MODELING PROSODY FOR LANGUAGE IDENTIFICATION ON READ AND SPONTANEOUS SPEECH

Jean-Luc Rouas[1], *Jérôme Farinas*[1], *François Pellegrino*[2] *and Régine André-Obrecht*[1]

[1]Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS INP UPS, France
[2]Laboratoire Dynamique Du Langage, UMR 5596 CNRS Univ. Lyon 2, France

ABSTRACT

This paper deals with an approach to Automatic Language Identification using only prosodic modeling. The actual approach for language identification focuses mainly on phonotactics because it gives the best results. We propose here to evaluate the relevance of prosodic information for language identification with read studio recording (previous experiment [1]) and spontaneous telephone speech. For read speech, experiments were performed on the five languages of the MULTEXT database [2]. On the MULTEXT corpus, our prosodic system achieved an identification rate of 79 % on the five languages discrimination task. For spontaneous speech, experiments are made on the ten languages of the OGI Multilingual telephone speech corpus [3]. On the OGI MLTS corpus, the results are given for languages pair discrimination tasks, and are compared with results from [4]. As a conclusion, if our prosodic system achieves good performance on read speech, it might not take into account the complexity of spontaneous speech prosody.

1. INTRODUCTION

During the last decade, the request for Automatic Language Identification (ALI) systems arose in several fields of application, and especially in Computer-Assisted Communication (Emergency Service, etc.) and Multilingual Man-Computer Interfaces (Interactive Information Terminal, etc.). More recently, content-based indexing of multimedia or audio data provided a new topic in which ALI systems are useful. However, current ALI systems are still not efficient enough to be used in a commercial framework. 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, an algorithm of language-independent extraction of rhythmic features is proposed and applied to model rhythm (Section 3). Meanwhile, an other algorithm, based on the automatically extracted fundamental frequency contours, computes statistics on these outlines in order to model each language's intonation (Section 3). The experiments and results are described in section 4.

2. MOTIVATIONS

2.1. Classifying languages according to rhythm

Languages can be clustered 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 [5]. However, more recent works provide an alternative framework in which these categories are replaced by a continuum [6]. Rhythmic differences between 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. We develop here a statistical approach, first introduced in [7] and now improved by considering stress features (Fundamental Frequency and Energy). This approach is based on a Gaussian modeling of the different *rhythm units* automatically extracted from a rhythmic segmentation in the languages.

2.2. Classifying languages according to intonation

Intonation can also be seen as an efficient cues for discriminating among languages. There is a linguistic grouping between languages using tone as a lexical marker and those that do not. For example, in Mandarin Chinese or Vietnamese, the use of changes in the tones assigned to syllables distinguish between lexical items. In English, stress is used at the sentence level and is used to determine the kind of the sentence, whether its declarative, interrogative, etc.

3. DESCRIPTION OF THE SYSTEM

The language identification system is based on the segmentation of the speech signal in “pseudo-syllable”. A pseudo-syllable is a language independent unit that is near the definition of the syllable and that can be automatically extracted. The extraction process of this unit is describe in 3.1, prosodic features extraction derived from this pseudo-syllable in 3.2 and the language identification task in 3.3. A synoptic of the system is displayed on figure 1.

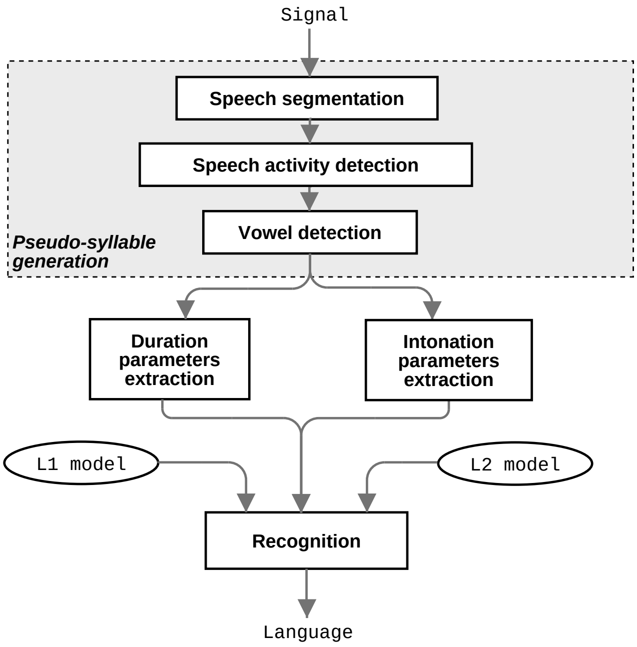
Signal
Speech segmentation
Speech activity detection
Vowel detection
Pseudo-syllable generation
Duration parameters extraction
Intonation parameters extraction
L1 model
L2 model
Recognition
Language

Figure 1. Overview of the system.

3.1. Pseudo-syllable unit

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 [7] the notion of pseudo-syllables derived from the most frequent syllable structure in the world, namely the CV structure [8].

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

- A language-independent speech segmentation algorithm [9] 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 results are short and long segments corresponding to transient and steady parts of the signal.
- A language-independent vowel detection algorithm (based on the Energy) [10]
- A speech activity detection algorithm that produces Silence, Non Vowel or Vowel labels on the detected segments. This algorithm, based on a spectral analysis of the signal, is described in [11]. It is applied in a language and speaker independent way without any manual adaptation phase.

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.

3.2. Features extraction

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

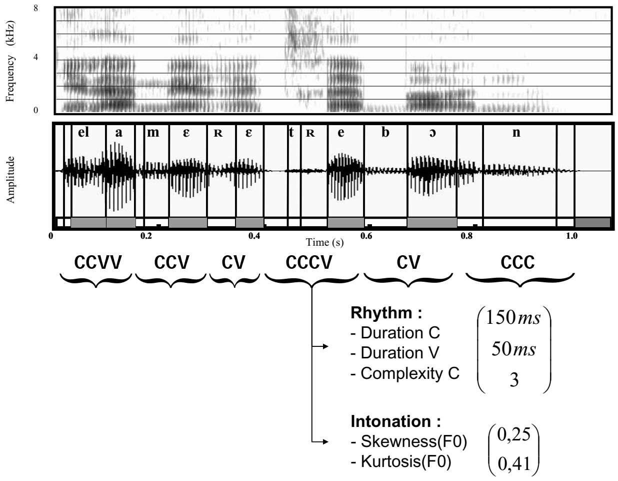

Figure 2. Extraction of prosodic features after the Pseudo-Syllable segmentation.

3.2.1. 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:

$$P_{.CCV.} = \{D_C \; D_V \; N_C\}$$

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

3.2.2. Fundamental frequency parameters

The fundamental frequency outlines are used to compute statistics inside of the same pseudo-syllable frontiers than those used for rhythm modeling, in order to model intonation on each pseudo-syllable. We choose to compute statistics until 4th order (mean, standard deviation, skewness and kurtosis) and a measurement of the accent location (maximum f0 location regarding to vocalic onset) and the normalized fundamental frequency bandwidth on each syllable.

3.3. Language identification system

Let be $L = \{L_1, L_2\}$ the set of language to identify. The problem is to find the most likely language L^* in the set L. Let be $O_\pi = \{\pi_1, \pi_2, ..., \pi_{n_p}\}$ the sequence of prosodic informations extracted from each pseudo-syllable.In a Bayesian approach, L^* is defined by the following equation:

$$L^* = arg \max_{1\leq i\leq 2} (Pr(L_i|O_\pi)) = arg \max_{1\leq i\leq N_L} (Pr(O_\pi|L_i)) \tag{1}$$

using Bayes rule and considering that *a priori* probabilities are equal, and $Pr(O_\pi|L_i)$ is obtained with the prosodic modeling.

The prosodic modeling uses Gaussian Mixture Models (GMM) on a set of 9 parameters extracted from each pseudo-syllable: Dc, Dv, Nc, F0 mean, F0 variance, F0 skewness, F0 kurtosis, the accent location, the F0 bandwidth. Considering that each pseudo-syllable is independent gives:

$$Pr(O_\pi|L_i) = \prod_{k=1}^{n_p} Pr(\pi_k|L_i) \tag{2}$$

and π_k is the vector formed by the 9 prosodic parameters for the pseudo-syllable k.

For each language a GMM is learned to characterize the π_k vector distribution, using EM algorithm with LBG initialization [12].

$$Pr(\pi_k|L_i) = \sum_{j=1}^{Q_i} N(\mu_j, \Sigma_j) \tag{3}$$

4. EXPERIMENTS

4.1. Language identification on read speech

Experiments were previously [1] made on the five languages of the MULTEXT database [2]: English, French, German, Italian and Spanish. The tests are made using 20 seconds read speech utterances. The identification rate is 79 % (Table 1).

	Eng	Fre	Ger	Ita	Spa
Eng	**62**	4	16	11	7
Fre	0	**100**	0	0	0
Ger	11	1	**86**	2	0
Ita	10	1	3	**62**	23
Spa	1	4	0	3	**91**

Table 1. Results for the language identification task on the MULTEXT corpus.

4.2. Language identification on spontaneous speech

Experiments are made on ten languages of the OGI Multilingual Telephone Speech Corpus (OGI MLTS) [3]: English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. The tests are made using the 45 seconds spontaneous speech utterances, and the results are displayed on Table 1.

5. DISCUSSION

On the read speech corpus, our system can achieve good performance (79 % of correct identification on five languages). The main confusion are between English and German (both stress timed languages), and Spanish and Italian.

On the spontaneous speech corpus, the discrimination is easier to achieve between languages which does not belong to the same rhythmic and intonation classes:

- English and German, stress-timed languages which does not use intonation as a lexical marker are well identified regarding to Japanese (mora-timed language), Mandarin and Vietnamese (which use intonation as a lexical marker), and Korean, Tamil, Farsi. But to discriminate English and German is not an easy task.
- In the same way, we can see that Mandarin (which uses intonation as a lexical marker) is discriminated from most of the languages, except Japanese and Vietnamese.

	English	German	French	Spanish	Mandarin	Vietnamese	Japanese	Korean	Tamil	Farsi
English	-	59.5	51.5	67.7	**75.0**	67.7	67.6	**79.4**	**77.4**	**76.3**
German	59.5	-	55.9	59.4	62.2	65.7	65.8	**71.4**	69.7	**71.8**
French	51.5	55.9	-	64.3	60.6	58.1	55.9	54.8	60.1	68.6
Spanish	67.7	59.4	64.3	-	**80.6**	62.1	62.5	**75.9**	65.4	66.7
Mandarin	**75.0**	62.2	60.6	**80.6**	-	50.0	50.0	**73.5**	**74.2**	**76.3**
Vietnamese	67.7	65.7	58.1	62.1	50.0	-	68.6	56.2	**71.4**	66.7
Japanese	67.6	65.8	55.9	62.5	54.1	68.6	-	65.7	59.4	66.7
Korean	**79.4**	**71.4**	54.8	**75.9**	**73.5**	56.2	65.7	-	62.1	**75.0**
Tamil	**77.4**	69.7	60.1	65.4	**74.2**	**71.4**	59.4	62.1	-	69.7
Farsi	**76.3**	**71.8**	68.6	66.7	**76.3**	66.7	66.7	**75.0**	69.7	-

Table 2. Results for the language pair identification task on ten languages.

As a comparison, we can cite [4] in which is described an language pair identification task on the same corpus, but using different features: the amplitude envelope modulation and a feature related to the fundamental frequency. The results we obtain are mainly above those obtained by Cummins et al. We can assess that our features take into account more information, especially when we have to discriminate stress-timed languages (as English and German) versus mora-timed languages which uses intonation as a lexical marker (like Japanese). But our discrimination performs worse when the task is to identify languages belonging to the same intonation family.

As a conclusion, our system can perform good performance on a read speech corpus, but we will have to develop more accurate tools in order to model the spontaneous speech prosody which seems to be too complex and with too much speaker variability for our features.

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] 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.

[2] E. Campione and J. Véronis, "A multilingual prosodic database," in *Proc. of ICSLP'98*, Sydney, Australia, Nov. 1998, pp. 3163–3166.

[3] Y.K. Muthusamy, R.A. Cole, and B.T. Oshika, "The ogi multilanguage telephone speech corpus," in *Proc. of ICSLP'92*, Alberta, October 1992.

[4] F. Cummins, F. Gers, and J. Schmidhuber, "Comparing prosody across many languages," Technical report idsia-07-99, Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Lugano, CH, 1999.

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

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

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

[8] 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, June 2000, pp. 93–96.

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

[10] 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.

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

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