Can Automatically Extracted Rhythmic Units Discriminate among Languages?

François Pellegrino[1], Jean-Hugues Chauchat[2], Ricco Rakotomalala[2] & Jérôme Farinas[3]

[1]DDL UMR 5596 CNRS – Univ. Lumière Lyon 2, Lyon, France
[2]ERIC Univ. Lumière Lyon 2, Lyon, France
[3]IRIT UMR 5505 CNRS – Univ. Toulouse 3, Toulouse, France
{pellegri;chauchat;rakotoma}@univ-lyon2.fr ; jfarinas@irit.fr

Abstract

This paper deals with rhythmic modeling and its application to language identification. 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 dedicated to rhythmic segmentation is described. Experiments are performed on read speech for 5 European languages. Several algorithms are compared. They show that salient features may be automatically extracted and efficiently modeled from the raw signal: a linear discriminant analysis of the extracted features results in a 80 % percent of correct language identification for the 5 languages, using 20 s duration utterances. Additional experiments reveal that the automatic rhythmic units convey also speaker specific features.

1. Introduction

As other supra segmental aspects of speech, rhythm modeling is challenging engineers for many years. Beside the problems related to the design of an efficient model, one of the main difficulties is to define *what* to model since it is intricately linked both to the segmental structure of the language (voiced/unvoiced sounds, etc.) and to its supra segmental dimension. In this paper, we address the problem of extracting rhythmic units relevant for language identification.

At present, the standard automatic language identification 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 [20]. 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, and especially rhythm, carry a substantial part of the language identity (Section 2). However, due to the numerous problems that arise when talking about automatic rhythm extraction (Section 3), experiments on this topic are seldom, and most of the previous ones aiming at language identification are based on hand-labeled data ([5],[7],[18]).

The approach presented here, introduced in [6], challenges the puzzle of automatic extraction of rhythmic features in a fully unsupervised language-independent approach (Section 4). Based on Vowel/Non-Vowel segmentation, it is subsequently exploited in a rhythm unit modeling for automatic language identification (Section 5). Several algorithms based on speech processing or data mining techniques are compared and the inter-speaker variability of rhythm is introduced.

2. Motivations

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

2.1. Language acquisition

Numerous works in psycholinguistics have shown the major role of rhythm in the early language acquisition process (e.g. [10]). Moreover, according to the frame-content theory [9], 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).

2.2. Language synthesis

Reaching a "natural voice" is one of the major challenges of language synthesis. At this moment, 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.

2.3. Language identification

Among others, Thymé-Gobbel and Hutchings point out the importance of rhythmic information in language identification [18]. 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. [5] have shown that newborn infants are sensitive to the rhythmic properties of languages. Other experiments based on consonant/vowel segmentation of eight languages established that acoustic parameters might be able to classify languages according to their rhythmic properties ([5],[16]).

3. Dealing with rhythm

3.1. 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 [1].
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 [4]. In this theory, rhythmic differences between languages are then mostly related to their syllable structure and the presence (or absence) of vowel reduction.

3.2. 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 [16], 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 [5], but the problem is far from being resolved. We propose a new approach, based on an automatic segmentation of speech into rhythmic units not strictly related to the syllabic parsing.

4. 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 may be based on a segmentation of speech according to Consonant/Vowel labels.
To reach that point, we take advantage from an algorithm formerly developed to model vowel systems in a language identification task [12]. The main features of this algorithm are reviewed hereunder.

4.1. 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 [2]. It results in a segmentation into short segments (bursts, but also transient parts of voiced sounds) and longer segments (steady parts of sounds).

4.2. Vowel detection

A segmental speech activity detection is performed to discard pauses (not related to rhythm) and a 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.

4.3. 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.

4.4. Rhythm modeling units: Pseudo-syllables

Modeling rhythm implies to select suitable units. We saw in Section 3 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 assessed in all the languages of the world. However, the segmentation of speech into syllables is typically a language-specific mechanism even if considering the sonority scale may allow deriving general rules [13]. At this moment, no automatic language independent algorithm is available.

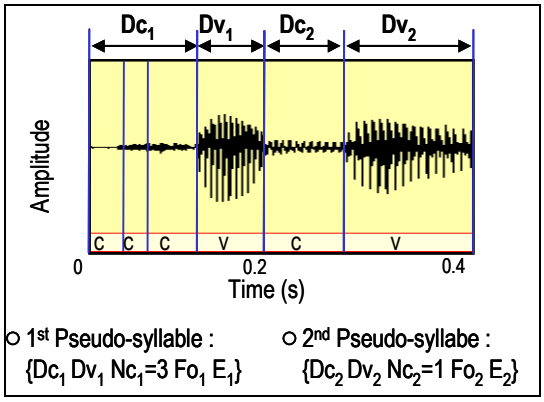

Figure 1: Example of a Vowel/Non-vowel segmentation. The speaker pronounced "… très bo(nne)*". Vertical lines are given by the segmentation algorithm.*

Consequently, we introduce the notion of Pseudo-Syllables (PS) derived from the most frequent syllable structure in the world, namely the CV structure ([19]). In our algorithm, the speech signal is parsed in sequential patterns matching the structure: .C^nV. (with *n* an integer that may be zero). This way, no boundary has to be placed between to consonantal segments.
For example, the parsing of the example displayed in Figure 1 results in the sequence of 2 pseudo-syllables .CCCV.CV. corresponding with the phonetic sequence /t R ɛ b ɔ/.

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.

5. Pseudo-syllabic modeling

5.1. Corpus

Experiments are performed on the MULTEXT multilingual corpus [3]. 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.
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 [3])

Language	Nb. Of speakers	Passages per speaker	Total duration (min.)	Average duration per passage (s)
English	10	15	44	17.6
French	10	10	36	21.9
German	10	20	73	21.9
Italian	10	15	54	21.7
Spanish	10	15	52	20.9
TOTAL	50	75	4h19	20.7

5.2. Pseudo-syllables description

A pseudo-syllable may be described as a sequence of segments characterized by their binary category (Consonant or Vowel) and their duration. It would result in a variable length description, which would be difficult to handle. For this reason, another description resulting in a constant length description for each pseudo-syllable has been derived:

For each pseudo-syllable, three basic parameters are computed, corresponding respectively with the total duration of the consonant cluster, the vowel duration and the complexity (in terms of number of segments) of the consonantal cluster. They are augmented with to parameters related to the stress structure of the rhythm: the Energy and the pitch of the vowel of the pseudo-syllable. These parameters are normalized according to their mean values along the sentence. With a .CCV. example, the description is then :

$$P_{CCV.}=\{(d_{C1}+d_{c2})\ d_V\ N_C\ E\ F_0\}$$

where C and V are binary labels with:

d_X the duration of the segment X,
N_C the number of segments in the consonantal cluster,
E the relative energy of the vowel (in dB) and
F_0 the relative pitch of the vowel.

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

5.3. Algorithms

Several algorithms have been compared to assess the relevancy of the pseudo-syllable segmentation. Some of the selected approaches are typically applied to speech processing while others are employed in data mining:

- Gaussian Mixture Model (GMM) [17]
- Multilayer Perceptron (MLP) [11]
- Decision Tree (DT) [14]
- Linear Discriminant Analysis (LDA) [8]

With the GMM, the identification is based on a maximum likelihood decision considering the pseudo-syllables as independent observations of the same stochastic process.
For the 3 other algorithms, means, variances and covariances and are computed for each parameter, resulting in 20 parameters per test item. Decision is then taken in this parameter space.

5.4. Language Identification

Experiments are performed on the five languages of the MULTEXT corpus. Since the amount of data is very limited (especially in terms of number of speakers), a cross-validation is performed: nine of the ten speakers are used for the training while the tenth one is classified according to the maximum likelihood criterion. This learning-testing procedure is applied for each speaker of the corpus.

5.4.1. Comparative Results

Table 2 displays the results of the different algorithms in the language identification task. Error rates are computed using the cross-validation approach for a total of 750 tests.

Table 2: Error rates (cross-validation) for each algorithm

Algorithm	Error rate (%)
DT	35 %
MLP	21 %
LDA	20 %
GMM	21 %

Except the Decision Tree that leads to 35 % of errors, all the algorithms reach about 80 % of correct identification. The first conclusion is that the automatic segmentation into rhythmic units is relevant for language identification. Using only this very basic cues (segmentation and durations) without any spectral features is already a relatively efficient way to discriminate among languages. However, the homogeneity of results with different approaches may signify that additional features will be necessary to improve those results.

The matrix of confusion resulting from the GMM modeling (14 Gaussian components per language) is given in Table 3. The best results are reached for Spanish while the worst error rates are for Italian, with about one third of the items confused either with Spanish or English.

Table 3: Matrix of confusion (in percent) with GMM. The average correct identification rate is 79 %.

Model / Item	EN	FR	GE	IT	SP
EN	68	2	12	13	5
FR	2	84	7	-	7
GE	13	-	85	2	-
IT	14	-	4	67	15
SP	3	3	-	3	91

5.4.2. Speaker specific or Language specific rhythm units?

One essential question that may be addressed using the rhythmic segmentation is to evaluate the inter-speaker and

inter-language variability of the rhythmic units. In order to evaluate the speaker-specific part of the variability, additional experiments have been performed using to cross-validation strategies:

In the first strategy (Random Selection), one tenth of the items are randomly selected without taking the speaker identity into account (thus, different sentences pronounced by the same speaker may be in both training and test corpora). The second strategy is based on a speaker selection for the cross-validation, resulting in no overlap at all between the training and the test corpora. Results are displayed in Table 4.

Table 4: Comparison of the error rates in function of the cross-validation strategy.

Algorithm	Random selection	Speaker selection
DT	25 %	35 %
MLP	16 %	21 %
LDA	15 %	20 %

Taking the identity of the speaker into account results in a decrease of the performances from 5 to 10 percent, depending on the algorithm. It means that the rhythmic units convey a significant part of the speaker identity as well as a language specific characterization. Thus, the pseudo-syllable may be considered as being correlated to the language rhythm and to the speaker own tempo.

6. Conclusion and perspectives

This paper develops one of the first approaches dedicated to rhythm identification of languages that is applied to a task more complex than paired language comparisons. The experiments performed with several algorithms on 5 languages produce relatively good results (up to 80% of correct identification rate for 20-second utterances). That shows the relevancy of the rhythmic units extracted from an automatic segmentation. 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), indicating that the temporal structure of the pseudo-syllables is quite language-specific.

Experiments show also that, at this moment, the pseudo-syllable modeling is dependent from the speaker-specific tempo. It means that speaker normalization may be necessary to improve the performances. Another improvement may be based on the use of acoustic correlates of sonority to reach a more accurate syllabic segmentation.

Finally, it is important to note that this segmentation is the first step toward a real rhythm modeling that will be based on the modeling of the sequences of rhythmic units.

7. 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*").

8. References

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

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

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

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

[5] Dominey, P.F; Ramus, F., 2000. 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.

[6] Farinas J.; Pellegrino F., 2001. Automatic Rhythm Modeling for Language Identification. *Proc. of Eurospeech '01*, Aalborg, Scandinavia, September 2001, 2539-2542.

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

[8] Fisher R., 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 179–188.

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

[10] Mehler, J.; Bertoncini, J.; Dupoux, E; Pallier, C, 1996. The role of suprasegmentals in speech perception and acquisition. In *Phonological Structure and Language Processing: Cross-linguistic Studies*, T. Otake and A. Cutler (Eds.). New York: Mouton de Gruyter, 145-169.

[11] Mitchell T., 1997. *Machine learning*. New York: McGraw Hill.

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

[13] Price, P.J, 1980. Sonority and syllabicity: Acoustic correlates of perception. *Phonetica*, 37, 327-343.

[14] Quinlan J., 1993. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann.

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

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

[17] Reynolds D., 1995. Speaker Identification and Verification using Gaussian Mixture Speaker Models. *Speech Communication*, 17, 1-2, 91-108.

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

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

[20] Zissman, M., 1996. Comparison of four approaches to automatic language identification of telephone speech. *Proc. IEEE Trans. On SAP*, 4, 1.