Automatic extraction of prosodic features for automatic language identification
Extraction automatique de paramètres prosodiques pour l'identification automatique des langues

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

[1] {jerome.farinas, rouas, obrecht}@irit.fr

Université Paul Sabatier ; Équipe SAMOVA, IRIT UMR 5505 ; F-31062 Toulouse Cedex 9

tél : +33 561557434 ; fax : +33 561556258

[2] Francois.Pellegrino@univ-lyon2.fr

DDL UMR 5596 - ISH ; 14, avenue Berthelot ; F-69363 Lyon Cedex 7

tél : +33 472726494 ; fax : +33 472726590

20 mai 2005

abstract

The aim of this study is to propose a new approach to Automatic Language Identification: it is based on rhythmic modelling and fundamental frequency modelling and does not require any hand labelled data. First we need to investigate how prosodic or rhythmic information can be taken into account for Automatic Language Identification. A new automatically extracted unit, the pseudo syllable, is introduced. Rhythmic and intonative features are then automatically extracted from this unit. Elementary decision modules are defined with gaussian mixture models. These prosodic modellings are combined with a more classical approach, a vocalic system acoustic modelling.

Experiments are conducted on the five European languages of the MULTEXT corpus: English, French, German, Italian and Spanish. The relevance of the rhythmic parameters and the efficiency of each system (rhythmic model, fundamental frequency model and vowel system model) are evaluated. The influence of these approaches on the performances of automatic language identification system is addressed. We obtain 91 % of correct identification with 21 s. utterances using all the information sources.

résumé

Le but de cette étude est de proposer une nouvelle approche pour l'identification automatique des langues, basée sur une modélisation du rythme, ne nécessitant pas de données étiquetées manuellement. Il faut tout d'abord savoir comment apporter des informations sur la prosodie, le rythme pour l'identification automatique des langues. Pour répondre à cette question nous avons introduit une nouvelle unité, la pseudo-syllabe, qui est automatiquement extraite. Des paramètres rythmiques et intonatifs sont alors calculés à partir de cette unité. Des modèles élémentaires pour chaque types de paramètres sont définis en utilisant des mélanges de lois gaussiennes. Ces modélisations de la prosodie sont couplées à une approche plus classique utilisant une modélisation acoustique des systèmes vocaliques. Les expériences sont menées sur les cinq langues européennes du corpus MULTEXT. L'intérêt des paramètres rythmiques, et l'efficacité de chaque système (modèle rythmique, modèle de la fréquence fondamentale et modèle vocalique) sont évalués. L'impact de ces approches sur les performances d'identification est analysé. Nous obtenons des résultats de 91 % d'identification correcte avec des fichiers de 21 secondes.

keywords

Automatic language identification, prosody, rhythm, fundamental frequency, Gaussian Mixture Models.

mots clefs

Identification automatique des langues, prosodie, rythme, fréquence fondamentale, mélange de lois gaussiennes.

1 introduction

Un système d'Identification Automatique des Langues (IAL) a pour but de déterminer l'identité de la langue parlée à partir d'un échantillon de parole. C'est un thème apparu il y a près de trente ans, mais c'est seulement depuis le début des années 1990 que la recherche s'est intensifiée dans ce domaine. Parmi les raisons de cette émergence, on peut citer :

– la croissance de la demande pour des Interfaces Homme-Machine,

– l'expansion des communications parlées dans un cadre multilingue.

L'ère actuelle est une ère de communication multilingue, qu'il s'agisse de communications entre humains ou entre humains et machines. Ce constat implique le développement d'applications capables de gérer plusieurs langues et/ou d'identifier une langue parmi d'autres. Une des applications possibles pour un tel système est la discrimination des langues en amont de systèmes de dialogue multilingue : le système a besoin de savoir quelle langue est parlée pour pouvoir comprendre et répondre à l'utilisateur. On peut distinguer deux approches possibles pour déterminer l'identité de la langue et dialoguer dans celle-ci : exécuter en parallèle autant de systèmes de reconnaissance de la parole que de langues offertes par le système de dialogue, ou bien utiliser un système dédié à l'identification de la langue, qui permet de lister rapidement les langues les plus probables qui sont ensuite départagées par les systèmes de reconnaissance de la parole adaptés aux langues pré-sélectionnées. La dernière approche, en considérant les contraintes de temps-réel d'un tel système, permet la prise en compte d'un nombre bien plus important de langues.

A ces enjeux applicatifs s'ajoutent des motivations linguistiques fortes liées d'une part à la perception de la parole et d'autre part à la notion de « distance linguistique ».

La notion de distance entre langues est au cœur de l'analyse linguistique depuis un siècle. Les systèmes d'IAL peuvent permettre d'apporter un nouvel éclairage à cette notion, en particulier en corrélant les distances linguistiques traditionnelles aux distances perceptuelles et automatiques. Ceci amènera éventuellement à repenser les typologies de langues et du même coup à expliciter les notions proches de langues et de dialectes.

L'identification automatique des langues est un domaine de recherche qui a connu de grandes avancées lors des campagnes d'évaluation NIST de 1993 à 1996 [1, §4]. Les systèmes d'IAL conçus à cette époque permettaient de discriminer une dizaine de langues sur de courts échantillons sonores (environ 45 s) avec un taux d'erreur de l'ordre de 10 % sur un corpus de parole téléphonique de 11 langues OGI-MLTS [2]. Depuis lors, les progrès réalisés sont principalement liés à l'utilisation d'une nouvelle paramétrisation acoustique (coefficients de type dérivée du cepstre) et l'amélioration des méthodes de classification [3, 4]. Parmi les informations les plus pertinentes pour identifier une langue, les informations phonétiques et phonotactiques (règles d'enchaînement des sons d'une langue) sont utilisées par ces systèmes et les informations prosodiques (intonation, rythme, accentuation) sont souvent négligées [5]. Les systèmes exploitant les sources d'information phonétiques et phonotactiques se basent en général sur les décodeurs acoustico-phonétiques : introduit par Lamel et Gauvain pour le français et l'anglais [6], ce type de système a été généralisé en mettant plusieurs décodeurs monolingues en parallèle par la suite [7]. Les modèles phonétiques et phonotactiques sont issus de la reconnaissance automatique de la parole, domaine bénéficiant de plusieurs décennies d'investissement intellectuel : ces modèles bénéficient d'un meilleur niveau des connaissances et de meilleures formalisations [8], [9]. Pourtant, des expériences perceptives [10, 11, 12] ont montré que l'oreille humaine permet d'identifier les langues à partir de leur seule prosodie mettant ainsi en avant le fort pouvoir discriminant de ces traits et l'intérêt manifeste de leur exploitation dans des systèmes d'IAL.

L'enjeu des systèmes automatiques est actuellement double : faciliter l'extension du système par l'ajout de nouvelles langues et utiliser le moins possible de données annotées manuellement. Par exemple, l'apprentissage des décodeurs acoustico-phonétiques nécessite de disposer d'un minimum d'étiquetage phonétique sur des corpora audio ; ce travail est long à produire et doit être réalisé par des spécialistes. Il est donc nécessaire d'exploiter le plus de sources d'informations possibles, tout en évitant celles qui demandent des ressources trop délicates à acquérir pour effectuer l'apprentissage des modèles sous jacents. D'où notre intérêt pour développer une approche nouvelle visant à extraire et à exploiter des paramètres basés sur le rythme, l'une des principales composantes de la prosodie, sans avoir recours à des données étiquetées manuellement. Après avoir expliqué plus en détail les motivations dans la section suivante, nous abordons les modélisations proposées, puis nous les évaluons par des expériences sur un corpus multilingue dans la section 5. Les résultats sont discutés dans la section 6.

2 motivations

D'un point de vue acoustique, la prosodie désigne les phénomènes liés à la variation dans le temps des paramètres de hauteur (liée à la fréquence fondamentale, fréquence de vibration des cordes vocales), d'intensité (liée à l'amplitude et à l'énergie) et de durée des sons. D'un point de vue perceptuel, la variation dans le temps de ces paramètres correspond à la perception de l'intonation des phrases, de l'accentuation et du rythme. Il s'agit de caractéristiques *supra-segmentales*, par opposition aux caractéristiques segmentales liées à la réalisation des phonèmes des langues.

Parmi ces informations, le rythme est indéniablement un paramètre caractéristique de la langue jouant un rôle important dans la communication parlée. On peut citer par exemple son rôle dans l'acquisition du langage [13] ou dans la théorie « Frame-Content » de l'origine du langage [14], cadre dans lequel la parole aurait émergé à partir du cycle d'oscillation mandibulaire fournissant un patron syllabique de type CV

(Consonne-Voyelle). La notion de classes rythmiques est issue, quant à elle, de la théorie de l'isochronie, introduite par Pike [15], développée par Abercrombie [16], puis par Ladefoged [17], qui, aux deux catégories initialement définies (langues syllabiques et langues accentuelles) ajoute la classe des langues de type moraïque, comme le japonais. Pour chacune de ces classes, syllabique, accentuelle et moraïque, la théorie de l'isochronie prévoit qu'un certain motif (syllabe, accentuation ou mora) se répète à intervalles de temps réguliers. Toutefois, des études plus récentes, basées sur la mesure de la durée des intervalles entre les accentuations à la fois pour les langues syllabiques et les langues accentuelles, laissent entrevoir un schéma alternatif dans lequel ces catégories discrètes sont remplacées par un continuum [18]. Les différences rythmiques entre les langues sont alors pour la plupart expliquées par la structure syllabique et la présence (ou l'absence) de réduction vocalique.

Ces controverses, si elles soulignent la difficulté à définir et à quantifier ce qu'est le rythme, signifient également que si l'on arrive à extraire des informations pertinentes, elles peuvent se révéler efficaces pour distinguer les langues. Par ailleurs, une étude récente [19] a montré que des paramètres basés sur le rythme syllabique, la durée syllabique, ainsi que des descripteurs de contours de l'amplitude et de la courbe mélodique, permettent de discriminer automatiquement des langues présentées par paires. D'autres modèles basés sur le calcul de statistiques à partir d'une segmentation manuelle en consonnes et voyelles pour huit langues, ont permis de faire émerger dans un espace de paramètres continu des groupes de langues distincts, correspondant globalement aux classes rythmiques traditionnelles [20]. Plusieurs études complémentaires ont confirmé l'importance des classes rythmiques dans l'identification des langues, à la fois sur le plan automatique (modélisation neuro-mimétique des séquences temporelles de consonnes et de voyelles [21]) et sur le plan de la perception (discrimination perceptuelle par des nourrissons [22]).

Les objectifs poursuivis dans les expériences rapportées ici sont multiples. Il s'agit en premier lieu d'extraire *automatiquement* et de manière entièrement non supervisée des informations rythmiques pertinentes pour l'IAL. Cela implique la définition d'une

unité rythmique et des paramètres adéquats, ainsi que la conception de modèles adaptés à ces paramètres. En relevant ce défi, les retombées attendues sont une amélioration des performances des systèmes d'IAL et une meilleure caractérisation du rythme des langues.

3 pseudo-syllabe et IAL

L'information nécessaire pour identifier la langue peut être segmentale ou prosodique (i.e. supra-segmentale). Les informations acoustiques, caractérisant la manière dont les locuteurs réalisent les phonèmes, relèvent de la partie segmentale. L'information prosodique est encodée dans les variations de fréquence fondamentale, d'intensité et de durée qui apparaissent au delà des segments. Bien que les informations segmentales et prosodiques puissent également refléter des contraintes linguistiques de haut niveau (telles que l'expression de la modalité des phrases), nous faisons l'hypothèse que cette information de haut niveau n'est pas nécessaire pour identifier la langue ou tout au moins pour faire émerger des classes rythmiques pertinentes pour identifier la langue.

De part sa nature universelle, la syllabe est une unité privilégiée pour la modélisation du rythme (même si le rôle de cette unité est sujet au débat [23]).

Malheureusement, segmenter le signal de parole en syllabes est une tâche spécifique à chaque langue, ce qui rend la conception d'un algorithme indépendant de la langue complexe.

Pour cette raison et compte tenu de notre expérience préliminaire en segmentation du signal de parole [24], nous avons introduit la notion de pseudo-syllabe [25]. Cette dénomination est liée au fait que cette unité est proche de la structure syllabique la plus fréquente au monde, à savoir la structure Consonne/Voyelle (CV) [26]. Nous avons basé notre système d'IAL sur cette unité.

3.1 description de la pseudo-syllabe

Pour obtenir une segmentation automatique en pseudo-syllabes, les signal de parole est tout d'abord découpé en une suite de zones quasi stationnaires appelées segments (§3.2.1) ; puis les segments de non parole sont écartés en utilisant une détection d'activité vocale (§3.2.2) et enfin les segments vocaliques sont détectés (§3.2.3). Un motif de pseudo-syllabe correspond à une structure $.C^nV.$ (où n est un entier qui peut être nul) ; V résulte éventuellement du regroupement de plusieurs segments vocaliques consécutifs, tandis que C correspond à un seul segment de parole non vocalique. Il est extrêmement rare de rencontrer deux segments VV adjacents représentant deux voyelles différentes : lorsqu'il existe deux ou plusieurs voyelles adjacentes, un court segment (de l'ordre de 20 ms) est détecté au niveau de la zone transitoire et est étiqueté C. La succession de deux voyelles donne donc au minimum la suite de segments VCV.

Par exemple, la séquence (CCVVCCVCVCCCVCVCCC) obtenue après le découpage et la détection de segments vocaliques (cf. figure 1), est partitionnée en cinq pseudo-syllabes : (CCV.CCV.CV.CCCV.CV). La séquence .CCC. finale, ne contenant pas de voyelles, est écartée.

3.2 extraction automatique de la pseudo-syllabe

Comme nous l'avons annoncé précédemment, la pseudo-syllabe est le résultat de trois traitements automatiques, que nous allons décrire ci-après. La figure 1 illustre le résultat de cette extraction. Leur but est pour l'essentiel de détecter les noyaux vocaliques.

3.2.1 segmentation de la parole

Notre approche, introduite en [27], se base sur la détection de ruptures dans le signal acoustique, supposé être décrit par une suite de zones quasi-stationnaires, estimées par des modèles auto-régressifs gaussiens (AR).

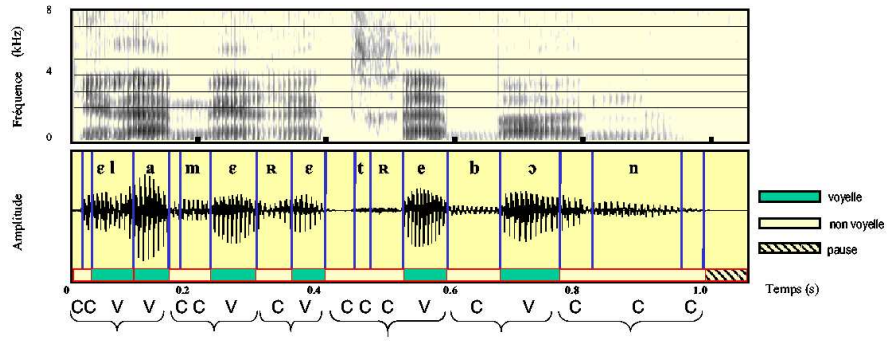


FIG. 1 – Exemple d'extraction de pseudo-syllabes sur la phrase « ...et la mer est très bonne ».

Deux modèles AR M_0et M_1 sont estimés à chaque instant à partir du signal acoustique (cf. figure 2). Le premier modèle est calculé sur une fenêtre de longueur croissante, débutant à l'instant de la rupture précédente. Il permet d'établir un modèle adaptatif du segment courant. Le modèle M_1est quant à lui estimé sur une fenêtre courte glissante : il permet d'établir un modèle de l'événement courant, c'est-à-dire de la trame de signal étudié. Lorsque la distance statistique entre ces deux modèles diverge au delà d'un seuil fixé, on considère que l'événement modélisé par M_1 ne correspond plus à la zone homogène modélisée par M_0 : il y a donc rupture. Le critère de rupture est calculé par un test statistique basé sur la divergence de Kullback qui mesure l'entropie mutuelle entre deux lois conditionnelles correspondant à deux modèles AR. La détection de la rupture est alors liée à un changement de pente de statistique.

Cet algorithme de détection a été mis à l'épreuve sur de nombreux corpus afin d'étudier sa robustesse face à différents environnements (parole propre, bruitée, téléphonique), face à diverses fréquences d'échantillonnage (ordre des modèles AR) et au changement de locuteurs [24]. Il est apparu qu'une version robuste de l'algorithme utilisant de simples modèles AR d'ordre 2, donnait d'excellents résultats en reconnaissance de parole [28]. Cette version est employée dans le cadre de cette étude.

Les unités détectées sont de nature infra phonétiques et peuvent être regroupées en trois classes :

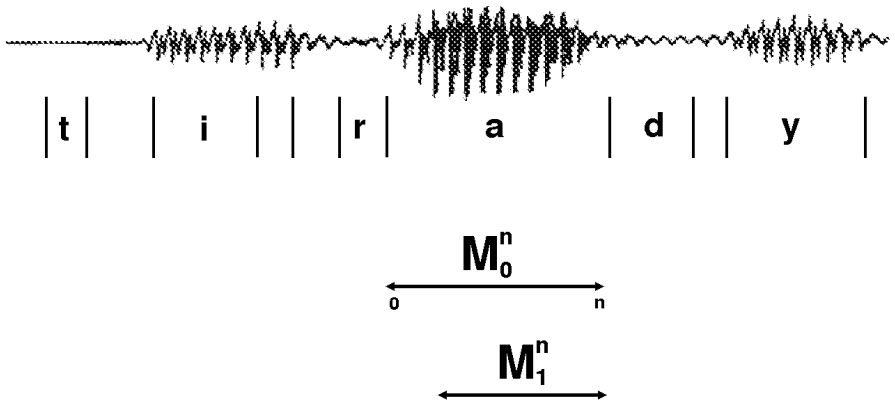


FIG. 2 – Localisation des fenêtres d'estimation des modèles M_0 et M_1 à l'instant n, l'instant 0 correspondant à la dernière frontière validée, l'instant n correspondant à la frontière en cours de validation ; la phrase prononcée est « [Il se garan]tira du... » (exemple extrait de [24]).

- des segments courts (de longueur inférieure à 20 ms) appelés segments événementiels. Ils correspondent à des gestes articulatoires brefs ou à des chevauchements de ces gestes (amortissement de la structure formantique lors de la fermeture du conduit vocal, explosion d'une occlusive...),
- des segments transitoires entre deux phonèmes,
- des segments quasi-stationnaires qui matérialisent la partie stable des sons, en particulier la partie centrale d'une voyelle, la barre de voisement d'occlusives voisées...

3.2.2 détection de l'activité vocale

Nous avons implémenté un détecteur d'activité basé sur une analyse statistique du premier ordre du signal temporel. On note N le nombre de segments issus de la segmentation automatique et $\{S_1, ..., S_N\}$ la suite de ces segments. Le seuil d'activité S_a est défini par :

$$S_a = \alpha \min_i \left(var\left(S_i\right)\right) \tag{1}$$

où le coefficient α a été expérimentalement fixé à 2,5 [29]. Les segments ayant une variance inférieure à S_a sont étiquetés comme étant des silences. On distingue les silences signalant une absence d'activité de parole (segments longs), des silences se produisant en cours de locution (les silences des occlusives, les pauses courtes). Un post-traitement permet de regrouper les segments de non-activité en cas de sur-segmentation : si plusieurs segments courts sont étiquetés comme étant des silences et ont une durée totale supérieure à 150 ms, on considère qu'il s'agit d'une zone de non activité.

3.2.3 localisation des voyelles

La détection des voyelles est basée sur une analyse fréquentielle du signal de parole. Le critère Rec (*Reduced Energy Cumulating*) [30], défini pour chaque trame t du signal à partir d'une analyse spectrale selon l'échelle de fréquence MEL, se présente de la manière suivante :

$$Rec(t) = \frac{E_{BF}(t)}{E(t)} \sum_{i=1}^{24} \alpha_i \left(E_i(t) - \overline{E}(t) \right)^+ \tag{2}$$

où t est le numéro de la trame du signal, $E_i(t)$ est l'énergie dans le i^{eme} filtre selon l'échelle MEL, $\overline{E}(t)$ est la moyenne de l'énergie sur les 24 bandes spectrales (échelle MEL) dans la bande spectrale 350-3500 Hz, α_i est le poids affecté au $i^{ième}$ filtre, $E(t)$ est l'énergie totale calculée pour une trame t :

$$E(t) = \sum_{i=1}^{24} \alpha_i E_i(t) \tag{3}$$

et $E_{BF}(t)$ est l'énergie dans la bande de fréquences inférieures à 1 KHz :

$$E_{BF}(t) = \sum_{i=1}^{10} \alpha_i E_i(t) \tag{4}$$

La grande majorité des pics de la fonction $Rec(t)$ correspond aux voyelles du signal. Les segments contenant un pic de la fonction sont étiquetés "noyau vocalique".

Nous avons évalué ce détecteur sur un corpus difficile, le corpus OGI-MLTS en calculant le taux d'erreur (VER : Vowel Error Rate) déterminé suivant la formule :

$$VER = \frac{suppressions + insertions}{total\,voyelles} * 100 \quad (5)$$

Le tableau 1 compare différentes méthodes d'extraction de voyelles. Le taux d'erreur moyen de notre détecteur est de 22,9 % ; il est correct comparé aux autres systèmes ; il est à noter qu'aucun apprentissage spécifique à une langue n'a été réalisé.

TAB. 1 – Comparaison de différents algorithmes de détection de voyelles

Algorithmes	Corpus	Type	Langues	VER
Pfitzinger et al., 1996 [31] (*)	PHONDATII	texte lu	allemand	12,9 %
	VERBMOBIL	parole spontanée	allemand	21,0 %
Fakotakis et al., 1997 [32]	TIMIT	texte lu	anglais	32,0 %
Pfau et Ruske, 1998 [33]	VERBMOBIL	parole spontanée	allemand	22,7 %
Howitt, 2000 [34]	TIMIT	texte lu	anglais	29,5 %
Pellegrino et Obrecht, 1997 [35]	OGI-MLTS	parole spontanée	français	19,5 %
			japonais	16,3 %
			coréen	28,5 %
			espagnol	19,2 %
			vietnamien	21,1 %
			moyenne	*22,9 %*

(*) : dans cette étude le taux d'erreur est exprimé en fonction du noyau syllabique et non pas explicitement les voyelles.

3.2.4 localisation de la pseudo-syllabe

A l'issue du processus de localisation automatique des noyaux vocaliques, les segments étiquetés "noyau vocalique" adjacents sont regroupés en un seul segment étiqueté V ; tous les autres segments de parole sont conservés et étiquetés C. Le regroupement des segments étiquetés V est réalisé de manière à corriger la sursegmentation induite par l'algorithme utilisé en 3.2.1. Il est extrêmement rare d'avoir deux segments adjacents étiquetés V et correspondants réellement à deux voyelles ; ils sont généralement séparés par une zone transitoire étiquetée C. Le signal de parole est désormais

une suite de segments étiquetés Consonne (C) ou Voyelle (V), qui se traduit en une suite de pseudo-syllabes dont les frontières sont placées systématiquement après les segments V.

3.3 un système d'identification des langues basé sur la pseudo-syllabe

L'introduction de la pseudo-syllabe permet d'utiliser trois types de connaissances pour identifier les langues :

- les connaissances acoustiques avec la caractérisation du noyau vocalique, cœur de la pseudo-syllabe ; des travaux antérieurs ont montré l'importance de l'espace acoustique vocalique dans la discrimination des langues [30].
- le rythme, directement lié à la durée de la pseudo-syllabe et aux durées relatives des segments qui la composent.
- l'intonation au travers des valeurs prises par la fréquence fondamentale et ses variations au cours de la pseudo-syllabe.

Cette approche induit comme échelle d'analyse temporelle des observations, l'échelle de la pseudo-syllabe. La pseudo-syllabe apparaît alors comme une unité acoustique, une unité prosodique et une unité rythmique, pour lesquelles il faut préciser paramétrisations et modélisations.

4 architecture générale du système

Le système d'identification des langues repose sur une approche statistique unifiée. Il opère en trois étapes : un prétraitement qui extrait des vecteurs de paramètres pertinents du signal de parole pour chaque type de connaissances étudiées, un module de décision statistique mettant en jeu de manière indépendante les modélisations des trois types de connaissances et une fusion des informations de décision. La figure 3 présente un schéma synoptique de ce système.

Les modélisations mises en place sont :

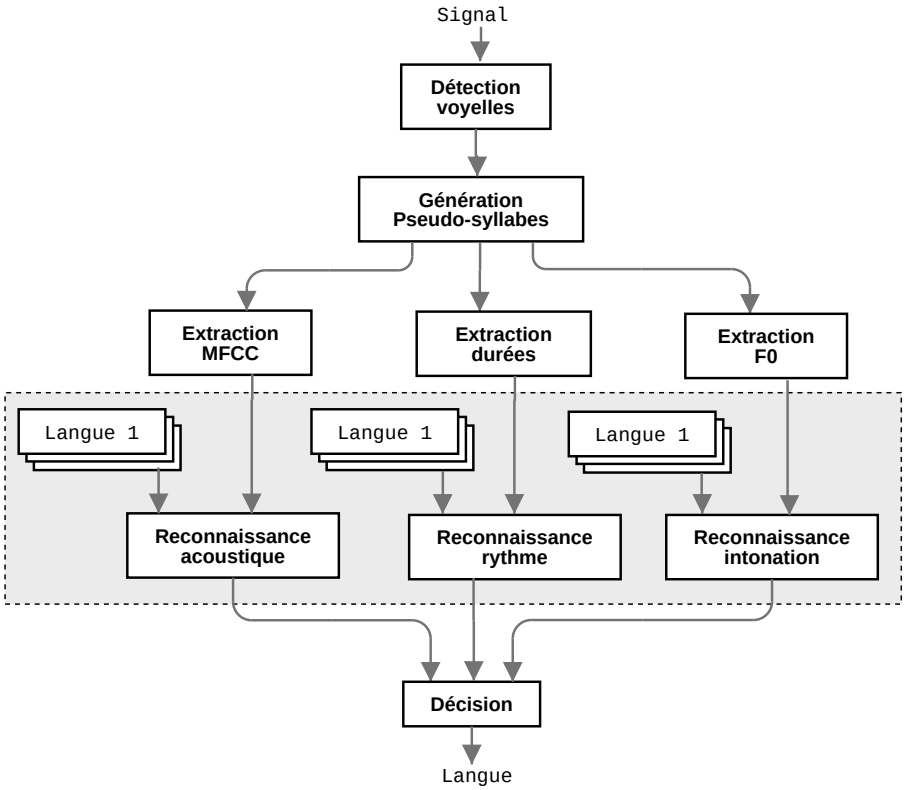


FIG. 3 – Schéma synoptique du système global d'identification des langues.

– une modélisation acoustique du système vocalique (cf. §4.2), visant à apporter un référentiel pour quantifier ultérieurement l'apport des modélisations prosodiques,
– une modélisation du rythme (cf. §4.3.1), basée sur l'exploitation de paramètres rythmiques dérivés de la segmentation en "pseudo-syllabes" et de sa composition,
– une modélisation de l'intonation (cf. §4.3.2), au travers de paramètres calculés à partir de la fréquence fondamentale.

Ces deux dernières modélisations fournissent le cadre statistique pour traiter l'information prosodique comme nous allons le préciser ci-après.

4.1 cadre statistique

Soit $L = \{L_1, L_2, ..., L_{N_L}\}$ l'ensemble des N_L langues à identifier ; le problème est de trouver la langue la plus probable L^* dans l'ensemble L. Deux types d'informations sont extraites du signal : les informations acoustiques et les informations prosodiques ; l'information prosodique se décomposant elle-même en deux composantes.

Si np représente le nombre de pseudo-syllabes extraites automatiquement du signal de parole à identifier, nous noterons :

- $V = \{v_1, v_2, ..., v_{np}\}$ la suite de np vecteurs représentant les paramètres acoustiques de chaque pseudo syllabe,
- $P = \{p_1, p_2, ..., p_{np}\}$, la suite des np vecteurs de paramètres liés à l'information prosodique.

Dans le cadre de l'approche bayésienne classique, la langue la plus probable L^* est définie par l'équation suivante :

$$L^* = arg \max_{1 \leq i \leq N_L} (Pr(L_i|V, P)) \tag{6}$$

qui peut-être également écrite sous la forme :

$$L^* = arg \max_{1 \leq i \leq N_L} (Pr(V|P, L_i)Pr(P|L_i)Pr(L_i)) \tag{7}$$

En considérant que les probabilités *a priori* des langues $Pr(L_i)$ sont identiques et que l'information acoustique est indépendante de l'information prosodique, l'équation 7 devient :

$$L^* = arg \max_{1 \leq i \leq N_L} (Pr(V|L_i)Pr(P|L_i)) \tag{8}$$

où $Pr(V|L_i)$ est obtenue à partir d'un modèle acoustique et $Pr(P|L_i)$ à partir d'un modèle prosodique.

Le modèle acoustique fait référence aux réalisations acoustiques des différents sons des différentes langues. Nous ne nous intéresserons par la suite qu'aux réalisations liées aux voyelles afin de se placer dans un espace acoustique homogène (cf. modélisation par approches différenciées [36]). Le modèle prosodique représente les différences de structures prosodiques entre les différentes langues, au travers des variations de rythme et d'intonation.

4.2 la modélisation acoustique du système vocalique

L'information acoustique est contenue dans le vecteur $V = \{v_1, v_2, ..., v_{np}\}$, où v_k est le vecteur acoustique représentant la $k^{ième}$ pseudo-syllabe du signal. Chaque vecteur v_k est un vecteur cepstral de type MFCC (Mel Frequency Cepstral Coefficients), calculé sur le segment vocalique de la pseudo-syllabe. Pour chaque segment vocalique, une fenêtre de 20 ms est centrée sur l'instant correspondant à la valeur maximale du critère Rec (équation (2)) : 8 MFCC sont extraits ainsi que les dérivées de ces coefficients calculées sur 5 fenêtres adjacentes centrées sur ce même instant. Pour s'affranchir des conditions d'enregistrement où le bruit est supposé convolutif, nous effectuons une classique soustraction cepstrale sur les MFCC [37]. Il s'en suit que v_k est un vecteur de dimension 16.

Afin de calculer la vraisemblance des observations acoustiques, à savoir :

$$Pr(V|L_i) \tag{9}$$

nous supposons que l'influence d'une voyelle sur les voyelles adjacentes est réduite, ce que nous traduisons d'un point de vue probabiliste par le fait que chaque vecteur acoustique qui compose V est indépendant des autres, conditionnellement à la langue. L'équation 9 devient alors :

$$Pr(V|L_i) = \prod_{k=1}^{np} p(v_k|L_i) \tag{10}$$

Pour chaque langue considérée, le modèle acoustique est modélisé par un mélange de lois gaussiennes, à savoir que la densité de probabilité acoustique d'un vecteur v_k, notée voc s'exprime sous la forme suivante :

$$voc(v_k|L_i) = \sum_{j=1}^{Q_i} \frac{\alpha_j^i}{(2\pi)^{d/2}\sqrt{\left|\Sigma_j^i\right|}} \exp\left(-\frac{1}{2}\left(v_k - \mu_j^i\right)^t \left(\Sigma_j^i\right)^{-1}\left(v_k - \mu_j^i\right)\right) = N(\mu^i, \Sigma^i, Q_i) \tag{11}$$

où Q_i est le nombre de composantes du mélange de lois gaussiennes, (μ_j^i, Σ_j^i) représentent les paramètres de la loi gaussienne j pour la langue L_i et d est la dimension de l'espace acoustique (en l'occurence d=16).

Les paramètres des lois gaussiennes sont appris en utilisant l'algorithme EM (Expectation-Maximisation). Le nombre Q_i de composantes du modèle est estimé de manière automatique en utilisant l'algorithme LBG-Rissanen (cf. [38] pour plus de détails).

4.3 la modélisation de la prosodie

Le modèle prosodique a pour objectif de capturer les différences qui existent au niveau des structures prosodiques des phrases. Pour cela, notre modèle prend en compte les paramètres liés à l'intonation, plus précisément des paramètres statistiques caractéristiques de la fréquence fondamentale (notés F), et des paramètres de durée extraits des pseudo-syllabes (notés Ψ) pour caractériser le rythme. Par conséquent :

$$Pr(P|L_i) = Pr(F, \Psi|L_i) = Pr(F|\Psi, L_i)Pr(\Psi|L_i) \tag{12}$$

où $Pr(\Psi|L_i)$ correspond au modèle lié au rythme des pseudo-syllabes, et $Pr(F|\Psi, L_i)$ au modèle lié à la fréquence fondamentale.

4.3.1 la modélisation du rythme

L'expression $Pr(\Psi|L_i)$ capture les informations sur la durée des pseudo-syllabes. Bien que le rythme puisse apporter des informations sur l'accentuation au niveau

syntaxique, nous ne nous intéressons ici qu'à sa caractérisation locale définie sur la "pseudo-syllabe".

Après détection automatique de chaque pseudo-syllabe, sont extraits sur chacune d'elles :

- la durée totale (en ms) des segments consonantiques, notée D_c,
- la durée totale (en ms) du segment vocalique, notée D_v,
- la complexité de la pseudo-syllabe mesurée par N_c le nombre de segments consonantiques.

La répartition des distributions de ces paramètres sur les cinq langues du corpus d'apprentissages se trouve en annexe 2.

Nous faisons l'hypothèse que chaque segment prosodique est indépendant des autres de manière à simplifier le modèle. Avec cette hypothèse d'indépendance, le modèle rythmique devient :

$$Pr(\Psi|L_i) = \prod_{k=1}^{np} Pr(\Psi_k|L_i) \tag{13}$$

où

$$\Psi_k = (D_c, D_v, N_c)_k \tag{14}$$

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

Pour chaque langue, le modèle rythmique est un modèle continu de mélange de lois gaussiennes dans l'espace réel de dimension 3. Pour initialiser l'algorithme d'apprentissage de type EM, nous avons utilisé soit l'algorithme standard LBG, soit l'algorithme LBG-Rissanen pour définir automatiquement le nombre optimal de composantes du mélange. Nous noterons $ryt(\Psi_k|L_i)$ la densité de probabilité rythmique associée à la langue L_i prise en Ψ_k.

4.3.2 la modélisation de l'intonation

Pour exprimer $Pr(F|\Psi, L_i)$, nous faisons l'hypothèse que d'une part, l'information rythmique est indépendante de l'information liée à l'intonation et que d'autre part, chaque segment est indépendant des autres. Cette hypothèse est fausse dans le cadre général d'une modélisation de la prosodie, étant donné son caractère supra-segmental. Néanmoins, dans une première approche qui vise à valoriser l'information liée à la pseudo-syllabe, il apparaît raisonnable de faire cette supposition. Sous cette hypothèse d'indépendance, le modèle de l'intonation permet d'écrire :

$$Pr(F|\Psi, L_i) = \prod_{k=1}^{np} Pr(F_k|L_i) \tag{15}$$

où F_k représente les paramètres qui sont extraits à l'intérieur des zones vocaliques de la pseudo-syllabe.

Les paramètres dérivent de l'extraction de la courbe de la fréquence fondamentale, effectuée en utilisant la combinaison de trois méthodes d'extraction : l'autocorrélation, l'AMDF et le peigne spectral (voir [39] pour un état de l'art sur les méthodes d'extraction de la fréquence fondamentale). Une interpolation quadratique est appliquée sur les valeurs de la fréquence fondamentale des zones voisées du signal, de manière à produire une valeur toutes les 10 ms quel que soit le voisement. Ne sont ensuite conservées que les valeurs prises sur les segments vocaliques des pseudo-syllabes pour calculer deux paramètres : le skewness et le kurtosis. Ces deux paramètres permettent de caractériser la distribution de la fréquence fondamentale sur les zones voisées (F_{voc}).

Le skewness, ou coefficient d'assymétrie, correspond au troisième moment central normalisé :

$$skewness(F_{voc}) = \frac{\sum_{i=1}^{N}(f0_i - \mu)^3}{N\sigma^3} \tag{16}$$

où $F_{voc} = \{f0_1, f0_2, ..., f0_N\}$ est la suite de fréquences fondamentales de la partie vocalique de la pseudo-syllabe considérée.

Le kurtosis, ou coefficient d'applatissement, correspond au quatrième moment central normalisé :

$$kurtosis(F_{voc}) = \frac{\sum_{i=1}^{N}(f0_i - \mu)^4}{N\sigma^4} \tag{17}$$

Nous espérons ainsi mesurer l'évolution de la fréquence fondamentale en distinguant les pseudo-syllabes accentuelles des autres, ainsi que le type d'accentuation au sein de cette unité.

A ces paramètres est ajouté le décalage (en ms) entre le maximum de la courbe mélodique et le début du segment vocalique. Cette variable permet de localiser "la place de l'accent" sur la pseudo-syllabe.

$$dist(F_{voc}) = t_{argmax(F_{voc})} - t_0 \tag{18}$$

où t_0 correspond au début du segment vocalique.

Ces trois paramètres caractérisent l'enveloppe de la fréquence fondamentale F_0 du segment vocalique de chaque pseudo-syllabe, et l'on écrit :

$$F_k = (skewness(F_{voc}), kurtosis(F_{voc}), dist(F_{voc})) \tag{19}$$

Pour chaque langue, un modèle de mélange de lois gaussiennes est appris pour caractériser la distribution des vecteurs F_k, en utilisant l'algorithme EM. Le nombre de composantes est déterminé en utilisant l'algorithme de LBG-Rissanen. Nous noterons $int(F_k|L_i)$ la densité de probabilité intonative associée à la langue L_i prise en F_k.

4.4 règle d'identification des langues

La règle de décision de notre système d'identification des langues est représentée par l'expression suivante :

$$L^* = arg \max_{1 \leq i \leq N_L} \left(\prod_{k=1}^{np} voc(v_k|L_i) int(F_k|L_i) ryt(\Psi_k|L_i) \right) \tag{20}$$

Une variante de cette règle de décision peut être envisagée en pondérant de manière empirique chacune des contributions (acoustique, rythmique et intonative), comme nous le verrons lors des expérimentations.

5 expériences

5.1 corpus et protocole expérimental

Les expériences ont été menées sur le corpus multilingue MULTEXT [40]. Les enregistrements sont extraits du corpus de parole EUROM1, réalisé à l'occasion du projet ESPRIT 2589 « Multi-lingual Speech Input/Output Assessment, Methodology and Standardisation » [41]. Les enregistrements audio sont de haute qualité (échantillonnage à 20 KHz, 16 bits) et effectués en chambre anéchoïque. Les enregistrements ont été contrôlés durant l'acquisition de manière à rejeter toute donnée bruitée ou toute erreur de lecture. MULTEXT reprend cinq des huit langues de EUROM1 (allemand, anglais, espagnol, français et italien). Les données correspondent au jeu de locuteurs "FEW TALKER SET" (comprenant dix locuteurs par langue : cinq femmes et cinq hommes) sur les passages lus de cinq phrases connectées par une structure sémantique cohérente. Notons qu'une même phrase est prononcée en moyenne par 4 locuteurs. Il est demandé à chaque locuteur de lire un extrait du passage et d'essayer d'avoir une intonation la plus naturelle possible. La durée de chaque passage est d'environ 20 s et la durée des enregistrements par langue est de 45 mn environ (cf. tableau 2 pour les détails de durée par langue).

Une procédure de validation croisée a été utilisée de manière à exploiter au mieux la taille limitée du corpus pour les expériences en identification des langues. L'apprentissage de tous les modèles est réalisé sur neuf locuteurs et les tests sont effectués sur le dernier locuteur. Ce processus est répété pour chaque locuteur et pour chaque langue. Les résultats présentés dans les tableaux suivants correspondent à la moyenne de ces résultats.

TAB. 2 – Nombre de passages par locuteurs et durée des enregistrements du corpus MULTEXT

Langue	Passages par locuteur	Durée totale (en mn)	Durée moyenne de chaque passage (en s)
allemand	20	73	21,9
anglais	15	44	17,6
espagnol	15	52	20,9
français	10	36	21,9
italien	15	54	21,7

Nous avons mesuré les pourcentages de textes en commun entre l'ensemble d'apprentissage et de test. En fonction des locuteurs considérés, ils varient entre 25,9 % et 33.4 % pour l'anglais, l'italien et l'espagnol, entre 11,1 % et 22,2 % pour le français et sont de 44,4 % pour l'allemand. Afin de vérifier que la présence d'un texte identique ne biaise pas les résultats, une série d'expériences supplémentaires a été menée en utilisant des sous-ensembles du corpus MULTEXT pour lesquels nous avons supprimé tous les textes en commun. Le tableau 3 détaille la constitution de ces sous-ensembles (à chaque passage de 5 phrases correspond une étiquette de type "lettre-chiffre", des exemples sont donnés en annexe pour l'anglais et le français) . Les sous-ensembles ont été réalisés en conservant deux locuteurs (un masculin et un autre féminin) pour le test, et en supprimant tous les passages en commun sur les autres locuteurs du sous-ensemble d'apprentissage. Les locuteurs restent disjoints entre les deux sous-ensembles. Le résultat des expériences est présenté en 5.4.

TAB. 3 – Passages, nb. de locuteurs et durée du jeu de données indépendantes du texte

Langue	Passages pour apprentissage (nb. loc.)	Passages pour test	Locuteurs du test	Durée pour apprentissage	Durée pour le test
allemand	g1 (5)	g2	bg, jm	29 mn	7 mn
anglais	o1 à o5 o6 à o0 p1 à p5 p6 à p0 q1 à q5 q6 à q0 (6)	r1 à r5 r6 à r0	fe, fh	24 mn	6 mn
espagnol	p0 à p4 p5 à p9 q0 à q4 q5 à q9 r0 à r4 r5 à r9 (6)	o0 à o4 o5 à o9	eb, nb	27 mn	8 mn
français	o0 à o9 p0 à p9 r0 à r9 (6)	q0 à q9	mh, sl	29 mn	7 mn
italien	g3 g4 g5 g6 g7 g8 (6)	g1, g2	a7, b7	30 mn	7 mn

5.2 évaluation séparée des modélisations en validation croisée

Afin d'évaluer l'apport de chaque type d'information dans la tâche d'identification des langues, une série d'expériences est effectuée en isolant successivement chaque composante du système.

5.2.1 la modélisation acoustique du système vocalique

L'information acoustique contenue dans le système vocalique telle qu'elle est extraite et modélisée dans cette étude, a été largement explorée [38]. Dans des travaux antérieurs et pour une paramétrisation très proche, nous avons obtenu sur le corpus OGI-MLTS [2] un taux d'identification de l'ordre de 70 % lorsqu'une langue doit être

reconnue parmi cinq. Par contre les tests d'identification étaient effectués sur 45 s de parole.

Nous avons refait cette expérimentation sur le corpus MULTEXT pour deux raisons :

- le module acoustique nous sert par la suite de référence, et nous permet d'évaluer l'apport de chaque type d'information ;
- nous cherchons à faire varier la longueur des signaux tests afin de savoir à partir de quelle durée, il est possible d'avoir un résultat fiable. Si la tâche d'identification des langues est réalisée en amont d'un système de reconnaissance de parole, il est important de réduire la longueur du signal de parole traité afin d'avoir une détection de la langue le plus rapidement possible.

La figure 4 présente les résultats de la tâche d'identification des langues sur les cinq langues, en faisant varier la durée des fichiers de test de 1 à 21 secondes. Le taux d'identification correcte est de 51 % avec une seconde de signal, et se stabilise très vite aux alentours de 70 % lorsque l'on augmente la durée des fichiers de test (à partir de 7s). Cette valeur correspond aux expériences menées antérieurement.

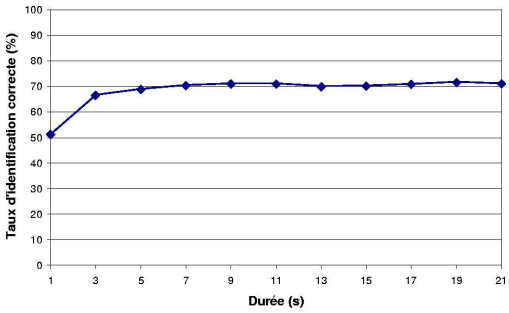


FIG. 4 – Taux d'identification correcte de la modélisation acoustique des segments vocaliques

L'examen de la matrice de confusion présentée dans le tableau 4 (obtenue avec 32 lois gaussiennes pour chacune des langues) correspondant aux énoncés de 21 s montre que le français et l'allemand sont bien discriminés. Il y a par contre beaucoup de confusions entre l'anglais et l'italien, ainsi qu'entre l'italien et l'espagnol. Ces erreurs proviennent en partie des erreurs de détection vocalique : l'italien possède

de nombreuses consonnes géminées, qui peuvent être détectées comme des voyelles et entraîner la constitution d'un segment vocalique et d'une pseudo-syllabe erronée. Nous retrouverons ce type de confusion dans les expériences suivantes.

TAB. 4 – Matrice de confusion du modèle acoustique des segments vocaliques pour les énoncés de 21 s (taux d'identification correcte : 70 %).

	anglais	français	allemand	italien	espagnol
anglais	44	0	0	38	18
français	0	92	1	1	6
allemand	2	0	96	2	0
italien	30	0	0	46	24
espagnol	5	10	0	13	72

5.2.2 la modélisation du rythme

Une expérience semblable est menée en utilisant seulement l'information rythmique, c'est à dire la durée des parties consonantiques et vocaliques ainsi que la complexité de la pseudo-syllabe.

Comme le montre la figure 5, les performances du modèle rythmique sont particulièrement intéressantes. Le taux d'identification correcte augmente graduellement avec la durée des énoncés de test pour arriver au meilleur résultat : 78 % avec des énoncés de 21 s. Avec des phrases de test plus courtes (moins de dix secondes), les résultats restent bons (de l'ordre de 70 %). Utiliser seulement la première seconde donne un taux d'identification correcte de 47 %. Ces résultats sont obtenues avec 16 lois gaussiennes pour les modèles d'anglais, allemand, italien et espagnol, et 18 lois pour le français.

La lecture de la matrice de confusion correspondant aux énoncés de 21 s (tableau 5) conduit aux mêmes observations que celles faites dans le cas acoustique seul. Les principales confusions se trouvent au niveau de l'anglais avec l'allemand et l'italien, ainsi que de l'italien avec l'espagnol. Ces erreurs proviennent en partie des erreurs de détection de pseudo-syllabe, comme nous l'avons déjà signalé. Mais il convient de signaler

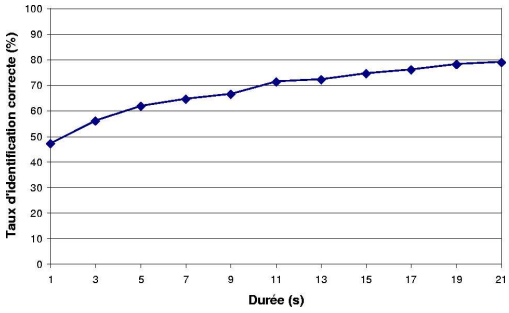


FIG. 5 – Taux d'identification correcte du modèle du rythme

TAB. 5 – Matrice de confusion du modèle du rythme pour les énoncés de 21 s (identifications correctes : 79 %).

	anglais	français	allemand	italien	espagnol
anglais	62	4	16	11	7
français	0	100	0	0	0
allemand	11	1	86	2	0
italien	10	1	3	62	23
espagnol	1	4	0	3	91

que l'on retrouve les résultats de classification par des paramètres rythmiques de Ramus [42], qui utilise une discrimination basée sur la proportion des durées vocaliques et l'écart type des durées des intervalles consonantiques.

5.2.3 la modélisation de l'intonation

Trois expériences sont réalisées afin de quantifier les apports de chaque composante de l'intonation en considérant trois types de vecteurs (cf. §4.3.2) :

- $F_k[skewness/kurtosis] = (skewness(F_{voc}), kurtosis(F_{voc}))$, système "skewness/kurtosis",
- $F_k[accent] = (dist(F_{voc}))$, système "place de l'accent",
- $F_k[skewness/kurtosis/accent] = (skewness(F_{voc}), kurtosis(F_{voc}), dist(F_{voc}))$,

Le tableau 6 permet une comparaison des résultats obtenus au cours de ces trois expériences, en utilisant les fichiers de 21 s sur les 5 langues. Le taux d'identification

TAB. 6 – Taux d'identification correcte en utilisant les paramètres basés sur la fréquence fondamentale sur les fichiers 21 s sur 5 langues.

Paramètres	Taux moyen
skewness/kurtosis	51,4 %
place de l'accent	55,7 %
skewness/kurtosis/accent	49,8 %
fusion pondérée	76 %

correcte est inférieur en utilisant les trois paramètres plutôt qu'en utilisant les paramètres skewness/kurtosis ou même le seul paramètre sur la place de l'accent.

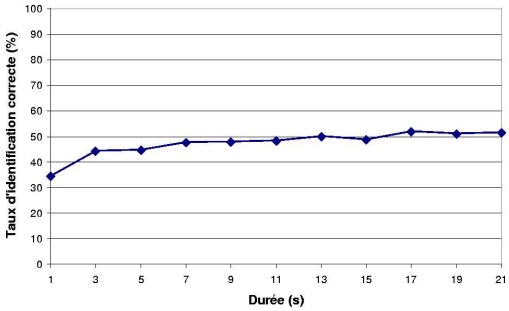


FIG. 6 – Taux d'identification correcte pour le système "skewness/kurtosis".

L'influence de la longueur des fichiers tests se fait naturellement plus ressentir sur le système utilisant simplement la "place de l'accent" que sur le système "skewness/kurtosis" comme le montrent les figures 6 et 7.

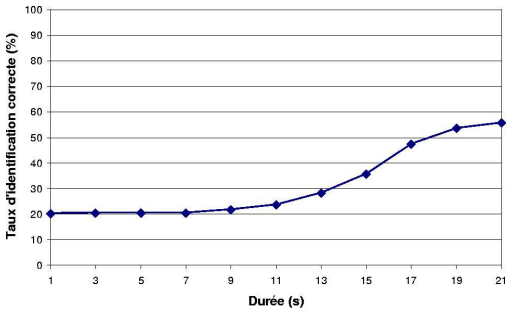


FIG. 7 – Taux d'identification correcte pour le système "place de l'accent".

Avec le système "skewness/kurtosis", le meilleur taux d'identification (51,4 %) est obtenu pour une longueur maximale de 21 s. On obtient 34 % d'identification correcte

avec seulement 1 s de signal. Avec le système "place de l'accent", le meilleur taux (55,5 %) est également obtenu avec la durée maximale. En revanche pour des durées de test inférieures à 11 s, les performances sont comparables au taux du hasard (20 %).

Au vu de ce comportement, une quatrième expérience est réalisée où la fusion des deux premiers systèmes, le système skewness/kurtosis et le système "place de l'accent", est faite en pondérant les scores de chacun d'eux. Le taux d'identification correcte (tableau 6) atteint 76 % pour les fichiers de 21 s sur 5 langues. La matrice de confusion obtenue avec la meilleure configuration de poids (ie. 0,7 pour le système "skewness/kurtosis" et 0,3 pour le système "place de l'accent") se trouve dans le tableau 7. Ces résultats représentent l'optimum de ce système, vu que les pondérations ont été optimisées sur le corpus de test. Une optimisation sur un corpus de développement et de test séparés donnerait des résultats inférieurs.

TAB. 7 – Matrice de confusion de la meilleure fusion entre le système "skewness/kurtosis" et le système "place de l'accent" pour les fichiers de 21 s.

	anglais	français	allemand	italien	espagnol
anglais	87	1	11	0	1
français	0	56	4	0	39
allemand	0	2	85	0	12
italien	1	4	8	65	21
espagnol	0	9	5	0	86

La figure 8 détaille la fusion des deux systèmes. Le meilleur taux d'identification correcte (84 %) est atteint pour une longueur de signal de 17 s. La longueur maximale (20 s) donne un résultat légèrement inférieur (78,2 %).

5.3 le système global

Le système global d'identification des langues est obtenu en concaténant les informations issues des trois types de modélisation :

– le modèle acoustique des segments vocaliques,

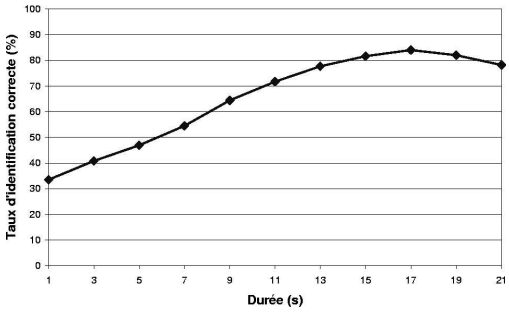


FIG. 8 – Fusion du système "skewness/kurtosis" et du système "place de l'accent".

- le modèle du rythme traitant les paramètres relatifs aux durées consonantique et vocalique de la pseudo-syllabe,
- le modèle de l'intonation exploitant les paramètres de la courbe mélodique du segment vocalique de la pseudo-syllabe.

selon la formule (cf. équation 20). En utilisant la version logarithmique de cette formule, c'est à dire en additionnant les log-vraisemblances (appelées scores) issues de ces trois modèles, nous avons envisagé de pondérer chacun d'eux. Cette version donne de meilleurs résultats : le meilleur compromis, trouvé expérimentalemant est obtenu en affectant comme poids 0,1 au score acoustique, 0,3 au score rythmique et, pour le modèle de l'intonation, 0,3 au score issu du système "skewness/kurtosis" et 0,3 pour le système "place de l'accent".

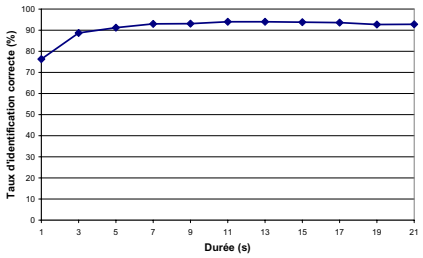


FIG. 9 – Résultats d'identification correcte pour 5 langues issus de la fusion des trois systèmes

La figure 9 présente les résultats corrects en identification automatique pour les 5 langues pour ce type de fusion. 75 % d'identification correcte sont obtenus dès la première seconde, puis le taux dépasse 90 % à partir de 5 s de parole. Le maximum se situe à 92,4 % pour 11 s de parole.

Le tableau 8 décrit la matrice de confusion correspondant à la meilleure combinaison (poids, longueur de signal). Les erreurs les plus importantes proviennent de la mauvaise discrimination entre l'italien et l'espagnol. Dans une moindre mesure, l'espagnol se confond avec l'anglais et le français. En dehors des commentaires avancés ci-dessus, signalons toutefois que ces confusions restent en concordance avec les regroupement en langues accentuelles (anglais, allemand) et langues à accent fixe (français, italien et espagnol) de Pike [15].

TAB. 8 – Matrice de confusion de la fusion globale pour des enregistrements de 21 s (Taux d'identifications correctes de 92,4 $\pm$ 2.3 %).

	anglais	français	allemand	italien	espagnol
anglais	92	0	1	0	7
français	0	99	0	0	1
allemand	0	0	100	0	0
italien	1	0	0	82	17
espagnol	0	8	0	1	91

5.4 expériences sur les sous-ensembles indépendants du texte

Le tableau 9 présente les résultats obtenus sans utiliser la validation croisée, sur le sous-corpus indépendant du texte lu décrit en 5.1. Dans cette section, du fait du faible nombre de fichiers, les nombres donnés dans les matrices de confusions ne sont plus exprimés en pourcentage, mais en nombre de fichiers. Néanmoins, traduit en pourcentage, on constate une dégradation puisque l'on passe de 92,4 $\pm$ 2,3 % à 85,9 $\pm$ 6,9 %. Cette dégradation s'accompagne d'un élargissement de l'intervalle de confiance puisque le nombre de tests effectués est quatre fois moins important que dans le cas de la validation croisée. Par ailleurs, cette procédure a également eu une

influence sur la taille du corpus d'apprentissage : la durée totale de l'ensemble d'apprentissage est passée de 4 h à 2h20. La dégradation des conditions expérimentales explique la différence de résultats observés.

En conclusion, cette dernière expérience confirme l'intéret de cette approche, bien que la taille limitée du corpus nuit à l'apprentissage des modèles. Le recouvrement entre les textes pour les ensembles d'apprentissage et de test ne semble pas être un biais important.

TAB. 9 – Matrice de confusion de la fusion globale (indépendamment du texte) pour des enregistrements de 20 s (Taux d'identifications correctes de 85,9 $\pm$ 6.9 %).

	anglais	français	allemand	italien	espagnol
anglais	80	0	0	20	0
français	0	100	0	0	0
allemand	0	0	100	0	0
italien	5	0	0	70	25
espagnol	5	10	0	5	80

6 discussion et conclusion

Nous avons dans ces travaux apporté une réponse au traitement automatique de la prosodie pour l'identification des langues. Nous avons proposé une modélisation multilingue du rythme et de l'intonation des langues en définissant d'une part une segmentation automatique des énoncés architecturée sous forme de pseudo-syllabes et d'autre part une extraction de paramètres adaptée à cette structure.

Pour valider cette aproche, nous avons mis en œuvre un système d'identification automatique des langues utilisant trois modélisations :

- une modélisation acoustique,
- une modélisation rythmique,
- une modélisation de l'intonation,

qui permettent d'atteindre respectivement 70 % d'identification correcte à partir de 7 secondes de signal, 78 % sur 21 secondes, et 76 % sur 21 secondes. Le modèle

acoustique (modèle de référence) produit de bons résultats dès les premières secondes de signal, les modèles prosodiques sont plus progressifs et nécessitent plus de temps pour pouvoir livrer des résultats convenables.

Cette série d'expériences menées sur les 5 langues du corpus MULTEXT démontre la pertinence des paramètres prosodiques pour l'identification des langues ainsi que celle de l'utilisation de la pseudo-syllabe et des paramètres dérivés. Nous obtenons 76 % d'identification correcte en utilisant des paramètres liés à l'intonation, et 79 % en utilisant des paramètres liés au rythme des langues sur des fichiers de 20 s.

Lors de la fusion des systèmes élémentaires, les performances en identification automatique des langues s'améliorent grandement et peuvent atteindre un optimum de 90 % d'identification correcte avec seulement 5 secondes de signal. Ce taux optimum peut dépasser 92 % avec des énoncés de 11 s. Les principales confusions se situent entre l'espagnol et l'italien, deux langues appartenant à la même classe rythmique.

Il est à noter que l'ajout d'une langue au système ne nécessitera que la collecte d'enregistrements de parole, sans ajouter de coûteuses segmentations du signal, puisqu'aucun étiquetage manuel n'est requis et que la pseudo-syllabe est un processus indépendant de la langue. Il faudrait néanmoins quantifier l'influence de la segmentation automatique en pseudo-syllabe, par rapport à une segmentation manuelle ou semi-automatique.

Une perspective intéressante est l'exploitation dynamique de la pseudo-syllabe : nous avons étudié dans ces expériences des paramètres statiques locaux liés à la structure interne de la pseudo-syllabe ; afin de prendre en compte le caractère supra-segmental de la prosodie, il faudrait modéliser l'enchaînement de plusieurs pseudo-syllabes et atteindre ainsi la structure temporelle des différents motifs rythmiques des langues. La modélisation par modèle de Markov cachés couplé à des modèles de mélanges de lois Gaussiennes, est un premier candidat. Restera ensuite à valider l'ensemble de cette approche sur les corpus multilingues téléphoniques (ou bruités).

7 Annexe 1 - Exemples de passages lus

En langue française :

Passage : O0

J'ai des problèmes avec mon adoucisseur d'eau : le niveau d'eau est toujours trop haut, ce qui fait que le trop-plein coule sans arrêt. Vous pourriez pas m'envoyer un réparateur mardi matin ? C'est le seul jour de la semaine que j'ai de libre. Si vous êtes d'accord, soyez gentil de me confirmer le rendez-vous par écrit.

Passage : O1

Passez-moi les réclamations, s'il vous plaît. On est venu réparer le tuyau d'arrivée d'eau, devant chez moi, et ça n'a pas tenu : ma cave est inondée. Quand j'ai téléphoné, on m'a répondu que toutes les équipes de dépannage étaient occupées pendant les deux semaines qui viennent. On peut vraiment pas faire confiance au Service des Eaux. Si j'ai bien compris, en attendant, ma cave va me servir de piscine.

En langue anglaise :

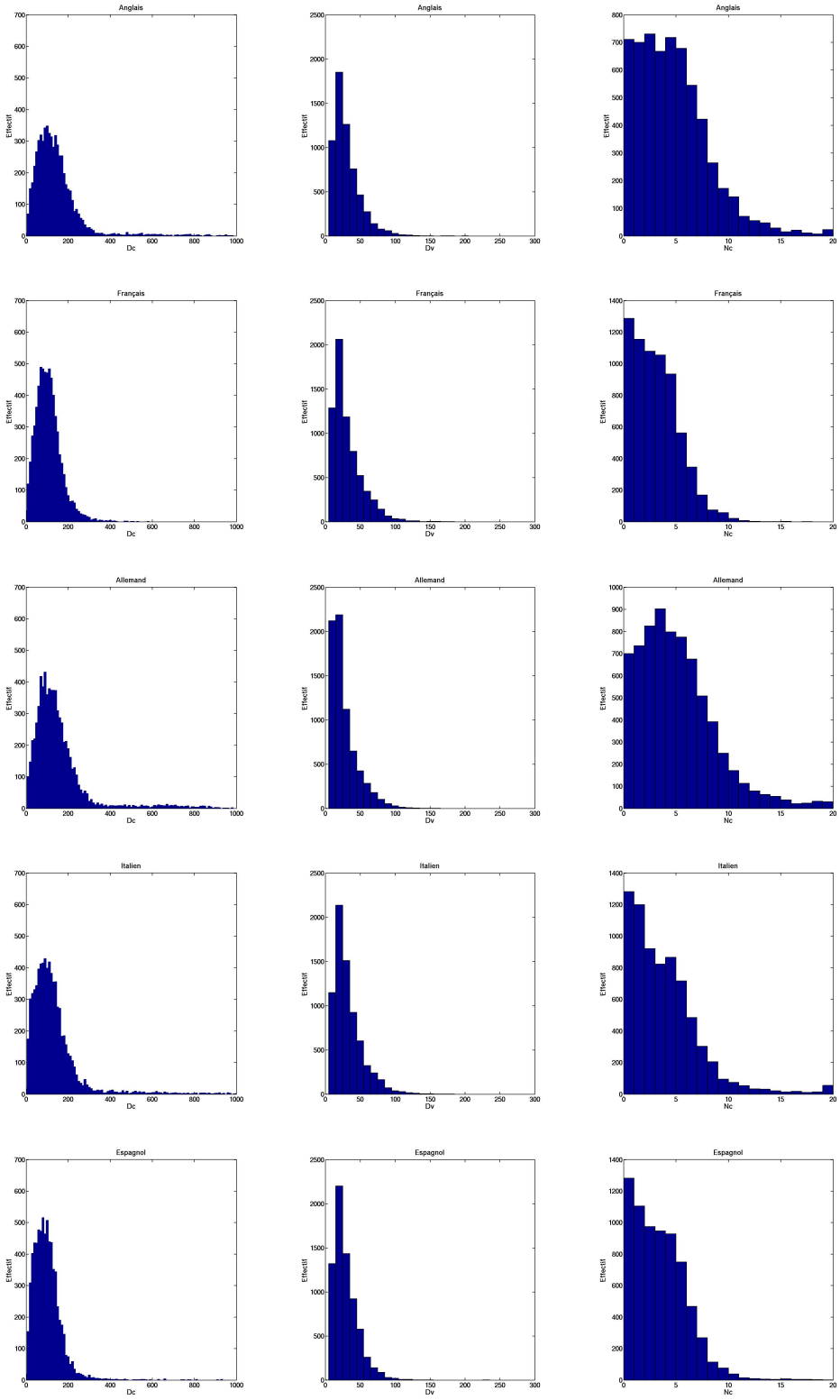
Passage : O0

Last week my friend had to go to the doctors to have some injections. She is going to the Far East for a holiday and she needs to have an injection against cholera, typhoid fever, hepatitis A, polio and tetanus. I think she will feel quite ill after all those. She is going to get them all done at once, at one session. I shan't feel sorry for her though !

Passage : O1

I have a problem with my water softener. The water-level is too high and the overflow keeps dripping. Could you arrange to send an engineer on Tuesday morning please ? It's the only day I can manage this week. I'd be grateful if you could confirm the arrangement in writing.

8 Annexe 2 - Histogrammes des distributions des valeurs de Dc Dv Nc sur les cinq langues du corpus d'apprentissage



9 remerciements

Cette recherche est financée par le programme EMERGENCE de la Région Rhône-Alpes, le programme *ACI Jeunes Chercheurs* du Ministère de la Recherche et le projet RAIVES du programme interdisciplinaire "Société de l'information" du CNRS.

Références

[1] M. A. Zissman and K. M. Berkling, "Automatic language identification," in *Speech Communication*, vol. 35, pp. 115–124, Elsevier Science, 2001.

[2] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The ogi multilanguage telephone speech corpus," in *International Conference on Speech and Language Processing*, vol. 2, pp. 895–898, Oct. 1992.

[3] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr., "Approaches to language identification using gaussian mixture models," in *4th International Conference on Spoken Language Processing*, vol. 1, (Denver, CO, USA), pp. 89–92, Sept. 2002.

[4] E. Singer, P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell, and D. A. Reynolds, "Acoustic, phonetic and discriminative approaches to automatic language identification," in *8th European Conference on Speech Communication and Technology* (ISCA, ed.), (Genève, Suisse), pp. 1345–1348, Sept. 2003.

[5] Y. K. Muthusamy, E. Barnard, and R. A. Cole, "Reviewing automatic language identification," *IEEE Signal Processing Magazine*, vol. 11, pp. 33–41, Oct. 1994.

[6] L. F. Lamel and J.-L. Gauvain, "Cross-lingual experiments with phone recognition," in *IEEE 18th International Conference on Acoustics Speech and Signal Processing*, (Minneapolis, USA), Apr. 1993.

[7] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, 1996.

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

[9] D. Matrouf, M. Adda-Dekker, L. Lamel, and J. Gauvain, "Language identification incorporating lexical information," in *5th International Conference on Spoken Language Processing*, pp. 181–184, 1998.

[10] J. J. Ohala and J. B. Gilbert, "Listeners' ability to identify languages by their prosody," in *Problèmes de prosodie : expérimentations, modèles et fonctions* (P. Léon and M. Rossi, eds.), vol. 2, pp. 121–131, Paris, France : Didier, 1979.

[11] J. F. Werker, J. H. V. Gilbert, K. Humphrey, and R. C. Tees, "Developmental aspects of cross-language speech production," *Child Development*, vol. 52, pp. 349–355, 1981.

[12] J. A. Maidment, "Language recognition and prosody : further evidence," in *Speech Hearing and Language*, vol. 1, pp. 131–141, University College London, 1983.

[13] J. Mehler, J. Bertoncini, E. Dupoux, and C. Pallier in *Phonological Structure and Language Processing : Cross Linguistic Studies* (T. Otake and A. Cutler, eds.), ch. The role of suprasegmentals in speech perception and acquisition, pp. 145–169, New-York, USA : Mouton de Gruyter, 1996.

[14] P. F. MacNeilage and B. L. Davis, "On the Origin of Internal Structure of Word Forms," *Science*, vol. 288, pp. 527–531, Apr. 2000.

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

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

[17] P. Ladefoged, ed., *A course in phonetics*. New York, USA : Harcourt Brace Jovanovich, 1975.

[18] R. M. Dauer, "Stress-timing and syllable-timing reanalysed," in *Journal of Phonetics*, vol. 11, pp. 51–62, Cambridge, UK : Academic Press, 1983.

[19] A. E. Thymé-Gobbel and S. E. Hutchins, "Prosodic features in automatic language identification reflect language typology," in *14th International Congress of Phonetics Sciences*, (San Francisco, CA, USA), pp. 29–32, Aug. 1999.

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

[21] P. F. Dominey and F. Ramus, "Neural Network Processing of Natural Language : 1. Sensitivity to Serial, Temporal and Abstract Structure in the Infant," *Language and Cognitive Processes*, vol. 15, no. 1, pp. 87–127, 2000.

[22] F. Ramus, M. D. Hauser, C. Miller, D. Morris, and J. Mehler, "Language Discrimination by Human Newborns and by Cotton-Top Tamarin Monkeys," *Science*, vol. 288, pp. 349–351, Apr. 2000.

[23] P. A. Barbosa, *Caractérisation et génération automatique de la structuration rythmique du français*. PhD thesis, Institut National Polytechnique, Grenoble, France, 1994.

[24] R. André-Obrecht, *Segmentation et parole ?* Habilitation à diriger les recherches, Université de Rennes - IRISA, Rennes, June 1993.

[25] J. Farinas and F. Pellegrino, "Comparison of two approaches to Language Identification ," in *7th International Conference on Speech Communication and Technology*, (Aalborg, Danemark), pp. 399–402, Sept. 2001.

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

[27] R. André-Obrecht, "A new statistical approach for automatic speech segmentation," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. 36, pp. 29–40, Jan. 1988.

[28] N. Suaudeau and R. André-Obrecht, "An efficient combination of acoustic and supra-segmental informations in a speech recognition system," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, (Adélaide,

Australie), Apr. 1994.

[29] F. Pellegrino, *Une approche phonétique en identification automatique des langues : la modélisation acoustique des systèmes vocaliques*. Thèse de doctorat, Université Paul Sabatier, Toulouse, France, Dec. 1998.

[30] F. Pellegrino and R. André-Obrecht, "From Vocalic Detection to Automatic Emergence of Vowel Systems," in *IEEE 22d International Conference on Acoustics Speech and Signal Processing*, (Munich, Allemagne), pp. 108–112, Apr. 1997.

[31] H. R. Pfitzinger, S. Burger, and S. Heid, "Syllable detection in read and spontaneous speech," in *4th International Conference on Spoken Language Processing*, vol. 2, (Philadelphia, PA, USA), pp. 1261–1264, 1996.

[32] N. Fakotakis, K. Georgila, and A. Tsopanoglou, "An continuous hmm text-independant speaker recognition system based on vowel spotting," in *5th European Conference on Speech Communication and Technology*, vol. 5, (Rhodes, Grèce), pp. 2247–2250, Sept. 1997.

[33] R. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection," in *IEEE 23rd International Conference on Acoustics Speech and Signal Processing*, vol. 2, (Seattle, WA, USA), pp. 945–948, May 1998.

[34] A. W. Howitt, "Vowel landmark detection," in *6th International Conference on Speech Communication and Technology*, (Budapest, Hongrie), Sept. 1999.

[35] F. Pellegrino and R. André-Obrecht, "An unsupervised approach to language identification," in *IEEE 24th International Conference on Acoustics Speech and Signal Processing*, vol. 2, (Phoenix, AR, USA), pp. 833–836, Mar. 1999.

[36] F. Pellegrino, J. Farinas, and R. André-Obrecht, "Comparison of Two Phonetic Aproaches to Language Identification," in *6th European Conference on Speech Communication and Technology*, (Budapest, Hongrie), pp. 399–402, Sept. 1999.

[37] C. Mokbel, D. Jouvet, and J. Monné, “Blind Equalization using Adaptative Filtering for improving Speech Recognition over Telephone,” in *4th European Conference on Speech Communication and Technology*, (Madrid, Espagne), pp. 1987–1990, 1995.

[38] F. Pellegrino and R. André-Obrecht, “Automatic Language Identification : an alternative approach to phonetic modeling,” in *Signal Processing*, vol. 80, pp. 1231–1244, Elsevier Science, jul 2000.

[39] J. Farinas, *"Une modélisation automatique du rythme pour l'identificationdes langues"*. PhD thesis, Université Toulouse III, Toulouse, France, Nov. 2002.

[40] E. Campione and J. Véronis, “A multilingual prosodic database,” in *5th International Conference on Spoken Language Processing*, (Sidney, Australie), pp. 3163–3166, Nov. 1998.

[41] D. Chan, A. Fourcin, D. Gibbon, B. Granström, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Velt, and J. Zeiliger, “EUROM : A Spoken Language Ressource for the E.U.,” in *4th European Conference on Speech Communication and Technology*, (Madrid, Espagne), 1995.

[42] F. Ramus in *De la caractérisation à l'identification des langues, actes de la 1ère journée d'étude sur l'identification automatique des langues* (F. Pellegrino, ed.), ch. La discrimination des langues par la prosodie : modélisation linguistique et étude comportementale, Editions de l'Institut des Sciences de l'Homme, 1999.