FUSION DE PARAMÈTRES RYTHMIQUES ET SEGMENTAUX POUR L'IDENTIFICATION AUTOMATIQUE DES LANGUES

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 INPT UPS, FRANCE
[2]Laboratoire Dynamique Du Langage, UMR 5596 CNRS Univ. Lyon 2, FRANCE
rouas@irit.fr, jfarinas@irit.fr, pellegri@univ-lyon2.fr, obrecht@irit.fr

ABSTRACT

This paper deals with an approach to Automatic Language Identification based on rhythmic modeling and vowel system modeling. Experiments are performed on read speech for 5 European languages. They show that rhythm and stress may be automatically extracted and are relevant in language identification: using cross-validation, 78% of correct identification is reached with 21 seconds utterances. The Vowel System Modeling, tested in the same conditions (cross-validation), is efficient and results in a 70% of correct identification for the 21 seconds utterances. Last, merging the output scores from the two models improves the results : with only 11 seconds test excerpts, the correct identification rate is over 80%.

1. INTRODUCTION

Depuis la dernière décennie, la demande en systèmes d'Identification Automatique des Langues (IAL) évolue vers différents champs d'applications, et particulièrement dans les communications assistées par ordinateur (e.g. services d'urgences) et les interfaces homme-machine multilingues (e.g. bornes interactives). Plus récemment, l'indexation par le contenu de documents multimédia ou sonores amène un nouveau thème pour lequel les systèmes d'IAL peuvent être utiles.

Les systèmes actuels sont basés sur la reconnaissance d'unités phonétiques. Identifier la langue nécessite l'utilisation de multiples reconnaisseurs phonétiques fonctionnant en parallèle. A cause de la multitude de langues possibles en entrée (qui peut aller de la dizaine à la centaine), le coût en temps de calcul devient prohibitif. L'emploi de paramètres alternatifs, calculés indépendamment de la langue, permettrait d'améliorer la rapidité du traitement.

Dans cet article, nous étudions une manière de prendre explicitement en compte la phonétique, tout en employant des informations alternatives également présentes dans le signal : la prosodie, et plus particulièrement le rythme qui sont connus pour contenir une part substantielle de l'identité de la langue.

Cependant, la modélisation de ces paramètres est toujours un problème ouvert, à cause de leur nature perceptuelle. Pour résoudre ce problème, un algorithme d'extraction des paramètres rythmiques, indépendant de la langue, est proposé et appliqué pour modéliser le rythme (Section 3).

Cet algorithme, couplé avec une Modélisation des Systèmes Vocaliques (MSV) est testé sur les cinq langues du corpus MULTEXT dans la section 4. La pertinence des paramètres rythmiques et l'efficacité de chaque système (Modèle de Rythme et Modèle de Système Vocalique) est évaluée. De plus, la possibilité de fusionner ces deux approches est examinée.

2. MOTIVATIONS

2.1. Pertinence des Paramètres Rythmiques

Le rythme est un paramètre caractéristique de la langue, il est important dans différentes activités en rapport avec la langue (e.g. apprentissage de la langue par les enfants, synthèse de la parole), et plus spécialement dans l'identification de la langue, que ce soit par l'homme ou par la machine.

Thymé-Gobbel et Hutchings ont montré l'importance des informations prosodiques pour les systèmes d'IAL [2]. Avec des paramètres basés sur le rythme syllabique, la durée syllabique et des descripteurs des contours de l'amplitude et de la courbe mélodique, ils ont obtenu des résultats prometteurs, et ont montré que ces paramètres prosodiques seuls permettent de distinguer entre différentes paires de langues avec des résultats comparables à ceux obtenus par les systèmes n'employant pas la prosodie.

Ramus et al. [3] montrent que les nouveau-nés sont sensibles aux propriétés rythmiques des langues.

D'autres expériences, basées sur une segmentation consonne/voyelle de huit langues, établissent que des paramètres dérivés peuvent être pertinents pour classer les langues suivant leurs propriétés rythmiques [4].

2.2. Classer les Langues Selon le Rythme

Les expériences reportées ici sont effectuées sur 5 langues européennes (anglais, français, allemand, espagnol et italien). D'après la littérature, le français, l'espagnol et l'italien sont des langues syllabiques (*syllable-timed*) tandis que l'anglais et l'allemand sont des langues accentuelles (*stress-timed*). Ces deux catégories émergent de la théorie de l'isochronie introduite par Pike et développée par Abercrombie [5].

Toutefois, de récentes études basées sur la mesure de la durée des intervalles entre les accentuations à la fois pour les langues syllabiques et les langues accentuelles fournissent un cadre alternatif dans lequel ces deux catégories

binaires sont remplacées par un continuum [6]. Les différences rythmiques entre les langues sont alors pour la plupart relatives à la structure syllabique et la présence (ou l'absence) de la réduction vocalique.

Les controverses sur le statut du rythme dans les langues du monde illustrent la difficulté de segmenter la parole en unités rythmiques. Même si des corrélations existent entre le signal de parole et le rythme linguistique, trouver une représentation pertinente semble difficile. Une autre difficulté surgit de la sélection d'un paradigme de modélisation efficace.

Nous développons ici une approche statistique, introduite dans [7] et améliorée en considérant des paramètres d'accentuation (F_0 et l'énergie). Elle est basée sur une modélisation par mélange de lois gaussiennes des différentes unités rythmiques extraites d'une segmentation automatique du signal de parole.

3. DESCRIPTION DU SYSTÈME

La Figure 1 présente une vue d'ensemble du système. Un algorithme de détection des voyelles indépendant de la langue est appliqué pour étiqueter le signal de parole en segments silence/voyelle/non voyelle.

Ensuite, le calcul des coefficients cepstraux pour les segments vocaliques permettent le calcul de Modèles de Systèmes Vocaliques spécifiques aux langues, pendant que les motifs rythmiques dérivés de la segmentation sont utilisés pour modéliser le rythme de chaque langue.

3.1. L'algorithme de Segmentation Voyelle/Non Voyelle

Cet algorithme, basé sur une analyse spectrale du signal, est décrit dans [8]. Il est appliqué indépendamment de la langue et du locuteur sans aucune phase d'adaptation manuelle. Ce processus fournit une segmentation du signal de parole en segments pause, non voyelle et voyelle, grâce à un détecteur d'activité vocale associé à un détecteur de segments vocaliques (taux de detection des zones vocaliques de 93,5 % en moyenne sur 5 langues).

Cependant, l'algorithme de segmentation peut séparer les parties voisées et non voisées d'un même phonéme. Il serait donc incorrect d'affirmer que cette segmentation est une exacte dichotomie entre les consonnes et les voyelles.

Toutefois, elle est indéniablement corrélée à la structure rythmique de la parole. Dans cet article, nous étudions l'hypothèse que cette corrélation peut permettre à un modèle statistique de discriminer les langues suivant leur structure rythmique.

3.2. Le Modèle de Système Vocalique

Chaque segment vocalique est représenté par un vecteur de 8 coefficients cepstraux extraits selon l'échelle de Mel (Mel Frequency Cepstral Coefficient) et de 8 delta-MFCC, plus l'énergie et la dérivée de l'énergie du segment. Ce vecteur de paramètres est complété avec la durée du segment concerné, résultant en un vecteur de 19 coefficients.

Une soustraction cepstrale permet à la fois de s'affranchir de l'effet du canal et des variations de locuteurs.

Pour chaque langue, un Modèle de Mélange de lois Gaussiennes (MMG) est appris en utilisant l'algorithme EM. Le nombre de composantes du mélange est calculé grâce à l'algorithme LBG-Rissanen [9]. Pendant les tests, la décision repose sur une procédure de maximum de vraisemblance.

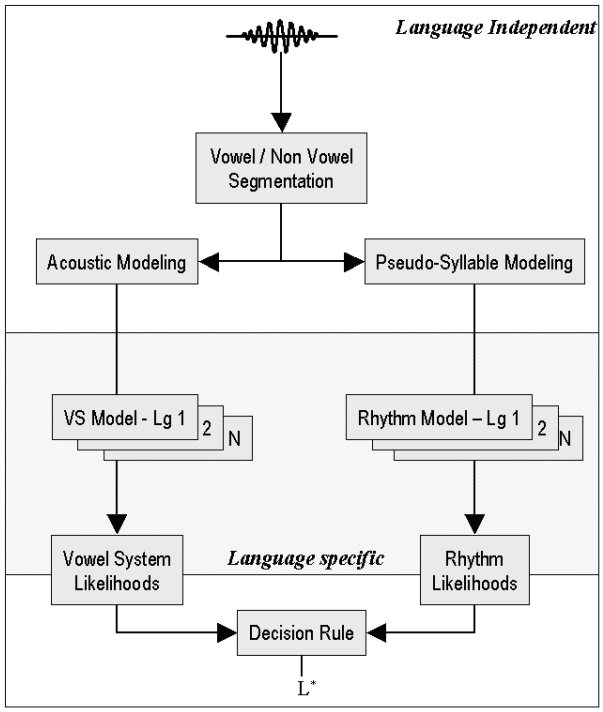

Figure 1 - Synopsis du système pour N langues.

3.3. Modélisation du Rythme

Choix des Unités Rythmiques : La syllabe est une unité privilégiée pour la modélisation du rythme. Malheureusement, segmenter le signal de parole en syllabes est une tâche typiquement spécifique à chaque langue. Aucun algorithme indépendant de la langue ne peut donc être appliqué.

Pour cette raison, nous avons introduit dans [7] la notion de pseudo-syllabe dérivée de la plus fréquente structure syllabique au monde, la structure CV [10].

Dans cet algorithme, le signal de parole est segmenté en motifs correspondants à la structure .C^nV. (où n est un entier qui peut être nul et V peut résulter du regroupement de plusieurs segments vocaliques consécutifs).

Par exemple, si l'algorithme de détection des voyelles donne la séquence (CCVVCCVCVCCCVCVCCC), elle est partitionnée en la séquence suivante de 5 pseudo-syllabes : (CCV.CCV.CV.CCCV.CV)

Description de la pseudo-syllabe : Pour chaque pseudo-syllabe, trois paramètres sont calculés, correspondant respectivement à la durée totale des segments consonantiques, la durée totale du segment vocalique et la complexité de la pseudo-syllabe (nombre de segments consonantiques).

Par exemple, la description pour une pseudo-séquence .CCV. est :

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

Où D_C est la durée totale des segments consonantiques, D_V est la durée totale du segment vocalique et N_C est le nombre de segments consonantiques contenus dans la pseudo-syllabe (ici, N_C = 2).

Additionnellement, deux paramètres relatifs aux structures accentuelles des langues (F_0 et l'énergie en dB) sont également considérés. Ils sont calculés sur les parties voisées de signal, c'est à dire les segments vocaliques : nous obtenons une valeur de F_0 et une valeur d'énergie par pseudo-syllabe, que l'on normalise sur l'ensemble de la phrase.

Notre hypothèse est que ces paramètres peuvent améliorer la discrimination des langues accentuelles.

Une segmentation rythmique aussi basique est évidemment limitée, mais elle fournit un point de départ pour modéliser le rythme qui ne requiert aucune connaissance des structures rythmiques des langues.

Modélisation statistique du rythme : Pour chaque langue, un Modèle de Mélange de lois Gaussiennes est appris, soit en utilisant l'algorithme standard LBG ou l'algorithme LBG-Rissanen qui fournit automatiquement le nombre optimal de composantes du mélange.

4. EXPÉRIENCES

4.1. Corpus

Les expériences sont faites sur le corpus MULTEXT [1]. Cette base de données contient des enregistrements de cinq langues européennes (anglais, français, allemand italien et espagnol), énoncés par 50 locuteurs différents (5 hommes et 5 femmes par langue).

Les données consistent en passages lus d'environ 5 phrases extraites du corpus EUROM1 (la durée moyenne de chaque passage est 20.8 secondes). Le contour de la fréquence fondamentale est aussi disponible. Une limitation est que les même textes sont prononcés en moyenne par 3.75 locuteurs, ce qui résulte dans une possible dépendance au texte des modèles.

A cause de la taille limitée du corpus, les expériences en identification des langues sont effectuées avec une procédure de validation croisée : 9 locuteurs sont employés pour l'apprentissage des modèles d'une langue pendant que le dixième est utilisé pour les tests. Cette procédure est répétée pour chaque locuteur, et pour chaque langue.

4.2. Modélisation du Rythme

Le tableau 1 résume les expériences effectuées avec les paramètres rythmiques. Les scores d'identification affichés sont moyennés selon plusieurs topologies de Modèles de Mélange de lois Gaussiennes et obtenus en utilisant la totalité des fichiers de test (environ 21 secondes).

Tableau 1 - Résultats moyens en validation croisée de la modélisation du rythme.

Paramètres	Taux d'identification moy.
D_V + D_C	64.8 %
D_V + D_C + N_C	70.0 %
D_V + D_C + N_C + E	75.0 %
D_V + D_C + N_C+ E + F_0	69.4 %

L'utilisation des paramètres de durée D_V et D_C résulte en un taux d'identification correcte de 64.8 %. L'utilisation de paramètres additionnels relatifs à la complexité de la structure de la pseudo-syllabe (N_C) et à l'accentuation (E) améliore significativement les résultats, permettant de réaliser 75 % d'identification correcte.

Au contraire, l'ajout du paramètre F_0 n'améliore pas les résultats. Ce résultat peut vouloir dire qu'une valeur statique de F_0 par pseudo-syllabe, même normalisée, n'est pas suffisamment représentative pour être utile.

Dans une autre expérience (voir Figure 2), l'influence de la durée des énoncés de test est observée. La modélisation est effectuée dans l'espace à quatre dimensions (D_V + D_C + N_C + E) qui est le plus performant.

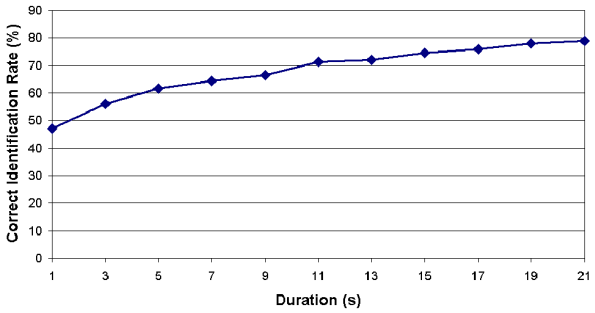

Figure 2 - Taux d'identification correcte en fonction de la durée des énoncés de test (Modèle rythmique).

Le taux d'identification augmente avec la durée des énoncés de test pour arriver au meilleur résultat : 78% avec des énoncés de 21 secondes.

Toutefois, même avec des phrases de test courtes (moins de dix pseudo-syllabes), les résultats sont bien supérieurs à la chance. De plus, utiliser uniquement la première seconde de la phrase donne un taux d'identification correcte de 47 % (à comparer avec la chance : 20%).

Comme on le voit sur le tableau 2 suivant, les similitudes rythmiques entre les différentes langues sont respectées [6] : on observe des confusions entre l'anglais et l'allemand, ainsi qu'entre l'italien et l'espagnol.

Tableau 2 - Matrice de confusion pour le modèle rythmique avec les énoncés de 21 secondes.

	EN	FR	GE	IT	SP
EN	**62**	4	16	11	7
FR	0	**100**	0	0	0
GE	11	1	**86**	2	0
IT	10	1	3	**62**	23
SP	1	4	0	3	**91**

Taux d'identification correcte = 79 %

4.3. Modélisation des Systèmes Vocaliques

Comme exposé sur la Figure 3, la modélisation du système vocalique est performante sur le corpus MULTEXT. Un taux d'identification correcte de 51 % est obtenu avec une seconde de signal. Augmenter la durée des énoncés de test permet d'obtenir 70 % d'identification correcte pour 21 secondes.

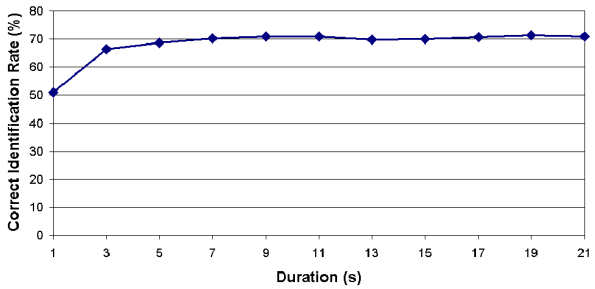

Figure 3 - Taux d'identification correcte en fonction de la durée des énoncés de test (Modèle de Système Vocalique).

Sur le tableau 3, on peut observer que des confusions ont lieu entre différentes langues (anglais/italien et italien/espagnol principalement). Ces regroupements sont en partie distincts de ceux obtenus avec le modèle rythmique. Une fusion entre les informations fournies par les deux modèles peut démontrer la complémentarité des approches.

Tableau 3 - Matrice de confusion pour le modèle vocalique avec les énoncés de 21 secondes.

	EN	FR	GE	IT	SP
EN	**44**	0	0	38	18
FR	0	**92**	1	1	6
GE	2	0	**96**	2	0
IT	30	0	0	**46**	24
SP	5	10	0	13	**72**

Taux d'identification correcte = 70 %

4.4. Fusion entre les Modèles Rythmiques et Vocaliques

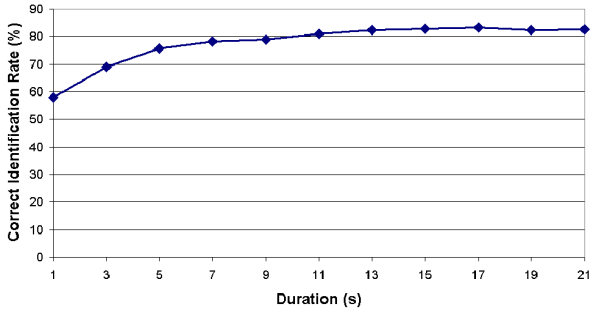

Figure 4 - Taux d'identification correcte en fonction de la durée des énoncés de test pour la fusion des deux approches.

Une simple fusion statistique est effectuée en additionnant les log-vraisemblances obtenues avec les modèles rythmiques et vocaliques.

Tableau 4 - Matrice de confusion pour la fusion des deux modèles avec les énoncés de 21 secondes.

	EN	FR	GE	IT	SP
EN	**67**	1	3	24	6
FR	0	**99**	0	0	1
GE	0	0	**100**	0	0
IT	22	0	0	**66**	12
SP	3	6	0	10	**81**

Taux d'identification correcte = 83 %

Les scores obtenus pour des énoncés de taille variable sont montrés sur la Figure 4. Le taux d'identification correcte dépasse les 80 % pour des énoncés de test d'une durée supérieure à 11 secondes. Fusionner les deux approches permet d'obtenir un maximum de 83 % d'identification correcte pour des énoncés de test de 21 secondes.

5. DISCUSSION

Nous proposons dans cet article deux algorithmes dédiés à l'identification automatique des langues. Les expériences, effectuées en validation croisée, montrent qu'il est possible de modéliser efficacement le rythme (78 % d'identification correcte) d'une manière qui ne requiert aucun savoir a priori sur la structure rythmique des langues. Le modèle de système vocalique permet d'obtenir 70 % d'identification correcte sur des fichiers relativement courts.

La fusion des deux approches améliore le taux d'identification jusqu'à 83 %. Plus precisément, les confusions entre les langues sont moindres, notamment pour la discrimination entre l'anglais et l'allemand.

Les résultats de la fusion montrent l'interêt d'employer de tels paramètres pour améliorer le temps de calcul : à partir de seulement 11 secondes de parole, le taux d'identification correcte dépasse les 80 %.

RÉFÉRENCES

[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., & Mehler, J., *Correlates of linguistic rhythm in the speech signal*, Cognition, 73(3), 265-292, 1999.

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

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

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

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

[9] Pellegrino, F. and André-Obrecht, R., *Automatic Language Identification: an Alternative Approach to Phonetic Modeling*, Signal Processing, 80, 2000.

[10] 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.

[11] Campione, E., and Véronis, J., *A multilingual prosodic database*, Proc. of ICSLP'98, Sidney, 1998.