



# Modeling Long and Short-term prosody for language identification

Jean-Luc Rouas

► **To cite this version:**

Jean-Luc Rouas. Modeling Long and Short-term prosody for language identification. International Speech Communication Association (ISCA. 9th European Conf. on Speech Communication and Technology (INTERSPEECH'2005 - EUROSPEECH), Lisboa, Portugal, 04/09/05-08/09/05, 2005, Lisboa, Portugal. 2005. <hal-00664989>

**HAL Id: hal-00664989**

**<https://hal.archives-ouvertes.fr/hal-00664989>**

Submitted on 31 Jan 2012

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Modeling Long and Short-term prosody for language identification

Jean-Luc Rouas

Institut de Recherche en Informatique de Toulouse  
118, Route de Narbonne  
31062 Toulouse Cedex 4  
jean-luc.rouas@irit.fr

## Abstract

This paper addresses the problem of modeling prosody for language identification. The main goal is to validate (or invalidate) some languages characteristics proposed by the linguists by the mean of an automatic language identification (ALI) system. In previous papers, we defined a prosodic unit, the pseudo-syllable. Static modeling has proven the relevance of the pseudo-syllable unit for ALI. In this paper, we try to model the prosody dynamics. This is achieved by the separation of long-term and short-term components of prosody and the proposing of suitable models. Experiments are made on seven languages and the efficiency of the modeling is discussed.

## 1. Introduction

The aim of automatic language identification is to recognize a language spoken by an unknown speaker, within a finite set of languages and for relatively short utterances (usually from 3 to 50 seconds).

In this paper, we investigate the efficiency of prosodic features for language identification, as they are known to carry a substantial part of the language identity (Section 2). However, modeling prosody is still an open problem, mostly because of the suprasegmental nature of the prosodic features. To address this problem, automatic extraction techniques of those features are studied and applied to ~~the modelings~~ proposed by linguists (section 4). Results obtained this way allowed us to create a characterization method of a prosodic unit adapted to language identification (section 5). Modeling prosodic features on this unit is then evaluated for language identification in order to validate the proposed unit. The experiments and results are described in section 6 and 7.

## 2. Motivations

~~To determine if prosodic features can characterize languages, I studied some linguistic theories about rhythmic and intonative properties of languages. Furthermore, I took support from perception experiments where human strategies are evoked.~~

### 2.1. Languages' rhythm

Languages' rhythm has been defined as an effect involving the isochronous (that is to say at regular intervals) recurrence of some type of speech unit [1]. Isochrony is defined as the property of speech to organize itself in pieces equal or equivalent in duration. Depending on the

unit considered, the isochrony theory allows to classify languages in three main sets:

- stress-timed languages,
- syllable-timed languages,
- mora-timed languages<sup>1</sup>.

Syllable-timed languages share the characteristic to have regular intervals between syllables, while stress-timed languages have regular intervals between stressed syllables, and for mora-timed languages, successive mora are quasi equivalent in terms of duration. This point of view has been made popular by Pike [2] and later by Abercrombie [3]. Distinction between stress-timed and syllable-timed languages is strictly categorical, languages cannot be more or less stress or syllable-timed. Despite its popularity among linguists, the rhythm class hypothesis is contradicted by several experiments (notably by Roach [4] and Dauer [5]). This forced some researchers (Beckman [6] for example) to slide from "objective" to "subjective" isochrony. True isochrony is described as a constraint, and the production of isochronous units is perturbed by phonetic, phonologic and grammatical rules of the languages. Some other researchers have concluded that isochrony is mainly a perceptual phenomenon (for example Lehiste [7]). Isochrony can then be seen as a concept relative to speech perception.

### 2.2. Languages' intonation

Three main groups of languages can be characterized regarding ~~to~~ their use of intonation:

- tone languages (as Mandarin Chinese for example),
- tone ~~accent~~ languages (as Japanese),
- other languages.

According to Cummins [8], distinction between languages using fundamental frequency alone had a moderate success. This can be explained in two ways:

- On one hand, we can imagine a discrimination based on the use of lexical tone (Mandarin) or not (English), but intermediate cases exist (Korean dialects) which are usually considered as representing transitory states between languages of one class and those of another.
- On the other hand, phenomenon linked to accents and intonation are less ~~usable~~. There are multiples

<sup>1</sup>a mora is a sub-unit of the syllable constituted by a short vowel and the preceding consonants

theories on utterance intonation that do not agree. The situation is made more complex by studies on the non-linguistic uses of intonation, as for example to express emotions. Several studies agree on a classification by degrees rather than separate classes.

### 2.3. Perceptual experiments

About prosodic features, several perceptual experiments try to shed the light on human abilities to distinguish languages keeping only rhythmic or intonation properties. The point is basically to degrade a speech recording by the meant of filtering or resynthesis to let only few indices to the subjects whom task is to identify the language. The subjects can either be naive or trained adults, infants or newborns, or even primates. For example, all the syllables are replaced by a unique syllable “/sa/” in Ramus’ experiments (see for example [9]). Other authors [10] propose different methods to degrade the speech signal in order to keep only the desired information (intensity, intonation or rhythm). From a general point of view, all those experiments show the notable human capacity to identify foreign languages after a short period of exposure.

All the experiments and the underlying theories show that humans are able to distinguish languages using their prosodic properties. The aim is to study if it is also possible, as humans do and in agreement with linguistic theories, to use prosodic features for automatic language identification.

## 3. Preprocessing

To automatically model the languages’ prosody, we use automatic processings to extract prosodic informations. Three baseline procedures conduct to relevant consonant, vocalic and silence segment boundaries:

- automatic speech segmentation in quasi-stationary segments [11],
- vocal activity detection,
- vowel localization [11].

We then described a syllable-like prosodic pseudo-unit. Syllable is a privileged unit for rhythm modeling. Nevertheless, automatic extraction of syllables (in particular the boundaries detection) is a difficult operation: the pronunciation quality and the speech rate are factors influencing directly the syllable segmentation [12]. Furthermore, to segment the speech signal in syllables is a language-specific task [13], no language-independent algorithm can be easily applied.

For this reason, we introduced the notion of pseudo-syllable [14]. The basic idea is to articulate the prosodic unit around primordial elements of the syllables: vowels, and to gather the neighboring consonants around those nuclei. We have decided to gather only the preceding consonants. This choice is explained in the fact that syllables boundaries detection is not an easy task in a multilingual framework, and that the most frequent syllables correspond to the consonant/vowel structure [5]. An example of this segmentation is shown on figure 1.

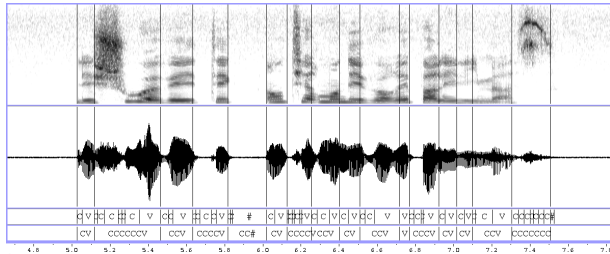



Figure 1: Pseudo-syllable segmentation 

## 4. Static Modeling

In previous papers [14], we have shown that the pseudo-syllable segmentation can be successfully used for language identification. Features characterizing durations and fundamental frequency variations are extracted from each pseudo-syllable and are used to learn the parameters of Gaussian mixtures for each language of the database.

With the rhythmic features, the correct identification rate is 67 % on the seven languages of the MULTTEXT corpus (see [14] for results with less languages). With the intonation features, the correct identification rate is 50 % (see [14] for results with less languages). Characterizing this unit with both rhythmic and intonation features allows to reach 70 % of correct identifications. The confusions observed of mainly across languages belonging to the same groups evoked in linguistic theories.

Nevertheless, the statistic models (Gaussian Mixture Models) we use to model pseudo-syllabic features are intrinsically static models. That doesn’t fit with the perceptive reality of prosody, which is by nature continue. We must use dynamic models to take into account this temporal aspect.

## 5. Dynamic modeling

According to Adami’s work (described in [16]), we used the features computed on each pseudo-syllable to label the fundamental frequency and energy trajectories. Two models are used to separate the long-term and short-term components of prosody. The long-term component characterizes prosodic movements over several pseudo-syllables while the short-term component represents prosodic movements inside a pseudo-syllable.

The processing used for coding long-term and short-term components are the same, the difference is only the units considered, which are in the first case the pseudo-syllables and in the second case the segments.

### 5.1. Fundamental frequency coding

The fundamental frequency processing is divided in two phases, representing the phrase accentuation and the local accentuation, as in Fujisaki’s work [17]:

- the baseline is extracted and labeled, as displayed on figure 2. This is done by finding all the local minimums of the  $F_0$  contour, and linking them. Then, this baseline is labeled in terms of U(p), D(own) and #(silence or unvoiced).
- then, this baseline is subtracted from the original contour, the resulting curve is called residue (fig-

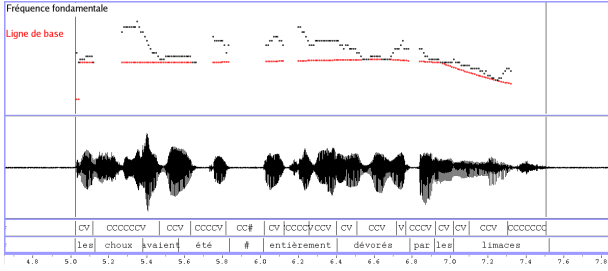


Figure 2: Extraction of the baseline

ure 3). This residue is then approximated on each considered unit (segments or pseudo-syllables) by a linear regression. The slope of the linear regression is used to label the  $F_0$  movement on the unit, according to three available labels (Up, Down and Silence).

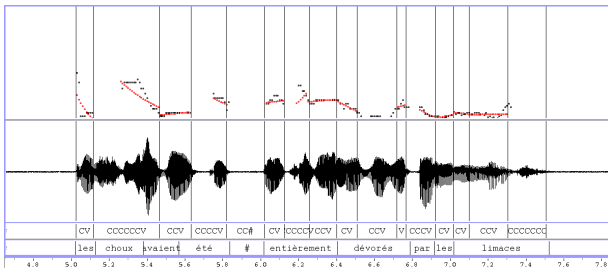


Figure 3: Extraction of the residue

### 5.2. Energy coding

The energy curve is approximated by linear regressions on each considered units (segments or pseudo-syllables) (figure 4). The process is the same as the one used for the residue coding. The labels are also the same, with three possibilities : Up, Down and Silence.

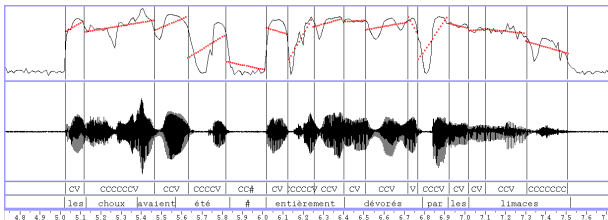


Figure 4: Approximation of the energy

### 5.3. Duration coding

Two duration coding are used regarding the considered unit.

- Considering pseudo-syllable units, 4 labels are used to characterize the contrasts between the vocalic and consonantic durations.
- For the segments units, 2 labels are used (short or long), regarding the nature of the segment (vocalic or consonantic).

## 5.4. Modeling

To model the prosodic variations, we use n-multigram language modeling [18], which can model recurrent patterns in the observation sequence. Unlike classical n-gram modeling, these patterns can have a variable length. The multigram modeling consist in finding the most likely segmentation in an observation sequence. This modeling is applied to the pseudo-syllable labels for the long-term model, and to the segments labels for the short-term model.

## 6. Experiments

All the experiments are made on the MULTEXT corpus [19] (a subset of EUROM1), which contains 5 languages (English, French, German, Italian and Spanish), and 10 speakers per language, balanced between male and female. ~~Kitazawa sent us~~ recordings of Japanese speakers he made using the same protocol than used for the MULTEXT corpus. The Japanese corpus contains 6 speakers, also balanced between male and female (see [20] for more details about the Japanese corpus). Mandarin Chinese recordings are also added to the original corpus, thanks to Komatsu [10].

The three theoretical rhythmic classes are represented in this corpus : English, German and Mandarin Chinese for stress-timed languages; French, Italian and Spanish for the syllable-timed languages, and Japanese for the mora-timed languages.

For the learning phase, we used 8 speakers (4 for Japanese), and 2 (one male and one female) were used for the tests. The test utterances are approximately 20 seconds long.

### 6.1. Long-term modeling

The sequences of labels computed on each pseudo-syllable are modeled by multigram models. The correct identification rate is 40 % on the MULTEXT corpus.

Table 1: Long term prosodic model ( $41,0 \pm 8,2\%$  (57/139))

	Eng	Ger	Man	Fra	Ita	Spa	Jap
Eng	7	-	3	1	4	-	5
Ger	5	12	2	-	1	-	-
Man	3	3	6	-	5	-	3
Fra	1	-	-	12	4	-	2
Ita	5	-	2	4	3	2	2
Spa	6	-	3	1	3	4	3
Jap	3	-	2	2	4	2	7

Those results show that only French and German are clearly identified. This model suffers from the ~~unsuitability of the~~ number of labels ~~and~~ the database size.

### 6.2. Short-term modeling

The sequences of labels computed on each segment are also modeled by multigram models. The identification rate obtained with this method is 63 %.

Those experiments allow us to hypothesize that the most characteristic prosodic elements of languages aren't

Table 2: Short term prosodic model ( $63,3 \pm 8,0$  % (88/139))

	Eng	Ger	Man	Fra	Ita	Spa	Jap
Eng	6	-	5	1	5	2	1
Ger	1	18	1	-	-	-	-
Man	4	2	12	-	2	-	-
Fra	-	-	-	16	1	2	-
Ita	1	-	1	2	13	2	1
Spa	2	-	-	1	7	9	1
Jap	-	-	1	-	5	-	14

pseudo-syllable sequences but sequences of elements constituting them.

### 6.3. Merging long and short-term components

The merging is addressed between the two systems described here-above. The merging technique is a weighted addition of the log-likelihoods. The identification rate obtained with this method is 71 %.

Table 3: Merging of short and long-term models ( $71,2 \pm 7,5$  % (99/139))

	Eng	Ger	Man	Fra	Ita	Spa	Jap
Eng	12	-	4	2	1	-	1
Ger	2	17	1	-	-	-	-
Man	1	1	18	-	-	-	-
Fra	1	-	-	14	4	-	-
Ita	-	1	1	2	10	3	3
Spa	1	-	1	1	6	9	2
Jap	-	-	-	-	1	-	19

Results show that most languages are well identified. Japanese is the only mora-timed language, and the only tone accent language in our corpus, therefore it seems natural that it is the most well identified language. Considering rhythmic classes (represented in different strength of grey in the matrix), we can see that most confusion are across languages of the same rhythmic family. The rhythmic classes identification rate is 89%

## 7. Conclusions and perspectives

These experiments shows that it is possible to identify languages using prosody alone. The dynamic modeling allows to reach 71% of correct identification on a seven language discrimination task. Results tend to confirm the existence of the rhythmic classes of the isochrony theory, as confusions are mainly across languages belonging to the same family.

This method gives promising results, but further experiments have to be made, with different kinds of data (spontaneous speech for example) and we need to test our system on many more languages to confirm the linguistic classes hypothesis.

## 8. References

[1] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in*

*Laboratory Phonology 7*, 2002.

- [2] P. Ladefoged, Ed., *The intonation of American English*. Michigan, USA: University of Michigan Press, 1945.
- [3] D. Abercrombie, Ed., *Elements of General Phonetics*. Edinburgh: Edinburgh University Press, 1967.
- [4] P. Roach, "On the distinction between "stress-timed" and "syllable-timed" languages," *Linguistic Controversies*, pp. 73–79, 1982.
- [5] R. M. Dauer, "Stress-timing and syllable-timing re-analysed," *Journal of Phonetics*, vol. 11, pp. 51–62, 1983.
- [6] M. E. Beckman, "Evidence for speech rhythms across languages," in *Speech perception, production and linguistic structure*, Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka, Eds., 1992, pp. 457–463.
- [7] I. Lehiste, "Isochrony reconsidered," *Journal of Phonetics*, vol. 5, pp. 253–263, 1977.
- [8] F. Cummins, "Speech rhythm and rhythmic taxonomy," in *Speech Prosody*, Aix-en-Provence, France, 2002, pp. 121–126.
- [9] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, no. 3, pp. 265–292, 1999.
- [10] M. Komatsu, T. Arai, and T. Sugawara, "Perceptual discrimination of prosodic types," in *Speech Prosody*, Nara, Japan, 2004, pp. 725–728.
- [11] F. Pellegrino and R. André-Obrecht, "Vocalic system modeling : A vq approach," in *IEEE Digital Signal Processing*, Santorini, July 1997.
- [12] I. Kopecek, "Syllable based approach to automatic prosody detection; applications for dialogue systems," in *Proceedings of the Workshop on Dialogue and Prosody*, Eindhoven, Pays-Bas, Sept. 1999.
- [13] H. R. Pfitzinger, S. Burger, and S. Heid, "Syllable detection in read and spontaneous speech," in *ICSLP*, vol. 2, Philadelphia, October 1996, pp. 1261–1264.
- [14] J.-L. Rouas, J. Farinas, and F. Pellegrino, "Automatic Modelling of Rhythm and Intonation for Language Identification," in *15th ICPHS*. Barcelona, Spain: Causal Productions Pty Ltd, 3-9 août 2003, pp. 567–570.
- [15] A. Adami, R. Mihaescu, D. A. Reynolds, and J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *ICASSP*, vol. 4, Hong Kong, China, 2003, pp. 788–791.
- [16] H. Fujisaki, "Prosody, information and modeling - with emphasis on tonal features of speech," in *Workshop on Spoken Language Processing*, Mumbai, India, January 2003.
- [17] S. Deligne and F. Bimbot, "Language modeling by variable length sequences: theoretical formulation and evaluation of multigrams," in *ICASSP*, 1995.
- [18] E. Campione and J. Véronis, "A multilingual prosodic database," in *ICSLP*, Sidney, 1998, <http://www.lpl.univ-aix.fr/projects/multext>.
- [19] S. Kitazawa, "Periodicity of japanese accent in continuous speech," in *Speech Prosody*, Aix en Provence, France, April 2002.