems that arise when talking about *automatic* rhythm extraction, most of the previous experiments aiming at language identification with rhythm are based on hand-labeled data ([2], [3], [4]). The approach presented here challenges the automatic extraction of rhythmic features in a fully unsupervised language-independent approach (Section 3). Based on a Vowel/Non-Vowel segmentation, it is subsequently exploited in a statistical rhythm modeling for automatic language identification (Section 4). Very promising results are obtained. Perspectives are explored in Section 5.

2. Addressing rhythm definition

2.1. The importance of rhythm

Rhythm is a characteristic of language that may be eventfully critical in different activities related to language:

- language acquisition

According to the frame-content theory [5], the rhythm, and especially the CV pattern (the frame), is closely related to the closed-open alternation of the mouth during speech production. According to MacNeilage & Davis, this cycle is provided by the mandibular oscillation and it may be the first step in the evolution and acquisition of speech, followed by the rise of the capacity to produce a sequence of frames filled with different consonants and vowels (the content). Moreover, several experiments have shown that newborn children pay a great attention to syllabic patterns (see [6] for a short review).

- language synthesis

In speech synthesis the notion of rhythm is most often related to the distinction between stressed vs. unstressed units. This distinction is important for the comprehension of stress-timed languages. However, this binary distinction does not define the fine timing distinctions of fluent speech, and does not match with syllable-timed family of languages.

- language identification

Among others, Thymé-Gobbel and Hutchings point out the importance of rhythmic information in language identification systems [4]. With parameters related to rhythm and based on syllable timing, syllable duration, and descriptors of amplitude patterns, they have obtained promising results, and proved that mere prosodic cues can distinguish between some language pair with results comparable to some non-prosodic systems.

Ramus et al. [7] show that newborn infants are sensitive to the rhythmic properties of languages. Other experiments based on a consonant/vowel segmentation of eight languages established that measured parameters might be able to classify languages according to rhythmic properties of languages [8].

2.2. Linguistic classes of rhythm

Experiments reported here focus on 5 European languages (English, French, German, Italian and Spanish). According to the literature, French, Spanish and Italian have syllable-timed rhythm while English and German have stress-timed rhythm. These two categories emerged from the theory of isochrony introduced by Pike and developed by Abercrombie [9]. But more recent works, based on the measurement of the duration of inter-stress intervals in both stress-timed and syllable-timed languages provide an alternative framework in which these two binary categories are replaced by a continuum [10]. Rhythmic differences between languages are then mostly related to their syllable structure and the presence (or absence) of vowel reduction.

2.3. Rhythmic units and patterns

The different works in linguistics or psycholinguistics reported above and the subsequent controversies on the status of rhythm in world languages illustrate dramatically the difficulty to segment speech into correct rhythmic units. Even

if correlates between speech signal and linguistic rhythm exist [8], reaching a relevant representation of it seems difficult. Another difficulty rises from the selection of an efficient modeling paradigm. At this moment, experiments based on neural networks show interesting trends [2], but the problem is far from being resolved. We propose a new approach, based on a Gaussian modeling of the different "rhythm units" automatically extracted from a pseudo rhythmic segmentation in the languages.

3. Rhythmic segmentation

Even if the existence of non-vocalic syllabic core is reported, most of the rhythmic patterns alternate Consonants and Vowels. Thus, automatic rhythm extraction necessitates a segmentation of speech according to Consonant/Vowel labels. To reach that point, we take advantage from an algorithm formerly used for model vowel systems in a language identification task [12]. The main features of this algorithm are reviewed hereunder.

3.1. Speech segmentation

In order to extract features related to the potential consonant cluster (number and duration of consonants), a statistical segmentation based on the "Forward-Backward Divergence" algorithm is applied. Interested readers are referred to [11] for a detailed study of the Forward Backward Divergence Algorithm. The algorithm results in a segmentation into short segments (bursts, but also transient parts of voiced sounds) and longer segments (steady parts of sounds).

3.2. Vowel detection

A segmental speech activity detection is performed to discard pauses (not related to rhythm) and the vowel detection algorithm locates sounds that match a vocalic structure *via* a spectral analysis of the signal [12]. It is applied in a language and speaker independent way without any manual adaptation phase.

3.3. Rhythm and automatic segmentation

The processing provides a segmentation of the speech signal in pause, non-vowel and vowel segments (see Figure 1). Due to the intrinsic properties of the algorithm (and especially the fact that transient and steady parts of a phoneme may be separated), it is somewhat incorrect to consider that this segmentation is exactly a Consonant/Vowel segmentation.

However, it is undoubtedly correlated to the rhythmic structure of the speech sound, and in this paper we investigate the assumption that this correlation enables a statistical model to discriminate languages according to their rhythm structure.

3.4. Rhythm modeling units: Pseudo-syllables

Modeling rhythm implies to select suitable units. We saw in Section 2 that they vary among the languages and that their intrinsic supra-segmental nature is not trivial to model.

The existence of syllables, even if this unit may not be the most salient in stress-timed languages, is confirmed in all the languages of the world. However, the segmentation of speech in syllables is typically a language-specific mechanism and thus no language independent algorithm can be derived, especially when syllable boundary occurs between consonants (e.g. in a CVC.CV occurrence as in the French word *parmi*)).

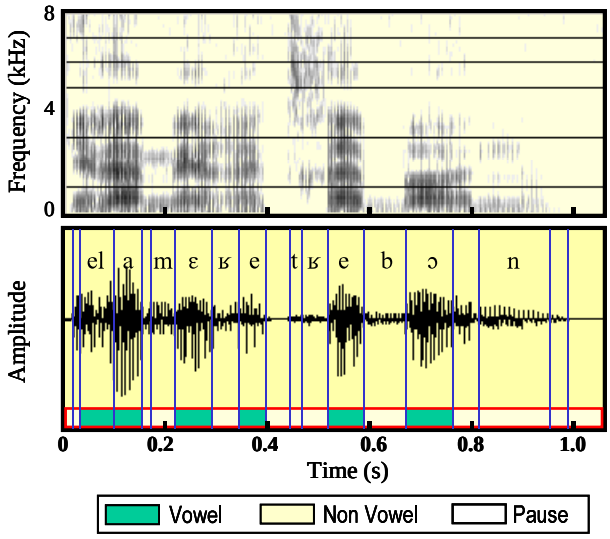

*Figure 1: Example of a Silence/Vowel/Non-vowel automatic segmentation. The speaker pronounces "*Et la mer est très bonne*". Vertical lines are given by the segmentation algorithm.*

For these reasons, we introduce the notion of Pseudo-Syllables (PS) derived from the most frequent syllable structure in the world, namely the CV structure ([13]). In our algorithm, the speech signal is parsed in patterns matching the structure: .C^nV. (with *n* an integer that may be zero).
For example, the parsing of the sentence displayed in Figure 1 results in the following sequence of 7 pseudo-syllables: (CCVV.CCV.CV.CCCV.CV.CCC). Then consecutive vowel segments are merged and clusters without vowels are discarded. So the example sequence becomes (CCV.CCV.CV.CCCV.CV).

We are aware of the limits of such a basic rhythmic parsing, but it provides an attempt to model rhythm that may be subsequently improved. However, it has the considerable advantage that neither hand-labeled data nor extensive knowledge on the language rhythmic structure is required.

4. Pseudo-syllabic modeling

4.1. Corpus

Experiments are performed on the MULTEXT multilingual corpus [14]. This database contains recordings from five European languages (French, English, Italian, German and Spanish), pronounced by 50 different speakers (5 male and 5 female per language). Data consist of read passages of about five sentences extracted from the EUROM1 speech corpus. They are augmented with the raw pitch contour and additional prosodic information (not considered here). A limitation is that the same texts are produced on average by 3.75 speakers, resulting in a possible partial text dependency of the models. Table 1 shows the duration of the corpus for each language.

Table 1: The MULTEXT Corpus (from Campione & Véronis [14])

Language	Passages per speaker	Total duration (min.)	Average duration per passage (s)
English	15	44	17.6
French	10	36	21.9
German	20	73	21.9
Italian	15	54	21.7
Spanish	15	52	20.9

4.2. Pseudo-syllables description

A pseudo-syllable is described as a sequence of segments characterized by their duration and their binary category (Consonant or Vowel). This way, each pseudo-syllable is described by a variable length matrix. For example, a .CCV. pseudo-syllable will give:

$$P_{CCV.} = \begin{pmatrix} C & C & V \\ d_{C1} & d_{C2} & d_{V1} \end{pmatrix} \qquad (1)$$

where C and V are binary labels and d_X is the duration of the segment X.

This variable length description is the most accurate, but it is not appropriate to a Gaussian Mixture Modeling (GMM). For this reason, another description resulting in a constant length description for each pseudo-syllable has been derived. For each pseudo-syllable, three parameters are computed, corresponding respectively with the total consonant cluster duration, the total vowel duration and the complexity of the consonantal cluster. With the same .CCV. example, the description is then :

$$P'_{CCV.} = \{(d_{C1} + d_{c2}) \quad d_V \quad N_C\} \qquad (2)$$

where N_C is the number of segments in the consonantal cluster (here, $N_C = 2$).

Even if this description is clearly non-optimal since the individual information on the consonant segments is loosed, it takes a part of the complexity of the consonant cluster into account.

4.3. GMM Modeling

GMM are used to model the pseudo-syllables which are represented in the three dimensional space described in the previous section. They are estimated using the EM algorithm initialized with the *k*-means algorithm.

Since the amount of data is very limited (especially in terms of number of speakers), a bootstrapping is performed (i.e. nine of the ten speakers are used for the training while the tenth one is classified according to the maximum likelihood criterion). The learning-testing procedure is iterated for each speaker of the corpus.

4.4. Language Identification

Language identification experiments are performed on the five languages of the MULTEXT corpus. Each passage of about 20 seconds is tested individually using the bootstrapping method described above. Three parameters are investigated in the language identification task.

4.4.1. *Influence of the GMM topology*

This first factor is the number N_{Gauss} of Gaussian components in the GMM (see Figure 2).

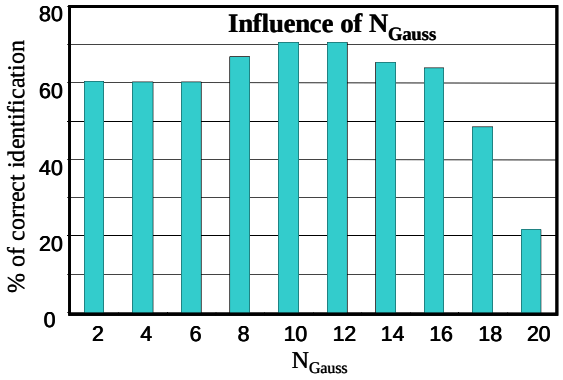

Figure 2: Evaluation of the influence of the number of Gaussian components in each model. This histogram displays mean values computed among 4 experiments.

It is worthy of note that with no more that 2 Gaussian components, an identification rate of 60 % is performed, underlying that the rhythmic modeling is relevant. The best average results (71 % correct) are reached with $N_{Gauss} = 12$, and a decrease is observed for more complex models. A more careful study of the results (not detailed here) has shown a great variation among the results obtained with these complex models. Several models reach very good performances (up to 81 % of correct identification), but the limited amount of data result in a significant instability.

4.4.2. *Influence of the complexity of the consonant cluster*

The second factor is N_{Cmax}, the maximum number of segments considered in a consonant cluster: if a very long sequence of consonant segments is detected in a waveform, it may result from an omission of the vowel detection algorithm. In that case the cluster does not represent correctly the rhythmic of the language.

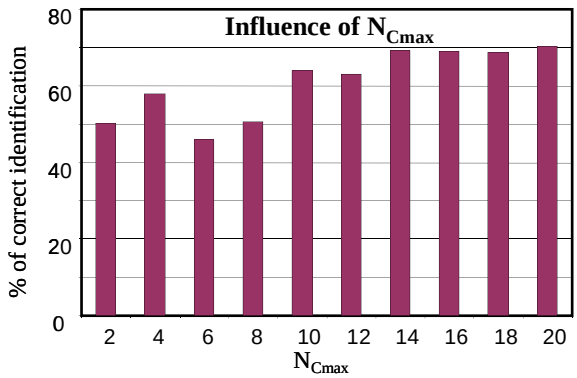

Figure 3: Evaluation of the influence of the maximum number of consonant segments considered in a consonant cluster. This histogram displays mean values computed among 4 experiments.

Experiments evaluate the influence of a maximum limit of complexity for the consonantal clusters. Figure 3 shows that the performances reach a flat level from N_{Cmax} equals 14 to 20

(clusters longer than 20 segments are very uncommon). It demonstrates that, even if over-segmentation results from the algorithm, the resulting segmentation is relevant and catches language rhythm variation.

4.4.3. Influence of the test duration

Table 2 presents the results of correct identification with a variation of the duration of the speech files taking into account. The results concern the $N_{Gauss} = 15$ and $N_{Cmax} = 14$ model. With only 5s of speech file, the results are much greater than chance. It shows that even with short excerpts, the rhythmic structure may be correctly modeled with GMM.

Table 2: Evolution of correct identification rates (in percent) in function of test duration (in seconds)

Duration / Language	5	10	15	20
EN	41	55	54	59
FR	72	80	82	83
GE	59	68	74	80
IT	34	41	49	55
SP	72	81	86	93
Mean	**56**	**65**	**69**	**74**

4.4.4. Matrix of confusion

The confusion matrix resulting from an experiment with $N_{Gauss} = 18$, $N_{Cmax} = 12$ and 20s test duration is shown in Table 3. Several results emerge from this experiment:

First, it proves that the pseudo-syllable modeling is able to take a significant part of the rhythmic structure of languages into consideration. The worst identification rate is for Italian (53 %), and even if it is far from the results reached for the other languages (ranging from 81 % for English to 100 % for Spanish), it is significantly different from chance.

Table 3: Matrix of confusion (in percent). The average correct identification rate is 81 %.

Mode / Item	EN	FR	GE	IT	SP
EN	**81**	1	9	9	1
FR	-	**84**	7	1	8
GE	9	-	**87**	4	-
IT	18	-	4	**53**	25
SP	-	-	-	-	**100**

5. Conclusion and perspectives

We propose one of the first approaches dedicated to rhythm language identification that is tested on a task more complex than paired language comparisons. The experiments done with 5 languages produce relatively good results (81% correct identification rate for 20-second utterances), that can be more easily compared to traditional non-prosodic systems. It is interesting to point out that the pseudo-syllable modeling manages to identify languages that belong to the same rhythmic family (e. g. syllable-timed rhythm for French, Italian and Spanish), showing that the temporal structure of the pseudo-syllables is quite language-specific. Problems reported with Italian and some English-German distinctions are arguments for testing more features, more related to stress and tone (parameters derived from energy and F_0). However, the limited amount of data, emphasized by the instability of the more complex Gaussian models, shows that it may be necessary to continue experiments with bigger corpora.

6. Acknowledgements

This research is supported by the EMERGENCE program of the *Région Rhône-Alpes* and the French *Ministère de la Recherche* (program ACI "*Jeunes Chercheurs*").

7. References

[1] Zissman, M., "Comparison of four approaches to automatic language identification of telephone speech", *Proc. IEEE Trans. On SAP*, Vol. 4, no. 1, 1996.

[2] Dominey, P. F., and Ramus, F., "Neural Network Processing of Natural Language: I. Sensitivity to Serial, Temporal and Abstract Structure in the Infant", *Language and Cognitive Processes*, 15(1), 87-127, 2000.

[3] Farinas, J. and André-Obrecht, R., "Identification automatique des langues : variations sur les multigrammes", *Proc. of JEP 2000*, Aussois, 2000.

[4] Thymé-Gobbel, A., and Hutchins, S. E., "Prosodic features in automatic language identification reflect language typology", *Proc. of ICPhS'99*, San Francisco, 1999.

[5] MacNeilage, P. F., and Davis, B. L., "On the Origin of Internal Structure of Word Forms", *Science*, 288:527:531, 2000.

[6] Mehler, J., and Dupoux, E., *Naître Humain*, Editions Odile Jacob, Paris, 1990.

[7] Ramus, F., Hauser, M. D., Miller, C., Morris, D. and Mehler, J., "Language discrimination by human newborns and by cotton-top tamarin monkeys", *Science*, 288, 349-351, 2000.

[8] Ramus, F., Nespor, M., & Mehler, J., "Correlates of linguistic rhythm in the speech signal", *Cognition*, 73(3), 265-292, 1999.

[9] Abercrombie, D., *Elements of General Phonetics*, Edinburgh University Press, Edinburgh, 1967.

[10] Dauer, R. M., "Stress-timing and syllable-timing reanalyzed", Journal of Phonetics, 11:51-62, 1983.

[11] André-Obrecht, R., "A New Statistical Approach for Automatic Speech Segmentation", *IEEE Trans. on ASSP*, vol. 36, n° 1, 1988.

[12] Pellegrino, F., and André-Obrecht, R., "An Unsupervised Approach to Language Identification", *Proc. of ICASSP'99*, Phoenix, 1999.

[13] Vallée, N., Boë, L.J., Maddieson, I. and Rousset, I., "Des lexiques aux syllabes des langues du monde – Typologies et structures", *Proc. of JEP 2000*, Aussois, 2000.

[14] Campione, E., and Véronis, J., "A multilingual prosodic database", *Proc. of ICSLP'98*, Sidney, 1998.