Automatic Modelling of Rhythm and Intonation for Language Identification

Jean-Luc Rouas[1] and **Jérôme Farinas**[1] and **François Pellegrino**[2]

[1] Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS INP UPS, France
jean-luc.rouas@irit.fr, jerome.farinas@irit.fr

[2] Laboratoire Dynamique Du Langage, UMR 5596 CNRS Univ. Lyon 2, France
francois.pellegrino@univ-lyon2.fr

Abstract

This paper deals with an approach to Automatic Language Identification using only prosodic modeling. The traditional approach for language identification focuses mainly on phonotactics because it gives the best results. Recent studies reveal that humans use different levels of perception to identify a language, in particular prosodic cues.

Among prosodic features, rhythm is known to carry a substantial information about language identity. Rhythm is produced by the periodicity of a pattern that can be syllable, which is a language specific unit. That is why we introduced the notion of "Pseudo-Syllable", derived from the most frequent syllable structure in the world, the Consonant-Vowel structure. In this paper, an automatic and language independent rhythmic units extraction algorithm is described: using a vowel detection algorithm, rhythmic units matching the Consonant-Vowel structure are segmented. Two models describing rhythm and intonation of each language are then learned using Gaussian Mixtures.

1 Introduction

Nowadays, the standard approach to automatic language identification considers a phonetic modeling as a front-end. The resulting sequences of phonetic units are then decoded according to language specific grammars [1]. This approach gives the best results but only marginal improvements have been achieved since 1996, and it seems crucial not to underestimate the relevance of alternative features also present in the signal to overcome the current limitations.

Recent studies (see [2] for a review) reveal that humans use different levels of perception to identify a language. Three major kinds of features are employed: segmental features (acoustic properties of phonemes), suprasegmental features (phonotactics and prosody) and high level features (lexicon). Beside phonetics and phonotactics, prosody is one of the most promising features to be considered for language identification, even if its extraction and modeling are not a straightforward issue. Actually, one of the main problems to address is what to model (Section 2).

Rhythm is known to carry a substantial information about language identity, and is useful for language identification by humans. Our assumption is that rhythm is produced by the periodicity of a pattern that can be syllable, which is a language specific unit. That is why we introduced the notion of "Pseudo-Syllable", derived from the most frequent syllable structure in the world, the Consonant-Vowel structure (Section 3).

In this paper, an automatic and language independent rhythm extraction algorithm is described: using a vowel detection algorithm , rhythmic units matching the Consonant-Vowel structure are segmented. Several parameters are extracted including consonantal and vowel durations, and cluster complexity. Other features related to pitch and intensity are also considered to model the languages tones. Two models describing rhythm and intonation of each language are then learned using Gaussian Mixtures (Section 4). Results are presented in section 5.

2 Motivations

2.1 About rhythm

Languages can be gathered in main rhythmic classes. According to the literature, Spanish is *syllable-timed* whereas English and German are *stress-timed*, and Japanese is *mora-timed*. These categories emerged from the theory of isochrony introduced by Pike and developed by Abercrombie [3].

However, 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 categories are replaced by a continuum [4]. Rhythmic differences be-

tween languages are then mostly related to their syllable structure and the presence (or absence) of vowel reduction. The 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, reaching a relevant representation seems to be difficult. Another difficulty rises from the selection of an efficient modeling paradigm.

2.2 About intonation

Intonation can also be seen as an efficient cue for discriminating among languages. There is a linguistic grouping between languages using tone as a lexical marker and those that do not. Some evidence have shown that intonation contours can be a part of a language's identity, even if there are universals [5] and speaker specific processes involved. Other experiments have shown that intonation patterns can help to discriminate among languages [6] and among dialects from the same language [7] and [8].

The approach we develop here was first introduced in [9] and improved by considering fundamental frequency features in [10]. Now, we validate this approach by widening the corpus with the Japanese language which will help us to have a closer look to linguistic categories.

3 System overview

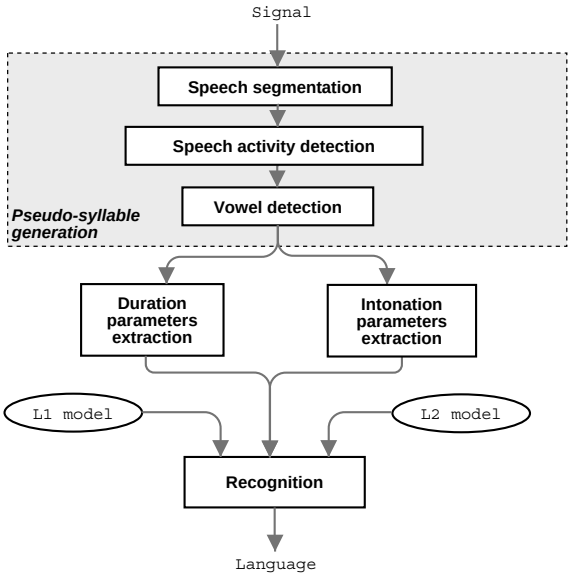

Figure 1: Overview of the system.

Syllable may be a first-rate candidate for rhythm modeling. Unfortunately, segmenting speech in syllables is typically a language-specific mechanism and thus no language independent algorithm can be derived. For this reason, we introduced in [9] the notion of pseudo-syllables derived from the most frequent syllable structure in the world, namely the CV structure [11].

A pseudo-syllable is a language independent unit that is near the definition of the syllable and that can be automatically extracted. A synoptic of the system is displayed on figure 1.

3.1 Pre-processing

The pseudo-syllable generation requires the following pre-processing steps:

- A language-independent speech segmentation algorithm [12] of the signal. This algorithm is based on the modeling of the speech signal with an autoregressive model. The changes in the coefficients of the autoregressive model are detected according to a distance measurement. The resulting short and long segments corresponds to transient and steady parts of the signal.
- A language-independent vowel detection algorithm (based on the Energy) [13].
- A speech activity detection algorithm that produces Silence, Non Vowel or Vowel labels on each of the detected segments. This algorithm, based on a spectral analysis of the signal, is described in [14]. It is applied in a language and speaker independent way without any manual adaptation phase.

3.2 Pseudo Syllable Extraction

A pseudo-syllable is articulated around the vocalic segment and consists in a C^nV pattern: n is an integer (that may be zero) and V may result from the merging of consecutive vowel segments. See an example of extraction in figure 2. To improve the robustness, we decided to discard any segment that last over 150 ms. Furthermore, pseudo-syllables containing only consonantal segments are discarded (as the last pseudo-syllable in the example).

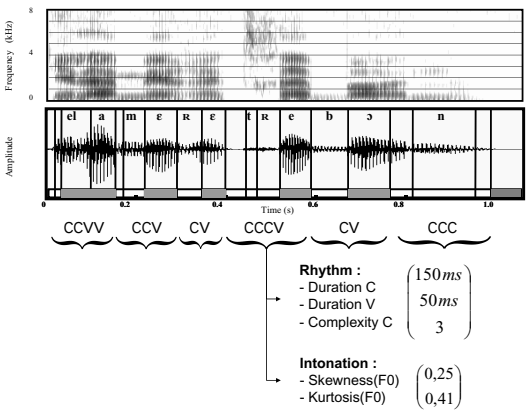

Figure 2: Extraction of prosodic features after the pseudo-syllable segmentation.

4 Features

Rhythmic and fundamental frequency statistics are extracted from each pseudo-syllable (Figure 2).

4.1 Rhythmic parameters

Three parameters are computed, corresponding respectively to the total consonant cluster duration, the total vowel duration and the complexity of the consonantal cluster. For example, the description for a .CCV. pseudo-sequence is:

$$\pi_{.CCV.} = \{D_C \; D_V \; N_C\} \tag{1}$$

where D_C is the total duration of the consonantal segments, D_V is the duration of the vowel segment and N_C is the number of segments in the consonantal cluster (here, $N_C = 2$). Such a basic rhythmic parsing is obviously limited, but provides a framework to model rhythm that requires no knowledge on the language rhythmic structure

4.2 Fundamental frequency parameters

The fundamental frequency outlines are used to compute statistics within the same pseudo-syllable frontiers than those used for rhythm modeling, in order to model intonation on each pseudo-syllable. The parameters used to characterize each pseudo syllable intonation are : a measurement of the accent location (maximum f0 location regarding to vocalic onset, noted α) and the normalized fundamental frequency bandwidth on each syllable (ΔF_0). The resulting feature vector is then, for each Pseudo-Syllable :

$$\pi_{F_0} = \{\alpha_{F_0} \; \Delta F_0\} \tag{2}$$

4.3 Modeling

Each pseudo-syllable is then characterized by a two vectors, one characterizing rhythmic units and the other characterizing intonation on each of these rhythmic units :

$$\pi_{.CVV.} = \{D_C \; D_V \; N_C\}, \pi_{F_0} = \{\alpha_{F_0} \; \Delta F_0\} \tag{3}$$

This vector is computed for each pseudo syllable of each sentence of the learning part of the corpus. For each language a two Gaussian Mixture Models are learned to characterize the language specific $\pi_{.CCV.}$ and π_{F_0} distributions, using the EM algorithm with LBG initialization [15].

4.4 Decision

Let be $L = \{L_1, ..., L_i, ..., L_{N_L}\}$ the set of language to identify. The problem is to find the most likely language L^* in the set L. Let be $S_\pi = \{\pi_1, ..., \pi_k, ..., \pi_{n_p}\}$ the sequence of prosodic informations extracted from each pseudo-syllable and π_k is the vector formed by the prosodic parameters for the pseudo-syllable k (equation (3)).

The probability that the observation π_k belongs to the language L_i is defined by the probability density function of the Gaussian mixture:

$$Pr(\pi_k|L_i) = \sum_{j=1}^{Q_i} \frac{a_j^i}{(2\pi)^{3/2}\sqrt{|\Sigma_j^i|}} \; e^{-\frac{1}{2}(\pi_k - \mu_j^i)^t \Sigma_j^{i\,-1}(\pi_k - \mu_j^i)} \tag{4}$$

where Q_i is the number of Gaussian mixtures and (μ_j^i, Σ_j^i) are the parameters of the Gaussian mixture j of the language L_i.

Assuming that each pseudo-syllable is independent, the probability that the sequence S_π belongs to the language L_i is:

$$Pr(S_\pi|L_i) = \prod_{k=1}^{n_p} Pr(\pi_k|L_i) \tag{5}$$

Using Bayes rule and considering that *a priori* probabilities are equal, the most likely language L^* is defined by the following equation:

$$L^* = arg \max_{1 \le i \le N_L} (Pr(L_i|S_\pi)) = arg \max_{1 \le i \le N_L} (Pr(S_\pi|L_i)) \tag{6}$$

5 Experiments

Experiments are made on the MULTEXT corpus [16] (a subset of EUROM1), which contains 5 languages (English, French, German, Italian and Spanish), and 10 speakers per language, balanced between male and female. Mr. Kitazawa recently 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 [17] for more details about the Japanese corpus).

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.

5.1 Rhythm Modeling

The matrix of confusion shows the relevance of our approach as rhythmic classes can be identified, like stress-timed languages (English and German) which are slightly confused. The simple rhythmic structure of Japanese which is mainly composed of CV syllables allows our system to easily discriminate this language among others. Rhythmic similarities can also be guessed between French, Italian and Spanish, which are said to be syllable-timed.

	Eng	Fre	Ger	Ita	Jap	Spa
Eng	**15**	3	1	8	-	3
Fre	-	**15**	-	-	-	4
Ger	5	-	**33**	2	-	-
Ita	-	3	1	**17**	-	9
Jap	-	0	-	-	**80**	-
Spa	-	4	-	4	-	**22**

Table 1: Matrix of confusion given by the rhythm model (Correct identification rate : 79.4 % (182/229)).

5.2 Fundamental frequency modeling

As shown in the matrix confusion below, most languages are well identified, except French and Italian. French and Spanish are confused with almost every language present in the database.

	Eng	Fre	Ger	Ita	Jap	Spa
Eng	**25**	1	4	-	-	-
Fre	3	**4**	6	3	-	3
Ger	6	2	**20**	4	1	7
Ita	2	-	8	**5**	1	14
Jap	2	-	-	1	**77**	0
Spa	3	1	6	4	2	**12**

Table 2: Matrix of confusion given by the fundamental frequency model (Correct identification rate : 62.4 % (143/229).

6 Conclusion

Experiments show that modeling rhythm and intonation is useful for language identification.

The rhythmic modeling manages to catch the rhythmic information and achieves a clear separation between languages which don't belong to the same hypothetic rhythmic class. But our rhythmic model don't catch the languages' rhythm but characterizes language-specific rhythmic units, so we need to model the sequences of these units to get a complete rhythm model.

The fundamental frequency model better characterizes languages which have well defined intonational rules like Japanese and English.

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 (or not) the linguistic classes hypothesis.

References

[1] M. A. Zissman and K. M. Berkling, "Automatic language identification," *Speech Communication*, vol. 35, no. 1-2, pp. 115–124, 2001.

[2] M. Barkat-Defradas, I. Vasilescu, and F. Pellegrino, "Stratégies perceptuelles et identification automatique des langues," *Traitement Automatique des Langues*, submitted.

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

[4] R. M. Dauer, "Stress-timing and syllable-timing reanalysed," *Journal of Phonetics*, vol. 11, pp. 51–62, 1983.

[5] D. R. Ladd, *The cognitive representation of speech*, ch. On intonational universals, pp. 389–397. 1981.

[6] E. Grabe, *Intonational Phonology: English and German*. PhD thesis, Max-Planck-Institut for Psycholinguistics and University of Nijmegen, 1998.

[7] R. Todd, "Speaker-ethnicity: attributions based on the use of prosodic cues," in *Proceedings of Speech Prosody 2002*, (Aix en Provence, France), April 2002.

[8] E. Grabe, B. Post, F. Nolan, and K. Farrar, "Pitch accent realisation in four varieties of british english," *Journal of Phonetics*, vol. 28, pp. 161–185, 2000.

[9] J. Farinas and F. Pellegrino, "Automatic rhythm modeling for language identification," in *Proc. of Eurospeech Scandinavia'01*, (Aalborg, Denmark), 2001.

[10] J. L. Rouas, J. Farinas, and F. Pellegrino, "Merging segmental, rhythmic and fundamental frequency features for language identification," in *Proc. of Eusipco'02*, (Toulouse, France), 2002.

[11] N. Vallée, L.-J. Boë, I. Maddieson, and I. Rousset, "Des lexiques aux syllabes des langues du monde : typologies et structures," in *XXIIIèmes Journées d'Etude sur la Parole*, (Aussois, France), pp. 93–96, June 2000.

[12] R. André-Obrecht, "A new statistical approach for automatic speech segmentation," *IEEE Trans. on ASSP*, vol. 36, no. 1, pp. 29–40, 1988.

[13] F. Pellegrino and R. André-Obrecht, "Automatic language identification: An alternative approach to phonetic modeling," *Signal Processing*, vol. 80, no. 7, pp. 1231–1244, 2000.

[14] F. Pellegrino and R. André-Obrecht, "An unsupervised approach to language identification," in *Proc. of ICASSP'99*, (Phoenix, Arizona), 1999.

[15] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transaction on Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[16] E. Campione and J. Véronis, "A multilingual prosodic database," in *Proceedings of ICSLP'98*, (Sidney), 1998. http://www.lpl.univ-aix.fr/projects/multext.

[17] S. Kitazawa, "Periodicity of japanese accent in continuous speech," in *Proceedings of Speech Prosody 2002*, (Aix en Provence, France), April 2002.