Automatic prosodic variations modelling for language and dialect discrimination

Jean-Luc Rouas

***Abstract*— This paper addresses the problem of modelling prosody for language identification. The aim is to create a system that can be used prior to any linguistic work to show if prosodic differences among languages or dialects can be automatically determined. In previous papers, we defined a prosodic unit, the pseudo-syllable. Rhythmic modelling has proven the relevance of the pseudo-syllable unit for automatic language identification. In this paper, we propose to model the prosodic variations, that is to say model sequences of prosodic units. This is achieved by the separation of phrase and accentual components of intonation. We propose an independent coding of those components on differentiated scales of duration. Short-term and long-term language-dependent sequences of labels are modelled by n-gram models. The performance of the system is demonstrated by experiments on read speech and evaluated by experiments on spontaneous speech. Finally, an experiment is described on the discrimination of Arabic dialects, for which there is a lack of linguistic studies, notably on prosodic comparisons. We show that our system is able to clearly identify the dialectal areas, leading to the hypothesis that those dialects have prosodic differences.**

***Index Terms*— Automatic language identification, prosody, read and spontaneous speech**

EDICS Category: SPE-MULT

I. Introduction

THE standard approach to automatic language identification (ALI) considers a phonetic modelling as a front-end. The resulting sequences of phonetic units are then decoded according to language specific phonotactic grammars [1].

Other information sources can be useful to identify a language however. 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 acoustics, phonetics and phonotactics, prosody is one of the most promising features to be considered for language identification, even if its extraction and modelling are not a straightforward issue.

In the NIST 2003 Language Verification campaign, most systems used acoustic modelling, using Gaussian Mixture Models adapted from Universal Background Models (a technique derived from speaker verification [3]), and/or phonotactic modelling (Parallel Phone Recognition followed by Language Modelling - PPRLM, see [1], [4]). While these techniques gave the best results, systems using prosodic cues have also been investigated, following research in speaker recognition [5], notably Adami's system [6]. More recently, systems using syllable-scale features have been under research, although their aim is to model acoustic/phonotactic properties of languages [7] or also prosodic cues [8].

Beside the use of prosody to improve the performances of ALI systems, we believe that there is a real linguistic interest in developing an automatic language identification system using prosody and not requiring any *a priori* knowledge (e.g. manual annotations). Hence, we will have the possibility of testing if prosodically unstudied languages can be automatically differentiated. The final aim of our studies is to automatically describe prosodic language typologies.

In this paper, we will describe our prosodic-based language identification system. In section II, we will recall the main linguistic theories about differences among languages. After reviewing the linguistic and perceptual elements that demonstrate the interest of prosody modelling for language identification, we will address the problem of modelling.

Indeed, modelling prosody is still an open problem, mostly because of the suprasegmental nature of the prosodic features. To address this problem, automatic extraction techniques of sub-phonemic segments are used (section III). After an activity detection and a vowel localisation, a prosodic syllable-like unit adapted to language identification is characterised. Results previously obtained by modelling prosodic features extracted on this unit are briefly described at the end of section III. We believe however that prosodic models should take into account temporal fluctuations, as prosody perception is mostly linked to variations (in duration, energy and pitch). That is why we propose a prosody coding which enables to consider the sequences of prosodic events in the same way as language models are used to model phone sequences in the PPRLM approach. The originality of our method lies in differentiating phrase accent from local accent, and modelling them separately. The method is described in section IV. The system is firstly tested on databases with languages that have known prosodic differences to assess the accuracy of the modelling. These experiments are carried out on both read (section V) and spontaneous (section VI) speech, using respectively the MULTEXT [9] and OGI-MLTS [10] corpora. Then, a final experiment is performed on Arabic dialects (section VII), for which we investigate if prosodic differences between those dialects can be automatically detected.

II. Prosodic differences among languages

The system described in this paper aims at determining to what extent languages are prosodically different. In this section, we will describe what the rhythmic and intonational properties of languages are and how humans perceive them.

The author is with L2F - INESC-ID, Rua Alves Redol, 9, 1000-029 Lisboa (email: rouas@l2f.inesc-id.pt, phone: +351213100314, fax: +351213145843).

I would like to thank François Pellegrino for providing helpful comments on this article and Mélissa Barkat for her help on the Arabic dialects experiments.

A. Rhythm

The rhythm of languages has been defined as an effect involving the isochronous (that is to say at regular intervals) recurrence of some type of speech unit [11]. Isochrony is defined as the property of speech to organise itself in pieces equal or equivalent in duration. Depending on the unit considered, the isochrony theory classifies languages in three main sets:

- stress-timed languages (as English and German),
- syllable-timed languages (as French and Spanish),
- mora-timed languages (as Japanese).

Syllable-timed languages share the characteristic of having 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 was made popular by Pike [12] and later by Abercrombie [13]. According to them, distinction between stress-timed and syllable-timed languages is strictly categorical, since 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 [14] and Dauer [15]). This forced some researchers (Beckman [16] for example) to shift 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 [17]). Isochrony can then be seen as a concept relative to speech perception.

The lack of an empirical proof of isochrony led Dauer [15] to propose a new rhythmic classification system. From her point of view, speakers do not try to keep equal inter-stress or inter-syllabic intervals, but languages are more or less stress or syllable-timed. Nespor [18] introduced the notion of rhythmically intermediate languages, which share properties associated with stress-timed languages and other associated with syllable-timed languages. As an example, she cites Polish – classed as stress-timed although it does not have vocalic reduction – and Catalan – syllable-timed but having vocalic reduction.

B. Intonation

Three main groups of languages can be characterised regarding their use of intonation:

- tone languages (as Mandarin Chinese),
- pitch-accent languages (as Japanese),
- stress-accent languages (as English and German).

According to Cummins [19], distinguishing languages using fundamental frequency alone had a moderate success. The explanation is twofold:

- On the 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 [20].
- On the other hand, phenomena linked to accents and intonation are less easy to handle. There are multiple 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 [21].

C. Perception

Over the last few decades, numerous experiments have shown the human capability for language identification [2]. Three major kinds of cues help humans to identify languages:

1) Segmental cues (acoustic properties of phonemes and their frequency of occurrence),
2) Supra-segmental cues (phonotactics, prosody),
3) High-level features (lexicon, morpho-syntax).

About prosodic features, several perceptual experiments try to shed light on human abilities to distinguish languages keeping only rhythmic or intonation properties. The method is to degrade speech recordings by filtering or re-synthesis to remove all segmental cues to the subjects whose task is to identify the language. The subjects are either naive or trained adults, infants or newborns, or even non-human primates. For example, all the syllables are replaced by a unique syllable "/sa/" in Ramus' experiments [22]. In other cases, processing of speech through a low-pass filter (cutoff frequency 400 Hz) is used to degrade the speech signal [23]. Other authors [24] 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 to some extent foreign languages after a short period of exposure.

III. Segmentation, vowel detection and pseudo-syllables

The starting point of recent work on rhythm is the rhythm description method proposed by Ramus in 1999 [22]. Following him, others have proposed different and more complex rhythm modelling methods (for example [11] and [25]). The weak point in many of those approaches is that they have only been applied after a manual segmentation of the speech signal. Consequently, their performance has only been assessed on relatively small corpora.

To overcome this limitation and to model the prosody of languages automatically, we use automatic processing to extract prosodic information. Three baseline procedures lead to relevant consonant, vocalic and silence segment boundaries:

- Automatic speech segmentation leading to quasi-stationary segments: This segmentation results from the "Forward-Backward Divergence" (DFB) algorithm [26], which is based on a statistical study of the signal in the temporal framework. The segmentation achieves an sub-phonemic segmentation where segments correspond to steady or transient parts of phonemes.
- Vocal activity detection: The vocal activity detection is based on a first order statistical analysis of the temporal signal [27]. The activity detection algorithm detects the less intense segments of the excerpt (in terms of energy)

and the other segments are classified as Silence or Activity according to an adaptive threshold.

- Vowel localisation: The vowel location algorithm is based on a spectral analysis (see [27] for more details). The fact that neither labelled data nor supervised learning are necessary constitutes the main advantage of this algorithm. The fact that no learning phase is necessary allows the algorithm to be used on different languages, even if no hand-labelled data is available. However, the consequence is that the algorithm is not optimised for any language even if it behaves correctly when compared to other systems [28].

This front-end processing results in a segmentation into vocalic, consonantal and silence segments. Labels "V", "C", or "#" are used to qualify each segments (Figure 1).

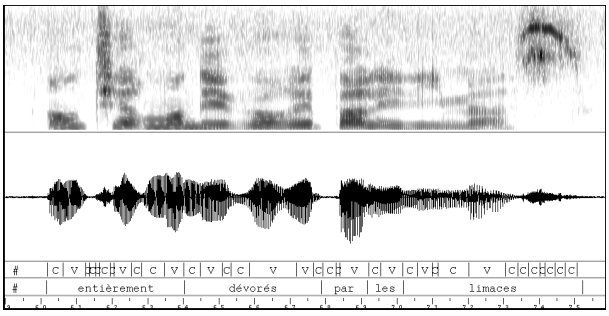

Fig. 1. Result of the automatic segmentation, vowel location and vocal activity detection on a French recording of the MULTEXT database. The sentence is: "les choux avaient été **entièrement devorés par les limaces**". "#" labels are for non-activity segments, "V" are for vocalic segments, other segments are labelled "C". Word boundaries are displayed for illustration purpose.

The syllable is a privileged unit for rhythm modelling. However, automatic extraction of syllables (in particular, boundary detection) is a controversial operation that is still in debate among phonologists: the pronunciation quality and the speech rate are factors influencing directly the syllable segmentation [29]. Furthermore, segmenting the speech signal in syllables is a language-specific task [30]. No language-independent algorithm can be easily applied.

We therefore used the notion of pseudo-syllable [31]. The basic idea is to articulate the prosodic unit around primordial elements of the syllables – vowels – and to gather the neighbouring consonants around those nuclei. We have decided to gather only the preceding consonants. This choice is explained by the fact that syllable boundary detection is not an easy task in a multilingual framework, and that the most frequent syllables correspond to the consonant/vowel structure [15]. An example of this segmentation is shown in Figure 2.

We showed in previous papers [31] that the pseudo-syllable segmentation can be successfully used for language identification. Features characterising 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 duration features, the correct identification rate is 67% for the seven languages of the MULTEXT corpus. This result is obtained using a mixture of 8 Gaussians. With the intonation features, the correct identification rate is 50%, using a GMM with 8 components. A fusion of the results obtained with both duration and intonation features allows to reach 70% correct. The confusions occur mainly across languages belonging to the same groups evoked in linguistic theories.

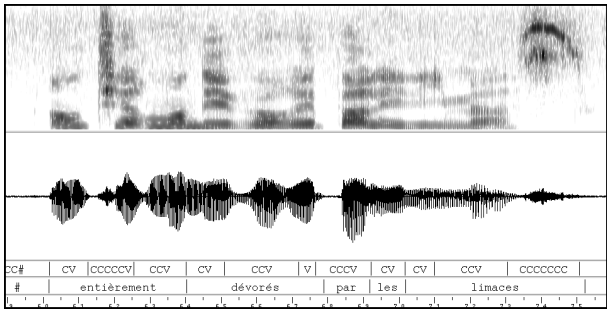

Fig. 2. Pseudo-syllable segmentation performed for the sentence "les choux avaient été **entièrement dévorés par les limaces**". Consecutive "C" segments are gathered until a "V" segment is found.

Identifying stress-timed languages (English, German and hypothetically Mandarin) versus syllable-timed languages (French, Italian and Spanish) versus mora-timed languages (Japanese), accuracy is 91% correct using both duration and intonation models.

Nevertheless, the statistical models (Gaussian Mixture Models) we use to model pseudo-syllabic features are intrinsically static models. This does not fit with the perceptive reality of prosody which is continue, dynamic and supra-segmental. We must use different models to take the temporal variations into account.

IV. MODELLING PROSODIC VARIATIONS

Following Adami's work [6], 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 characterises prosodic movements over several pseudo-syllables while the short-term component represents prosodic movements inside a pseudo-syllable. An overview of the system is displayed in Figure 3. Fundamental frequency and energy are extracted from the signal using the SNACK Sound toolkit [32].

A. Fundamental frequency coding

The fundamental frequency processing is divided into two phases, representing the phrase accentuation and the local accentuation, as in Fujisaki's work [33]:

1) Baseline computing & coding: The baseline extraction consists in finding all the local minima of the F_0 contour, and linking them. Then, the baseline is labelled in terms of U(p), D(own) and #(silence or unvoiced).

To find local minima, the sound file is automatically divided into "sentences," defined here as intervals between silent segments of duration over 250 ms. The linear regression of the F_0 curve is then computed on each sentence. Then each part of the curve under the linear regression is used to find a unique minima (Figure 4).

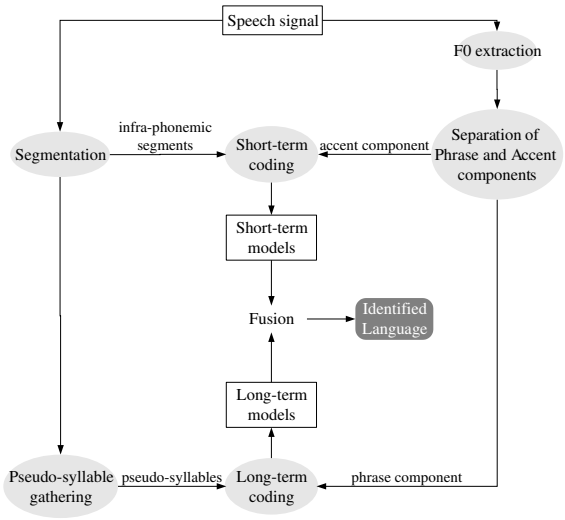

Fig. 3. Description of the system.

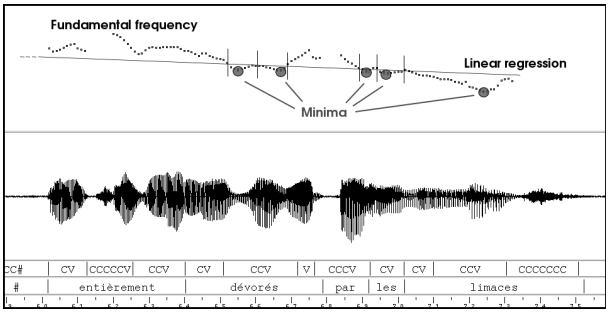

Fig. 4. Finding local minima for the sentence "les choux avaient été **entièrement dévorés par les limaces**". Local minima under the linear regression are displayed as a big dot.

Successive minima are linked by linear approximation. An example of a resulting baseline curve is displayed in Figure 5. The slope of the regression is used to label the baseline. We use one label per pseudo-syllable. Labels are 'U' for a positive slope and 'D' for a negative slope. Unvoiced pseudo-syllables (less than 70% voiced in duration) are labelled '#'. In this example, the label sequence corresponding to the sentence is:

U.U.U.U.D.D.D.D.D.D.#

2) Residue approximation & coding: The baseline is subtracted from the original contour. The resulting curve is called residue (Figure 6). This residue is then approximated for each segment 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). The example sentence is labelled in the following way:

D.D.#.#.#.#.#.D.D.U.D.D.D.D.U.U.D.-
U.#.#.D.D.U.D.U.D.D.U.D.#.#.#.#.#.#

B. Energy coding

The energy curve is approximated by linear regressions for each considered units (sub-phonemic segments or pseudo-

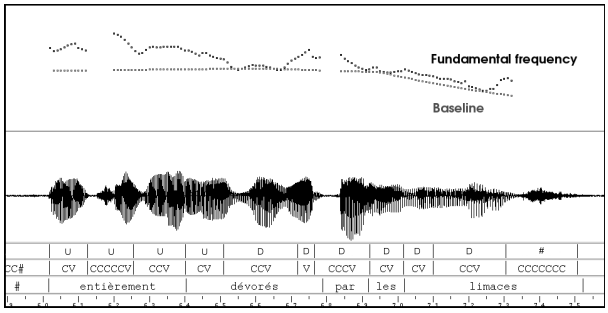

Fig. 5. Extraction of the baseline for the sentence "les choux avaient été **entièrement dévorés par les limaces**". Previously found local minima are linked with a straight line.

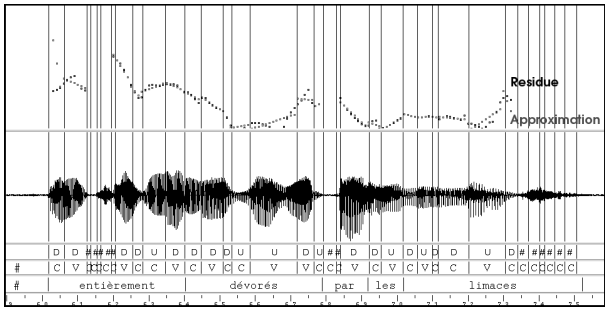

Fig. 6. Approximation of the residue for the sentence "les choux avaient été **entièrement dévorés par les limaces**". The residue is approximated by a linear regression for each segment.

syllables) (Figure 7). The process is the same as the one used for the residue coding. As there is no segment with no energy, only two labels are used: Up and Down. In this example, the sequence of labels is:

U.D.D.U.U.U.D.U.D.U.D.U.U.D.U.D.D.-
D.U.U.U.D.D.U.D.U.D.D.D.U.U.D.D.D.D

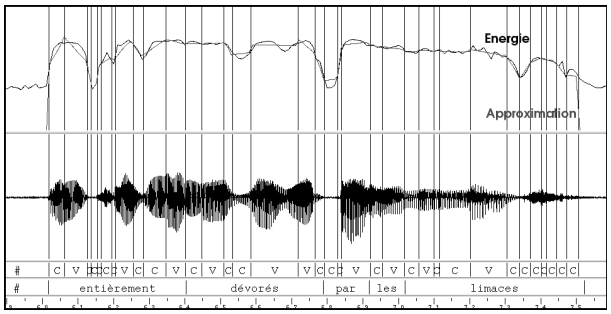

Fig. 7. Approximation of the energy for the sentence "les choux avaient été **entièrement dévorés par les limaces**". The energy is approximated by a linear regression for each segment.

C. Duration coding

Duration labels are computed only for the sub-phonemic segment units. The labels are assigned considering the mean duration of each kind of segment (vocalic, consonantic or silence). If the segment to label is a vocalic segment with a duration above the mean vocalic segments duration computed on the learning part of the corpus, it is labelled "l" (long), If

the current vocalic segment duration is below the mean, the "s" (short) label is then used. The duration labels generated on the example sentence are:

```
s.l.s.s.s.s.s.s.s.l.s.l.s.s.l.s.-
s.s.s.l.s.l.s.s.s.s.l.s.s.s.s.s.s
```

D. Modelling

To model the prosodic variations, we use classical n-gram language modelling provided by the SRI language modelling toolkit [34]. For each system – long- and short-term – each language is modelled by a n-gram model during the learning procedure. During the test phase, the most likely language is picked according to the model which provides the maximum likelihood. For the long-term models, this modelling is applied at the pseudo-syllable level and n-grams are learnt using baseline labels, eventually combined with energy labels coded at the pseudo-syllable scale. The short-term models are learnt using the sub-phonemic segmentation, and using the residue labels, optionally combined with energy and duration labels. For each segment, the label is then composed of three symbols. For the example sentence, we have:

```
DUs.DDl.#Ds.#Us.#Us.#Us.#Ds.DUs.-
DDs.UUs.DDl.DUs.DUl.DDs.UUs.UDl.-
DDs.UDs.#Us.#Us.DUl.DDs.UDl.DUs.-
UDs.DUs.DDs.UDl.DDs.#Us.#Us.#Ds.-
#Ds.#Ds.#Ds
```

Several lengths for the n-gram models have been tested (from 3- to 5-grams), but as the best results are obtained with 3-grams, only the results obtained using 3-grams models are displayed.

V. Experiments on read speech

The first experiments are made on the MULTEXT corpus [9] (a subset of EUROM1), which contains 5 languages (English, French, German, Italian and Spanish), and 10 speakers per language, balanced between male and female. We have 10 20-second files per speaker. This baseline corpus has been extended with recordings of Japanese speakers [35]. Mandarin Chinese recordings are also added to the original corpus, thanks to Komatsu [24].

The three theoretical rhythmic classes are represented in this corpus : English, German and Mandarin Chinese are stress-timed languages; French, Italian and Spanish are syllable-timed languages, and Japanese is a mora-timed language. Moreover, Mandarin Chinese is a tone language and Japanese is a pitch-accent language.

For the learning phase, we used 8 speakers (4 for Japanese), and 2 (one male and one female) were used for the tests. We have 80 learning files per language ($\approx$ 26 minutes) and 20 test files per language ($\approx$ 6 minutes) – except for French (19 files) – resulting in a total of 139 test files.

A. Long-term modelling

The sequences of labels computed on each pseudo-syllable are modelled by n-gram models. We investigated different combinations of labels to see which features are more useful (Table I). The best performance is obtained using only the baseline labels. With these labels, the correct identification rate is 50%. The confusion matrix is drawn in Table II with identified languages in columns and references in rows (confidence intervals are provided in the legend). Table II show that French, Spanish and German are clearly identified.

TABLE I

Experiments with long-term labels (139 files)

Features	# Labels	Model	% correct
Baseline	3	3-grams	50.3% $\pm$ 8.3
Energy	3	3-grams	26.6% $\pm$ 7.4
Baseline + Energy	9	3-grams	29.4% $\pm$ 7.6

TABLE II

Long term prosodic model (correct=50.3 $\pm$ 8.2% (70/139 files)). Results are displayed as percentages.

ref\id	Eng	Ger	Man	Fre	Ita	Spa	Jap
Eng	25	-	-	35	5	20	15
Ger	5	85	-	5	-	-	5
Man	5	5	10	5	25	25	25
Fre	-	-	-	69	5	26	-
Ita	5	-	15	5	40	20	15
Spa	-	-	-	-	15	80	5
Jap	5	10	-	5	15	20	45

B. Short-term modelling

The sequences of labels computed on each sub-phonemic segment are also modelled by n-gram models. Contributions of different features are shown in Table III. The best results are obtained using the combination of residue, energy and duration labels. The identification rate is 80%.

Detailed results using this configuration are displayed in Table IV. These results allow us to hypothesise that the most characteristic prosodic elements of languages are not pseudo-syllable sequences but sequences of elements constituting them.

TABLE III

Experiments with short-term labels (139 files)

Features	# labels	Model	% correct
Duration	2	3-grams	39.6% $\pm$ 7.9
Residue	3	3-grams	52.5% $\pm$ 8.2
Energy	3	3-grams	56.8% $\pm$ 8.3
Residue+Energy+Duration	18	3-grams	79.8% $\pm$ 6.6

TABLE IV

Short term prosodic model (correct=79.8 $\pm$ 6.6% (111/139 files)). Results are displayed as percentages.

ref\id	Eng	Ger	Man	Fre	Ita	Spa	Jap
Eng	65	-	15	-	10	10	-
Ger	-	95	5	-	-	-	-
Man	-	20	80	-	-	-	-
Fre	-	-	-	90	5	5	-
Ita	10	-	-	-	85	5	-
Spa	-	-	-	-	45	55	-
Jap	-	-	-	-	10	-	90

C. Merging long and short-term components

The merging of the two systems described previously is addressed. The merging technique is a simple addition of the log-likelihoods. The identification rate obtained with this method is 83%, which is considerably better than the 70% obtained with static modelling (section III).

TABLE V
MERGING OF SHORT AND LONG-TERM MODELS (CORRECT=83.5 ± 6.1% (116/139 FILES)). RESULTS ARE DISPLAYED AS PERCENTAGES.

ref\id	Eng	Ger	Man	Fre	Ita	Spa	Jap
Eng	70	-	15	-	5	10	-
Ger	-	95	5	-	-	-	-
Man	-	-	100	-	-	-	-
Fre	-	-	-	95	5	-	-
Ita	10	-	-	-	80	5	5
Spa	-	-	-	-	45	55	-
Jap	-	-	-	-	10	-	90

Despite the poor performance of the long-term model, merging the long and short term results allows to improve the identification rate. Results show that most languages are well identified.

The language classes evoked in section II have an influence on the correct identification rates: Mandarin (i.e. the only tone language of the corpus) is the most clearly identified language. Japanese, the only mora-timed language, and the only pitch-accent language in our corpus, is also well identified. The main confusions are between Italian and Spanish, both belonging to the syllable-timed and stress-accent groups. Considering rhythmic classes (represented in different strength of grey in the matrix), we can see that most confusions are within languages of the same rhythmic family.

Consequently, a rhythmic classes identification experiment was performed using the same data with the fusion of the short and long-term models. English, German and Mandarin are part of the stress-timed group; French, Italian and Spanish are in the syllable-timed group, and Japanese constitutes in itself the mora-timed group. The rhythmic classes identification rate is 94% (Table VI).

TABLE VI
RHYTHM CLASSES IDENTIFICATION TASK (CORRECT=93.5 ± 4.1% (130/139 FILES)). RESULTS ARE DISPLAYED AS PERCENTAGES.

reference\identified	Stress-timed	Syllable-timed	Mora-timed
Stress-timed	90	10	-
Syllable-timed	2	98	-
Mora-timed	10	-	90

On this data, our system manages well to classify languages according to their prosodic properties.

VI. EXPERIMENTS ON SPONTANEOUS SPEECH

The same experiments have been made on a spontaneous speech corpus, the OGI Multilingual Telephone Speech Corpus (OGI-MLTS) corpus [10]. This corpus is composed of telephone speech recorded in ten languages: English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. Experiments are made on 6 languages (English, French, German, Japanese, Mandarin and Spanish) using spontaneous speech utterances of 45 seconds each, one file per speaker. The data organization is displayed in Table VII.

TABLE VII
DATA ORGANISATION SUMMARY - OGI

Language	Learning part			Test part		
	# files	Duration (s.)		# files	Duration (s.)	
		Total	Mean		Total	Mean
English	89	2456.69	27.60	18	878.09	48.78
French	88	2438.94	27.71	15	692.29	46.15
German	88	2469.20	28.05	19	909.64	47.87
Japanese	90	2492.78	27.69	19	884.72	46.56
Mandarin	75	1952.62	26.03	18	805.66	44.75
Spanish	88	2553.12	29.01	13	618.14	47.55
Overall	518	14363.30	27.73	102	4788.57	46.94

The tests here are made using the same tuning as for the read speech experiments – same features used (baseline labels for the long-term model, and the combination of residue, energy and duration labels for the short-term model), same n-gram configuration (3-grams). Results are displayed in Tables VIII and IX for the long and short-term models respectively.

The long-term model only manages to reach performances slightly better than chance (on a six-way classification, the chance level is $\approx 16\%$).

The short-term model achieves a performance of 40% of correct identifications. The best recognised languages are English (55%) and Japanese (53%). We can observe that German is mainly confused with English.

The performances of the system for the rhythm classes identification task is given in Table X.

TABLE VIII
LONG TERM PROSODIC MODELS (CORRECT=20.6 ± 8.0% (21/102 FILES)). RESULTS ARE DISPLAYED AS PERCENTAGES.

ref\id	Eng	Ger	Man	Fre	Spa	Jap
Eng	33	27	6	28	-	6
Ger	26	11	16	21	16	10
Man	28	-	11	17	5	39
Fre	20	20	7	20	20	13
Spa	15	7	8	8	31	31
Jap	26	16	5	16	16	21

TABLE IX
SHORT TERM PROSODIC MODELS (CORRECT=40.2 ± 9.5% (41/102 FILES)). RESULTS ARE DISPLAYED AS PERCENTAGES.

ref\id	Eng	Ger	Man	Fre	Spa	Jap
Eng	55	17	6	5	17	-
Ger	42	37	-	11	10	-
Man	17	33	33	11	-	6
Fre	33	33	-	20	14	-
Spa	38	-	-	16	38	8
Jap	10	10	-	11	16	53

The global identification rate is 69%. The identification rates for each class are respectively 85% for the stress-timed languages, 46% for the syllable-timed languages and 53% for the mora-timed language. As expected from language

TABLE X

RHYTHM CLASSES IDENTIFICATION TASK (CORRECT=68.6 $\pm$ 9.1% (70/102 FILES)). RESULTS ARE DISPLAYED AS PERCENTAGES.

reference \ identified	Stress-timed	Syllable-timed	Mora-timed
Stress-timed	85	11	4
Syllable-timed	50	46	4
Mora-timed	16	31	53

identification results, syllable-timed languages are the least recognised. The identification mistakes are made principally on the French language, responsible for 10 errors – that is to say 66% of the syllable-timed languages identification error. This may be explained by the different varieties of French (Canadian and European) encountered in the corpus.

It is however quite difficult to directly transpose our approach, designed on read speech, to spontaneous speech. The main reason may be the intrinsic variability of the spontaneous data. This variability can be linked to the great variations of speech rate observed within each language (see [36] for a study of automatic speech rate estimation on read and spontaneous speech). Figure 8 displays the accumulated number of vowels per second for the English parts of the MULTEXT and OGI-MLTS corpus. Spontaneous speech leads to much more variation in terms of number of vowels per second than read speech. The Table XI shows the standard deviations of the speech rate approximation on spontaneous and read speech for different languages (45 s. spontaneous speech utterances of the OGI-MLTS corpus and 20 s. read utterances from the MULTEXT corpus). These standard deviation values show that there is a much greater dispersion of spontaneous speech rate within each language than for read speech. These intra-language variations can explain why our models achieve a poor performance on the test part of the corpus.

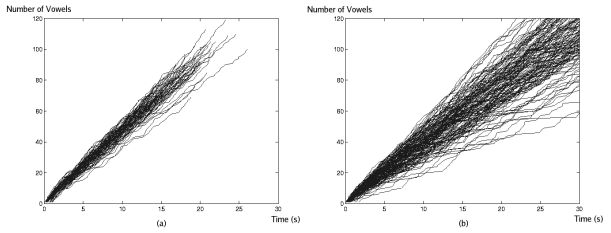

Fig. 8. (a) Number of detected vowels as a function of time on English read speech, (b) Number of detected vowels as a function of time on English spontaneous speech

TABLE XI

STANDARD DEVIATIONS OF SPEECH RATE (ESTIMATION) FOR DIFFERENT LANGUAGES ON READ AND SPONTANEOUS SPEECH

Lang.	Eng	Fre	Ger	Ita	Jap	Man	Spa	All
Read	0.36	0.35	0.37	0.56	0.42	0.49	0.40	0.50
Spont.	0.77	0.93	0.80	n.d.	0.98	0.77	0.81	0.85

VII. EXPERIMENTS ON ARABIC DIALECTS

We believe nonetheless that our system can be applied, at least on groups of languages, to investigate if the groups considered can be automatically distinguished prosodically. We carried out experiments with different Arabic dialects to show how the features described in this paper help to classify them.

The corpus ARABER has been recorded at the *Dynamique du Langage* Laboratory in Lyon, France. It consists of semi-spontaneous recordings (comments on an image book) from 40 speakers from Maghreb, Middle-East and an intermediate area (Tunisia, Egypt). The mean duration for each speaker's recording is 5 minutes, in 40 files of 7.6 seconds. All the data have been used both for learning and testing, via a cross-validation procedure.

The quality of the recordings made this corpus interesting because it is intermediate between the studio-recorded read speech corpus MULTEXT and the telephone spontaneous speech of the OGI corpus.

As research on the prosody of Arabic dialects is emerging, very few observations are available to hypothesise how much they differ. At this point, Arabic dialects are classified according to inter-comprehension between speakers, resulting in 3 main areas: Occidental (Moroccan, Algerian), Intermediate (Tunisian, Egyptian) and Oriental (Lebanese, Jordanian, Syrian). The aim of this study is to see, with an ALI system that does not require any kind of annotation, if it is possible to identify the hypothesised dialectal areas using only prosody.

The prosodic ALI system – using the combination of the long and short-term models – has been applied to this corpus without any tuning procedure: same features (baseline labels for the long term model and the combination of residue, energy and duration features for the short-term model), same n-gram configuration. Experiments are made according to a cross-validation procedure applied on each speaker: learning is done using all speakers except one who is used for the test. This is repeated until all speakers have been used for the test. Results of the cross-validation tests are displayed in Table XII.

TABLE XII

ARABIC DIALECTAL AREAS IDENTIFICATION (CORRECT= 98.0 $\pm$ 0.6% (1563/1592 FILES). RESULTS ARE DISPLAYED AS PERCENTAGES.

identified \ reference	Occidental	Intermediate	Oriental
Occidental	99.5	-	0.5
Intermediate	1.8	96.0	2.2
Oriental	1.0	0.3	98.7

The system performs very well on this data, which shows that prosodic differences may be important between the hypothesised dialectal areas of Arabic. Further studies are needed in order to identify the exact prosodic differences between dialectal areas, and if there are differences among dialects of a same area.

VIII. CONCLUSIONS AND PERSPECTIVES

Experiments on read speech show that our system is able to automatically identify languages using prosody alone. Differences between languages seem more characterised by micro-prosodic events (short term) than macro-prosodic ones. The experiments show that variations of fundamental frequency, energy and segment duration that occur within a syllable

are more characteristic. The dynamic modelling allows to reach 83% of correct identification on a seven language discrimination task. Results tend to confirm the existence of automatically identifiable rhythmic classes (accuracies above 90% for rhythm class identification).

Considering spontaneous speech, accuracies are lower. This can be partly explained by the greater inter-speaker variability due to the number of different prosodic realizations allowed by spontaneous speech, especially in terms of changes in speech rate. The results remain however interesting for the language class identification experiment.

An applicative example has been shown with the experiments on Arabic dialects, where the speech quality is intermediate between read and spontaneous. This experiment has shown that there exist some automatically detectable prosodic differences between hypothesised dialectal areas of Arabic. A careful study with the help of linguists is needed in order to define precisely what are those differences and where they appear. The next experiment will be to test if the dialects can be automatically clustered using our system.

The main advantage of our prosodic ALI system lies in the fact it does not require any manual annotations (especially phonetic annotations which are very time-consuming). Hence, the system can be directly applied to unknown data, and be used to evaluate if the prosodic differences between languages or dialects can be automatically detected. Using this system can help linguists to verify that further investigation is needed, leading to new collaborations between the automatic processing and the linguistic communities.

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 and I. Vasilescu, "From perceptual designs to linguistic typology and automatic language identification: Overview and perspectives," in *Eurospeech*, Aalborg, Denmark, 2001.

[3] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixtures speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, January 1995.

[4] T. J. Hazen and V. W. Zue, "Segment-based automatic language identification," *Journal of the Acoutical Society of America*, 1997.

[5] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "The supersid project: Exploiting high-level information for high-accuracy speaker recognition," in *ICASSP*, vol. 4, Hong Kong, April 2003, pp. 784–787.

[6] 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.

[7] D. Zhu, M. Adda-Decker, and F. Antoine, "Different Size Multilingual Phone Inventories and Context-Dependent Acoustic Models for Language Identification," in *InterSpeech*, Lisbon, September 2005.

[8] T. Matrin, B. Bake, E. Wong, and S. Sridharan, "A syllable-scale framework for language identification," *Computer Speech and Language*, vol. 20, pp. 276–302, 2006.

[9] E. Campione and J. Véronis, "A multilingual prosodic database," in *ICSLP*, Sidney, 1998, pp. 3163–3166.

[10] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The ogi multilanguage telephone speech corpus," in *ICSLP*, October 1992, pp. 895–898.

[11] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in Laboratory Phonology 7*, 2002.

[12] P. Ladefoged, Ed., *The intonation of American English*. Michigan, USA: University of Michigan Press, 1945.

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

[14] P. Roach, "On the distinction between "stress-timed" and "syllable-timed" languages," *Linguistic Controversies*, pp. 73–79, 1982.

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

[16] 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.

[17] I. Lehiste, "Isochrony reconsidered," *Journal of Phonetics*, vol. 5, pp. 253–263, 1977.

[18] M. Nespor, "On the rhythm parameter in phonology," in *Logical Issues in Language Acquisition*, I. Roca, Ed., 1990, pp. 157–175.

[19] F. Cummins, "Speech rhythm and rhythmic taxonomy," in *Speech Prosody*, Aix-en-Provence, France, 2002, pp. 121–126.

[20] L. Hyman, "Word-prosodic typology," UC Berkeley, Tech. Rep., 2005.

[21] D. R. Ladd and R. Morton, "The perception of intonation emphasis: continuous or categorical?" *Journal of Phonetics*, vol. 25, pp. 313–342, 1997.

[22] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, no. 3, pp. 265–292, 1999.

[23] S. Frota, M. Vigario, and F. Martins, "Language discrimination and rhythm classes: evidence from portuguese," in *Speech Prosody*, Aix-en-Provence, France, 2002.

[24] M. Komatsu, T. Arai, and T. Sugawara, "Perceptual discrimination of prosodic types," in *Speech Prosody*, Nara, Japan, 2004, pp. 725–728.

[25] A. Galves, J. Garcia, D. Duarte, and C. Galves, "Sonority as a basis for rhythmic class discrimination," in *Speech Prosody*, Aix en Provence, France, April 2002.

[26] R. André-Obrecht, "A new statistical approach for automatic speech segmentation," *IEEE Transanctions on Acoustics, Speech and Signal Processing*, vol. 36, no. 1, pp. 29–40, 1988.

[27] F. Pellegrino and R. André-Obrecht, "Vocalic system modeling : A vq approach," in *IEEE Digital Signal Processing*, Santorini, July 1997, pp. 427–430.

[28] F. Pellegrino, "Une approche phonétique en identification automatique des langues : la modélisation acoustique des systèmes vocaliques," Ph.D. dissertation, Université Paul Sabatier, Toulouse, France, Dec. 1998.

[29] 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, pp. 89–93.

[30] H. Pfitzinger, S. Burger, and S. Heid, "Syllable detection in read and spontaneous speech," in *4th ICSLP*, vol. 2, Philadelphia, October 1996, pp. 1261–1264.

[31] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. André-Obrecht, "Rhythmic unit extraction and modelling for automatic language identification," *Speech Communication*, vol. 47, no. 4, pp. 436–456, 2005.

[32] K. Sjölander. The snack sound toolkit. [Online]. Available: http://www.speech.kth.se/snack/

[33] H. Fujisaki, "Prosody, information and modeling - with emphasis on tonal features of speech," in *ISCA Workshop on Spoken Language Processing*, Mumbai, India, January 2003.

[34] A. Stolcke, "Srilm - an extensible language modeling toolkit," in *ICSLP*, 2002, pp. 901–904. [Online]. Available: http://www.speech.sri.com/projects/srilm/

[35] S. Kitazawa, "Periodicity of japanese accent in continuous speech," in *Speech Prosody*, Aix en Provence, France, April 2002, pp. 435–438.

[36] F. Pellegrino, J. Farinas, and J.-L. Rouas, "Automatic Estimation of Speaking Rate in Multilingual Spontaneous Speech ," in *Speech Prosody 2004 , Nara, Japon*, 23-26 mars 2004, pp. 517–520.

Jean-Luc Rouas Jean-Luc Rouas has obtained a PhD in Computer Science at IRIT (Computed Science Research Institute of Toulouse, France) and Paul Sabatier University (Toulouse, France) in 2005. His thesis work is named "Languages' characterisation and identification". He had a one-year post-doctoral position at INRETS (French National Institute for Safety and Transports). He is now at INESC, Laboratório de Sistemas de Lingua Falada (L2F) in Lisbon, Portugal. His research interests include signal analysis, speech processing, prosody.