MERGING SEGMENTAL, RHYTHMIC AND FUNDAMENTAL FREQUENCY FEATURES FOR AUTOMATIC LANGUAGE IDENTIFICATION

Jean-Luc ROUAS[1], Jérôme FARINAS[1], François PELLEGRINO[2]
[1]Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS INPT UPS, FRANCE
[2]Laboratoire Dynamique Du Langage, UMR 5596 CNRS Univ. Lyon 2, FRANCE
Jean-Luc.Rouas@irit.fr, Jerome.Farinas@irit.fr, Francois.Pellegrino@univ-lyon2.fr

ABSTRACT

This paper deals with an approach to Automatic Language Identification based on rhythmic and fundamental frequency modeling. Experiments are performed on read speech for 5 European languages. They show that rhythm can be automatically extracted and is relevant in language identification: using cross-validation, 79% of correct identification is reached with 21 s. utterances. The fundamental frequency modeling, tested in the same conditions (cross-validation), produces 50% of correct identification for the 21 s. utterances. The Vowel System Modeling gives an identification rate of 70% for the 21 s. utterances. Last, merging the three models slightly improves the identification rate.

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 the standard up to date approach, sequences of phonetic units (provided by a phonetic modeling system) are decoded according to language-specific statistical grammars [1]. This approach, initiated at the beginning of the 90s, is still the most efficient one. However, only marginal improvements have been performed for five years, and it seems crucial to propose new approaches. In this paper, we investigate the way to explicitly take phonetics into account and to take advantage from alternative features also present in the signal: prosodic features, and especially rhythmic features, are known to carry a substantial part of the language identity (Section 2). However, their modeling is still an open problem, mostly because of the 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 fundamental frequency contours, computes statistics on these outlines in order to model intonation (Section 4). These algorithms, coupled with a Vowel System Model (VSM) are tested on the five languages of the MULTEXT corpus in section 5. The relevance of the rhythmic parameters and the efficiency of each system (Rhythmic Model, Fundamental Frequency Model and Vowel System Model) are evaluated. Then, the possibility of merging these three approaches is addressed.

2 MOTIVATIONS

2.1 Relevance of Rhythm

Rhythm is a critical characteristic of language in different activities (e.g. child language acquisition, language synthesis), and especially in both human and computer language identification. Among others, Thymé-Gobbel and Hutchings pointed out the importance of prosodic information in language identification systems [2]. With parameters related to rhythm and based on syllable timing, syllable duration, and descriptors of amplitude patterns, they have obtained promising results, and proved that mere prosodic cues can distinguish between some language pair with results comparable to some non-prosodic systems. Ramus et al. [3] show that newborn infants are sensitive to the rhythmic properties of languages. Other experiments based on a consonant/vowel segmentation of eight languages established that derived parameters might be relevant to classify languages according to their rhythmic properties [4].

2.2 Relevance of Intonation

Intonation is a seldom employed parameter in language identification whereas the extraction of Fundamental Frequency (F0) is usual in speech analysis. In fact, few experiments using intonation have already been made in language identification [5] and the correct identification rate reached was near 30 % on 10 languages. However, linguistic studies show that languages should be discriminated by their intonation patterns [6]. Therefore, we proposed that modeling intonation might be more effi-

cient if we keep the pseudo-syllable timing issued from the rhythm model for the description of each language's intonation patterns.

2.3 Classifying languages according to rhythm and intonation

Experiments reported here focus on 5 European languages (English, French, German, Spanish and Italian). According to the literature, French, Spanish and Italian are *syllable-timed* while English and German are *stress-timed*. These two categories emerged from the theory of isochrony introduced by Pike and developed by Abercrombie [7]. 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 two binary categories are replaced by a continuum [8]. 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. Another difficulty rises from the selection of an efficient modeling paradigm. We develop here a statistical approach, first introduced in [9] and now improved by considering stress features (Fundamental Frequency and Energy). It is based on a Gaussian modeling of the different *rhythm units* automatically extracted from a rhythmic segmentation in the languages.

3 DESCRIPTION OF THE SYSTEM

A synopsis is displayed in Figure 1. A language-independent vowel detection algorithm is applied to label the speech signal in Silence/Non Vowel/Vowel segments. The rhythmic pattern is derived from this segmentation, as we try to obtain a syllable-like segmentation. Then, this segmentation is used to compute statistics over the fundamental frequency outlines for each syllable. Afterward, computation of cepstral coefficients for the vowel segments leads to language-specific Vowel System Models (VSM) while the rhythmic pattern derived from the segmentation is used to model the rhythm of each language.

3.1 The Vowel/Non Vowel segmentation algorithm

This algorithm, based on a spectral analysis of the signal, is described in [10]. It is applied in a language and speaker independent way without any manual adaptation phase. This processing provides a segmentation of the speech signal in pause, non-vowel and vowel segments. Due to the intrinsic properties of the algorithm (and especially the fact that transient and steady parts of a phoneme may be separated), it is somewhat incorrect to consider that this segmentation is exactly a Consonant/Vowel segmentation. However, it is undoubtedly correlated to the rhythmic structure of the speech sound, and in this paper, we investigate the assumption that this correlation enables the definition of a statistical model to discriminate languages according to their rhythmic structure.

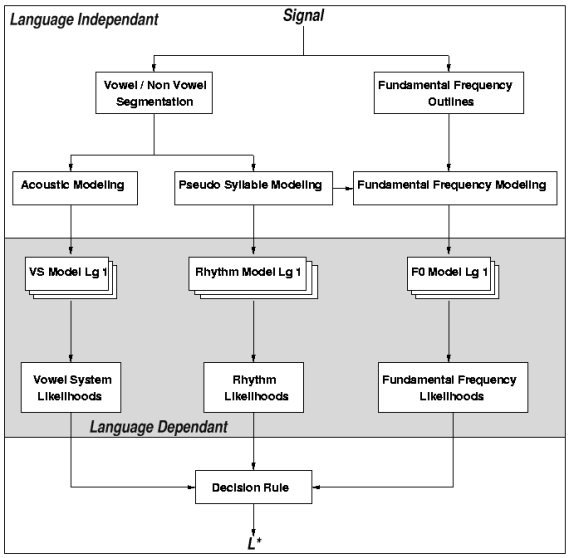

Figure 1 - Synopsis of the system for N languages.

3.2 Vowel System Modeling

Each vowel segment is represented with a set of 8 Mel-Frequency Cepstral Coefficients (MFCCs) and 8 delta-MFCCs, augmented with the Energy and delta Energy of the segment. This parameter vector is extended with the duration of the underlying segment providing a 19-coefficient vector. A cepstral subtraction performs both blind removal of the channel effect and speaker normalization. For each recording sentence, the average MFCC vector is computed and subtracted from each coefficient. For each language, a Gaussian Mixture Model (GMM) is trained using the EM algorithm. The number of components of the model is computed using the LBG-Rissanen algorithm [11]. During the test, the decision lays on a Maximum Likelihood procedure.

3.3 Rhythm Modeling

3.3.1 Rhythmic units

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 [12]. In the algorithm, speech signal is parsed in patterns matching the structure: .CnV. (where n is an integer that may be zero and V may result from the merging of consecutive vowel segments). For example, if the vowel detection algorithm produces the sequence (CCVVCCVCVC-

CCVCVCCC), it is parsed in the following sequence of 5 pseudo-syllables: (CCV.CCV.CV.CCCV.CV)

3.3.2 Pseudo-syllable description

For each pseudo-syllable, 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, $N_C = 2$). Additionally, one parameter related to the stress structure of the language (Energy in dB, normalized among the sentence) is also considered. Our hypothesis is that this parameter may improve the discrimination of stress-timed languages. Such a basic rhythmic parsing is obviously limited, but provides a framework to model rhythm that requires no knowledge on the language rhythmic structure

3.3.3 Statistical Rhythm modeling

For each language, a GMM is trained, either by using the standard LBG algorithm or the LBG-Rissanen algorithm to provide the optimal number of Gaussian components.

3.4 Fundamental Frequency Modeling

3.4.1 Extraction of fundamental frequency contours

To extract the outlines, we used a tool called "MESSIGNAIX" developed by the Laboratoire Parole et Language in Aix-en-Provence (France) [13]. This tool allows to extract the fundamental frequency outlines with three different approaches: Average Magnitude Difference Function (AMDF), spectral comb, autocorrelation and an overall of the three. We computed the contours with the overall method, using spline interpolation to obtain values every 10ms, even on the unvoiced segments.

3.4.2 Features Extraction

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 the intonation of each pseudo-syllable. We choose to compute statistics until 4th order (mean, standard deviation, skewness and kurtosis), in order to describe the variations of intonation within a pseudo-syllable.

3.4.3 Statistical Modeling

For each language, as for rhythm modeling, a GMM is trained by using the standard LBG algorithm or the LBG-Rissanen algorithm to provide the optimal number of Gaussian components.

4 EXPERIMENTS

4.1 Corpus

Experiments are performed on the MULTEXT corpus [14]. This database contains recordings from five European languages (English, French, German, Italian and Spanish), pronounced by 50 different speakers (5 males and 5 females per language). Data consist of read passages of about five sentences extracted from the EUROM1 speech corpus (the mean duration of each passage is 20.8 seconds). The raw pitch contour of the signal is also available. A limitation is that the same texts are produced by a mean of 3.75 speakers. This leads to a possible partial text dependency of the models. Due to the limited size of the corpus, language identification experiments are performed using a cross-validation procedure: 9 speakers are used to train the models of one language and the tenth speaker is used to perform the test. This procedure is iterated for each speaker, and for each language.

4.2 Rhythm Modeling

Table 1 summarizes the experiments performed with the rhythm parameters. The identification scores displayed are averaged among several GMM topologies and obtained using the whole duration of the test excerpts (about 21 seconds).

Parameters	Mean Identification Rate
$D_V + D_C$	65 %
$D_V + D_C + N_C$	75 %
$D_V + D_C + N_C + E$	79 %

Table 1 - Results in cross-validation experiments with rhythm modeling.

The use of duration parameters D_V and D_C results in a 64.8 % of correct identification. The use of additional parameters related to the complexity of the pseudo-syllable structure (N_C) and to the stress (E) significantly improves the results, reaching 79 % of correct identification.

4.3 Fundamental Frequency Modeling

Table 2 summarizes the experiments performed with the statistical F0 parameters.

Parameters	Mean Identification Rate
Skewness + Kurtosis	53 %
Mean+Var+Skew+Kurt	47 %

Table 2 - Results in cross-validation experiments with fundamental frequency modeling.

These results show that keeping the pseudo-syllable segmentation is relevant for modeling fundamental frequency patterns. Using only high order moments seems to better describe the variations of intonation, as far as the best results (53 %) are obtained with only skewness and kurtosis. Using all parameters including mean and variance decreases the identification rate to 47 %.

4.4 Vowel System Modeling

The Vowel System modeling results in an identification rate of 70 % with 21 seconds of signal.

4.5 Integrating Segmental, Rhythm and Fundamental Frequency Modeling

A simple statistical merging is performed by adding the log-likelihoods of both the Rhythm model, the F0 model, and the Vowel System model for each language. Merging the results given by the three models only results in a small 5% improvement, increasing the identification rate to 84 %. Anyway, if only a little improvement is observed, at least no degradation results from the statistical merging.

Parameters	Mean Identification Rate
Rhythm+F0	82 %
F0+MFCCs	74 %
Rhythm+MFCCs	83 %
Rhythm+F0+MFCCs	84 %

Table 3 - Fusion of Rhythm, F0 and Cepstral parameters.

5 DISCUSSION

We propose in this paper three algorithms dedicated to automatic language identification. Experiments, performed with cross-validation, show that it is possible to achieve an efficient rhythmic modeling (78% of correct identification) in a way that requires no a priori knowledge of the rhythmic structure of the processed languages. Besides, the F0 model gives 53 % of correct identification, whereas 70 % of correct identification is obtained with the Vowel System Model.

With these read data, merging the three approaches results in a slight improvement. However, rhythm features may be less sensible to spectral degradation and then be more useful with lower quality data, meanwhile the F0 model might be more efficient with spontaneous speech. Furthermore, the fusion method used here is very simple and requires an extremely low computational cost. Using more complex methods should increase the identification rate.

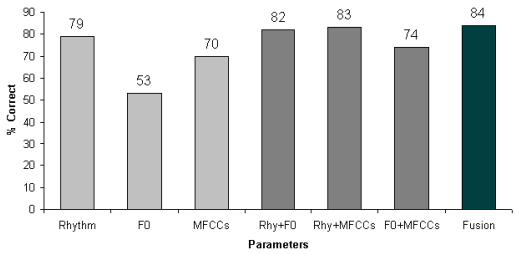

Figure 2 - Results obtained for each parameter, and merging of the three approaches.

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*).

References

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

[2] Thymé-Gobbel, A., and Hutchins, S. E., "Prosodic features in automatic language identification reflect language typology", *Proc. of ICPhS99*, San Francisco, 1999.

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

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

[5] Itahashi S., Kiuchi T., and Yamamoto M., "Spoken Language Identification using Fundamental Frequency and Cepstra", *Eurospeech99*, Bucarest, Hungary, September 1999.

[6] Hirst, D. and DiCristo, A., *Intonation Systems. A Survey of Twenty Languages*, Cambridge University Press, Cambridge, 1998.

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

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

[9] Farinas, J. and Pellegrino, F., "Automatic Rhythm Modeling for Language Identification", *Proc. Of Eurospeech Scandinavia'01*, Aalborg, 2001.

[10] Pellegrino, F., and André-Obrecht, R., "An Unsupervised Approach to Language Identification", *Proc. of ICASSP99*, Phoenix, 1999.

[11] Pellegrino, F. and André-Obrecht, R., "Automatic Language Identification: an Alternative Approach to Phonetic Modeling", *Signal Processing*, 80(7), 1231-1244, 2000.

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

[13] Espesser R., "MES : un environnement de traitement du signal", *XXIes JEP*, Avignon, p.447, 1996. `http://www.lpl.univ-aix.fr/projects/mes_signaix`

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