De la caractérisation…

…à l'identification des langues

Sélection de conférences données
lors de la 1ère journée d'étude
sur l'identification automatique des langues,
Lyon, 19 janvier 1999

avec le soutien

de l'Association Francophone de la Communication Parlée

du Groupe Parole du GDR-PRC I3

Edité par F. Pellegrino
Institut des Sciences de l'Homme de Lyon

Table des matières

Introduction

Le 19 janvier 1999 se tenait à Lyon un colloque intitulé *Identification Automatique des Langues : de la caractérisation à l'identification des langues*. Cette manifestation, organisée par le Groupe Francophone de la Communication Parlée[1] et le groupe Parole du GDR-PRC I3 (Information – Interaction – Intelligence), réunissait pour la première fois en France des scientifiques, français et étrangers, autour du thème de l'Identification Automatique des Langues (IAL). Ce domaine, dont l'essor aux Etats-Unis date du début des années 90, n'a encore été que relativement peu abordé en France, alors que des enjeux importants, tant scientifiquement qu'économiquement, y sont rattachés. Ce colloque tenait donc à la fois de la table ronde entre chercheurs directement concernés par les travaux récents en IAL et du séminaire d'ouverture à l'attention de personnes désirant se familiariser avec ce domaine pluridisciplinaire émergent. Les chercheurs et les ingénieurs présents, issus de différentes communautés scientifiques fortement liées à la communication parlée, sont à l'origine de cet ouvrage, par leurs communications et les débats qui s'en suivirent : qu'ils en soient chaleureusement remerciés.

Cet ouvrage rassemble une sélection des communications données par les spécialistes présents, qu'ils soient linguistes, professionnels du traitement automatique de la parole, ou des sciences cognitives.

Les articles rassemblés au sein d'une première partie, permettront au lecteur de se familiariser avec les enjeux de l'Identification automatique des langues. Un état de l'art des méthodes actuelles ainsi qu'un article abordant les classifications et les typologies des langues du monde sous un éclairage linguistique accompagnent cette introduction. Ce vaste panorama est accompagné d'une réflexion sur les performances atteintes à ce jour, et sur les limites qui semblent se dégager des recherches menées depuis le début des années 90.

La seconde partie est consacrée à la description de plusieurs systèmes mettant en œuvre des approches originales développées récemment en France. Nous verrons ainsi que l'introduction de modélisations de type phonétique ou lexicale, à côté de l'approche phonotactique traditionnellement employée, offre sans doute un potentiel important dans la conception des systèmes à venir.

Si les deux premières parties de cet ouvrage faisaient un point sur la situation actuelle dans le domaine de l'Identification, les deux dernières parties sont spécifiquement tournées vers les perspectives offertes par de nouvelles approches, tirant profit de sources d'informations jusqu'à présent peu exploitées.

---

[1] Le GFCP est un groupe spécialisé de la Société Française d'Acoustique (SFA) et de l'International Speech Communication Association (ISCA). Il s'est depuis transformé en Association Franphone de la Communication Parlée (AFCP).

Nous verrons ainsi au cours de la troisième partie comment des traits linguistiques peuvent être exploités dans des systèmes automatiques. Un premier article introduit brièvement une approche basée sur la détection de traits phonétiques rares dans les langues du monde. Une expérience pilote d'identification dialectale basée sur des caractéristiques phonétiques et phonologiques de la langue arabe est proposée par la suite. Elle est complétée par une réflexion sur les difficultés inhérentes aux approches pluridisciplinaires mêlant expertise humaine et modélisation automatique.

La quatrième et dernière partie est quant à elle consacrée à la modélisation de la prosodie en vue de l'identification des langues. La prosodie, qui comprend aussi bien des aspects rythmiques que mélodiques, est un facteur dont l'importance dans la communication parlée est fondamentale. Cependant, de par sa nature, son étude est complexe, et à l'heure actuelle, aucune modélisation n'est pleinement satisfaisante. Deux études, basées respectivement sur une approche statistique et sur une approche neuro-mimétique sont présentées et mises en perspective avec des études comportementales.

La conclusion générale de cet ouvrage dresse un bilan du colloque et des perspectives qu'il a fait émerger.

1ère Partie
Etat de l'art

Une introduction à l'identification automatique des langues

François Pellegrino[1], Régine André-Obrecht[2]

[1] Laboratoire Dynamique Du Langage (UMR CNRS 5596)
[2] Institut de Recherche en Informatique de Toulouse (UMR CNRS 5505)

Francois.Pellegrino@univ-lyon2.fr – obrecht@irit.fr

Abstract

An introduction to the interdisciplinary field of automatic language identification is provided. Both theoretical and applied stakes are addressed, demonstrating the increasing importance of this topic in the automatic speech processing area. This overview finishes with a short historical review of the first studies initiated at the beginning of the seventies since they investigated linguistic features that still provide the core of the current systems.

Résumé

Cet article constitue une introduction au champ pluridisciplinaire de l'identification automatique des langues. Les enjeux, tant scientifiques qu'applicatifs, qui font de ce thème un domaine en pleine expansion du traitement automatique de la parole sont abordés. Cette introduction est complétée par un historique des premières études menées dès le début des années 70, et qui se basaient déjà sur des caractéristiques des langues qui sont encore au cœur des systèmes actuels.

1. Introduction

Identifier automatiquement la langue parlée par un locuteur est une tâche ardue. Comme dans d'autres domaines du traitement automatique de la parole, la science-fiction a véhiculé certains clichés à l'origine de faux espoirs : sorti des salles de cinéma, C-3PO, le robot spécialiste du protocole de la Guerre des Etoiles, se révèle un interprète décevant. Si l'époque où une machine sera capable de reconnaître automatiquement les milliers de langues parlées sur notre planète apparaît lointaine, la possibilité d'identifier une langue parmi un petit ensemble de candidats est envisagée depuis plusieurs décennies : Les première recherches menées aux Etats-Unis en Identification Automatique des Langues (IAL) datent du début des années 70. Ce thème est resté relativement confidentiel pendant près de vingt ans, principalement pour cause de manque de ressources : ressources humaines d'une part (la reconnaissance automatique de la parole drainant la plupart des chercheurs) et ressources matérielles d'autre part (aucune base de données multilingue n'était pleinement disponible pendant longtemps).

Heureusement, l'évolution de la technologie s'accélérant, le traitement automatique de la parole explore aujourd'hui des thèmes auparavant délaissés. Ainsi, depuis le début des années 1990, l'effort de recherche s'est réellement intensifié en IAL, et ce thème est devenu un enjeu majeur pour la décennie à venir. Parmi les multiples raisons de ce regain d'intérêt, on peut citer :

- la croissance de la demande pour des interfaces Homme-Machine,

- l'explosion des communications dans un cadre multilingue,
- l'augmentation des performances des systèmes de reconnaissance automatique de la parole et
- l'enregistrement et la mise à disposition de la communauté scientifique de corpus multilingues.

Si l'IAL est devenu dès lors un axe de recherche incontournable aux Etats-Unis, L'Europe, et la France en particulier, sont plutôt restés à l'écart de ce phénomène, puisqu'une seule étude française sur l'IAL a été publiée avant 1995 [Lamel 94]. Le mouvement commence pourtant à s'amorcer et, à l'heure actuelle, plusieurs équipes françaises participent au mouvement.

L'objectif poursuivi est de concevoir un système qui, à partir d'un énoncé prononcé par un locuteur, détermine la langue qu'il a employée. A cette formulation élémentaire correspond une réalité bien plus complexe, puisque l'IAL peut se décliner et s'envisager dans un nombre assez important d'applications, où les conditions sont très variables, du nombre de locuteurs au nombre de langues, et de la longueur de l'énoncé à ses conditions d'enregistrements.

2. Les enjeux applicatifs

L'ère actuelle est une ère de communication multilingue, que ce soit entre humains (au sein des grandes mégalopoles ou par téléphone interposé), ou entre humains et machines (IHM). Ce constat implique le développement d'applications capables de gérer plusieurs langues et/ou d'identifier une langue parmi d'autres. Ces systèmes d'IAL peuvent être envisagés dans une tâche d'assistance au dialogue humain (DH) ou au sein d'IHM.

2.1. Assistance au dialogue humain

L'exemple le plus célèbre de situation de DH multilingue, cité par Muthusamy [Muthusamy 94] est assez significatif des besoins à venir : aux Etats-Unis, les numéros des services d'urgence sont centralisés et accessibles en appelant le 911. La nécessité d'avoir un standard téléphonique disposant d'un service d'interprètes efficace est depuis longtemps une réalité dans ce pays multi-ethnique, à tel point que ATT a mis en place un service, (nommé *ATT Language Line*) chargé de diriger chaque appel vers le correspondant qui pourra comprendre la langue employée. Ce service d'interprètes gère 140 langues, et l'aiguillage des appels est à l'heure actuelle réalisé de manière entièrement manuelle : lorsqu'un appel arrive à un standardiste et qu'il ne s'agit pas d'une langue qu'il connaît, il le renvoie vers un autre standardiste en fonction de la langue – ou du type de langue – qu'il croit avoir reconnu. Ainsi l'appel peut transiter par plusieurs standardistes pendant un temps assez long, comme le rapporte Muthusamy[1]. Cet exemple montre à quel point il peut être important, dans le cadre d'une intervention d'urgence, de pouvoir identifier une langue rapidement. Un système d'IAL permettrait de confronter l'avis du standardiste avec la décision de la machine, et si cela ne se

[1] En s'exprimant en Tamoul, Muthusamy a attendu plusieurs minutes avant d'obtenir un correspondant le comprenant, après avoir été dirigé – sans succès – vers trois interprètes d'Asie du Sud-Est. La situation a été débloquée uniquement lorsqu'il a prononcé le nom anglais 'Tamil'.

révèle pas plus efficace, on peut imaginer un système automatique qui donne à l'interprète une liste des langues potentielles correspondant à l'appel.

Si l'exemple du 911 est le plus marquant, il n'est pas pour autant unique, et dans bien des cas, l'IAL permettrait d'apporter une assistance au DH. Que ce soit dans un aéroport international, dans un grand hôtel ou pour un standard de réservation de billets pour une coupe de monde de football, le dialogue serait facilité si les gens pouvaient s'exprimer dans leur propre langue.

2.2. IAL et Interfaces Homme-Machine

L'IAL a un rôle croissant à jouer au sein des IHM, que ce soit pour permettre leur utilisation dans des pays plurilingues ou dans un cadre international. Si l'on reprend l'exemple de l'Espagne, un système de réservation de billets de train par téléphone doit être capable de recevoir un appel en castillan, en catalan, en basque ou en aranais, de même que tout système de dictée vocale par exemple. Ces contraintes linguistiques sont courantes à travers le monde, même si elles nous sont peu familières dans un pays où une seule langue officielle subsiste. Une autre application des IHM en plein développement consiste à mettre en des lieux publics (aéroports, gares, mais aussi offices du tourisme) des bornes d'information à disposition des voyageurs. Si, à l'heure actuelle, ces bornes peuvent être activées en deux ou trois langues en appuyant sur un bouton, l'extension du nombre de langues traitées se satisfera mieux d'un système d'identification de la langue. On retrouve aussi ce type d'utilisation dans le cadre de systèmes de synthèse multilingue et de traduction automatique lorsque plusieurs langues sont acceptées en entrée.

2.3. Indexation de documents multilingues

L'indexation automatique de documents à partir de leur contenu est en passe de devenir un enjeu incontournable de notre société où la masse d'information diffusée est chaque jour plus importante. Ce thème dépasse largement le cadre du traitement de la parole puisque, dans la plupart des cas, il s'agit réellement de traiter des flux multimédia, où les canaux vidéo et audio sont logiquement liés. Que ce soit pour procéder à l'indexation du colossal patrimoine que constituent les bandes vidéo enregistrées de part le monde, ou pour réaliser du filtrage en temps réel à partir des émissions disponibles simultanément sur un récepteur, les technologies de la parole sont mises à contribution (segmentation en Bruit/Parole/Musique, recherche de mots clefs, identification du locuteur ou de la langue,…).

Plus précisément, l'IAL intervient, tout comme l'identification automatique du locuteur, pour :

- détecter les changements de langue (resp. de locuteur) au cours du temps,
- identifier la langue (resp. le locuteur) présent(e) sur la bande audio à chaque instant.

3. Les enjeux scientifiques

Un corollaire implicite de l'IAL est la notion de distance entre langues, et de distance entre locuteurs et langues. De manière générale, on considère que les locuteurs utilisant un système automatique multilingue s'expriment dans leur langue naturelle. En effet, dès lors que l'on s'exprime dans une langue étrangère L2, le système phonologique employé est intermédiaire entre ceux de L1 et L2, et les structures morpho-syntaxiques sont aussi teintées

d'un « accent » plus ou moins fort. Un des premiers thèmes scientifiques que l'on peut envisager pour des systèmes d'IAL est donc l'enseignement des langues étrangères. Dans un laboratoire de langue, il n'est pas toujours possible à l'enseignant d'écouter chaque étudiant et de le corriger ; cette tâche peut être partiellement prise en compte par un système automatique : le locuteur, en se perfectionnant, doit s'éloigner du modèle de sa langue maternelle et se rapprocher de celui de la langue L2. Pour qu'un tel système soit efficace, il est nécessaire que la distance qu'il calcule ait une réalité perceptive, c'est-à-dire que la distance perceptive soit corrélée à la distance automatique. Ce paradigme intervient aussi pour la validation croisée de modèles phonologiques et de modèles automatiques. Nous avons vu au cours de la première partie que des modèles phonologiques de systèmes vocaliques assez nombreux existent, et l'une des validations expérimentales possibles consisterait à vérifier que les distances entre les systèmes prédits sont conformes à ce qui est réellement observé. Une telle mise en correspondance de systèmes automatiques et phonologiques peut se révéler très fructueuse à la fois pour les informaticiens et les linguistes. Un autre thème de recherche intervenant aussi en IAL consiste à étudier la caractérisation et la modélisation des langues. En particulier, la recherche d'unités inter-langues [Corredor Ardoy 97] rejoint les préoccupations des linguistes et des phonologues tout autant que des cogniticiens s'intéressant à l'universalité du langage.

D'autres motivations scientifiques existent pour l'IAL, mais ce qui apparaît actuellement est plutôt une sous-exploitation manifeste des systèmes automatiques puisqu'ils sont orientés vers la prise de décision en faveur d'une langue ou d'une autre et non vers une exploitation scientifique du processus de décision.

4. Les enjeux militaires

Un autre domaine d'application essentiel est le domaine militaire. Le US *Department of Defense* est à l'origine des premières recherches menées en IAL au sein des laboratoires Texas Instruments au début des années 70. Depuis, de nombreux autres projets ont vu le jour, aux Etats-Unis comme en Europe (Projet DGA 95/118 : discrimination automatique multilingue). Cet intérêt est pleinement motivé par l'éventail des applications entrevues par les militaires, que ce soit dans le cadre de la communication ou de ce qu'il est commun d'appeler « l'intelligence » ou le renseignement militaire.

Pour ce qui est de la communication militaire, les enjeux sont proches des enjeux commerciaux : la communauté multilingue concernée peut être réduite (exemple de la Force d'Action Rapide franco-germanique) ou à l'inverse plus étendue (pays de l'OTAN, intervention de casques bleus de l'ONU) ; les applications peuvent être des IHM (systèmes en service dans plusieurs pays) ou des assistances au dialogue humain (mise en présence de militaires de différents pays).

Le renseignement militaire est, pour sa part, demandeur de systèmes plus fins à la fois capables d'identifier la langue parlée par un locuteur *a priori* non coopératif et de fournir des renseignements sur ce locuteur : il existe des langues parlées sur des espaces couvrant des millions de kilomètres carrés, et une localisation plus précise de l'origine de la personne peut-être souhaitée. Dans ce cas, il est nécessaire d'opérer une modélisation plus fine de chaque langue, et la prise en compte de caractéristiques dialectales requiert alors des connaissances linguistiques poussées. Une des différences principales avec les applications civiles demeure que, dans le cas le plus général, on n'est pas assuré que la langue parlée par le locuteur fasse

partie des langues apprises par le système : cela implique que le système intègre une décision de rejet. Il est bien évident qu'un système ne peut pas reconnaître toutes les langues du monde mais il peut déjà être extrêmement intéressant d'identifier la famille à laquelle une langue appartient. Un système automatique effectuant cette présélection est parfaitement envisageable, mais là aussi, cela implique d'intégrer de manière efficace des connaissances linguistiques fondamentales (quelles familles employer, quels aspects modéliser…). Une dernière application repose là encore sur la notion de distance entre langues puisqu'il s'agit de confirmer si un locuteur donné s'exprime réellement dans la langue qu'il annonce ou plutôt dans une autre. Toutes ces applications nécessitent une exploitation scientifique poussée des résultats fournis par les systèmes en IAL ou en identification des familles linguistiques, et les préoccupations des militaires rejoignent donc à la fois les enjeux des linguistes et ceux des informaticiens.

Comme cela a été précisé au début de ce paragraphe, les premières recherches en IAL ont été entreprises dans les années 70 sous l'impulsion du DoD. Nous allons maintenant étudier les approches envisagées durant les années 70 et les années 80 de manière à dégager les termes actuels de la problématique de l'IAL.

5. Un survol historique (1973-1989)

Dans [Muthusamy 93], l'auteur signale qu'au cours de ces deux décennies, seulement 14 articles sur le sujet ont été publiés. Ce chiffre est à comparer avec les dizaines d'articles qui, depuis le début des années 90, ont été édités dans les différents congrès et revues sur la communication parlée.

5.1. La Genèse

- **Texas Instruments**

Dès 1973, Texas Instruments consacre un effort de recherche soutenu à l'Identification Automatique des Langues. Ces recherches, menées jusqu'en 1980, visent à identifier une langue parmi sept langues candidates (non précisées) à partir d'enregistrements de parole lue [Leonard 80]. L'approche choisie est d'établir des modèles statistiques de motifs phonétiques caractéristiques de chaque langue ainsi que de leur fréquence d'occurrence. L'obtention des motifs pertinents repose sur une sélection manuelle par les experts de TI au cours d'une première phase d'analyse du signal. La phase suivante consiste à implanter des algorithmes d'extraction semi-automatique de ces motifs. Cette approche « experte » implique une bonne connaissance des langues étudiées, et effectivement, lorsque cette connaissance fait défaut, les taux d'identification chutent. Les améliorations des différentes versions ont surtout porté sur la détection des motifs caractéristiques à partir du signal (seuil sur les fréquences d'occurrence, utilisation de critères d'entropie).

- **House et Neuberg**

En 1977, une autre étude en IAL est publiée par House et Neuberg. Il s'agit de modéliser par un réseau de Markov les séquences de macro-classes phonétiques (plosives, fricatives, voyelles...) rencontrées dans chaque langue [House 77]. Ces travaux, menés à partir de transcriptions phonétiques de textes (et non de parole) écrits en huit langues, ont inspiré par la suite la plupart des études menées dans les années 90 et intégrant des considérations

phonotactiques. Le fait que le système ne traite pas de parole réelle ne permet évidemment pas à cette époque de savoir si cette approche est efficace ou non.

- **Discussion**

Les deux voies expérimentées au cours des années 70 ont donc été la modélisation acoustique d'une part, et la modélisation phonotactique d'autre part. Les expériences montrent que l'IAL peut se baser sur ces deux approches, mais aucune tentative intégrant les deux aspects n'a été publiée. D'autre part, les conditions expérimentales (étiquetage manuel ou semi-automatique, transcription phonétique et non signal réel) ne permettent pas encore de préciser les performances que l'on peut espérer de tels systèmes.

5.2. Les Années 80

- **Li et Edwards**

Les principes de modélisation stochastique de séquences de sons vont être repris par Li et Edwards sur des enregistrements de parole lue par des locuteurs masculins de cinq langues [Li 80]. A partir de six classes acoustico-phonétiques (noyaux syllabiques, consonnes fricatives voisées…) trois modèles statistiques (à base de chaînes de Markov) seront développés. Le premier modélise la succession des segments dans chaque langue alors que les deux autres modélisent, dans un cas, la succession de syllabes, et dans l'autre les relations intra-syllabiques. La méthode employée permet en fait de modéliser les séquences de consonnes de chaque langue. Il est intéressant de noter que le système obtenu discrimine efficacement les langues à structure syllabique simple (deux langues asiatiques tonales) des langues européennes caractérisées par des séquences consonantiques plus longues.

Au cours des années 80, les travaux de Li et Edwards ont été les seuls à exploiter une modélisation markovienne. Les autres études menées utiliseront des approches à base de fonctions de décision polynomiales, de règles ou de quantification.

- **Cimarusti et Ives**

Cimarusti et Ives s'intéressent en 1982 à l'identification de neuf langues à partir d'une analyse LPC (Linear Predictive Coefficients) du signal acoustique [Cimarusti 82]. Une centaine de paramètres sont extraits (coefficients d'autocorrélation, coefficients cepstraux, fréquences des formants...). Un classificateur polynomial est optimisé de manière itérative pour les données d'apprentissage puis appliqué sur les données de test. Les résultats obtenus sont bons, mais le faible nombre de locuteurs (tous masculins) ne garantit pas que le système soit indépendant du locuteur.

A partir de ces travaux, Ives développe en 1986 un autre système à base de règles issues du seuillage de 50 des paramètres précédemment utilisés. Les règles qui émergent sont au nombre de 9 et elles sont en partie de nature prosodique (basées sur la fréquence fondamentale F_0 du signal, la variance du second formant F_2, le nombre de voyelles, la densité d'énergie spectrale dans des filtres...).

Les résultats obtenus [Ives 86] sont là encore bons, mais difficiles à évaluer car l'indépendance des corpus de test et d'apprentissage n'est pas clairement précisée.

- **Foil et Goodman**

Les travaux que nous venons de citer ont tous été réalisés sur des corpus de données enregistrés en laboratoire, dans des conditions « propres ». Foil s'intéresse pour sa part à de la parole enregistrée à la radio en trois langues, dans des conditions de bruit bien réelles.

Le système conçu se base à l'origine sur l'identification prosodique (intonation et rythme) des langues [Foil 86]. Le système obtenu par classification bayésienne à partir de F_0 et de l'enveloppe du signal se révèle décevant, et une autre approche est alors développée : par quantification vectorielle, Foil obtient 10 vecteurs formantiques caractéristiques de chaque langue, puis il opère une classification sur les données de test, calculant ainsi une distorsion pour chaque langue.

Ces travaux sont poursuivis par Goodman en 1989, qui améliore la robustesse de chaque élément en développant un nouvel algorithme de suivi de formants, en ajoutant un module de décision voisé/non voisé plus performant et en utilisant une distance euclidienne pondérée pour la classification [Goodman 89]. Il faut ajouter à cela qu'il conserve 60 vecteurs caractéristiques par langue au lieu des 10 retenus par Foil.

- **Discussion**

Si les travaux entrepris dans les années 70 avaient privilégié les approches phonétiques et phonotactiques, on assiste dans les années 80 à une diversification des paramètres discriminants testés. Chaque approche, qu'elle soit basée sur des contraintes phonotactiques, sur des paramètres spectraux ou sur la prosodie, atteint des résultats intéressants bien qu'il soit difficile de juger des conditions expérimentales. On peut noter que les paramètres prosodiques ne sont pas efficacement utilisés dans l'approche bayésienne de Foil mais qu'une approche à base de règles [Ives 86] peut se révéler efficace.

5.3. Vers les systèmes actuels

Le bilan général des années 70 et 80 est contrasté. Des recherches avec des données extrêmement disparates et des méthodes variées ont été menées, mais le domaine de l'IAL n'a pas atteint sa maturité : les systèmes ne sont pas complètement détaillés, les protocoles expérimentaux ne sont pas uniformisés, et aucune approche ne se dégage comme étant une référence.

C'est sur ce constat en demi-teinte que débutent les années 90. En fait, la situation va changer de manière radicale en deux ou trois ans sous l'action conjuguée de plusieurs facteurs, principalement liés à la Reconnaissance Automatique de la Parole (RAP). En effet, au cours des années 80, l'amélioration des systèmes de RAP a été prodigieuse ; on a assisté à la mise sur le marché des premiers systèmes efficaces, les serveurs vocaux opérationnels ont fait leur apparition et, vers 1990, les IHM sont sorties des laboratoires et ont pénétré le monde réel… Cette situation aura plusieurs effets sur l'IAL. Tout d'abord, les enjeux applicatifs deviennent plus pressants puisque dorénavant les systèmes d'IHM sont opérationnels ; il devient urgent d'envisager de les doter d'une capacité multilingue. Ensuite, la plupart des équipes de recherche souhaitent appliquer le savoir-faire acquis au cours des années 80 en RAP à d'autres domaines. On assiste alors à un regain d'intérêt pour des domaines considérés jusqu'alors comme

« secondaires » comme l'identification du locuteur[2] ou de la langue. Dès lors, la dynamique se met en marche : sous l'impulsion conjuguée de la demande applicative et de l'offre scientifique, des corpus de données sont enregistrés [Muthusamy 92]. Il n'en fallait pas plus pour que le domaine de l'IAL émerge comme un thème majeur du TAP, et que les modélisations markoviennes ou neuromimétiques, massivement employées en RAP, s'imposent en IAL.

6. Bibliographie

[Cimarusti 82] D. Cimarusti & R. B. Ives, "Development of an Automatic Identification System of Spoken Languages: Phase 1", *Proc. of ICASSP '8*2, Paris, pp. 1661-1663, (1982)

[Corredor-Ardoy 97] C. Corredor-Ardoy, J.L. Gauvain, M. Adda-Decker & L. Lamel, "Language Identification with Language-Independent Acoustic Models", *Proc. of Eurospeech '9*7, Rhodes, pp. 55-58, (1997)

[Foil 86] J. T. Foil, "Language Identification using Noisy Speech", *Proc. of ICASSP '86*, Tokyo, pp. 861-864, (1986)

[Goodman 89] F. J. Goodman, A. F. Martin & R. E. Wohlford, "Improved Automatic Language Identification in Noisy Speech", *Proc. of ICASSP '89*, Glasgow, pp. 528-531, (1989)

[House 77] A. S. House & E. P. Neuberg, , "Toward Automatic Identification of the Language of an Utterance. I. Preliminary Methodological Considerations", *Journal of the Acoustical Society of America* 62, Vol. 3, pp. 708-713, (1977)

[Ives 86] R. B. Ives, "A Minimal Rule AI Expert system for real-Time Classification of Natural Spoken Languages", *Proc. of 2 nd Artificial Intelligence Advanced Computer Technolog*y, Long Beach, pp. 337-340, (1986)

[Lamel 94] L. F. Lamel & J.L. Gauvain, "Language Identification using Phone-Based Acoustic Likelihood", *Proc. of ICASSP '9*4, Adelaide, pp. 293-296, (1994)

[Leonard 80] R. G. Leonard, *Language Recognition Test and Evaluatio*n, Technical Report RADC-TR- 80-83, RADC/Texas Instruments Inc., Dallas, 1980

[Li 80] K. P. Li & T. J. Edwards, "Statistical Models for Automatic Language Identification", *Proc. of ICASSP '8*0, Denver, pp. 884-887, (1980)

[Muthusamy 93] Y. K. Muthusamy, *A Segmental Approach to Automatic Language Identificatio*n, Ph. D. Thesis, Oregon Graduate Institut of Science & Technology, (1993)

[Muthusamy 94] Y. K. Muthusamy, E. Barnard & R. A. Cole, "Reviewing Automatic Language Identification", *IEEE Signal Processing Magazin*e, 10/94, pp 33-41, (1994)

[2] Dans le cas de l'identification du locuteur, le mouvement a été plus précoce que pour la langue.

Un état de l'art de l'Identification Automatique des Langues

François Pellegrino[1], Régine André-Obrecht[2]

[1] Laboratoire Dynamique Du Langage (UMR CNRS 5596)
[2] Institut de Recherche en Informatique de Toulouse (UMR CNRS 5505)

Francois.Pellegrino@univ-lyon2.fr – obrecht@irit.fr

Abstract

This paper provides a state of the art of automatic language identification. Most of the approaches developed during the last ten years are briefly described. Results are summarized the main tendencies are analyzed from the most recent works.

Résumé

Cet article présente un état de l'art de l'identification automatique des langues. Un panorama des travaux entrepris depuis une dizaine d'années est proposé. Chaque système est succinctement décrit, puis une synthèse des résultats obtenus est proposée, ainsi qu'une analyse des approches employées.

1. Introduction

Cet article fait suite à la « Présentation du cadre de l'identification automatique des langues » proposée précédemment et il constitue une introduction aux travaux récents présentés dans ces actes. Si les travaux menés sur ce thème au cours des décennies 1970 et 1980 furent rares, il n'en est pas de même depuis 1990. Ce domaine est en effet en pleine expansion, en particulier du fait de la nécessité d'étendre les capacités des interfaces homme-machine aux environnements multilingues. La disponibilité, à partir de 1992, du corpus de parole téléphonique OGI MLTS (cf. Section 2) aura permis à de nombreuses équipes de recherche d'évaluer des approches issues pour la plupart de la reconnaissance automatique de la parole : un système « standard », à base de modèles phonétiques et phonotactiques, s'est dégagé de ces travaux, et des recherches visant à intégrer des informations supplémentaires (prosodie,…) ont été menées. La Section 3 de ce document propose un panorama des méthodes employées dans les systèmes actuels, complété par une discussion sur les tendances qui s'en dégagent (Section 4).

2. Les corpus de données

Si les systèmes expérimentaux développés au cours des décennies 70 et 80 ont été validés sur des corpus spécifiques et parfois insuffisamment décrits dans la littérature, les systèmes récents reposent pour la plupart sur des corpus internationalement reconnus. Ces corpus ont été enregistrés à l'initiative d'organismes publics comme le NIST (National Institute of Standards and Technology) aux Etats-Unis ou à la demande des laboratoires eux-mêmes, et la communauté scientifique mondiale dispose actuellement de bases de données et de corpus d'évaluation conséquents.

On peut regrouper ces corpus en deux grands types, selon qu'ils sont constitués de données enregistrées en haute qualité (studio anéchoïque, microphone hi-fi) ou d'appels enregistrés via le canal téléphonique (distorsion, coupure à 3,5 kHz). Le Tableau 1 présente un aperçu des corpus multilingues couramment utilisés en IAL. La taille du corpus (en nombre de locuteurs ou en durée) ne figure pas dans ce tableau puisque sa signification intrinsèque est faible et qu'il serait nécessaire de prendre en compte d'autres facteurs (équilibrage du corpus entre les langues, équilibrage du nombre de locuteurs masculins et féminins, …).

Nous n'allons pas décrire ici la totalité des corpus cités (nous renvoyons le lecteur désireux d'obtenir de plus amples renseignements sur les différentes données aux références données en notes) mais uniquement les corpus EUROM_1, OGI MLTS et CALLFRIEND. Le premier a longtemps été le seul corpus enregistré en studio pour un nombre important de langues, tandis que le second a été le standard utilisé par les systèmes d'IAL lors des campagnes d'évaluation du NIST ; CALLFRIEND est en passe de devenir un corpus de référence pour les campagnes futures.

Nom du corpus	Nombre de langues	Conditions d'enregistrement	Type de parole	Transcriptions (type – quantité)
CALLFRIEND[1]	12 (15)	Téléphone	Conversation	-
CALLHOME[2]	6	Téléphone	Conversation	Orthographique partielle
EUROM_1[3]	11	Studio	Lue	Phonétique totalité
GlobalPhone[4]	9	Studio	Lue	Orthographique totalité
IDEAL[5]	4	Téléphone	Mixte (spontanée / lue)	-
OGI 22 languages[6]	22	Téléphone	Spontanée / contrainte	-
OGI MLTS[7]	11	Téléphone	Spontanée / contrainte	Phonétique partielle

Tableau 1 – Principaux corpus multilingues disponibles

[1] cf. http://www.ldc.upenn.edu/ldc/catalog/html/package/cf.html

[2] cf. http://www.ldc.upenn.edu/ldc/catalog/html/package/ch.html

[3] cf. http://www.icp.grenet.fr/Relator/multiling/eurom1.html

[4] [Schultz 97]

[5] [Lamel 98]

[6] cf. http://www.cse.ogi.edu/CSLU/corpora/22lang/

[7] cf. http://www.cse.ogi.edu/CSLU/corpora/mlts.html

2.1. EUROM_1

Cette base de données a été développée dans le cadre du contrat européen ESPRIT SAM. La version finale doit rassembler des enregistrements en 11 langues européennes. Il s'agit d'enregistrements de parole dite « de laboratoire », c'est-à-dire de données lues, recueillies dans un studio anéchoïque et échantillonnées à 20 kHz. La base est composée pour chaque langue de phrases, de mots isolés et de logatomes (successions de séquences Voyelle-Consonne-Voyelle) prononcés par 30 locuteurs masculins et 30 locuteurs féminins. La totalité du corpus a été phonétiquement étiquetée par des experts. EUROM_1 constitue donc un matériau de choix pour les études phonétiques et phonologiques nécessaires en IAL.

2.2. OGI MLTS [Muthusamy 92]

Cette base de données est la référence dans le monde de l'IAL. Elle a été utilisée successivement avec 6, 9, 10 puis 11 langues pour les campagnes de test du NIST jusqu'en 1995. Contrairement à EUROM_1, il s'agit ici de parole téléphonique échantillonnée à 8 kHz dans une ambiance le plus souvent bruitée. Les locuteurs ont fourni des réponses plus ou moins contraintes à plusieurs questions. Voici un extrait du protocole d'enregistrement proposé par OGI :

1. Quelle est votre langue natale ?
2. Quelle langue parlez-vous la plupart du temps ?
3. Enumérez les chiffres de zéro à dix, SVP.
4. Récitez les sept jours de la semaine, SVP.
5. Parlez-nous du climat de la ville où vous habitez.
6. Décrivez la pièce d'où vous nous appelez.
7. …

Chaque réponse est enregistrée durant un temps donné (10 secondes pour la question 5 par exemple). A cela s'ajoute l'enregistrement d'une minute de parole spontanée pour chaque locuteur, enregistrée sous le nom de « story ». On obtient finalement pour chaque locuteur environ deux minutes de parole. Cette base de données est particulièrement intéressante car elle permet de se rapprocher des conditions d'utilisation réelles de systèmes d'IAL : milieu bruité (parfois même très bruité), canal téléphonique, présence de pauses et d'hésitations dans les énoncés, parole spontanée. Par contre, certains aspects peuvent se révéler plutôt gênants ; en particulier, pour le corpus dit « français », de nombreux locuteurs ont un fort accent de français québécois. Dans ce cas précis, il s'agit plus d'un corpus francophone que français. Il est vraisemblable que cet aspect se retrouve pour d'autres langues (anglophone, hispanophone, …).

De plus, nous n'avons pas encore évoqué les transcriptions phonétiques du corpus. Elles sont de deux types, soit sous forme de classes phonétiques majeures (voyelle, fricative, silence ou closion, explosion, ou autres consonnes) soit sous forme phonétique classique. L'étiquetage phonétique classique est disponible pour six langues[8] pour un nombre de « story » variable allant de 64 (japonais) à 210 (anglais). Il a été réalisé manuellement par des experts. L'étiquetage en classes majeures est quant à lui disponible pour toutes les langues pour une partie du corpus (environ 8 minutes par langue) mais il est réalisé de manière semi-automatique : la procédure d'étiquetage automatique est corrigée manuellement par un expert.

[8] Allemand, anglais, espagnol, hindi, japonais et mandarin.

L'étiquetage obtenu se révèle parfois inexact ou ambigu (cas de bruits non produits par le locuteur et étiquetés comme des occlusives ou cas de phonèmes successifs regroupés au sein d'un même segment).

Ces quelques remarques étant formulées, OGI MLTS est la base de données sur laquelle les performances de la plupart des systèmes sont évaluées à l'heure actuelle, et elle représente à ce titre une contribution majeure au domaine de l'IAL. La plupart des tests sont réalisés (sur recommandation du NIST) avec le signal nommé « story », limité aux 45 premières secondes.

2.3. LDC CALLFRIEND

Le dernier corpus que nous allons décrire ici est aussi l'un des plus récents puisque le projet CALLFRIEND a débuté en 1996. Alors qu'OGI proposait un protocole permettant d'obtenir de la parole quasi-spontanée au moyen de questions, l'approche employée par le LDC (Linguistic Data Consortium) pour collecter de la parole continue « naturelle » est encore plus simple : pour chacune des langues du corpus, 60 conversations totalement spontanées ont été enregistrées, avec une durée allant de 5 à 30 minutes.

Le corpus comprend aussi des informations sur chaque interlocuteur (sexe, age, niveau d'éducation, numéro de téléphone) ainsi que sur chaque appel (qualité de la transmission, nombre d'interlocuteurs). Il s'agit dans tous les cas d'appels locaux et les locuteurs s'expriment toujours dans leur langue natale.

Ce corpus bénéficie de l'expérience dont on dispose aujourd'hui sur l'enregistrement des corpus multilingues, et on peut s'attendre à ce qu'un meilleur contrôle des enregistrements ait été réalisé. Par contre, aucune transcription n'est disponible, et le fait que plusieurs locuteurs soient enregistrés ensemble ajoute des difficultés supplémentaires dans une tâche d'IAL. En 1996, la campagne de test du NIST a principalement portée sur ce corpus, et il est donc en passe de devenir incontournable. De ces trois corpus, celui qui a été le plus employé ces dernières années en IAL est OGI MLTS, comme le montrera le tableau récapitulatif des principaux travaux entrepris au cours des années 90 (Tableau 2).

3. Un panorama des systèmes actuels

Si l'on s'interrogeait au début des années 70 sur les paramètres discriminants pertinents pour distinguer les langues entre elles, les expériences menées jusqu'en 1989 ont apporté quelques éléments de réponse. Depuis cette date en effet, la majeure partie des systèmes s'appuient comme nous allons le voir sur une modélisation phonotactique du langage. De manière à peine exagérée, on peut dire que la modélisation des sons du langage sert uniquement à transformer l'énoncé d'un espace continu acoustique en une suite de symboles discrets. Certains systèmes tirent cependant avantage du décodage acoustico-phonétique, soit en en exploitant explicitement un score d'identification de la langue, soit en optimisant globalement le décodage phonétique et le modèle phonotactique qui le suit.

On peut considérer que les systèmes actuels relèvent principalement de deux approches méthodologiques : la modélisation statistique (3.1), basée sur la recherche de la langue la plus vraisemblable par rapport à des modèles et la modélisation neuro-mimétique (3.2) qui généralise la notion de règle de manière très performante. D'autres approches, moins influencées par la RAP que les précédentes, sont également développées. On peut citer les expériences de S. Itahashi [Itahashi 95] basées sur une modélisation purement prosodique, celles de K. P. Li [Li

94], basées quant à elle sur une méthode d'identification du locuteur (3.3) ou encore celle de F. Pellegrino influencée par les typologies phonologiques des langues.

3.1. Les approches statistiques

3.1.1. Rensselaer Polytechnic Institute, New York, Etats-Unis

Les travaux présentés en 1991 dans [Savic 91] sont caractéristiques du passage des années 80 aux années 90 : alors que le système présenté – basé sur une modélisation markovienne de chaque langue et sur des paramètres issus de la détection de la fréquence fondamentale F_0 – est novateur, les informations sur le corpus utilisé demeurent incomplètes. Pour chaque langue (elles sont au moins 4 dans le corpus), un réseau de Markov ergodique à 5 états est appris (premier module) et, à partir d'un algorithme de détection de F_0, la distribution fréquentielle et les variations de la fréquence fondamentale sont évaluées (second module). Un classificateur (non décrit) est utilisé pour prendre en compte les résultats issus des deux modules. Il est difficile d'évaluer l'apport de chacun d'entre eux puisque les résultats ne sont pas communiqués dans l'article.

3.1.2. LIMSI, France

Les travaux débutés au LIMSI par L. Lamel et J.L. Gauvain s'appuient en grande partie sur la connaissance acquise en modélisation phonétique sur les réseaux de Markov en RAP. Les expériences ont été menées à l'origine sur les 10 langues du corpus initial OGI MLTS [Lamel 94].

A partir des données d'apprentissage, un large modèle ergodique est appris pour chaque langue ; chaque modèle élémentaire représente une unité au niveau phonétique et non une macro-classe phonétique comme dans les travaux précédents [House 77, Savic 91]. L'originalité de l'algorithme est qu'il effectue une optimisation des modèles en prenant en compte la vraisemblance conjointe du signal et de la suite d'unités décodées alors que dans la plupart des systèmes, cette modélisation des séquences décodées (modèle phonotactique) ne modifie pas les modèles acoustico-phonétiques (il s'agit alors d'un post-traitement). Cette approche sera reprise entre autres par Marc Zissman (cf. 3.1.10). Les résultats obtenus au LIMSI, portant sur des données de 10 secondes de durée, sont proches de 60 % d'identification correcte [Lamel 94].

Actuellement, deux axes de recherche sont privilégiés au LIMSI en IAL. Le premier traite de l'intégration de modèles de langage intégrant une modélisation lexicale [Jardino 96, Matrouf 98] et le second recherche une représentation phonétique unifiée par regroupement de phonèmes dépendants des langues en un modèle de décodage acoustico-phonétique global [Corredor-Ardoy 97]. Les résultats les plus récents obtenus par ces approches sont présentés dans ces actes par D. Matrouf et al. et M. Adda-Decker et al.

3.1.3. Ensigma Ltd, Angleterre

Les travaux rapportés dans [Tucker 94] ont été réalisés sur 3 langues (Anglais, Hollandais et Norvégien) issues du corpus EUROM_1.

L'approche adoptée repose sur deux ensembles de modèles. Le premier ensemble, indépendant du langage (IL) est obtenu à partir de modèles estimés sur TIMIT (donc en anglais américain). Le second modèle, dépendant du langage repose sur une ré-estimation des modèles IL avec les données de chaque langue, après un alignement de chaque phrase sur les modèles

IL. La modélisation permet de constituer ainsi un ensemble d'unités spécifiques pour chaque langue.

En phase de test, la décision finale est prise en tenant compte des scores générés lors des décodages effectués avec les différents ensembles de phonèmes, ainsi que d'une statistique calculée sur la fréquence d'occurrence de ces phonèmes. A partir de données de test de 10 secondes de durée, le taux d'identification correcte obtenu est de 90 %.

3.1.4. ATT Bell Labs - Etats-Unis (Kadambe - Hieronymous)

Les travaux menés au laboratoire ATT Bell par S. Kadambe et J. L. Hieronymous ont eux aussi porté sur 3 langues, extraites du corpus OGI MLTS. L'approche présentée (Figure 1) est classique : il s'agit de calculer un score acoustico-phonétique à partir de Modèles de Markov Cachés (MMC) puis d'utiliser la suite de phonèmes ainsi générée en entrée d'un système à base de grammaire N-gramme pour générer un score phonotactique. La reconnaissance phonétique est effectuée par un système développé au laboratoire Bell, basé sur des MMC Continus à Durée Variable, tandis que le score phonotactique est généré par un modèle trigramme. Les résultats obtenus atteignent 91 % d'identification correcte avec des phrases de test de 50 secondes [Kadambe 94].

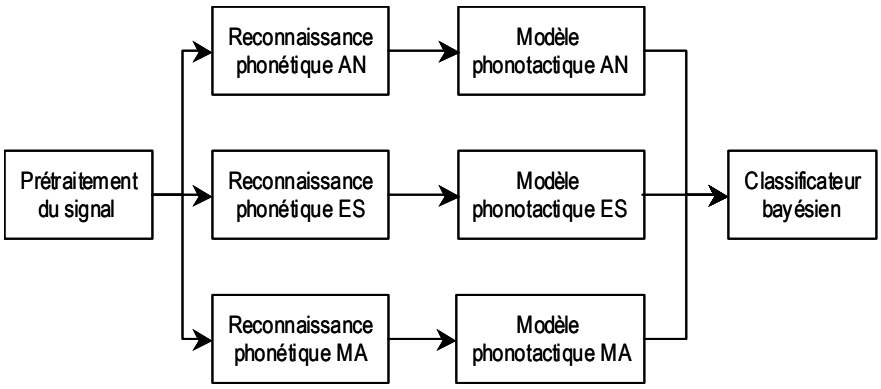

Figure 1 – Schéma bloc du système de ATT Bell Labs (d'après [Kadambe 94]) AN = anglais, ES = espagnol et MA = mandarin

D'autres travaux ont ensuite été entrepris en intégrant pour chaque langue un modèle lexical à base de transducteurs et en utilisant une procédure de normalisation des scores phonotactiques par la vraisemblance de la meilleure séquence phonétique (sans contraintes phonotactiques). Les performances du système se trouve alors nettement amélioré puisqu'il atteint 98 % d'identification correcte sur 5 langues du corpus OGI.

Il est intéressant de noter que si la reconnaissance phonétique se passe mal, il n'est pas possible de compenser son erreur ; elle se répercute au niveau du modèle phonotactique en pénalisant la langue parlée. C'est pour cette raison que la plupart des autres auteurs ont choisi de connecter un modèle phonotactique pour chacune des langues à identifier en sortie de chacun des décodeurs phonétiques (cf. Figure 3).

3.1.5. BBN Systems and Technologies, Etats-Unis

Les travaux menés par M. A. Lund et H. Gish [Lund 95] s'articulent eux aussi autour d'un système de reconnaissance phonétique complété par un modèle de langage. Les tests ont été effectués avec 9 langues issues du corpus OGI MLTS. L'originalité de l'approche réside dans le fait que l'effort de recherche porte entièrement sur le modèle phonotactique : le décodage phonétique utilise en effet un réseau de Markov développé uniquement sur la langue anglaise, sous forme de MMC à 3 états représentant les 50 phonèmes retenus. Cela revient à projeter les données acoustiques de toutes les langues dans l'espace phonétique de la langue anglaise. Deux approches sont étudiées pour les modèles de langage, toutes deux basées sur des grammaires bigrammes :

- La première consiste à générer des grammaires, non seulement sur les phonèmes mais aussi sur des unités plus longues, appelées pseudo-mots (extraites par un algorithme dynamique basé sur la similarité des séquences de phonèmes),
- La seconde approche, itérative, est une méthode d'application de l'algorithme EM aux grammaires bigrammes sur les phonèmes. Les résultats fournis ne portent que sur des expériences de discrimination entre deux langues (Langue 1 vs Langue 2), et en moyenne le score obtenu est de 92,4 %.

3.1.6. ATT Bell Labs - Etats-Unis (Ramesh - Roe)

Nous trouvons dans [Ramesh 94] un autre travail développé au sein des laboratoires Bell par R. Ramesh et D. B. Roe. L'étude réalisée vise à démontrer l'apport d'une modélisation par mots-clefs dans une tâche d'Identification des Langues dans un cadre restreint puisqu'il s'agit d'enregistrements ayant trait à une tâche d'opérateur téléphonique dans 4 langues.

Les données utilisées durant la phase de développement ont été enregistrées chez ATT sur une ligne téléphonique numérique. Les tests finals sont faits avec des données OGI MLTS dans les mêmes langues. Un réseau de Markov est utilisé, avec une modélisation explicite des 30 à 40 mots-clefs retenus par langue, et une modélisation du reste de la parole en unités représentées par des réseaux gauche-droite de 5 à 10 états. Pour ne pas détériorer les résultats en passant du canal numérique (apprentissage) au canal analogique (test), une normalisation cepstrale et une modélisation du canal (par un modèle MMC adapté sur une phase de « silence ») sont étudiés. Les résultats mentionnés font état de 96 % d'identification correcte avec des séquences de test de durée de 5 à 10 secondes (il s'agit des énumérations de chiffres).

3.1.7. Université d'Aalborg, Danemark

En introduisant la notion de polyphonèmes (communs à plusieurs langues) et de monophonèmes (spécifiques à chaque langue), P. Dalsgaard et O. Andersen [Dalsgaard 94, Andersen 97] aboutissent à un réseau décrivant les 4 langues de leur corpus (issues de EUROM_0) avec 114 phonèmes modélisés par des MMC à trois états. Une première phase d'identification, menée par algorithme de Viterbi, aboutit classiquement à la reconnaissance de la suite de phonèmes la plus probable (au sens de Viterbi). Une deuxième phase est alors appliquée, par modification des poids de chaque phonème (initialement équi-pondérés) en tenant compte des matrices de confusion inter-phonèmes établies durant l'apprentissage pour chaque langue : on augmente ainsi l'importance relative des phonèmes les plus discriminants. Avec des phrases de test de 2 minutes, le score d'identification obtenu est de 88,1 %.

3.1.8. Université de Tokyo, Japon

Nous voici encore en présence d'un système conjuguant une étape de reconnaissance phonétique et un modèle de grammaire N-gramme [Kwan 95]. 5 langues ont été retenues dans le corpus OGI MLTS, et plusieurs approches ont été testées. La première étude porte sur la structure à utiliser en phase de reconnaissance phonétique, à savoir s'il est plus efficace d'employer plusieurs systèmes de reconnaissance en parallèle, chacun modélisant une langue, (comme le système présenté Figure 1) ou un système unique modélisant l'ensemble des langues, et construit à partir de tous les phonèmes, communs à toutes les langues ou spécifiques (Figure 2).

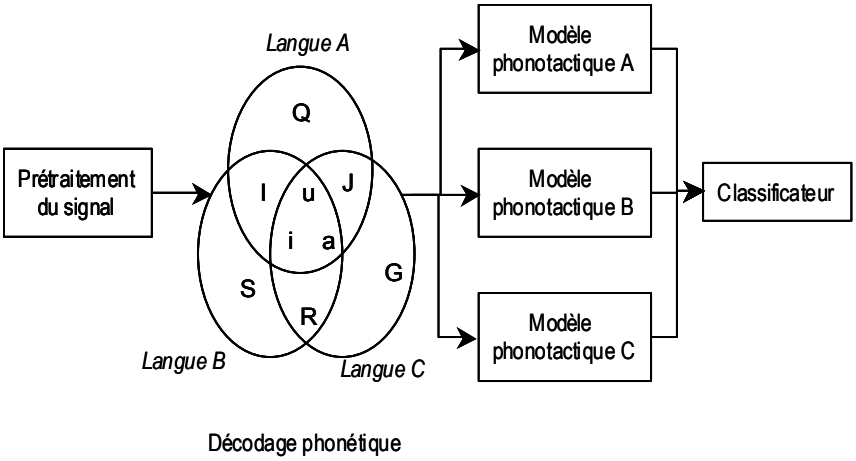

Figure 2 – Synopsis du Système MPR (Mixed Phoneme Recognition) d'après [Kwan 95].

La seconde étude porte sur le modèle de langage, Kwan et Hirose étudiant l'efficacité de modèles unigrammes et bigrammes pour chaque paire de langues. Les meilleurs résultats en discrimination (Langue 1 vs Langue 2) sont obtenus avec le système de décodage MPR et les modèles unigrammes (78 % d'identification correcte en moyenne). Selon les auteurs, la dégradation – inattendue – des résultats en utilisant des grammaires bigrammes s'explique par le manque de données pour certaines langues.

3.1.9. MIT, Etats-Unis, (Hazen - Zue)

Il est proposé dans [Hazen 97] un système d'IAL basé sur le système de reconnaissance phonétique SUMMIT [Zue 90]. A ce module de base (indépendant de la langue) s'ajoutent un module phonotactique (fondé sur une grammaire trigramme) et un module prosodique. Ce dernier prend en compte, d'une part F_0 au niveau segmental (modèle statistique gaussien) et d'autre part la durée des segments issus de SUMMIT (là encore un modèle statistique gaussien par classe phonétique). Les expériences rapportées ont tendance à montrer que le module prosodique est inefficace, et que le modèle trigramme est performant puisque le taux d'identification correcte atteint 78,1 % avec les phrases de test de 45 secondes pour 11 langues du corpus OGI MLTS.

3.1.10. MIT, Etats-Unis, (Zissman)

Un autre système développé au MIT est décrit dans [Zissman 96]. Parmi plusieurs approches, celle offrant les meilleurs résultats est basée sur une architecture hybride entre décodage acoustico-phonétique spécifique à chaque langue et décodage commun. Le système de base (PRLM : Phone Recognition followed by Language Modeling) repose non pas sur le seul système de décodage SUMMIT mais sur des modules acoustico-phonétiques développés avec HTK[9]. En sortie de ce seul décodeur, un modèle n-gramme est appliqué pour chacune des langues à identifier. L'identification est donnée par la vraisemblance phonotactique la plus élevée. Cette approche effectuant une projection aveugle dans un espace phonétique *a priori* peut se révéler inefficace si les caractéristiques phonétiques des langues à identifier sont trop éloignées du décodeur employé. Pour éviter cet effet, Zissman utilise plusieurs décodeurs acoustico-phonétiques de manière à optimiser la couverture de l'espace acoustico-phonétique et à permettre de rattraper les éventuelles erreurs de décodage d'une langue.

Cette approche (Parallel PRLM) est proposée dans le cas où l'on ne dispose pas d'un décodeur acoustico-phonétique pour chacune des langues traitées. Ce système, implanté avec 6 décodeurs (allemand, anglais, espagnol, hindi, japonais et mandarin) obtient 80 % d'identification correcte avec les 11 langues du corpus OGI MLTS (Figure 3). Sur la figure, les décodeurs phonétiques sont numérotés de 1 à 6 et les langues à reconnaître de A à K. Il y a un modèle phonotactique par langue à reconnaître pour chacun des décodeurs phonétiques, soit un total de 66 modèles phonotactiques dans le cas présent. En intégrant plusieurs améliorations (modélisation acoustico-phonétique dépendante du sexe, modélisation de la durée, …), le score d'identification correcte atteint 89 % avec les phrases de test de 45 secondes.

Ne disposant pas d'un décodeur acoustico-phonétique pour chacune des 11 langues, Zissman n'a pas testé l'approche consistant à optimiser conjointement les unités du décodeur et les modèles de langage (approche proposée dans [Lamel 94]).

[9] *Entropic*, la société commercialisant HTK, a depuis été rachetée par Microsoft. A l'heure actuelle, tous les produits de la gamme HTK ont été retirés de la vente. Pour plus d'informations, consulter le serveur http://www.entropic.com

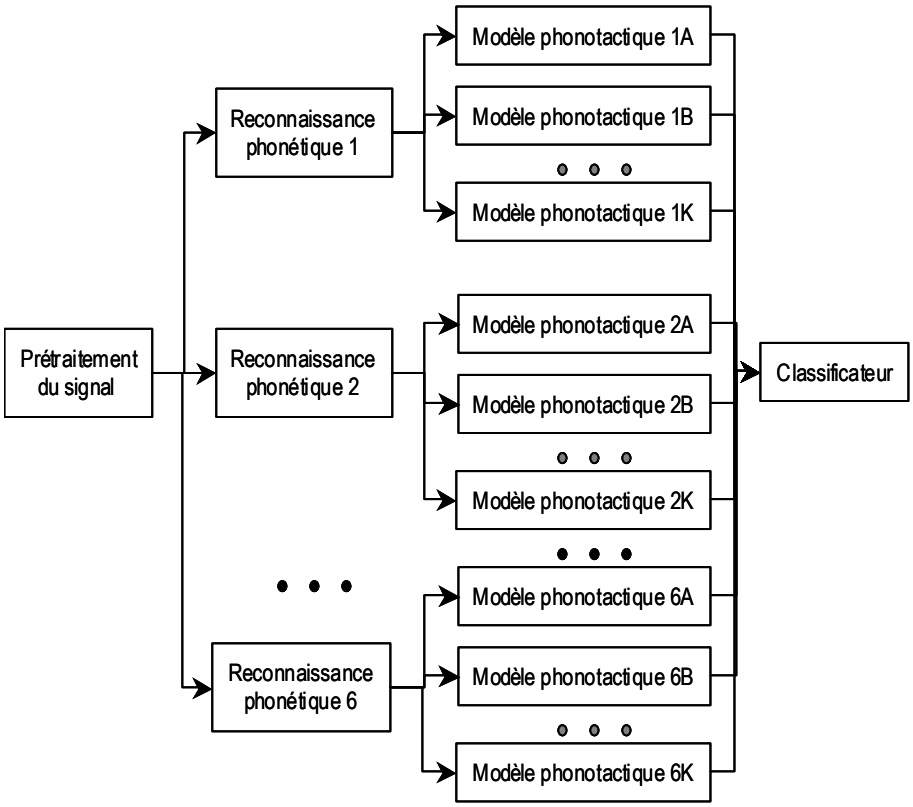

Figure 3 – Schéma du système Parallel PRLM (d'après [Zissman 96]) ; 3 des 6 décodeurs phonétiques sont représentés et pour chacun d'entre eux, 3 des 11 modèles phonotactiques sont indiqués.

3.1.11. OGI, Etats-Unis, (Yan - Barnard)

Le système développé à OGI [Yan 96] repose sur la même architecture que le Parallel PRLM de Zissman. 6 décodeurs à base de MMC sont utilisés pour fournir les séquences phonétiques aux 66 modèles de langage. L'effort de recherche a principalement porté sur l'optimisation de ces derniers et sur le classificateur final (réseau de neurones). Le modèle phonotactique s'articule en trois modules :

- une grammaire bigramme classique prenant en compte le contexte passé du phonème (modèle backward),
- une grammaire bigramme prenant en compte le contexte futur du phonème (modèle forward),
- un modèle de durée.

L'objectif est bien évidemment d'approcher les performances que l'on peut attendre d'un modèle trigramme sans en avoir la complexité et surtout le besoin considérable en données. Avec cette méthode et une classification par réseau de neurones, Yan atteint 86,7 % d'identification correcte sur les 11 langues avec les énoncés de 45 secondes de durée.

3.1.12. Technical University of Ilmenau, Allemagne

Les travaux entrepris par J. Navrátil et W. Zühlke [Navrátil 97] visent aussi à optimiser les modèles de langage utilisés en IAL. A partir d'un unique décodeur phonétique implanté sous HTK avec les phonèmes anglais, des modèles de langage optimisés par deux approches sont appris :

- La première approche est basée sur une grammaire bigramme tenant compte des deux contextes précédant chaque phonème par l'intermédiaire du calcul de classes d'équivalences à partir de ces deux contextes,
- la seconde approche est basée sur des arbres de décision binaires.

Les expériences montrent que les deux modèles ainsi établis améliorent les performances par rapport à un modèles bigramme classique et que l'utilisation conjointe des deux modèles est encore plus efficace : le taux d'identification obtenu avec les phrases de 45 secondes de 9 langues du corpus OGI MLTS s'élève à 90,6 %.

3.1.13. IRIT, France

Le système développé à l'IRIT repose sur une modélisation acoustico-phonétique ne nécessitant aucune donnée étiquetée pour l'apprentissage. Il est inspiré des typologies phonologiques développées en linguistique. Deux modèles de mélanges de lois gaussiennes segmentaux sont évalués pour chaque langue. Le premier modèle prend en compte la totalité des segments de parole du corpus d'apprentissage obtenus par l'algorithme de divergence forward-backward [André-Obrecht 88] alors que le second modélise le système vocalique de la langue à partir des segments vocaliques détectés automatiquement dans le signal. L'utilisation conjointe des deux modèles en identification de 5 langues du corpus OGI MLTS permet d'atteindre 91 % d'identification correcte (énoncés de 45 secondes) avec les locuteurs masculins et sans utiliser aucune information phonotactique.

3.2. Les approches neuro-mimétiques

3.2.1. OGI, Etats-Unis, (Berkling - Barnard)

E. Barnard mène avec K. Berkling des travaux basés sur une modélisation acoustico-phonétique par réseaux de neurones [Berkling 95]. Ces recherches visent à réduire la quantité de données nécessaires lors de l'apprentissage en regroupant les phonèmes. Un premier algorithme est développé pour extraire des unités acoustiques « mixtes » (à mi-chemin entre phonèmes et classes phonétiques majeures) pertinentes en IAL. Le système s'articule ensuite autour de deux modules :

- un réseau de neurones est chargé de reconnaître la suite d'unités prononcées parmi les 60 unités issues de l'algorithme ,
- un classificateur (linéaire ou neuronal) calcule des statistiques sur les fréquences d'occurrence des différentes unités et suites d'unités pour chacune des langues.

Les meilleurs résultats sont de 59 % d'identification correcte avec 6 langues du corpus OGI MLTS.

3.2.2. OGI, Etats-Unis, (Muthusamy)

Y. K. Muthusamy, auteur d'un fameux article sur l'IAL [Muthusamy 94] a développé un système neuronal à deux niveaux [Muthusamy 93] : 9le premier effectue la segmentation du

signal en 7 classes majeures. Le chemin trouvé par le réseau neuronal est optimisé par un algorithme de Viterbi, 9le second prend en charge l'identification de la langue à partir du chemin trouvé et d'un grand nombre de critères statistiques (fréquences d'occurrence de motifs bigrammes et trigrammes, …). Il s'agit là encore d'un réseau neuronal. Les meilleurs résultats obtenus par Muthusamy sont de 62,4 % d'identification correcte sur les phrases de 45 secondes de 10 langues du corpus OGI MLTS.

3.3. Les autres approches

3.3.1. Un système d'identification prosodique

S. Itahashi propose un système d'identification des langues basé uniquement sur la prosodie [Itahashi 95]. Il ne s'agit pas là de la première tentative ([Ives 86], [Foil 86], [Hazen 97]) mais assurément d'une des plus réussies. Après avoir opéré une détection voisé/non voisé, Itahashi effectue l'approximation de la courbe de F_0 par des fonctions linéaires par morceaux afin d'extraire 21 paramètres statistiques (liés aux variations de F_0 et d'énergie) de chaque zone voisée. A partir d'une analyse discriminante utilisant une distance de Mahalanobis, les résultats obtenus avec les locuteurs masculins de 6 langues du corpus OGI MLTS sont remarquables puisqu'ils atteignent 63,3 % d'identification correcte avec 20 secondes de parole.

3.3.2. Un système d'identification du locuteur

L'approche proposée par K. P. Li en 1994 est bien différente des autres approches contemporaines et de ses travaux des années 80 [Li 80, Li 94]. Il s'agit en effet d'appliquer des techniques de reconnaissance du locuteur à l'IAL. Le processus, constitué de deux étapes, suppose que la langue à identifier est celle du locuteur le plus proche dans la base d'apprentissage. La première phase consiste à extraire les noyaux syllabiques du signal acoustique avec un réseau de neurones, puis à calculer 75 paramètres acoustiques et spectraux. La seconde étape consiste alors à déterminer le locuteur le plus proche. Cette approche se révèle pertinente puisque les résultats obtenus sur 10 langues du corpus OGI MLTS sont de l'ordre de 78 % d'identification correcte pour les phrases de 45 secondes de durée.

4. Discussion

Nous allons dans cette section récapituler les résultats obtenus par les différents systèmes et étudier les tendances générales qui s'en dégagent.

4.1. Tableau récapitulatif

Le Tableau 2 donne, pour chaque système cité dans ces pages, plusieurs informations sur son architecture et sur les résultats qu'il obtient.

Dans le tableau, *n. p.* indique *non précis*é, MMC Modèle de Markov Caché, MMG Modèle de Mélange de Gaussiennes et RN Réseau Neuromimétique.

Référence	Reconnaissance phonétique	Modèle phonotactique	Prosodie	Corpus	Nombre de langues	Résultats	
						Durée	%
Savic 91	MMC	-	oui	*n. p.*	*n. p.*	*n. p.*	*n. p.*
Muthusamy 93	RN (1)	RN	-	OGI MLTS	10	45 s	62,4
Berkling 94	RN	-	-	OGI MLTS	3	45 s	74,2
Dalsgaard 94	MMC (1)	-	-	EUROM 0	4	2 min.	88,1
Kadambe 94	MMC	N-gramme	-	OGI MLTS	3	45 s	91
Lamel 94	MMC	N-gramme	-	OGI MLTS	10	10 s	59,7
Li 94	RN	-	-	OGI MLTS	10	45 s	78
Ramesh 94	MMC	Modèles de mots	-	OGI MLTS[10]	4	10 s	96
Tucker 94	MMC	N-gramme	-	EUROM 1	3	10 s	90
Itahashi 95	-	-	oui	OGI MLTS[11]	6	20 s	63,3
Kwan 95	MMC (1)	N-gramme	-	OGI MLTS	5	45 s	78
Lund 95	MMC (1)	N-gramme	-	OGI MLTS	9	*n. p.*	*n. p.*
Hieronymous 96	MMC	N-gramme Modèle lexical	-	OGI MLTS	5	45 s	98
Yan 96	MMC (6)	N-gramme	-	OGI MLTS	11	45 s	86,7
Zissman 96	MMC (6)	N-gramme	-	OGI MLTS	11	45 s	89
Andersen 97	MMC	N-gramme	-	OGI MLTS	3	45 s	83,7
Corredor-Ardoy 97	MMC (1)	N-gramme	-	IDEAL	4	10 s	91
Hazen 97	MMC (1)	N-gramme	oui	OGI MLTS	11	45 s	78,1
Navrátil 97	MMC (1)	N-gramme Arbre binaire	-	OGI MLTS	9	45 s	90,6
Matrouf 98	MMC	N-gramme Modèle lexical	-	IDEAL[12]	4	5 s	92
Pellegrino 98	MMG Global MMG vocalique	-	-	OGI MLTS	5	45 s	91

Tableau 2 – Récapitulatif des études en IAL citées

La description de chaque système comporte les champs suivants :

Référence : il s'agit de l'article d'où sont tirées les informations suivantes.

Reconnaissance phonétique : lorsque le système repose sur un décodage acoustico-phonétique, son type (gaussien, markovien ou neuronal) est mentionné. Lorsqu'il y a un chiffre entre parenthèses, il s'agit du nombre de décodeurs mis en parallèle dans l'expérience ; si ce chiffre n'est pas mentionné, il y a un décodeur par langue.

[10] Il s'agit d'une modélisation par mots-clefs, seules les énumérations de chiffres sont prises en compte.

[11] Seuls les locuteurs masculins sont pris en compte.

[12] Seuls les enregistrements de parole spontanée sont considérés.

Modèle phonotactique : ce champ indique le type du modèle phonotactique utilisé (généralement N-gramme).

Prosodie : ce champ indique si un modèle prosodique est employé dans le système.

Corpus : on trouve ici le nom du corpus sur lequel les résultats indiqués sont obtenus.

Nombre de langues : il s'agit bien évidemment du nombre de langues sur lesquelles le test a porté.

Résultats : on trouve ici le résultat en pourcentage d'identification correcte dans une tâche d'identification des langues n'intégrant pas de décision de rejet. La durée des stimuli utilisés pour le test est également précisée.

4.2. Tendances générales en IAL

Sur les 21 études succinctement décrites dans ce chapitre, 17 utilisent une modélisation statistique markovienne ou multigaussienne pour effectuer un décodage acoustico-phonétique, 2 effectuent ce décodage via des réseaux neuromimétiques et 2 autres systèmes reposent sur des critères différents (modélisation statistique de la prosodie dans [Itahashi 95] et modélisation neuronale du locuteur dans [Li 94]).

La plupart des systèmes s'articulent en deux modules, le premier effectuant un décodage acoustico-phonétique de manière à fournir une ou plusieurs séquences d'unités phonétiques discrètes en entrée d'un second module, généralement basé sur une grammaire statistique, qui modélise alors les contraintes phonotactiques de la langue (12 systèmes sur 19). Les systèmes qui n'exploitent pas ce type d'information obtiennent généralement de moins bons résultats.

4.2.1. Modélisation acoustico-phonétique

Si l'on étudie plus précisément les systèmes de décodage acoustico-phonétique, la tendance actuelle (95-98) privilégie l'usage d'un unique décodeur, commun à toutes les langues. Il peut être construit à partir des unités phonétiques d'une seule langue [Lund 95, Hazen 97, Navrátil 97] ou en faisant émerger un ensemble d'unités couvrant l'espace phonétique de toutes les langues [Berkling 95, Kwan 95, Corredor-Ardoy 97]. L'approche inverse, visant à obtenir un ou plusieurs modèles phonétiques pour chacune des langues, et ce sans recourir à des données étiquetées, est également étudiée [Pellegrino 98]. L'efficacité des modèles repose alors sur le choix d'espaces de modélisation adaptés aux différentes classes de sons (modèles différentiés pour les systèmes vocaliques et consonantiques par exemple).

Une approche intermédiaire consiste à utiliser plusieurs décodeurs dépendants d'une langue en parallèle, même s'ils ne correspondent pas aux langues à identifier [Yan 96, Zissman 96]. L'objectif est alors d'augmenter la robustesse de l'ensemble en recombinant plusieurs scores phonotactiques pour chaque langue plutôt qu'en en calculant un seul. Cette tendance fait reposer l'identification quasiment intégralement sur les modèles de langage, puisqu'en sortie du (ou des) décodeur(s) acoustico-phonétique, aucun score d'identification n'est généré. Le principal avantage de cette méthode est de compenser l'absence de système de décodage acoustico-phonétique performant pour certaines langues et son efficacité justifie à elle seule l'usage qui en est fait. Cela dit, un tel décodage du signal est probablement réducteur, et il sous-exploite « l'identité phonétique » de chaque langue : en opérant une *projection* de données acoustiques d'une langue *X* dans l'espace phonétique de la langue *Y*, on perd une partie de l'information qui peut se révéler capitale, en particulier lorsque l'on réduit la durée des stimuli. Le fait d'utiliser plusieurs décodeurs Y_1 à Y_M permet certes de diminuer ces pertes (par analogie avec du traitement

d'antennes), mais, comme le souligne Zissman, le choix des décodeurs est alors crucial puisqu'il conditionne les performances du système.

4.2.2. Modélisation phonotactique et lexicale

Si la modélisation des contraintes phonotactiques des langues par des grammaires de type n-gramme est largement employée en IAL, la tendance actuelle vise à faire émerger des contraintes de plus haut niveau en modélisant des unités plus longues. Il peut s'agir des mots les plus courants [Hieronymous 96, Matrouf 98] ou des séquences les plus courantes, qu'elles représentent des mots ou non. L'objectif est bien évidemment d'intégrer des traits morpho-lexicaux ou même rythmiques (les séquences modélisées peuvent s'apparenter à un modèle de rythme syllabique), de manière à exploiter au mieux l'information présente dans le signal. Si l'on observe l'évolution des performances obtenues depuis 1996, on se rend compte qu'on semble cependant atteindre une limite ou tout au moins un palier : il est probable que le pouvoir discriminant des modèles phonotactiques soit fortement limitée par la précision du décodage acoustico-phonétique. On retombe ainsi sur le problème à notre avis crucial de la modélisation phonétique. La possibilité de prendre en compte de nouvelles informations (bien évidemment lexicales mais aussi prosodiques) semble donc inéluctable.

4.2.3. Modélisation prosodique

La modélisation de la prosodie en IAL demeure problématique. Si des expériences montrent de manière indéniable qu'il est possible d'extraire des informations pertinentes du niveau prosodique (cf. les articles de P. Dominey, P. Paulin et F. Ramus dans ces mêmes actes), leur modélisation pose un problème non trivial. Les approches basées sur une modélisation segmentale du F_0 et de l'énergie se révèlent assez logiquement décevantes [Hazen 97], mais la voie reste ouverte à des modélisations supra-segmentales [Itahashi 95], inspirées par exemple des travaux entrepris en synthèse de la parole [Hirst 98].

4.2.4. Décision de rejet

L'intégration d'une décision de rejet est depuis longtemps une priorité en identification automatique du locuteur : il est préférable que le système ne rende pas de décision plutôt qu'il en retourne une erronée. Cette constatation bien connue est parfaitement transposable à l'IAL, et les chercheurs en ont pleinement conscience. En effet, plusieurs travaux récents [Kwan 97, Parris 97] portent sur ce point précis, et il est prévisible que cette tendance s'accentue car l'intégration d'une décision de rejet efficace à un système est un préalable à son utilisation hors d'un laboratoire. Cela nécessite cependant, tout comme en identification du locuteur, d'évaluer des modèles du « monde », c'est-à-dire des langues non présentes dans le corpus.

5. Conclusion

Le panorama des recherches menées en IAL depuis environ cinq ans permet de constater que ce domaine est en passe d'atteindre sa maturité. L'émergence d'une approche standard (modélisation phonétique + phonotactique) ainsi que la disponibilité de corpus rigoureux permettent d'évaluer les apports relatifs des différentes méthodes proposées. On constate de plus que la tendance est à la diversification des traits caractéristiques pris en compte. Que ce soit en adaptant la reconnaissance de la parole grand vocabulaire (filtrage lexical) ou en cherchant à modéliser des informations phonologiques (structure du système vocalique) ou prosodiques

(rythme ou mélodie), plusieurs systèmes récents tirent profit de plusieurs niveaux d'analyse. En effet, il est raisonnable de penser que la fusion de classificateurs modélisant des informations complémentaires permettra de caractériser de manière plus robuste les langues du monde.

6. Bibliographie

[Andersen 97] O. Andersen & P. Dalsgaard, "Language-Identification Based on Cross-Language Acoustic Models and Optimised Information Combination", *Proc. of Eurospeech '9*7, Rhodes, pp. 67-70, (1997)

[André-Obrecht 88] R. André-Obrecht, "A New Statistical Approach for Automatic Speech Segmentation", *IEEE Trans. on ASS*P, Vol. 36, n° 1, pp 29-40, (1988)

[Berkling 95] K. M. Berkling, T. Arai & E. Barnard, "Theoretical Error Prediction for a Language Identification System using Optimal Phoneme Clustering", *Proc. of Eurospeech '9*5, Madrid, pp. 351-354, (1995)

[Cimarusti 82] D. Cimarusti & R. B. Ives, "Development of an Automatic Identification System of Spoken Languages: Phase 1", *Proc. of ICASSP '8*2, Paris, pp. 1661-1663, (1982)

[Corredor-Ardoy 97] C. Corredor-Ardoy, J.L. Gauvain, M. Adda-Decker & L. Lamel, "Language Identification with Language-Independent Acoustic Models", *Proc. of Eurospeech '9*7, Rhodes, pp. 55-58, (1997)

[Dalsgaard 94] P. Dalsgaard & O. Andersen, "Application of Inter-Language Phoneme Similarities for Language Identification", *Proc. of ICSLP '9*4, Yokohama, pp. 1903-1906, (1994)

[Hazen 97] T. J. Hazen, & V. W. Zue, "Segment-based automatic language identification", *Journal of the Acoustical Society of Americ*a, Vol. 101, No. 4, pp. 2323-2331, April, (1997)

[Hieronymous 96] J. Hieronymous & S. Kadambe, "Spoken Language Identification Using Large Vocabulary Speech Recognition", *Proc. of ICSLP '9*6, Philadelphia, (1996).

[Hirst 98] D. J. Hirst, A. Di Cristo & R. Espesser (sous presse) "Levels of representation and levels of analysis for intonation" dans *Prosody : Theory and Experiment*s, Edited by M. Horne, Kluwer Academic Publishers, Dordrecht, (1998)

[House 77] A. S. House & E. P. Neuberg, , "Toward Automatic Identification of the Language of an Utterance. I. Preliminary Methodological Considerations", *Journal of the AcousticalSociety of America* 62, Vol. 3, pp. 708-713, (1977)

[Itahashi 95] S. Itahashi & L. Du, "Language Identification Based on Speech Fundamental Frequency", *Proc. of Eurospeech '9*5, Madrid, pp. 1359-1362, (1995)

[Ives 86] R. B. Ives, "A Minimal Rule AI Expert system for real-Time Classification of Natural Spoken Languages", *Proc. of 2 nd Artificial Intelligence Advanced Computer Technolog*y, Long Beach, pp. 337-340, (1986)

[Jardino 96] M. Jardino, "Multilingual Stochastic N-Gram Class Language Models", *Proc. of ICASSP '9*6, Atlanta, pp. 161-164, (1996)

[Kadambe 94] S. Kadambe & J. L. Hieronymous, “Spontaneous Speech Language Identification with a Knowledge of Linguistics”, *Proc. of ICSLP '9*4, Yokohama, pp. 1879-1882, (1994)

[Kwan 95] H. Kwan & K. Hirose, “Recognized Phoneme-Based N-Gram Modeling in Automatic Language Identification”, *Proc. of Eurospeech '9*5, Madrid, pp. 1367-1370, (1995)

[Kwan 97] H. Kwan & K. Hirose, “ Use of Recurrent for Unknown Language Rejection in Language Identification System”, *Proc. of Eurospeech '9*7, Rhodes, pp. 63-66, (1997)

[Lamel 94] L. F. Lamel & J.L. Gauvain, “Language Identification using Phone-Based Acoustic Likelihood”, *Proc. of ICASSP '9*4, Adelaide, pp. 293-296, (1994)

[Lamel 98] L. F. Lamel, M. Adda-Decker, C. Corredor, J.J. Gargolf & J.L. Gauvain, , “A Multilingual Corpus for Language Identification”, *Proc. of 1 st International Conference on Language Resources & Evaluatio*n, Granada, pp. 1118-1122, (1998)

[Leonard 80] R. G. Leonard, *Language Recognition Test and Evaluatio*n, Technical Report RADC-TR- 80-83, RADC/Texas Instruments Inc., Dallas, 1980

[Li 80] K. P. Li & T. J. Edwards, “Statistical Models for Automatic Language Identification”, *Proc. of ICASSP '8*0, Denver, pp. 884-887, (1980)

[Li 94] K. P. Li, “Automatic Language Identification using Syllabic Spectral Features”, *Proc. of ICASSP '9*4, Adelaide, pp. 297-300, (1994)

[Lund 95] M. A. Lund & H. Gish, “Two Novel Language Model Estimation Techniques for Statistical Language Identification”, *Proc. of Eurospeech '9*5, Madrid, pp. 1363-1366, (1995)

[Matrouf 98] D. Matrouf, M. Adda-Decker, L. F. Lamel & J. L. Gauvain, Language identification incorporating lexical information”, *Proc. of ICSLP'9*8, Sidney, pp. 181-184, (1998)

[Muthusamy 92] Y. K. Muthusamy, R. A. Cole & B. T. Oshika, “The OGI Multilingual Telephone speech Corpus”, *Proc. of ICSLP '9*2, Banff, pp. 895-898, (1992)

[Muthusamy 93] Y. K. Muthusamy, *A Segmental Approach to Automatic Language Identificatio*n, Ph. D. Thesis, Oregon Graduate Institut of Science & Technology, (1993)

[Muthusamy 94] Y. K. Muthusamy, E. Barnard & R. A. Cole, “Reviewing Automatic Language Identification”, *IEEE Signal Processing Magazin*e, 10/94, pp 33-41, (1994)

[Navrátil 97] J. Navrátil & W. Zuhlke, “Phonetic-Context Mapping in Language Identification”, *Proc. of Eurospeech '9*7, Rhodes, pp. 71-74, (1997)

[Parris 97] E. S. Parris, H. Lloyd-Thomas, M. J. Carey & J. H. Wright, “Bayesian Methods for Language Verification”, *Proc. of Eurospeech '9*7, Rhodes, pp. 59-62, (1997)

[Pellegrino 98] F. Pellegrino, *Une approche phonétique en identification automatique des langues : la modélisation acoustique des systèmes vocalique*s, thèse de 3 ème cycle, Univ. Paul Sabatier, Toulouse, (1998)

[Ramesh 94] P. Ramesh & D. B. Roe, “Language Identification with Embedded Word Models”, *Proc. of ICSLP '9*4, Yokohama, pp. 1887-1890, (1994)

[Savic 91] M. Savic, E. Acosta & S. K. Gupta, “An Automatic Language Identification System”, *Proc. of ICASSP '9*1, Toronto, pp. 817-820, (1991)

[Schultz 97] T. Schultz & A. Waibel, “Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets”, *Proc. of Eurospeech '9*7, Rhodes, pp. 371-374, (1997)

[Tucker 94] R. C. F. Tucker, M. J. Carey & E. S. Parris , “Automatic Language Identification using Sub-Words Models”, *Proc. of ICASSP '9*4, Adelaide, pp. 301-304, (1994)

[Yan 96] Y. Yan, E. Barnard & R. A. Cole, “Development of An Approach to Automatic Language Identification based on Phone Recognition”, *Computer Speech and Languag*e, Vol. 10, n° 1, pp 37-54, (1996)

[Zissman 96] M. A. Zissman, “Comparison of Four Approaches to Automatic Language Identification of Telephone Speech”, *IEEE Trans. on Speech and Audio Processin*g, Vol. 4, n° 1, pp 31-44, (1996)

[Zue 90] V. Zue, J. Glass, D. Goodine, H. Leung, M. Phillips, J. Polifroni & S. Seneff, “Recent Progress on the SUMMIT System”, *3 rd DARPA Speech and Natural Language Worksho*p, pp. 380-384, (1990)

Typologies des structures sonores des langues du monde

Tendances et diversité

Nathalie Vallée & Louis-Jean Boë

Institut de la Communication Parlée
UPRESA 5009 INPG/Université Stendhal, BP 25 38040 Grenoble Cedex

{vallee, boe}@icp.inpg.fr

Abstract

Automatic Language Identification (ALI) implies large fundamental knowledge about linguistic taxinomy. These issues have recently motivated a large amount of works and databases about linguistic universals, particularly phonological universals. Representative databases of the world languages are necessary not only to describe linguistic systems, but also to test the validity of sound structure typologies.

Until now, taxinomies and language relationships have been built from lexical criteria, and not on a phonological basis. We think that ALI needs new typologies based on sound structures, since two languages can be supposed to be more easily discriminated if their phonological systems differ significantly. It is therefore necessary to elaborate sound structure typological databases, to allow the use of very distant/very close language specimens for the evaluation of ALI.

Résumé

L'identification automatique de la langue (IAL) à partir d'un échantillon de parole renvoie à des connaissances et à des problèmes fondamentaux de taxinomie linguistique. Ces questions connaissent actuellement un renouveau d'intérêt, comme en témoigne l'avancée des travaux sur les universaux linguistiques – et en particulier phonologiques – associée à l'élaboration et à la diffusion de bases de données représentatives des langues du monde. Outre leur intérêt descriptif, ces bases de données ont permis de mettre à l'épreuve les typologies des structures sonores avancées jusque-là.

Dans un premier temps, nous insistons sur le fait que les taxinomies et les parentés ont été élaborées sur des critères lexicaux indépendants des structures sonores. Pour l'IAL nous montrons la nécessité de mettre en place de nouvelles typologies des langues sur la base de ces structures. Nous faisons l'hypothèse, raisonnable, que si deux langues possèdent des systèmes vocaliques et consonantiques voisins, elles sont moins bien discriminables (automatiquement) que si leurs systèmes sont très différents. Les connaissances typologiques nous semblent être une étape difficilement contournable pour la constitution de bases de données destinées à l'évaluation de procédures en IAL. Grâce aux typologies des structures sonores, il sera ainsi possible de disposer des échantillons contenant des langues a priori très différentes et/ou très voisines. Il s'agit ainsi de mettre en évidence ce que la connaissance des systèmes vocaliques et consonantiques peut apporter à l'IAL.

1. De l'intérêt des typologies

L'IAL [Muthusamy 1994], à partir d'un échantillon de parole, renvoie à des connaissances et à des problèmes fondamentaux de taxinomie linguistique. Ces questions connaissent actuellement un renouveau d'intérêt, comme en témoigne l'avancée des travaux sur les universaux linguistiques et en particulier phonologiques associée à l'élaboration et à la diffusion de bases de données représentatives des langues du monde.

La taxinomie vise à proposer un ordre à l'intérieur duquel seront classés les éléments d'un ensemble. C'est un des programmes fondateurs de toute science. Les botanistes, les zoologistes ont fait partie des premiers spécialistes de ces grandes opérations de structuration. Avec de tels travaux se posent toute une série de questions de méthode. Où commence la variété ? Quelles sont les limites d'une espèce ? Sur quels critères doit-on rassembler les éléments d'une même famille ? Au-delà de la complexité et de la diversité des éléments il faut savoir trouver la ressemblance, les traces d'une même organisation : maîtriser les différences, se détacher des caractères extérieurs pour accéder de plus en plus aux structures de moins en moins factuelles.

La typologie linguistique est actuellement en pleine maturation, mais nous sommes encore loin du stade qu'avait atteint la botanique au XVII^{e} siècle, probablement parce que les problèmes fondamentaux de taxinomie affluent dès l'élaboration des bases de données représentatives des langues du monde. Il existe un très grand nombre de langues parlées à la surface du globe (entre 5000 et 8000 selon les sources), dont la plupart sont encore peu ou mal connues : on ne dispose pas de leurs descriptions systématiques que ce soit au niveau phonétique, phonologique, morphologique, lexical, syntaxique, sémantique… De plus, selon les sources, une même langue peut avoir des dénominations différentes. Il n'existe pas non plus de critères irréfutables permettant de faire la différence entre ce que l'on peut considérer comme une langue, un dialecte, un parler. Dans l'état actuel des recherches, on est encore loin de disposer d'un vaste échantillon des descriptions linguistiques pour les langues du monde (si la notion d'échantillon a ici un sens), alors que nombre d'entre elles sont en voie de disparition rapide et avec elles, des éléments fondamentaux de la diversité linguistique de notre espèce [Ladefoged 1995, 1996].

2. La recherche des parentés linguistiques

C'est par la comparaison du vocabulaire[1] que se sont ébauchés les premiers travaux, qu'un raccourci historique[2] attribue abusivement[3] à Sir William Jones [Jones 1786]. Ce dernier n'en est pas moins un personnage central dans la recherche des parentés. En effet, la conquête de

[1] Platon remarquait déjà dans *Le Cratyle* des ressemblances entre le phrygien et le grec.

[2] « Comme chacun sait, tout commence à Calcutta : le 2 février 1786, un juge anglais, Sir William Jones, lit devant la *Société Asiatique* une communication qui contenait ce remarquable passage : «The sanscrit language…" » [Boltanski, 1995].

[3] Cf. « Toute une tradition, principalement anglo-saxonne (!), place au point de départ des études indo-européennes le célèbre Anglais William Jones. […] Les études en question avaient presque deux cents ans à l'époque. […] À vrai dire, Jones est un homme charnière. D'un côté, il est le dernier de ce que l'on peut appeler les "précurseurs", c'est-à-dire tous ces savants qui ont vu la parenté des langues indo-européennes […] D'un autre côté […] ses remarques pertinentes, décisives, ont donné le branle aux travaux qui prennent leur essor en Europe à partir du tout début du XIX^{e} siècle » [Sergent 1995: 25).

l'Inde par les Anglais avait permis de recueillir une masse de documents. Dès 1786, le juge Jones avait communiqué dans son « Troisième discours anniversaire » à la Société Asiatique de Calcutta, dont il était le fondateur, son hypothèse d'apparentement du sanscrit avec le grec, le latin et probablement le celte, le gotique[4] et le perse, et que ces langues étaient issues d'une langue commune peut-être disparue :

> « *La langue sanscrite, quelque ancienne qu'elle puisse être, est d'une étonnante structure ; plus complète que le grec, plus riche que le latin, elle l'emporte par son raffinement exquis, sur l'une et l'autre de ces langues, tout en ayant avec elles, tant dans les racines des mots que dans les formes grammaticales, une affinité trop forte pour qu'elle puisse être le produit d'un hasard ; si forte même, en effet qu'aucun philologue ne pourrait examiner ces langues sans acquérir la conviction qu'elles sont en fait issues d'une source commune, laquelle, peut-être, n'existe plus. Il y a du reste une raison similaire, quoique pas tout à fait aussi contraignante, pour supposer que le gotique et le celtique, s'ils ont été mêlés par la suite avec un parler différent, n'en descendent pas moins de la même origine que le sanscrit ; on pourrait ajouter en outre à cette famille le vieux perse, s'il y avait lieu ici de débattre de quelque façon des antiquités persanes* ».

C'est une étape importante des travaux de linguistique historique. En 1813, le savant anglais Thomas Young [Young 1913] propose le terme de langue « indo-européenne » pour caractériser ce groupe linguistique, très étendu géographiquement :

> *« [un ensemble de langues] dont les éléments lexicologiques, morphologiques, syntaxiques présentent pour la plupart entre eux des ressemblances de nature telle qu'ils peuvent se ramener à l'unité, en supposant, pour chaque groupe d'éléments comparés, qu'il procède d'évolutions divergentes à partir de formes originelles disparues » [Sergent 1995].*

Les études comparatives s'inscrivent dans le courant scientifique des recherches typologiques en anatomie, botanique, zoologie des XVII^e^ et XVIII^e^ siècles. Il s'agit alors d'établir la parenté des langues, de reconstituer une « langue mère », d'établir ainsi des parentés génétiques entre des familles linguistiques.

Mais ce n'est qu'au XIX^e^ siècle que s'établit la *grammaire comparée*, une approche scientifique complètement nouvelle pour expliquer les similitudes surprenantes mises en évidence entre des langues très éloignées dans l'espace et dans le temps.

Les travaux des comparatistes s'étendent sur deux générations. Les fondateurs en sont Rask [Rask 1818], qui approfondit la parenté structurale des langues germaniques, et Bopp [Bopp 1816] qui recherche l'origine même des langues à travers le sanscrit. Il est le premier à avoir « retiré des rapprochements du sanscrit avec les langues de l'Europe un ensemble de doctrines » [Meillet 1936 : 457]. Bopp assimile la langue à un *organisme* vivant auquel il donne le sens linguistique de structure.

Dans ce courant va s'inscrire toute une série de travaux sur les lois des « mutations phonétiques » déjà exposés par Rask : correspondance entre [p t k] du grec et du latin avec

---

[4] Le gotique est la langue germanique la plus ancienne ; on la connaît par une traduction de la bible datant du IV^e^ siècle.

[f θ h] de l'allemand, de [p t k] ou [p t x] de l'allemand avec [b d g] du grec et du latin… À partir des dialectes germaniques Grimm [Grimm 1822] propose « le premier exemple et le premier modèle des "lois phonétiques" sur la connaissance desquelles repose la linguistique historique moderne » [Meillet 1934 : 461-462].

Schleicher [Schleicher 1861-1862] – botaniste avant d'être linguiste et phonéticien –, utilise systématiquement la technique de reconstruction de toutes les formes phonétiques (et non plus orthographiques) d'un même mot, puis à remonter à des formes hypothétiques grâce à des correspondances entre les langues. La phonétique occupe dans son œuvre une place importante, il est peut-être le premier à utiliser les sons et leurs processus articulatoires. Schleicher est très explicitement influencé par le darwinisme dont il utilise les concepts pour présenter les familles linguistiques[5].

3. Les typologies sonores

Les études typologiques des structures sonores ont commencé à la fin du XIX^e^ siècle avec les travaux de Baudouin de Courtenay [Baudouin 1894], mais c'est Troubetzkoy, une figure de proue du Cercle Linguistique de Prague, qui ouvre l'ère des taxinomies phonologiques en 1939 [Troubetzkoy 1939] :

> « *J'ai mis au net tous les systèmes vocaliques que je connaissais par cœur (34 en tout) et j'ai essayé de les comparer les uns aux autres […]. Les résultats sont extrêmement curieux […]. Tous les systèmes se réduisent à un petit nombre de type et peuvent être représentés par un schéma symétrique […]. Plusieurs lois "de la formation des systèmes" se laissent dégager sans peine […]. Je crois que les lois empiriques acquises ainsi seront d'une grande importance, particulièrement pour l'histoire de la langue et sa reconstruction […]. Elles devront être applicables à toutes les langues, aussi bien aux langues mères (Ursprachen) reconstruites théoriquement qu'aux divers stades de développement des langues historiquement attestées.* » (note du 19-IX-1928).

Les travaux de Troubetzkoy seront poursuivis en relation avec les études sur l'ontogenèse du langage [Jakobson 1941] et la définition des traits phonétiques [Hockett 1955]. Par la suite, les travaux les plus marquants vont s'effectuer aux USA dans une démarche descriptive associée à une recherche sur les universaux. Après le travail d'archivage publié dans l'*International Journal of American Linguistics* depuis 1944, et celui des missionnaires formés par le *Summer Institute of Linguistics*, la quête des universaux est marquée par :

• les travaux de Greenberg dans les années 1950 et la *Conference on Language Universals* à New-York en 1961 ;

• la discussion des propositions de Chomsky au *Symposium on Universals in Linguistic Theory* à Austin, en 1967.

Ces deux événements ont fortement contribué à faire reconnaître le caractère institutionnel de ce champ de recherche avec pour objectifs la mise en évidence des structures de base communes à toutes les langues actuellement connues et la description de leur évolution.

[5] Cf. *Die darwinische Theorie und die Sprachwissenschaft*, 1865. Schleicher est bien au fait des théories de Darwin, *L'origine des espèces* (1859) a été traduit en Allemagne en 1860 (voir l'entrée *Linguistique évolutioniste* du *Dictionnaire du Darwinisme et de l'évolution*, vol 2, 2645-2656, sous la dir. de Patrick Tort, Paris: PUF, 1996).

The Language Universal Project [1967-1976] a permis la constitution des *Stanford Phonology Archives* [Greenberg 1978] à partir desquelles sont nés des travaux d'importance concernant les classifications typologiques et l'étude des universaux phonologiques [Sedlack 1969] [Crothers 1978] [Maddieson 1986] [Vallée 1994] [Schwartz 1997]. Depuis Troubetzkoy, les données ont sans cesse été complétées et améliorées. Les interrogations sur le matériau abondent et diffèrent selon les auteurs. Toutefois les travaux typologiques ont davantage porté sur les systèmes vocaliques. Cet état de fait est probablement à relier au nombre plus important de consonnes dans les systèmes et à la multiplicité des paramètres de classement, rendant plus ardu l'émergence des types. Citons parmi les taxinomies consonantiques les travaux de [Hockett 1955] [Hagège 1982] [Maddieson 1986] [Lindblom 1988] [Laver 1994] [Stefanuto 1996] [Vallée 1998].

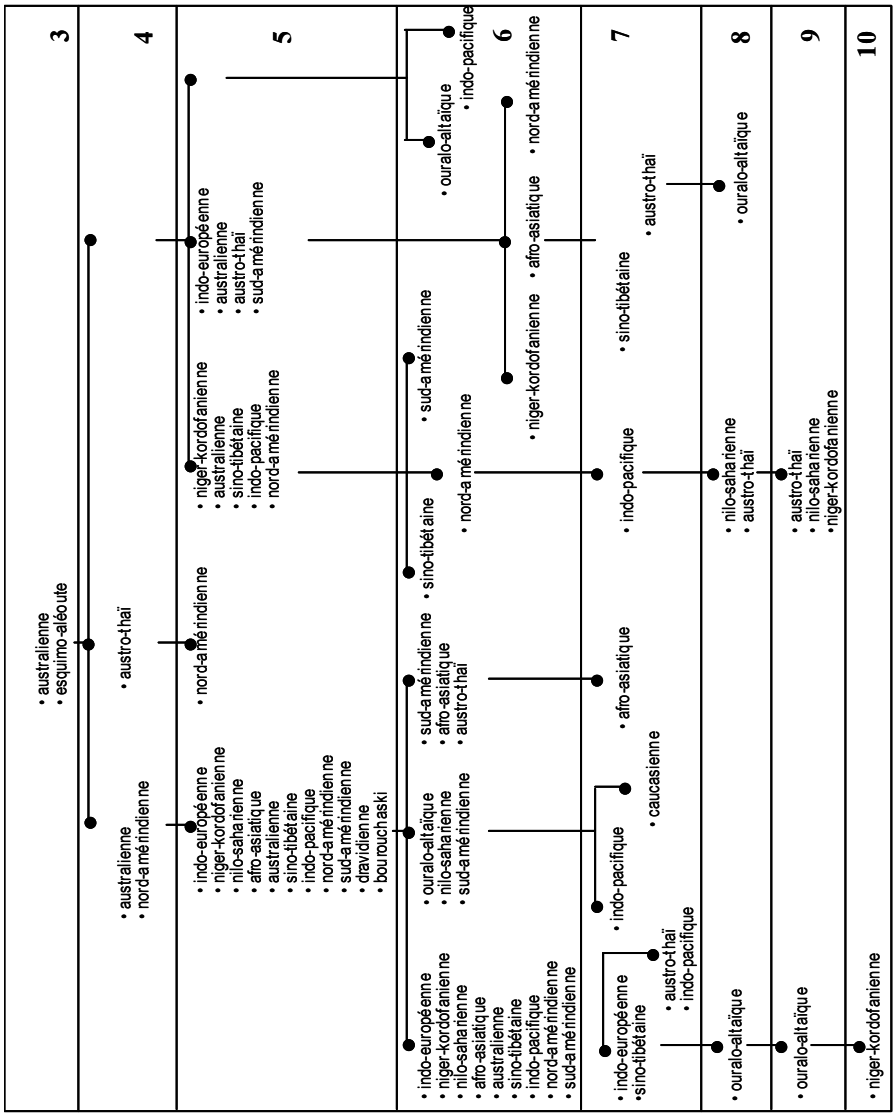

Figure 1. Les structures sonores identiques (de 3 à 10 voyelles) ne correspondent pas aux familles linguistiques.

La plupart des études typologiques permettent parfois d'observer une association entre certains traits phonologiques et certaines aires géographiques (comme c'est le cas par exemple pour les langues du continent africain). Cependant, les types phonologiques et la grande majorité des tendances structurelles mises en évidence à partir d'inventaires représentatifs des langues du monde s'étendent au-delà des parentés génétiques et jusqu'à l'ensemble des familles linguistiques (Figure 1).

Après une présentation détaillée des données, nous proposons des typologies à partir desquelles sont mises à jour les grandes tendances et la diversité des systèmes sonores des langues du monde.

4. Les données

La constitution d'un échantillon « représentatif » de l'inventaire phonologique des langues dans leur état actuel (synchronie), est devenu un outil indispensable pour examiner le contenu et la structure des systèmes sonores des langues rencontrées à la surface du globe.

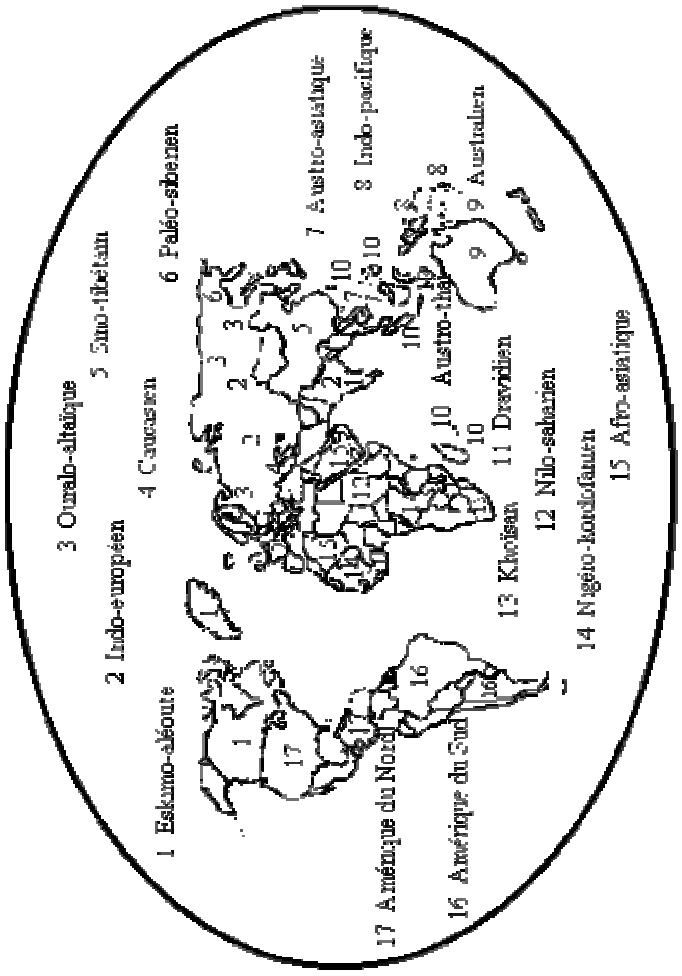

Figure 2. Carte de la répartition par famille des langues d'UPSID$_{317}$.

Ce n'est que depuis une quinzaine d'années que l'on dispose de bases de données phonologiques « représentatives » des langues du monde. Constituées à partir des archives de *Stanford*, les bases UPSID (UCLA *Phonological Segment Inventory Database*), dont la première version publiée contient 317 systèmes phonologiques [Maddieson 1986], et la plus récente 451 [Maddieson 1990], contiennent des données phonologiques génétiquement pondérées et descriptivement harmonisées d'au moins 5% de l'ensemble des langues du monde (Figure 2 et Table 1)[Maddieson 1986 : 5–6].

Famille (Nombre de langues)	Langues
indo-européen (23)	grec, irlandais, breton, allemand, norvégien, lithuanien, russe, bulgare, français, espagnol, roumain, farsi, pashto, kurde, hindi-urdu, bengalais, cachemirien, cinghalais, albanais, arménien, népalais, konkani, ormuri
ouralo-altaïque (28)	aïnou, khanty, mari, komi, finnois, hongrois, saami, nenets, nganasan, turc, azerbaïdjanais, chouvache, yakoute, kirghiz, bashkir, selkup, tuva, khalkha, even, nanai, mandchou, coréen, japonais, youkaghir, monguor, moghol, ouzbek, dagur
austro-asiatique (14)	mundari, kharia, khasi, vietnamien, sedang, cambodgien, parauk, sre, brao, khmu?, nicobarais, pacoh, kur, bruu
austro-thaï (39)	thaïlandais, lakkia, yay, soui, kam, po-ai, lungchou, atayal, javanais, malgache, cham, sama, batak, tagalog, sa'ban, chamorro, rukai, tsou, adzera, roro, kaliai, iai, hawaiien, tigak, lenakel, lue, ivatan, mor, pohnpeien, tiruray, lai, kwaio, paiwan, iban, gelao, tetun, fidjien, irarutu, maranao
sino-tibétain (21)	mandarin, taishan, hakka, changzhou, xiamen, fuzhou, bai, tamang, dafla, burmese, lahu, jingpho, ao, chin, bodo, karen, mien, newari, hmong, phlong, naxi
caucasien (7)	géorgien, kabardien, lak, rutul, bats, archi, avar
autres familles euro-asiatiques (7)	nivkh, kète, tchouktchi, koryak, itelmen, basque, bourouchaski
dravidien (6)	telugu, kota, kouroukh, koya, tulu, brahoui
niger-congo (55)	moro, kadugli, kpelle, bisa, bambara, dan, wolof, diola, temne, dagbani, senadi, tampulma, bariba, ewe, akan, igbo, ga, lelemi, efik, birom, tarok, amo, beembe, ogbia, ejagham, zoulou, teke, doayo, gbeya, azande, aizi, mumuye, klao, aghem, kpan, kohumono, yorouba, bobo-fing, noni, gwari, ewondo, jomang, sango, bete, konyagi, mbum, isoko, fe?fe?, ndut, ijo, lua, alladian, mambila, mba-ne, dogon
nilo-saharien (23)	songhaï, kanouri, maba, fur, maasai, luo, nubien, nyangi, ik, sebei, tama, temein, nera, tabi, mursi, lugbara, yulu, berta, kunama, koma, daju, dinka, nyimang
afro-asiatique (26)	arabe, tigre, amharique, socotri, néo-araméen, chleuh, tamasheq, somali, awiya, iraqw, beja, kullo, dizi, kefa, hamer, hausa, angas, margi, ngizim, kanakuru, kera, dahalo, kotoko, dangaleat, tera, lame
khoisan (4)	hadza, sandawe, !xu, nama
na-dené (7)	haida, tlingit, navajo, hupa, chipewyan, ahtna, eyak
Amérique du Nord (58)	percé, klamath, maidu, wintu, zoque, tzeltal, totonac, k'ekchi, mixe, huave, mazahua, mazatec, mixtec, chatino, tseshaht, kwakw'ala, quileute, lushootseed, papago, luiseno, hopi, yaqui, picuris, karok, pomo, diegueno, achumawi, yana, shasta, tol, zuni, acoma, ojibwa, tonkawa, wiyot, seneca, wichita, dakota, yuchi, tunica, alabama, wappo, nahuatl, kawaiisu, amuzgo, bella coola, tlapanec, chehalis, tsimshian, caddo, huasteco, yucatec, shuswap, miwok, jacaltec, cherokee, kiowa, chinantec

Amérique du Sud (66)	itonama, bribri, pirahã, cayapa, paez, ocaina, muinane, caraïbe (galibi), kaingang, apinaye, amahuaca, epena pedee, tacana, axluxlay, abipon, nambiquara, arabela, auca, quechua, jaqaru, tehuelche, wapishana, caraïbe insulaire, amuesha, campa, guajiro, moxo, guarani, siriono, guahibo, ticuna, barasano, siona, jivaro, cofan, mapudungu, resigaro, yagua, cayuvava, huari, hixkaryana, yucuna, jebero, iranxe, cubeo, tarascan, japreria, panare, bororo, andoke, warao, akawaio, paya, shiriana, bakairi, cuna, maxakali, qawasqar, movima, saliba, guambiano, camsa, cacua, trumai, ache, iate
eskimo-aléoute (3)	aléoute, inuit, yupik
australien (25)	garawa, yanyuwa, waray, murinhpatha, maung, tiwi, burarra, nunggubuyu, alawa, malakmalak, bardi, wik-munkan, désert occidental, arrernte, gugu-yalanji, ya, diyari, bandjalang, kalkatungu, yidiny, dyirbal, ngiyambaa, mbabaram, ngarinjin, yolngu
papou (39)	andamanese, asmat, kwoma, sentani, nimboran, iwam, selepet, gadsup, yagaria, kewa, chuave, dani, wantoat, dadibi, fasu, suena, dera, kunimaipa, yareba, koiari, taoripi, nasioi, rotokas, nambakaengo, angaatiha, wahgi, yawa, usan, baining, yessan-mayo, woisika, alamblak, amele, kiwai, waris, vanimo, savosavo, makian, ekari

Table 1. Les langues d'UPSID$_{451}$ classées par famille avec l'indication du nombre de langues retenues dans la base.

Réparties sur les cinq continents (Figure 2), dix-huit familles de langues (Table 1) sont représentées dans la version la plus récente. Est inclus dans l'inventaire au moins un dialecte par groupe de langues (sous-famille) sur la base d'une distance génétique de séparation des langues d'au moins 1500 ans, *« a long enough period for substantial independant developments to occur in the phonological patterns of any two languages belonging to the same larger family. »* [Maddieson 1991a : 348]. C'est une marge d'assurance plus large que celle estimée par [Ruhlen 1987 : 6] : *« under most circumstances 500 years is probably sufficient to seriously impair mutual intelligibility, and 1000 years will usually obliterate it entirely. »* Cette distance de 1500 ans correspond à la séparation entres langues germaniques de la branche nordique (islandais, danois, suédois, norvégien) et langues germaniques de la branche occidentale (allemand, néerlandais, anglais, frison, anglo-frison), pour lesquelles on estime un certain degré d'indépendance : entre les langues issues d'une même souche la distance génétique, bien qu'elle reste difficile à apprécier, est d'autant plus grande que leur séparation est plus ancienne.

Le nombre de locuteurs ne rentre en aucune manière dans la sélection des langues de l'échantillon : *« The size of extent populations of speakers of languages is an accident of political and social history that is quite irrelevant to the questions relating to the structure of human languages »* [Maddieson 1986 : 158].

Comme nous l'avons vu, Troubetzkoy avait fait appel, de mémoire, à 34 langues, Sedlak a travaillé sur 150 langues tirées des *Stanford Phonology Archives* [Sedlak 1969] et Crothers en 1978 sur 209 langues [Crothers 1978]. UPSID, enrichie à 451 langues, est actuellement une des rares sources d'étude des systèmes phonologiques des langues du monde qui permette de dresser un inventaire des types de systèmes et de repérer les objets fondamentaux porteurs d'informations sur les sons du langage humain et sur la formation des systèmes sonores. Pour reprendre le propos de Maddieson [Maddieson 1991b], UPSID doit être considérée comme une fenêtre par laquelle il est possible d'entrevoir un état actuel des langues du monde.

Au niveau de la répartition par famille, la première version d'UPSID présente 17 familles de langues plus trois langues isolées : basque, bourouchaski, aïnou alors que la deuxième

comporte 18 familles et quatre langues isolées qui sont nivkh (paléo-sibérien), basque, bourouchaski et kète. Dans la deuxième version, sont constatés également le glissement de la famille indo-pacifique dans le phylum papou, l'intégration de l'aïnou dans la famille de langue ouralo-altaïque, le kète comme langue isolée et non plus rattachée à la famille paléo-sibérienne, le tchouktchi-kamchatkan comme une famille à part entière (3 langues : tchouktchi, itelmen, koryak) dans $UPSID_{451}$, tout comme le na-déné, intégré dans la famille amérindienne dans $UPSID_{317}$ et qui figure comme famille dans la version plus récente représentée par les langues : eyak, hupa, chipewyan, navajo, haïda, ahtna et tlingit. Le passage d' $UPSID_{317}$ à $UPSID_{451}$ est bien plus qu'une simple addition : 192 langues ont été ajoutées et 58 autres ôtées d'$UPSID_{317}$. Une partie des langues conservées a été modifiée (prise en compte de nouvelles références). Ces changements relèvent de deux types : le premier porte sur le nombre de phonèmes (exemple, 47 systèmes ont vu soit baisser, soit augmenter leur nombre de voyelles), le deuxième correspond à une différence de transcription API des unités du système. Malgré un certain nombre de modifications, l'analyse typologique d'$UPSID_{451}$ [Abbadeni 1996] [Alcantara 1998] n'apporte pas de changements fondamentaux dans les grandes tendances des systèmes phonologiques [Maddieson 1986], [Vallée 1994], [Stefanuto 1996].

L'essor des techniques numériques a favorisé la constitution et l'exploitation de banques de données. Nous avons implanté à l'ICP différentes versions d'UPSID pour nourrir les recherches typologiques sur les consonnes et les voyelles. Pour chaque langue, identifiée par un numéro dans la base, on dispose de sa famille et sous-famille, du nombre et de la liste des voyelles et de leur position sur une grille 2D en fonction des catégories, nombre et liste des consonnes et leur lieu d'articulation dans le conduit vocal en fonction du mode, nombre et liste des diphtongues. Dans la phase d'implantation, nous avons éclaté la liste des voyelles et celle des consonnes en plusieurs catégories. Voyelles et consonnes sont codées à la fois en symbole API[7] (177 symboles servent à coder les voyelles, 654 les consonnes), pour l'exploitation phonétique, et le codage numérique pour les recherches systématiques.

L'exemple de système fourni ici (Figure 3) présente, pour le *cachemirien*, une langue indo-européenne du groupe indo-aryen, les 8 voyelles brèves, les 7 voyelles longues, les 6 nasales brèves et 7 nasales longues.

[7] Alphabet Phonétique International

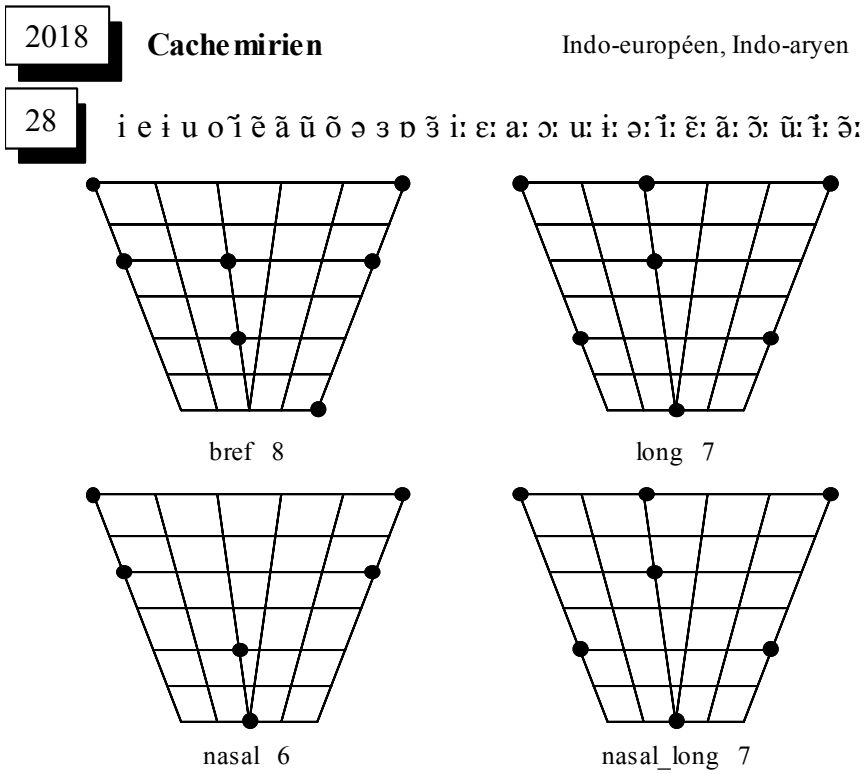

Figure 3. Un exemple de langue dans la base de données vocalique implantée à l'ICP, à partir des données de [Maddieson 1990] (UPSID$_{451}$).

5. Systèmes consonantiques : de la taxinomie aux tendances universelles

5.1. Éléments de taxinomie

La classification traditionnelle des consonnes repose sur les lieux et les modes d'articulation. Ces paramètres constituent bien évidemment les dimensions élémentaires de notre étude typologique. UPSID$_{451}$ présente un ensemble de 920 phonèmes dont 654 segments consonantiques, répartis sur 13 lieux d'articulation (Figure 4) et une quinzaine de modes. On dénombre 153 plosives, 61 implosives, occlusives éjectives et clicks non affriqués, 134 fricatives, fricatives éjectives, 155 affriquées, affriquées éjectives et clicks affriqués, 95 vibrantes, battues et approximantes centrales et latérales, 51 nasales et 5 consonnes de type *h sounds*.

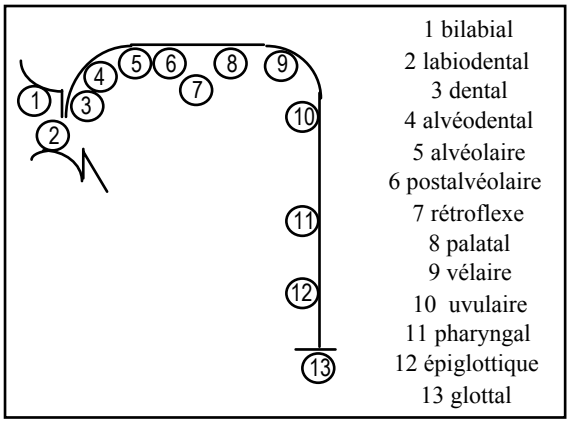

Figure 4. Lieux d'articulation consonantiques ($UPSID_{451}$).

Les modes ont été regroupés sous 7 catégories classant ainsi les consonnes en plosives (dont implosives, éjectives, occlusives glottales), nasales, fricatives (dont celles éjectives et *h sounds*), affriquées (dont éjectives), approximantes, vibrantes/battues, clicks.

Nous avons élaboré notre typologie avec une détermination des lieux plus détaillée que celle de Maddieson, [Maddieson 1986] en composant avec celles proposées par Creissels [Creissels 1994] et Ladefoged & Maddieson [Ladefoged 1996]. Bien que désignée par un mode, nous avons conservé la classe des rétroflexes comme le proposent Maddieson [Maddieson 1986] et Ladefoged [Ladefoged 1996] et l'API de 1996. Une répartition de cette catégorie de consonnes, qui recouvrent plusieurs lieux (alvéodental, alvéolaire, prépalatal...), nécessiterait une investigation des sources.

Soulignons que la comparaison avec les taxinomies antérieures n'est toutefois pas aisée car toutes diffèrent dans la préparation des données et les méthodes de classement.

5.2. Distributions

Les systèmes phonologiques ont le plus souvent entre 18 et 23 consonnes, 50% ont entre 16 et 24 consonnes, 70% entre 10 et 25 ; minimum 6 pour le rotokas (langue indo-pacifique, 11 phonèmes), mode à 22, et maximum 95 (dont 48 clicks) pour le !xũ (famille khoisan, 141 phonèmes) (Figure 5). Si on détaille les moyennes, nous calculons par langue 7.8 plosives, 4.1 fricatives, 3.3 nasales, 2.9 approximantes, 2 affriquées, 0.6 éjective, 0.52 vibrante ou battue, 0.25 implosive, 0.2 click.

Le Tableau 2 donne la répartition des consonnes (654 au total) par lieux d'articulation. Les plus fréquentes, tous modes confondus, sont les alvéodentales (15.3%), suivies des bilabiales (14.3%) puis des vélaires (12.6%). Le regroupement des consonnes dentales, alvéodentales, alvéolaires, postalvéolaires et de l'ensemble des rétroflexes en une catégorie *coronale* permet de mieux apprécier l'écrasante proportion de celle-ci : 44.5% dans $UPSID_{317}$. Bien que cette catégorie regroupe une grande diversité d'articulations [Ladefoged 1996], ce choix est tout à fait acceptable pour l'élaboration de notre typologie. En effet, le trait de lieu ne sert que rarement à distinguer phonologiquement des consonnes de ce type, sauf pour les affriquées, et en tout cas pas entre coronales antérieures [Keating 1990].

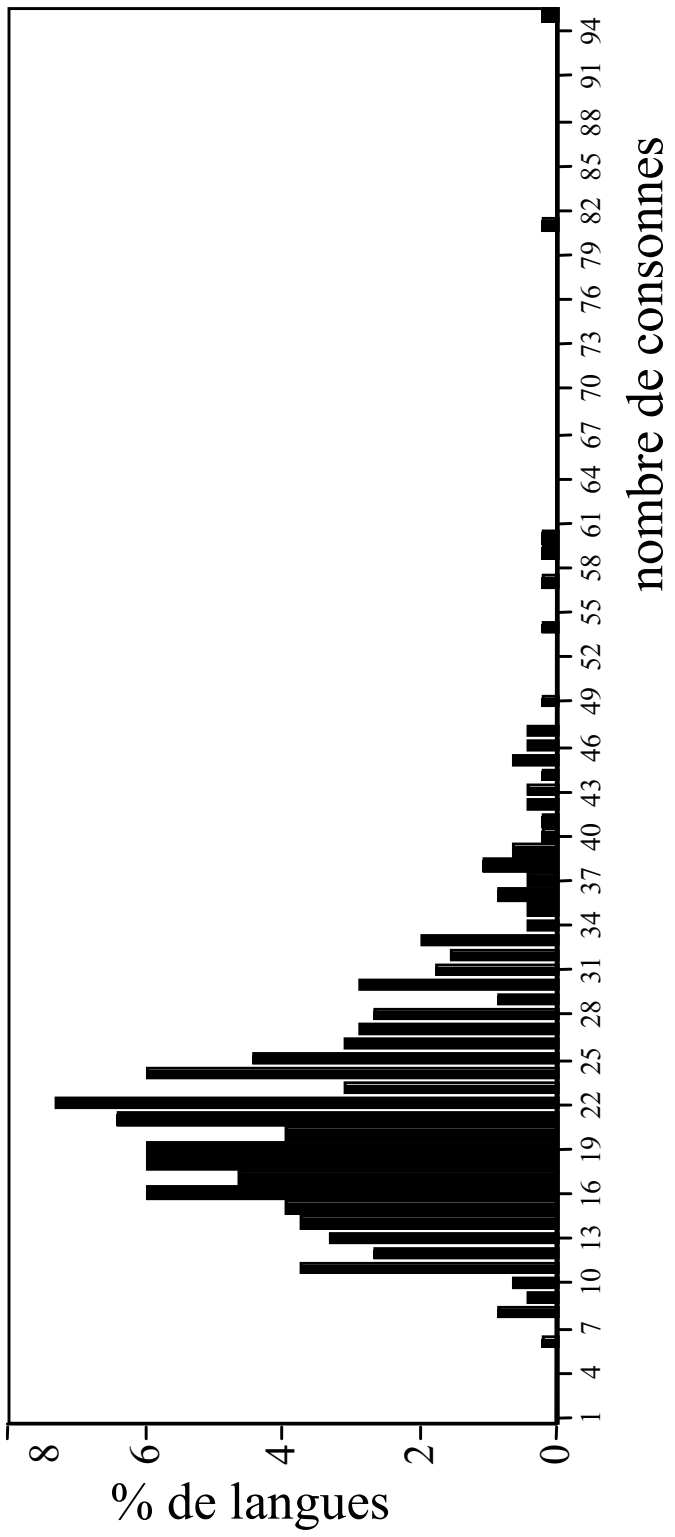

Figure 5. Histogramme des consonnes ($UPSID_{451}$).

alvéodental	15,3%
bilabial	14,3%
vélaire	12,6%
palatal	10%
apico-alvéolaire	9,9%
glottal	7,6%
lamino-postalvéolaire	6,7%
labiovélaire	5,1%
dental	5%
lamino-alvéolaire	3,4%
labiodental	3,3%
rétroflexe	2,9%
uvulaire	1,7%
autres	<1%

Table 2. Fréquences d'occurrences des lieux d'articulation consonantiques ($UPSID_{317}$).

5.3. Systèmes

La répartition des consonnes d'$UPSID_{317}$ par mode montre que certains sont plus exploités que d'autres : plosives 38.6%, fricatives 20.2%, nasales 14.6%, approximantes 13%, affriquées 9.6%, vibrantes ou battues 3.9%. Parmi les plus fréquentes arrivent en tête les occlusives coronales (type /t/), orales et sourdes (97.5%), les nasales bilabiales et coronales (types /m/ et /n/, présentes aussi dans plus de 9 langues sur 10), suivies de /k/, /j/, /p/ rencontrées dans plus de 80% des langues, /w/ et /s/ dans 2 langues sur 3, /d/, /b/, /h/, dans plus de 60%, /l/, /g/, /ŋ/, /ʔ/ dans la moitié des langues.

5.3.1 Taille des systèmes de lieux et modes d'articulation

L'élaboration d'une typologie par lieux et modes des systèmes consonantiques fait apparaître une forte corrélation entre la taille des systèmes de lieux et les différents modes d'articulation (Tableau 3).

- Dans la plupart des langues, les plosives et les nasales mettent en jeu respectivement de 3 à 6, et de 1 à 6 lieux. Le mode plosif est le seul universellement exploité par les systèmes sonores. Il est aussi le seul à utiliser systématiquement les oppositions de lieux : les langues en distinguent au moins trois.
- Les fricatives et les approximantes constituent les catégories qui recrutent le plus grand nombre d'oppositions de lieux : le mode fricatif est beaucoup plus largement distribué entre les différentes tailles de systèmes. Si le jeu des oppositions de lieux des fricatives est souvent plus complexe que celui des plosives, il est important de préciser qu'il n'est pas exploité par 8% des langues de l'échantillon – alors que toutes possèdent des plosives.

- Dans la catégorie des *trills, flaps* et *taps* (vibrantes et battues), les systèmes ne présentent que rarement une opposition de lieu, et en aucun cas ils n'en distinguent plus de 2. Cette même tendance se retrouve pour les affriquées.

Plus généralement :

• Les langues exploitent systématiquement l'opposition de lieu pour les consonnes plosives ; près de la moitié les répartissent dans des systèmes à 4 lieux.

• L'opposition de lieu est rare pour les affriquées, exceptionnelle pour les vibrantes/battues.

• Les approximantes sont de préférence réparties sur 3 lieux d'articulation.

• Les fricatives se répartissent plus largement dans diverses tailles de systèmes de lieux.

	1 lieu	2 lieux	3 lieux	4 lieux	5 lieux	6 lieux	7 lieux	8 lieux
trills, flaps, taps	67,2%	4,7%						
affriquées	44,8%	20,8%	6,9%	0,3%				
plosives			28,4%	43,2%	22,4%	5,9%		
nasales	2,5%	31,5%	30,6%	25,9%	4,5%	2,2%		
approximantes	8,2%	26,5%	49,8%	8,2%	2,5%	0,3%	0,3%	
fricatives	3,4%	18,3%	25,9%	21,1%	14,2%	7,3%	1,9%	0,9%
Nb d'oppositions de lieux	0	1	3	6	10	15	21	28

Table 3. Taille des systèmes de lieux et leur représentativité dans $UPSID_{317}$.
Le nombre d'oppositions de lieux varie selon les modes ; de 0 (c'est-à-dire 1 lieu par système) à 28 (8 lieux).
Les valeurs présentées ne tiennent pas compte du trait de voisement : 28,4% des 317 langues ont un système de plosives à 3 lieux, 43,2% possèdent des plosives réparties sur 4 lieux, etc.

5.3.2 Systèmes de lieux par modes

La Table 4 rend compte du contenu des systèmes de lieux par mode d'articulation. Sont représentées les répartitions des plosives, nasales, fricatives et affriquées par nombre de lieux d'articulation.

Il n'existe pas de langues sans plosives, ni même n'en ayant qu'une seule. Lorsqu'un système les répartit sur 2 lieux, l'un est coronal, l'autre bilabial ; sur 3 lieux : coronal, bilabial et vélaire. C'est à 4 qu'apparaît le lieu glottal ; si le système répartit ses consonnes entre 5 lieux d'articulation, il recrute les régions bilabiale, coronale, vélaire, uvulaire et glottale.

Une seule consonne nasale dans un système, elle est de type coronal. S'il en possède 2, elles sont de type bilabial et coronal. Réparties sur 3 lieux, s'ajoute une nasale vélaire. Puis à 4, apparaît le lieu palatal.

Si les fricatives d'un système n'exploitent qu'un lieu d'articulation, il s'agit de la région alvéolaire donc d'une coronale ; 2 lieux : alvéolaire et glottal ; à 3 lieux apparaissent les fricatives labiodentales ; à 4 est exploitée la région palatale et à 5 lieux apparaissent les fricatives vélaires.

Les affriquées, quant à elles, se répartissent dans les régions d'articulation des coronales, quelle que soit la taille du système de lieu (1 à 3).

lieux	plosives		nasales		fricatives		affriquées	
0	0%	–	3,5%	–	7%	–	33%	–
1	0%	–	1,8%	coronal	6.6%	coronal	42%	coronal
2	0,2%	coronal bilabial	31%	coronal bilabial	18%	coronal glottal	20%	coronal coronal 2
3	31%	coronal bilabial vélaire	32%	coronal bilabial vélaire	28%	coronal glottal labiodental	4.9%	coronal coronal 2 coronal 3
4	43%	coronal bilabial vélaire glottal	27%	coronal bilabial vélaire palatal	20%	coronal glottal labiodental coronal 2	–	–
5	21%	coronal bilabial vélaire glottal uvulaire	3,5%	coronal bilabial vélaire palatal coronal 2	11%	coronal glottal labiodental coronal 2 vélaire	–	–

Table 4. Les catégories consonantiques (plosives, nasales, fricatives, affriquées) d'$UPSID_{451}$, classées par nombre de lieux d'articulation avec le pourcentage de langues correspondant (par exemple, 31% des 451 langues présentent un système de plosives réparties sur 3 lieux, 43% sur 4 lieux etc. ; 0 lieu correspondant à une absence de catégorie). Ne sont portés dans cette table que les systèmes les plus répandus.

De la répartition des systèmes de lieux par catégories nous retenons les tendances suivantes :

• Comme pour les systèmes vocaliques [Vallée 1994], le système de lieux à n éléments contient le système à n-1 lieux consonantiques.

• Lorsqu'il n'y a qu'un lieu, quelle que soit la catégorie (plosives, fricatives ou autres), il s'agit de consonnes coronales (lieux 4 et 5).

• 95% des langues contrastent sur 3 lieux à l'intérieur des catégories (plus d'une langue sur 3 pour les nasales).

• Une taille de système domine chaque catégorie : 43% des langues ont 4 lieux pour les plosives et 74% ont 3 ou 4 lieux) ; 32% des langues répartissent leur nasales sur 3 lieux (59% sur 3 et 4) ; 28% présentent 3 lieux pour les fricatives (48% sur 3 et 4) ; 43% n'ont qu'un lieu d'articulation pour les affriquées.

• Les combinaisons à 5 lieux sont présentes dans seulement 36% des langues, essentiellement dans les plosives orales (97 langues).

• Les combinaisons à 6, 7 ou 8 lieux sont peu répandues (15% des langues). On les rencontre essentiellement dans les systèmes de fricatives.

Nos résultats confirment donc que les plosives sont de loin les phonèmes consonantiques « vedettes », omniprésentes dans les langues et que, comptabilisées avec les nasales pour chaque langue, leur nombre est beaucoup plus important que le nombre de fricatives (Figure 6), à une douzaine d'exceptions près.

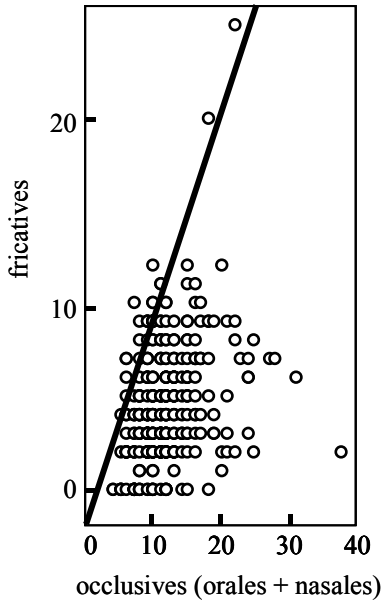

Figure 6. Nombre de fricatives en fonction du nombre d'occlusives dans $UPSID_{317}$.

5.3.3 Rapport sourdes/sonores

Dans la catégorie des plosives orales, les sourdes dominent largement les sonores (64%) : on relève 30% de sourdes en plus dans les bilabiales, 50% de sourdes en plus dans les coronales, 60% dans la catégorie vélaire. Font exception à cette tendance, les langues du continent africain – familles niger-kordofanienne, nilo-saharienne et afro-asiatique : elles présentent en moyenne 1,5 fois plus de plosives bilabiales sonores que de sourdes.

Ce rapport sourdes/sonores est encore plus écrasant pour les fricatives : les langues présentent 5 fois plus de /s/ que de /z/, 3,2 fois plus de /ʃ/ que de /ʒ/ et presque 2 fois plus de /f/ que de /v/. Quel que soit le lieu d'articulation, le trait sourd concerne 72% des fricatives.

Les affriquées sont également majoritairement sourdes (74%).

6. Consonnes vs. voyelles

Pour l'ensemble des langues la corrélation entre le nombre de consonnes et le nombre de voyelles n'est pas significative (Figure 7).

Seule régularité, chaque langue possède plus de consonnes que de voyelles, à 12 exceptions près : exemple, le pawaian (famille austro-thaï) avec 12 voyelles et 10 consonnes ; et l'apinaye (famille sud-amérindienne, groupe macro-ge), 17 voyelles et 13 consonnes.

La répartition des langues par nombre de voyelles plus nombre de consonnes (Figure 8) présente un minimum à 11 (le rotokas, langue papoue), un pic à 25, et un maximum à 119 (le !xũ, langue khoïsan), l'essentiel des langues se situant entre 16 et 49 C+V.

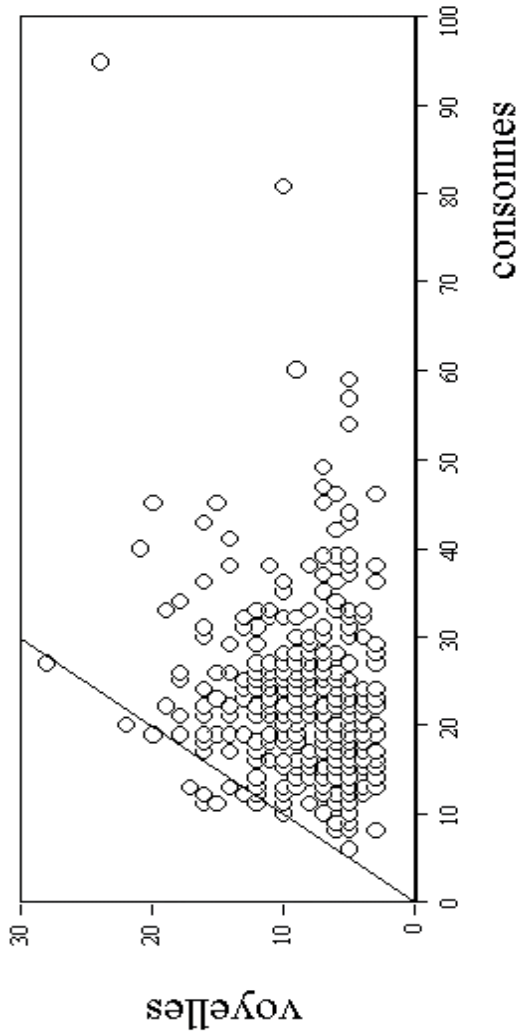

Figure 7. Nombre de voyelles en fonction du nombre de consonnes dans les systèmes d'UPSID$_{451}$.
(La droite correspond aux systèmes pour lesquels le nombre de consonnes est égal à celui des voyelles)

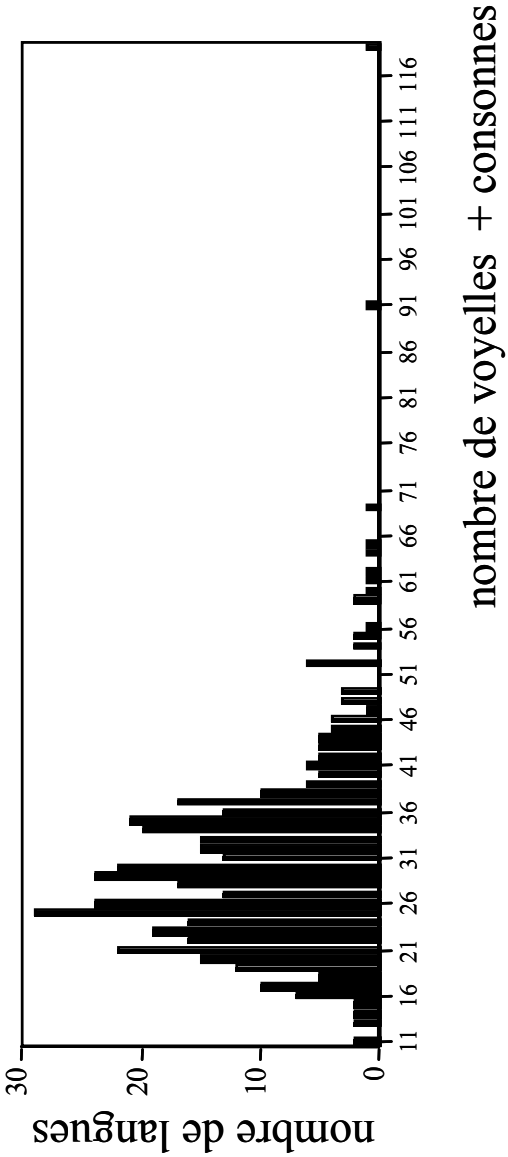

Figure 8. Histogramme du nombre de voyelles + nombre de consonnes ($UPSID_{451}$)

7. Une typologie et des universaux vocaliques

L'analyse menée à partir d'UPSID atteste que le potentiel linguistique, phonologique, des langues du monde puise dans un ensemble de 37 possibilités vocaliques et que parmi ces possibilités, les langues ne choisissent pas leur unités distinctives sur des critères arbitraires (Figures 10a & 10b).

Les systèmes vocaliques recrutent 3 à 28 phonèmes mais deux tiers d'entre eux ont entre 5 et 7 voyelles (Figure 9).

La comparaison des systèmes les plus fréquents (Figures 10a & 10b) met évidence un ordre d'apparition des voyelles dans les systèmes. Les 3 « vedettes » /i a u/ sont présentes dans

97% des langues. S'y ajoute la voyelle antérieure /ˈeˈ/ dans le système à 4 le mieux représenté. Le système à 5 /i ˈeˈ a ˈoˈ u/ est de loin le plus « populaire » dans les langues du monde. C'est le cas dans 3 des 4 grands groupes linguistiques (eurasien, américain, australien), alors que c'est le système périphérique à 7 /i e ɛ a o ɔ u/ qui est majoritaire en Afrique [Maddieson 1991b].

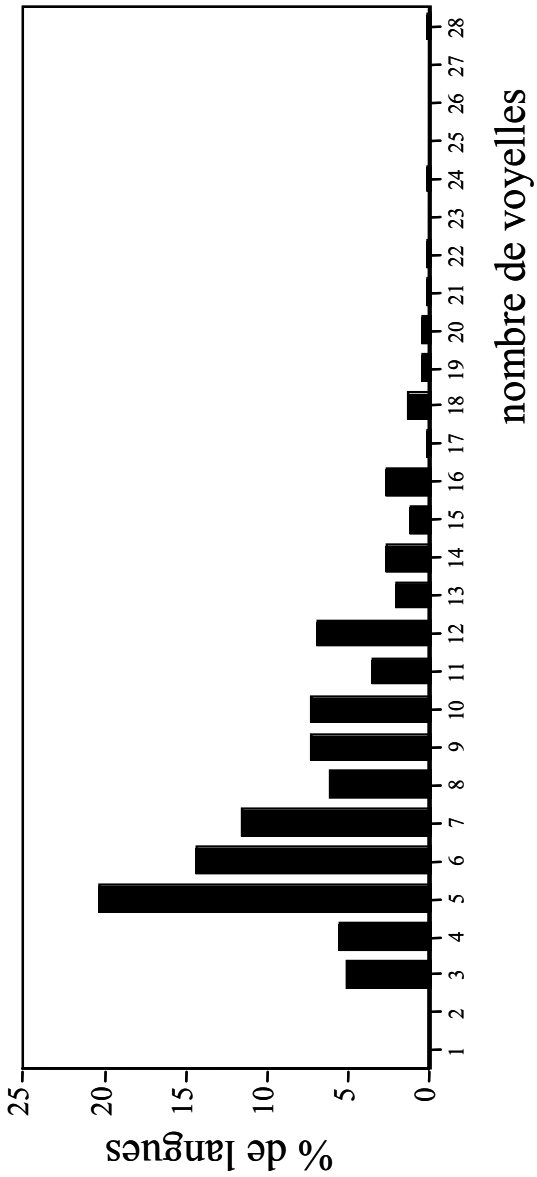

Figure 9. Histogramme du nombre de voyelles par langues ($UPSID_{451}$).

Parmi les systèmes pairs à 6 et 8 voyelles, les plus fréquents possèdent une voyelle centrale de type /ə/ qui pourrait renvoyer à une autre dimension liée probablement aux principes intrinsèques de la réduction vocalique [Schwartz 1997]. À 9, le système préféré est, comme y le système à 7, constitué de voyelles situées à la périphérie. Dans ces systèmes, la symétrie entre voyelles d'avant et voyelles d'arrière est une tendance forte. Dans les cas d'asymétrie (30% des

systèmes d'UPSID) les périphériques antérieures sont très souvent plus nombreuses que les postérieures (seulement 9% des langues ont plus de voyelles périphériques à l'arrière).

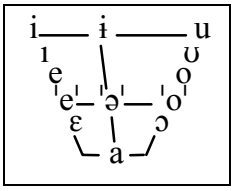

Figure 10a. Qualités vocaliques.

Nb. de voyelles *(Nb. de langues)*	**Systèmes vocaliques et % d'occurrences (UPSID 317 langues)**				
3 (20)	60				
4 (32)	37,5	18,7	6,3	6,3	
5 (134)	65,7	4,5			
6 (72)	30,6	13,9	12,5	6,9	
7 (49)	40,8	10,2	10,2	6,1	4
8 (34)	17,6	8,8	8,8	5,9	
9 (35)	20	14,3	6		

Figure 10b. Occurrences des systèmes vocaliques par nombre de voyelles, de gauche à droite : du plus fréquent au moins répandu ($UPSID_{451}$). La voyelle centrale qui relève probablement d'une autre dimension est représentée par un cercle transparent.

Au-delà de 9 timbres de base, il est clair que les langues puisent leurs unités distinctives dans d'autres dimensions. En effet, les systèmes de plus de 9 voyelles, vraisemblablement saturés, diminuent leur nombre de timbres de base en ajoutant un système « parallèle » (système secondaire vs. primaire) dans une autre dimension : la nasalité en général, ou la quantité. Cette relation tendancielle a été proposée par Vallée [Vallée 1994] qui a quantifié la relation entre nombre de voyelles par langue et timbres de base. Les grandes tendances mises en évidence

dans la structure des systèmes primaires se retrouvent dans les systèmes secondaires : préférences pour le système à 5, la dispersion périphérique et symétrique, la voyelle centrale de type /ə/ qui, comme dans les systèmes primaires, ne semble pas « interagir » avec les autres voyelles du système.

Les langues qui peuvent avoir des systèmes vocaliques identiques n'appartiennent pas forcément à la même famille linguistique. Les langues d'une même famille n'ont pas forcément le même système vocalique. Ces résultats sont, en fait, orthogonaux avec le classement en familles à l'origine de la constitution de la base (cf. Figure 1).

8. Les Diphtongues

Elles n'ont pas encore été systématiquement analysées, mais on peut d'ores et déjà noter :

• que 90 % des langues n'en possèdent pas,

• que les diphtongues sont présentes dans les langues qui possèdent un système vocalique d'au moins 5 éléments,

• que lorsqu'une langue en possède, elles sont au nombre de 1 à 8, mises à part deux notables exceptions à 22 et à 25,

• que 80% des diphtongues possèdent deux segments vocaliques qui ont des apertures différentes,

• et enfin, le caractère décroissant de l'aperture (type /ai/ ou /au/) pour 50% d'entre elles.

9. Différences et ressemblances entre les langues

Si l'on considère l'ensemble des 920 phonèmes possibles, on constate que 854 d'entre eux n'apparaissent que dans 1 à 10% des langues, que 23 d'entre eux sont présents dans 11% à 20% des langues, et que seuls 1% d'entre eux apparaissent dans 91% à 100% des langues (Figure 11). Ces données montrent à quel point il est possible d'identifier les langues à partir d'un petit nombre de sons (voir aussi [Hombert 1997]).

En ce qui concerne les catégories, il n'existe pas dans UPSID de langues sans plosives ou même n'en ayant qu'une. Le minimum de plosives observé est de 3, le maximum atteint 20 ; près d'une langue sur cinq en possède 6. Les affriquées sont le plus souvent contrastées sur 1 ou 2 lieux, jamais sur plus de 4 et on note 149 langues qui ne possèdent pas cette catégorie de consonnes. Les vibrantes/battues sont très majoritairement seules dans leur catégorie, et les approximantes sont souvent dispersées sur 3 lieux (158 langues). On relève donc que 100% des langues possèdent des plosives, 97% des consonnes nasales, 93% des fricatives, 66% ont des affriquées, 66% des approximantes et 63% des vibrantes/battues.

On note que, voyelles et consonnes confondues, les phonèmes les plus fréquents dans toutes les langues sont /m k i a j p u w b h g ŋ/.

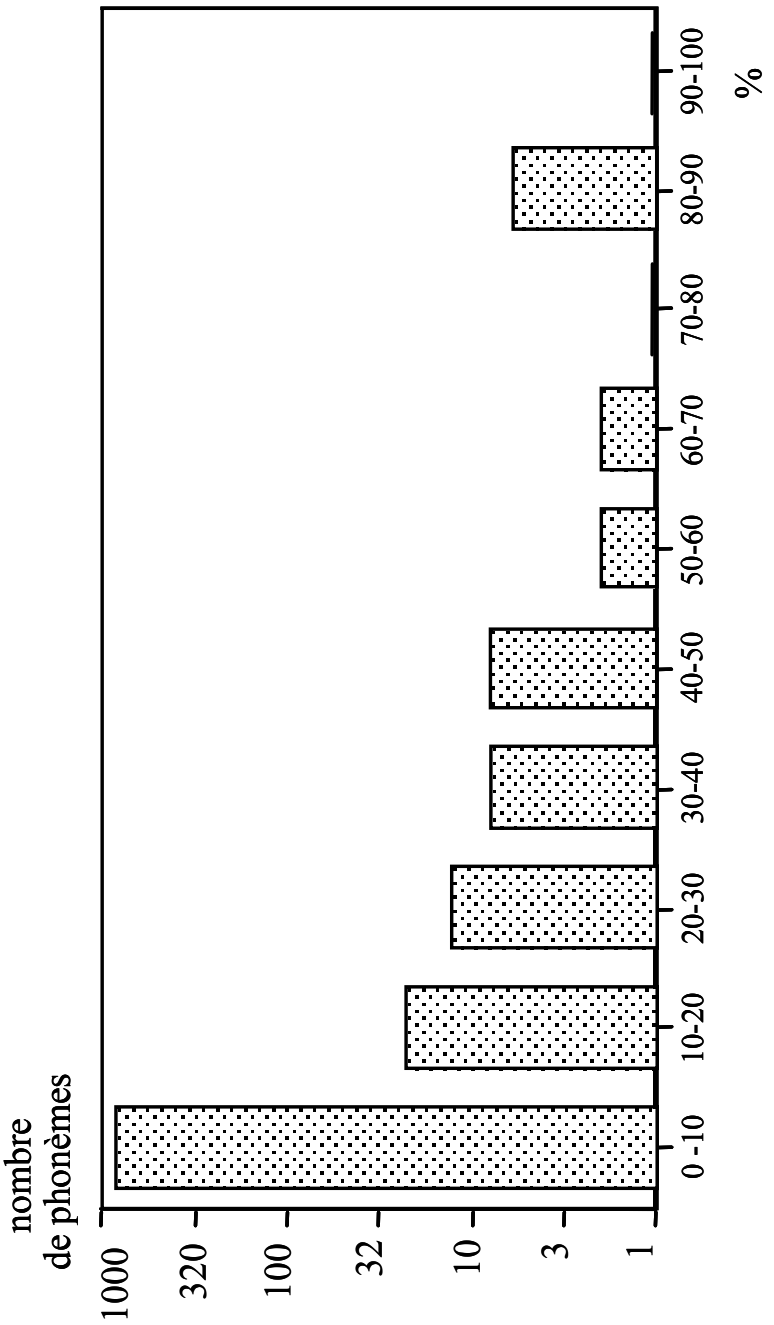

Figure 11. Nombre de phonèmes (échelle logarithmique) et pourcentage d'apparition dans les langues.

Figurent dans $UPSID_{451}$, 307 types de systèmes vocaliques et 271 langues qui ont un système vocalique unique. Rappelons que Vallée [Vallée 1994] répertorie 219 types de systèmes dans $UPSID_{317}$. La Table suivante présente les 5 systèmes vocaliques d'$UPSID_{451}$ les plus fréquents, soit des systèmes à 3, 5, 6 et 7 voyelles :

nombre de voyelles	système vocalique	nombre de langues
5	/i ˈeˈ a ˈoˈ u/	44
6	/i ˈeˈ a ˈoˈ u ˈəˈ/	12
5	/ɛ u i ɔ a/	11
7	/i e ɛ a ɔ o u/	11
3	/i a u/	11

Table 5. Les cinq systèmes vocaliques les plus fréquents ($UPSID_{451}$).

Ci-dessous, la liste des langues ayant l'un de ces 5 systèmes les plus fréquents :

/i ˈeˈ a ˈoˈ u/

ahtna ; arabela ; arménien ; avar ; beja ; berta ; bourouchaski ; chuave ; cuna ; dahalo ; espagnol ; fidjien ; hadza ; hausa ; huasteco ; jingpho ; kadugli ; kaliai ; kefa ; kewa ; kiwai ; koiari ; kota ; kullo ; kwaio ; k'ekchi ; monguor ; mor ; movima ; mursi ; nera ; nubien ; roro ; rotokas ; savosavo ; tetun ; tlingit ; wappo ; wintu ; wiyot ; yana ; yaqui ; yareba ; yawa.

/i ˈeˈ a ˈoˈ u ˈəˈ/

achumawi ; dagbani ; dera ; guambiano ; iban ; iwam ; lenakel ; makian ; sama ; socotri ; tarok ; ya.

/ɛ u i ɔ a/

aïnou ; burarra ; chin ; ekari ; khasi ; maung ; mongol ; russe ; suena ; zoulou ; zuni ;

/i e ɛ a ɔ o u/

birom ; dangaleat ; efik ; irarutu ; maba ; mba-ne ; nyangi ; nyimang ; temein ; tunica ; youlou.

/i a u/

aléoute ; arrernte ; désert occidental ; diyari ; dyirbal ; gugu-yalanji ; inuit ; kalkatungu ; ngiyambaa ; tsimshian ; yidiny.

Il est donc possible de dresser une liste des langues qui partagent exactement le même système vocalique. Ainsi la Table 6 livre 271 langues possédant un système de voyelles qui n’est partagé avec aucune autre langue (systèmes uniques) ; 38 ont des systèmes qu'elles partagent avec une autre langue et ce sont des systèmes à 3, 4, 5, 6, 8, 10, 12, 14, 18 voyelles ; 15 langues partagent leurs systèmes avec 2 autres langues et ce sont des systèmes à 4, 5, 6, 7, 12 voyelles etc. Enfin 44 langues partagent leur système avec 43 autres langues et ce sont des systèmes à 5 voyelles.

Nombre de langues	Nombre de langues identiques	Nombre de voyelles
271	0	de 3 à 28
38	1	3 4 5 6 8 10 12 14 18
15	2	4 5 6 7 12
8	3	3 9
10	4	5 7
6	5	12
14	6	6 10
33	10	3 5 7
12	11	6
44	43	5

Table 6. Nombre de langues possédant des systèmes identiques, avec une indication sur le nombre de voyelles par système ($UPSID_{451}$).

10.Conclusion et perspectives

L'approche que nous avons tracée ici – qui vise à l'étude systématique des ressemblances et des différences entre les langues –, outre son intérêt linguistique, ouvre de nouvelles perspectives pour l'Identification Automatique des Langues. En effet, les connaissances typologiques nous semblent être un a priori difficilement contournable pour la constitution de bases de données destinées à l'évaluation de procédures d'IAL. Grâce aux typologies des structures sonores, il est ainsi possible de construire des bases de données contenant des langues a priori très différentes et/ou a priori très voisines et non plus sélectionnées au hasard des disponibilités, c'est-à-dire « tirées dans le noir ».

C'est dans cette approche qu'un travail pilote a été mené dans le cadre d'une convention avec la DGA [Boë 1998]. Les données ainsi constituées ont pu servir à Pellegrino [Pellegrino 1998] pour le choix des langues (en l'occurrence les cinq langues suivantes : coréen, français, espagnol, japonais et vietnamien) et l'élaboration d'un système de reconnaissance sans apprentissage, à partir de la connaissance a priori des systèmes phonologiques respectifs. C'est une voie prometteuse qui associe les connaissances phonologiques et les techniques développées en reconnaissance automatique et qui permet – qui plus est – une interprétation linguistique des matrices de confusion entre les langues.

11.Remerciements

Nos remerciements à Ian Maddieson qui a mis à notre disposition $UPSID_{451}$ et nous a consacré beaucoup de son temps pour le suivi de son implantation. Un grand merci à Noureddine Abbadeni, Valérie Alcantara, Florence Mistrulli, Élisabeth Pinto pour leurs contributions.

Cette présentation a été élaborée dans le cadre de la convention DGA/DRET N° 95/118 : Discrimination Multilingue Automatique, et une partie a été financée par une action de la Région Rhône-Alpes (ARASSH) et du GIS des Sciences de la Cognition.

12.Références

[Abbadeni 1996] Abbadeni N. *Tendances universelles et diversité des structures sonores des langues du monde.* DEA Sciences du Langage, Université Stendhal, Grenoble, (1996)

[Alcantara 1998] AlcantaraV. *Le trait de labialité dans les systèmes vocaliques d'UPSID : de la typologie à la modélisation articulatoire. Application à la correction phonétique.* TER de maîtrise Sciences du Langage, Université Stendhal, Grenoble, (1998)

[Baudoin 1894] Baudouin de Courtenay J. An Attempt at a Theory of Phonetic Alter Nations. *A Baudoin de Courtenay Anthology*, Bloomington, Indiana, 144-212, édition de 1972 (article publié en polonais en 1894), (1894)

[Boë 1998] Boë L.J., Vallée N., Belrhali R. *Typologies des langues à partir des données vocaliques et consonantiques. Discrimination multilingue automatique.* Convention DGA n° 95/118, 459 p (1998).

[Boltanski 1995] Boltanski J.E. *La linguistique diachronique*. Paris : PUF, (1995)

[Bopp 1816] Bopp F. *Vergleichende Grammatik des Sanskrit, Zend, Griechischen, Lateinischen, Litauischen, Gotischen und Deutschen.* Berlin (1816).

[Creissels 1994] Creissels D. *Aperçus sur les structures phonologiques des langues négro-africaines*, 2ème éd., ELLUG, Grenoble, 322p, (1994)

[Crothers 1978] Crothers J. Typology and Universals of Vowel Systems. *Universals of Human Language,* J.H. Greenberg Ed., 93-152, Stanford University Press, Stanford, (1978)

[Greenberg 1978] Greenberg J.H., Ferguson C.A., Moravcsik E.A. (Ed.) *Universals of Human Languages: Method and Theory, Phonology, Word Structure, Syntax.* Stanford Univ. Press, California, (1978)

[Grimm 1822] *Deutsche Grammatik.* 2nd édition du volume 1 de 1819, Götttingen (1822).

[Hagège 1982] Hagège C. *Les structures des langues.* Que sais-je, PUF, Paris. 2e édition 1986, (1982)

[Hockett 1955] Hockett C.F. *A Manual of Phonology.* Waverly Press, Baltimore, 246 p., aussi Publications in Anthropology and Linguistics, Indiana University, Bloomington, (1955)

[Hombert 1997] Hombert J.M., Maddieson I. Linguistic approaches to Automatic Language Recognition. *XVI^e Congrès International des Linguistes,* Paris (1997).

[Jakobson 1941] Jakobson R. *Kindersprache, Aphasie, und allgemeine Lautgestze.* Uppsala Universitets Årsskrift 1942, 1-83. Republié dans Jakobson R, (1962), Selected writings I, Mouton, The Hague, 328-401, (1941)

[Jones 1786] Jones Sir W. Troisième discours anniversaire : On the Hindus. Reproduit dans *The Collected Works of Sir William Jones III*, 1807, John Stockdale, Londres, 23-46 (1786).

[Keating 1990] Keating P.A. Coronal places of articulation. *UCLA WPP 74,* 35-60, (1990)

[Laver J. 1994] Laver J. *Principles of Phonetics*. Cambridge University Press, Cambridge, (1994)

[Ladefoged 1995] Ladefoged P. The sounds of disappearing languages. *The newsletter of The Acoustical Society of America*, 5, 1, 1-6, (1995)

[Ladefoged 1996] Ladefoged P., Maddieson I. *The Sounds of World's Languages*. Blackwell publishers, Oxford, (1996)

[Lindblom 1988] Lindblom B., Maddieson I.Phonetic universals in consonant systems. *Language, Speech and Mind*, ed. by Hyman L.H. & Li C.N., London, New-york, 62-78, (1988)

[Maddieson 1986] Maddieson I. *Patterns of Sounds.* 2nd Ed. Cambridge University Press, Cambridge, (1986)

[Maddieson 1990] Maddieson I., Precoda K. Updating UPSID. *UCLA Working Papers in Phonetics 74*, 104-111, (1990)

[Maddieson 1991a] Maddieson I. Investigating Linguistic Universals. *XII^e Congrès International des Sciences Phonétiques*, Aix-en-Provence, France, Vol. 1/5, 346-354, aussi UCLA Working Papers in Phonetics 78, 26-37, (1991)

[Maddieson 1991b] Maddieson I. Testing the Universality of Phonological Generalizations with a Phonetically Specified Segment Database: Results and Limitations. *Phonetica 48*, 193-206, (1991)

[Meillet 1922] Meillet A. *Introduction à l'étude comparative des langues indo-européennes*. Hachette, Paris (1922).

[Muthusamy 1994] Muthusamy Y., Barnard E. & Cole R. Reviewing Automatic Language Identification. *IEEE Signal Processing*, oct., 33-40 (1994).

[Pellegrino 1998] Pellegrino F. *Une approche phonétique en identification automatique des langues : la modélisation acoustique des systèmes vocaliques.* Thèse en informatique : traitement automatique de la parole, Université Paul Sabatier, Toulouse (1998).

[Rask 1918] Rask R.K. *Undersøgelse om det nordiske eller islandske sprogs oprindelse.* Copenhague (1818).

[Ruhlen 1987] Ruhlen M.*A Guide to the World's Languages*. Volume 1. Classification. Edward Arnold, London, (1987)

[Schleicher 1861-1862] Schleicher A. *Compendium der vergleichenden Grammatik der indogermanischen Sprachen,* Weimar, (1861-1862).

[Schwartz 1997] Schwartz J.L., Boë L.J., Vallée N., Abry C. Major Trends in Vowel System Inventories. *J. of Phonetics* 25, 233-253, (1997)

[Sedlak 1969] Sedlak, P. Typological Considerations of Vowel Quality Systems. *Working Papers on Language Universals 1*, Stanford University, 1-40, (1969)

[Sergent 1995] Sergent, B. *Les Indo-Européens. Histoire, langue, mythes*. Bibliotèque Scientifique Payot, Paris (1995).

[Stefanuto 1996] Stefanuto M. *Typologie des lieux d'articulation des langues du monde*. TER Sciences du Langage, Université Stendhal, Grenoble, (1996)

[Troubetzkoy 1939] Troubetzkoy N.S. *Grundzüge des Phonologie.* Travaux du Cercle Linguistique de Prague 7, 272p., traduit en français par Cantineau J. (1970) sous le titre Principes de phonologie, Klincksieck, Paris, 394p, (1939)

[Vallée 1994] Vallée N. *Systèmes vocaliques : de la typologie aux prédictions.* Thèse de Doctorat Sciences du Langage, Université Stendhal, Grenoble, (1994)

[Vallée 1998] Vallée N., Boë L.J., Stefanuto M. *Les systèmes consonantiques. Des tendances universelles à l'ontogenèse.* XXII^{emes} Journées d'Étude sur la Parole, Martigny, Suisse, 15-19 juin, 241-244, (1998)

[Young 1913] Young T. Mithradates, oder allgemeine Sprachenkunde. *The Quarterly Review 10*, 250-292 (1813).

2ème Partie

Approches automatiques en identification

Identification automatique de la langue par téléphone

D. Matrouf[1,2], M. Adda-Decker[1], J.-L. Gauvain[1], L.Lamel[1]

[1]LIMSI-CNRS, BP 133, 91403 Orsay Cedex, FRANCE
[2]LIA, Université d'Avignon, France

{madda, lamel, gauvain}@limsi.fr – Driss.Matrouf@lia.univ-avignon.fr

Résumé

Dans cette contribution nous présentons nos travaux récents en identification automatique de la langue à travers le téléphone. Différentes approches sont présentées et discutées. Une approche acoustico-phonétique et lexicale a été mise en œuvre et testée pour 4 langues (corpus IDEAL). L'introduction des N mots les plus fréquents dans chaque langue a permis de réduire le taux d'erreur d'environ 20% en relatif. Ceci montre l'importance de l'information lexicale pour un système d'identification automatique de la langue. Pour une tâche de 11 langues (corpus OGI) une approche phonotactique a été implémentée et testée avec différentes configurations de prétraitement acoustico-phonétique. En particulier ces expériences montrent l'intérêt de modèles acoustiques multilingues.

Abstract

In this contribution we describe our recent progress in automatic language identification (LID) on telephone speech. Different approaches are presented and discussed. An acoustic-phonetic approach incorporating lexical information has been implemented and tested on a four language task (IDEAL). Adding the N most frequent words of each language into the LID system yields a relative error reduction of about 20% relative. This result underlines the importance of lexical information for automatic LID. A phonotactic approach has been developed and evaluated on an 11 language task using the OGI corpus. Within the framework of the phonotactic approach different acoustic-phonetic preprocessing configurations have been experimented with. In particular a comparison of multilingual and language-dependent acoustic phone models is carried out. The multilingual models are shown to improve LID results.

1. Introduction

Afin d'identifier une langue de manière automatique à partir d'un signal de parole, différentes possibilités de modélisation ont été explorées par les chercheurs. On admet généralement que l'information utile à l'identification de la langue se trouve répartie aux différents niveaux de représentation de la langue : acoustique, phonétique, phonotactique, prosodique, lexical... Les systèmes d'identification automatique se limitent le plus souvent à une modélisation qui reste proche du signal. Une comparaison des principales méthodes a été présentée par Zissman [Zissman 96]. Une revue plus large peut être trouvée dans [Pellegrino 98].

Dans cette contribution nous comparons deux familles d'approches pour l'identification de la langue : une approche acoustico-phonétique qui est fondée sur des modèles acoustiques de

phones dépendants de la langue et une approche phonotactique qui fait simplement appel à des modèles phonotactiques dépendants de la langue. Avec l'approche acoustico-phonétique nous introduisons pour chaque langue connue du système, en plus de l'ensemble de phonèmes dépendants de la langue, un ensemble des N mots les plus fréquents, et évaluons l'apport de cette information supplémentaire. Cette approche fait donc appel à de nombreuses connaissances, ce qui permet d'espérer de bonnes performances en identification automatique. En contrepartie l'extension à de nouvelles langues est pour le moins coûteux, voire impossible par manque de ressources spécifiques pour ces nouvelles langues. Le deuxième type d'approche présente l'avantage de permettre une extension facile à une nouvelle langue. Il suffit d'un simple corpus de parole à partir duquel on estime des modèles phonotactiques dépendants de la langue via un module de prétraitement acoustico-phonétique (un système de décodage acoustico-phonétique quelconque). Nous étudions le lien entre les résultats d'identification et ce système de décodage acoustico-phonétique.

2. Approche acoustico-phonétique et lexicale

2.1. Méthode

Les modèles spécifiques à la langue mis en jeu sont de nature acoustico-phonétique, $f(x|\phi, l)$ (modèle de Markov caché correspondant à une séquence de phonèmes ϕ pour la langue l) et de nature phonotactique, $\Pr(\phi|l)$ (modèle bigramme sur les séquences de symboles de phonèmes de la langue l). Ainsi, si l'on dispose de modèles acoustico-phonétiques et phonotactiques pour chacune des langues considérées, on obtient la solution optimale suivante :

$$\arg\max_{l} \sum_{\phi} f\left(x|\phi, l\right) \Pr\left(\phi|l\right)$$

En utilisant l'approximation de Viterbi le problème peut être considérablement simplifié grâce à l'usage de la programmation dynamique :

$$\arg\max_{l} \max_{\phi} \left(f\left(x|\phi, l\right) \Pr\left(\phi|l\right) \right)$$

L'approche est schématisée pour 3 langues dans la figure 1.

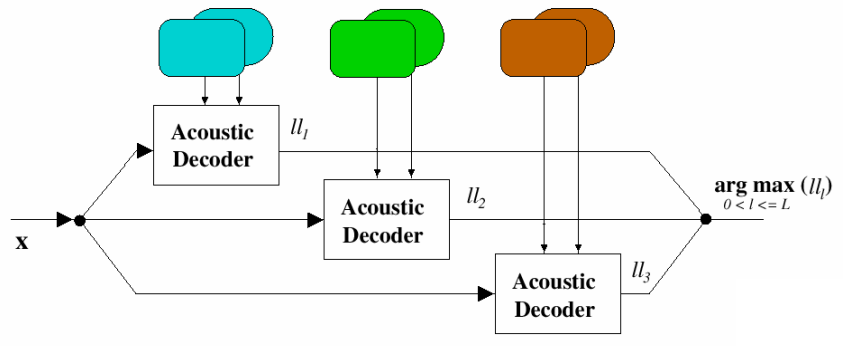

Figure 1 – *Système d'identification utilisant pour chaque langue (parmi L = 3 dans le schéma) des modèles acoustiques de phones et un bigramme de phones qui peut être augmenté des N mots les plus fréquents.*

2.2. Corpus

Le corpus utilisé correspond au corpus IDEAL [Lamel 98]. D'autres travaux de recherche en identification de la langue utilisant ce corpus ont été publiés dans [Matrouf 98, Corredor-Ardoy 97].

IDEAL est un large corpus téléphonique comprenant 4 langues (français, anglais, allemand et espagnol). IDEAL a été conçu pour la recherche en identification automatique de la langue. Le contenu de IDEAL est similaire au corpus OGI [Muthusamy 92], mais les locuteurs sont des autochtones, qui appellent de leur pays.

langue	#appels	#hommes	#femmes	#heures
Allemand	257	109	148	15,8
Anglais	258	109	149	14,8
Espagnol	253	114	139	17,9
Français	259	129	130	13,1

Tableau 1–Résumé des données d'apprentissage IDEAL utilisées pour les expériences avec l'approche acoustico-phonétique et lexicale.

Le corpus comprend de la parole lue et de la parole spontanée. On peut distinguer 3 types de données pas locuteur :

- des **informations** générales concernant l**'appel** non utilisées dans ces travaux.
- de la **parole lue et préparée** avec des phrases et des suites de chiffres et de nombres à lire, ou des réponses à des questions simples concernant la date et l'heure par exemple.
- de la **parole spontanée**, en réponse à des questions diverses ("décrivez l'endroit où vous habitez, décrivez votre maison de rêve...").

La parole spontanée représente environ 15% du corpus. Le corpus comprenant plus de 300 appels par langue, environ 250 appels sont réservés pour l'apprentissage des modèles et quelques 50 appels différents pour le test.

2.3. Résultats expérimentaux

Des modèles acoustico-phonétiques (HMM de phones contexte-indépendants) sont appris pour chaque langue à partir de la totalité des 250 appels en apprentissage. Pour différentes valeurs de N (mots les plus fréquents) des modèles de langage (bigrammes de phones et de N mots) sont estimés. N varie de 0 (approche acoustico-phonétique pure) à N=500.

Suivant le type d'énoncé (texte de journal, spontané, chiffres...) les N mots vont permettre une couverture plus ou moins importante du test. Afin de mesurer l'influence du type de parole (lue et préparée ou spontanée) sur les résultats d'identification, nous avons conçu différents ensembles de test à partir du corpus de test. Pour la parole lue et préparée un premier jeu de test **lu & préparé** contient tous les énoncés de cette partie du corpus : des phrases de textes de journaux, des transcriptions lues de demande d'informations, des adresses, des dates... Un

deuxième jeu de test **nombres** se limite alors aux énoncés faisant intervenir majoritairement des nombres : dates, heures, cartes de crédit, numéros de téléphone, somme d'argent... Pour la parole spontané un seul jeu de test **spontané** a été utilisé incluant la totalité de cette partie du corpus. Des modèles phonotactiques spécifiques sont estimés pour chaque type de test **spontané, lu & préparé, nombres** à partir des 200 appels d'apprentissage.

Les résultats sont montrés par les courbes **spontané, lu & préparé, nombres** dans la figure 2 en fonction du nombre N des mots les plus fréquents rajoutés pour le modèle de langage.

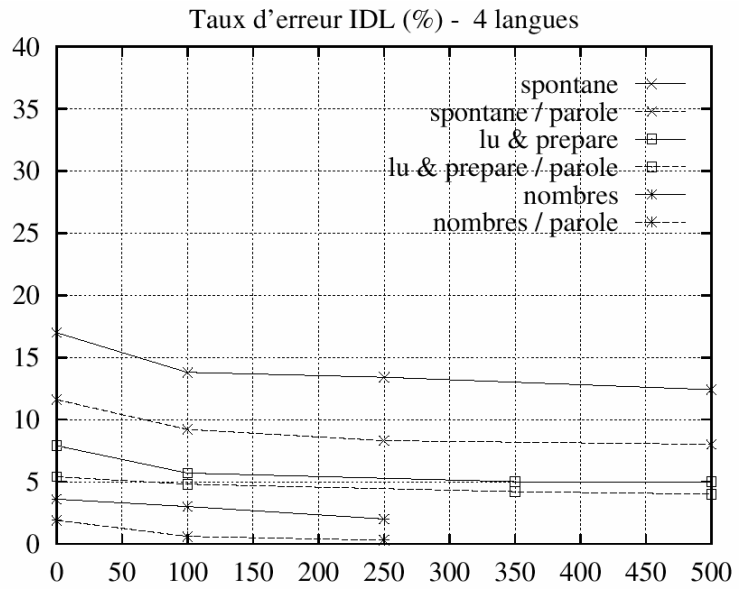

Figure.2–Taux d'erreur d'identification sur des segments de ***5 sec.*** en fonction de N *(*N = *0; ...; 500*), N étant le nombre de mots utilisés dans le modèle phonotactique. Les courbes notées *spontan*é, *lu & préparé*, e*t nombres* (phrases lues incluant majoritairement des nombres) permettent de mesurer l'impact du style de parole et du contenu du test sur le taux d'erreur. Les courbes marquées */parole* correspondent à des segments de parole (excluant le silence en début et fin).

Pour les 3 jeux de test seuls les segments de 5 secondes sont utilisés. La durée utile de parole est donc au plus 5 secondes par énoncé, la durée complémentaire correspondant à du silence ou du bruit. Nous avons décidé de mesurer l'influence de ces silences sur les résultats d'identification en comparant les résultats obtenus à ceux où seule la parole utile est utilisée. Celle-ci a donc été localisée auparavant par alignement de la transcription avec le signal. Ces résultats supplémentaires sont signalés en rajoutant **/parole** à chaque type de test. Les résultats obtenus permettent de constater plusieurs choses :

- L'approche acoustico-phonétique permet d'obtenir de bons résultats pour des segments très courts (5 sec.).

- La modélisation des N mots les plus fréquents a permis d'améliorer les taux d'identification de manière significative avec un gain relatif supérieur à 25% dans toutes les configurations de test.
- Les résultats varient de manière significative suivant le contenu du test. Des résultats proches de 100% sont obtenus pour le jeu de test **nombres** alors que pour la parole spontanée le taux d'erreur reste supérieur à 10% dans la meilleure configuration.
- Les résultats montrent qu'à l'instar des systèmes de reconnaissance, les systèmes d'identification de la langue ont beaucoup plus de difficultés avec la parole spontanée qu'avec la parole lue.
- Les segments de silence perturbent les résultats d'identification, particulièrement pour la parole spontanée et les nombres.

Afin de remédier à la sensibilité observée en présence de silence ou bruit le système d'identification doit faire appel à un module de détection de la parole.

3. Approche phonotactique

3.1. Méthode

Dans le cas où il n'existe pas de données d'apprentissage transcrites pour une ou plusieurs langues, l'approche acoustico-phonétique devient impossible. On fait alors appel à l'approche phonotactique.

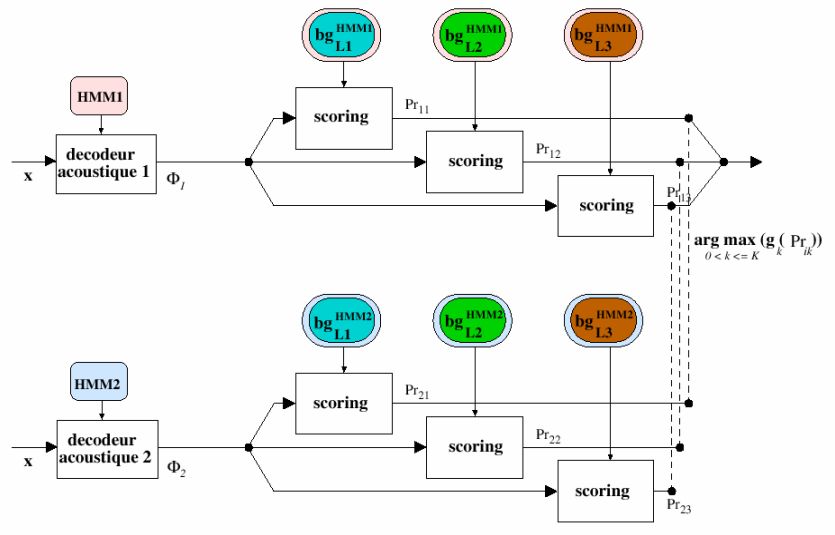

Figure 3 – Schéma d'un système d'identification fondé sur l'approche phonotactique avec 2 décodeurs acoustiques en parallèle et 3 langues à identifier.

On peut transcrire le signal x de la langue l à l'aide des modèles acoustiques de phones d'une langue quelconque k (prétraitement acoustico-phonétique) :

$$\phi_k = \arg\max_{\phi}\left(f\left(x|\phi, k\right)\Pr(\phi|k)\right) \qquad (1)$$

ou bien, en ignorant les contraintes phonotactiques de la langue k :

$$\phi_k = \arg\max_{\phi} f(x|\phi, k) \qquad (2)$$

A partir de ϕ_k de la langue l on peut estimer le modèle phonotactique $\Pr(\phi_k | l)$.Le problème d'identification de la langue revient alors à l'équation suivante :

$$l^* = \arg\max_{l} \Pr(\phi_k | l) \qquad (3)$$

Si l'on dispose de modèles acoustiques pour K langues, on peut considérer que l'observation est (ϕ_1, …, ϕ_2, …, ϕ_K) (c'est-à-dire le résultat des K décodeurs pour le signal x), le problème d'identification de la langue se ramène alors à l'équation suivante si l'on suppose les différentes suites de phones indépendantes :

$$l^* = \arg\max_{l} \prod_{k=1}^{K} \Pr(\phi_k | l) \qquad (4)$$

3.2. Corpus

Pour les modèles acoustiques nous utilisons exclusivement le corpus IDEAL. Pour l'estimation des modèles phonotactiques dépendants de la langue (mais aussi du décodeur acoustique qui fournit la suite de phones ϕ_k) le corpus OGI-TS (Oregon Graduate Institute Multi-Language Telephone Speech) est utilisé [Muthusamy 92]. Ce corpus comprend 11 langues (français, anglais, allemand, espagnol, japonais, coréen, mandarin, tamil, farsi, vietnamien, hindi) avec plus de 100 appels par langue provenant de locuteurs allogènes (d'une origine différente de celle de la population autochtone) vivant aux états-Unis. Par appel il y a 4 énoncés du type **parole préparée** ("prononcez les jours de la semaine, les chiffres de 0 à 10")...) et 6 énoncés spontanés ("décrivez l'endroit d'où vous appelez")...). Seuls les énoncés spontanés sont utilisés pour les modèles de langage phonotactiques.

Les tests utilisent uniquement les données spontanées avec des durées des segments variables de 10 à 45 secondes.

3.3. Résultats expérimentaux

Différentes combinaisons de décodeurs acoustico-phonétiques ont été évaluées, allant de la configuration avec un seul décodeur acoustico-phonétique dépendant d'une langue jusqu'à 5 décodeurs acoustico-phonétiques en parallèle. Les décodeurs acoustico-phonétiques dépendants de la langue sont ceux développés pour l'approche acoustique avec le corpus IDEAL.

Un ensemble de modèles acoustiques de phones multilingues a été développé par classification automatique des modèles de phones dépendants de la langue [Boula de Mareüil 2000, Corredor-Ardoy 98]. Ce système est appelé IL (indépendant de la langue) dans la figure 4.

Des configurations avec 1, 4 ou 5 systèmes en parallèle ont été utilisées. Les résultats de cette figure sont donnés en fonction de la longueur du test (de 10 à 45 secondes) avec 11 langues du corpus OGI-TS (français, anglais, allemand, espagnol, japonais, coréen, mandarin, tamil, farsi, vietnamien, hindi). Alors que le taux d'erreur est encore proche de 20% avec le meilleur système sur des segments de 10 secondes, ce taux chute à environ 10% pour des segments de 45 secondes.

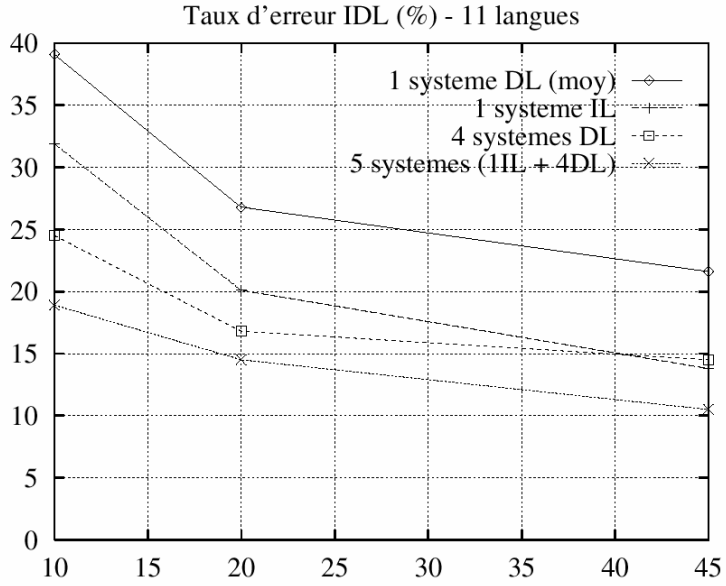

Figure 4–Taux d'erreur d'identification avec 11 langues en fonction de la longueur du test (de 10 à 45 sec.). Les courbes présentées correspondent à différentes combinaisons de décodeurs acoustiques : systèmes dépendants de la langue (DL, développés pour l'approche acoustique avec le corpus IDEAL) et indépendant de la langue (IL, obtenu par classification automatique).

Les courbes DL et IL correspondent à une configuration où un seul décodeur acoustico-phonétique, soit dépendant de la langue (DL), soit indépendant de la langue (IL), est utilisé. On peut remarquer que le décodeur utilisant les modèles multilingues est significativement meilleur que si on utilise un seul décodeur dépendant de la langue.

La figure montre la performance moyenne (1 système DL (moy)) après avoir testé séparément les 4 décodeurs dépendants de la langue. De manière générale les 4 systèmes en parallèle (4 systèmes DL) permettent d'améliorer les taux d'identification par rapport au décodeur commun.

Pour les segments de 45 secondes le système acoustique commun produit des résultats équivalents à la configuration avec 4 systèmes acoustiques (dépendants de la langue) en parallèle. Le passage à 5 systèmes (5 systèmes(1IL+4DL)) produit une amélioration significative de 14,5% à 10,5%.

Nous donnons dans le Tableau 2 la matrice de confusion observée pour les 11 langues. On peut remarquer que le nombre de segments par langue est faible.En examinant les résultats par langue on ne peut pas distinguer un comportement clairement différent entre les langues représentées par les modèles acoustico-phonétiques et celles qui ne l'étaient pas. On peut noter que le français, l'allemand, le tamil et le farsi sont identifiés à 100%. Pour l'allemand, le farsi, le japonais et le mandarin aucune fausse alerte n'a été faite.

Pour l'approche phonotactique nous avons montré l'intérêt d'utiliser un ensemble de modèles acoustiques de phones indépendant de la langue lors du décodage acoustique. La combinaison de plusieurs décodeurs en parallèle permet d'améliorer encore les taux d'identification. Des résultats comparatifs ont été obtenus en utilisant un ou plusieurs décodeurs, sur des segments de test de longueur variable. Les résultats sont d'autant meilleurs que les segments à identifier sont longs (de 10 à 45 sec.). L'extension à une nouvelle langue est

beaucoup plus simple ici qu'avec l'approche acoustique pour laquelle la mise au point d'un décodeur acoustico-phonétique dépendant de la langue est nécessaire.

45s	#segm	fr	an	al	es	ja	ko	ma	ta	fa	hi	vi
fr	15	100										
an	19		94,7		5,3							
al	18			100								
es	14	7,1	7,1		78,6				7,1			
ja	15	6,7				86,7						6,7
ko	8						87,5				12,5	
ma	11	9,1					9,1	81,9				
ta	12								100			
fa	13									100		
hi	11		9,1				9,1				72,7	9,1
vi	16	6,2							6,2			87,5

Tableau 2–Résultats sur le corpus OGI-11. Matrice de confusion entre les 11 langues pour des segments de durée de 45s avec 5 décodeurs acoustico-phonétiques.

4. Conclusion

Nous avons présenté nos travaux en identification de la langue utilisant deux familles d'approches. La première, l'approche acoustico-phonétique, nécessite des corpus étiquetés phonétiquement pour chaque langue, afin de pouvoir estimer des modèles acoustico-phonétiques pour chaque langue. L'ajout des N mots les plus fréquents dans le modèle de langage (bigramme de phones augmenté de N mots) permet une réduction du taux d'erreur de 25% en relatif. Sur une tâche de 4 langues un corpus de test de parole spontanée contenant des segments de 5 secondes donne un taux d'erreur proche de 10%. Ce taux chute près de 5% pour un test de parole lue et préparée et près de 2% pour un corpus de test incluant majoritairement des chiffres et des nombres. Les erreurs observées ici sont dues surtout au silence. L'approche acoustico-phonétique donne de bons résultats sur des segments relativement court (5 secondes), mais l'extension à une nouvelle langue est souvent problématique à cause des ressources spécifiques à la langue requises.

L'approche phonotactique est plus facile à adapter à de nouvelles langues dans la mesure où de simples corpus de parole sont suffisants. Cette approche a été utilisée pour développer un système d'identification pour 11 langues (corpus OGI-TS). Les résultats d'identification, mauvais sur des courts segments de parole, s'améliorent avec la durée des segments de test et des taux d'erreurs proche de 10% sont obtenus avec des segments de 45 secondes. Nous avons montré l'intérêt d'utiliser un ensemble de modèles acoustiques de phones multilingue lors du décodage acoustique. La combinaison de plusieurs décodeurs en parallèle permet d'améliorer encore les taux d'identification.

5. Références

[Lamel 98] L. Lamel, G. Adda, M. Adda-Decker, C. Corredor-Ardoy, J.J. Gangolf, J.L. Gauvain, 'A Multilingual Corpus for Language Identification," *in Proc. of 1st International Conference on Language Resources and Evaluatio*n, 1, pp. 1115-1122, Grenade, mai 1998.

[Muthusamy 92] Y.K. Muthusamy, R.A. Cole, B.T. Oshika (1992),"The OGI Multi-Language Telephone Speech Corpus," *in Proc. of International Conference on Speech and Language Processin*g, 2, pp. 895-898 Banff, octobre 1992.

[Boula de Mareüil 2000] P. Boula de Mareüil, C. Corredor-Ardoy, D. Matrouf, M. Adda-Decker, "Classement automatique de phonèmes dans un cadre multilingue & application à l'identification de la langue", ailleurs dans ces actes.

[Lamel 94] L.F. Lamel, J.L. Gauvain, "Language Identification Using Phone-based Acoustic Likelihoods", *in Proc. of IEEE-ICASS*P, Adelaide 1994.

[Matrouf 98] D. Matrouf, M. Adda-Decker, L. Lamel, and J.L. Gauvain, "Language Identification Incorporating Lexical Information", *in Proc. of International Conference on Speech and Language Processin*g, Sydney 1998.

[Corredor-Ardoy 98] C. Corredor-Ardoy, P. Boula de Mareüil, M. Adda-Decker, L. Lamel, J.L. Gauvain, "Classement automatique de phonèmes dans un cadre multilingue", in actes *XXIIièmes Journées d' études sur la Parol*e, pages 75-78, Martigny, Suisse, juin 1998.

[Corredor-Ardoy 97] C. Corredor-Ardoy, J.L. Gauvain, M. Adda-Decker, L. Lamel, "Language identification with language-independent acoustic models", *in Proc. of European Conference on Speech Technology, EuroSpeec*h, 3, pages 1423-1426, Rhodes, septembre 1997. 9

[Lamel 93] L. Lamel, J.L. Gauvain, "Identifying Non-Linguistic Speech Features," *in Proc. of European Conference on Speech Technology, EuroSpeech* 1, pp. 23-28, Berlin, septembre 1993.

[Lowe 94] S. Lowe, A. Demedts, L. Gillick, M. Mandel, B. Peskin, "Language Identification via Large Vocabulary Speaker Independent Continuous Speech Recognition," *in Proc. of ARPA Human Language Technology Worksho*p, pp. 437-441, Plainsboro, mars 1994.

[Schultz 97] T. Schultz, A. Waibel, "Fast Bootstrapping of LVCSR Systems with Multilingual Phone Sets" , *in Proc. of European Conference on Speech Technology, EuroSpeec*h, 1, pp. 371-374, Rhodes, septembre 1997.

[Zissman 96] M.A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech," *IEEE Transactions on SA*P, 4(1), Jan. 1996.

[Zissman 97] M.A. Zissman, "Predicting, Diagnosing and Improving Automatic Language Identification Performance,", *in Proc. of European Conference on Speech Technology, EuroSpeec*h, 1, pp. 51-54, Rhodes, septembre 1997.

[Pellegrino 98] F. Pellegrino, "Une approche phonétique en identification automatique de la langue : la modélisation acoustique des systèmes vocaliques", Thèse de l'Université Paul Sabatier, Toulouse 1998.

Classement automatique de phonèmes dans un cadre multilingue et application à l'identification automatique de la langue

P. Boula de Mareüil[+], C. Corredor-Ardoy[*], D. Matrouf[1,2] M. Adda-Decker[1]

[1] LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
[2] LIA Avignon, FRANCE

{madda, matrouf}@limsi.fr, pboula@elan.fr, Driss.Matrouf@lia.univ-avignon.fr

Résumé

Nous décrivons un algorithme de classement automatique visant à regrouper les phonèmes de différentes langues et nous donnons une interprétation linguistique des résultats. Les données acoustiques sont des vecteurs de paramètres cepstraux, et chaque phonème de chaque langue est représenté par un modèle de Markov caché. Le classement utilise une mesure de similitude entre phonèmes, calculée à partir des vecteurs acoustiques et des modèles de Markov. L'application du classement à six langues: français, anglais, allemand, espagnol (corpus IDEAL), italien et portugais (corpus SPEECHDAT) de deux corpus différents a confirmé la robustesse de la méthode à travers des conditions acoustiques différentes. Un regroupement des 235 phonèmes initiaux en 90 classes préserve assez bien l'information phonématique. L'algorithme de classement a été appliqué avec succès à l'identification de la langue.

Abstract

An approach for automatic multi-lingual phoneme classification using agglomerative hierarchical clustering is described. The used similarity measure is based on the likelihood between the acoustic frames and the Hidden Markov Models. The method was applied to French, English, German, Spanish (IDEAL corpus), as well as to Italian and Portuguese (SPEECHDAT corpus). The phone cluster analysis showed that, despite an acoustic mismatch between both corpora, the approach remains robust. For 90 clusters, the obtained classes correspond, to a large extent, to well defined linguistic groups. A qualitative analysis of the results is given. 90 classes preserves the phonemic information quite well. This algorithm was successfully used for automatic language identification.

1. Introduction

La recherche d'une typologie des phonèmes constitue un des enjeux de l'étude de la parole [Haudricourt 67, Vallée 96] dans un but d'enseignement [Delattre 65] ou de traitement automatique de la langue [Berkling 94, Berkling 96, Corredor-Ardoy 97]. Dans un cadre

---

+ P. Boula de Mareüil a travaillé au LIMSI-CNRS pendant le développement de la méthode proposée dans cet article. Il travaille maintenant chez Elan Informatique, Toulouse.

[*] C. Corredor-Ardoy a travaillé au LIMSI-CNRS pendant le développement de la méthode proposée dans cet article. Il travaille maintenant chez Bouygues Telecom. 51,Avenue de l'Europe. 78944 Vélizy Cedex.

monolingue, l'algorithme de classement hiérarchique que nous proposons, peut aider à l'analyse des classes phonétiques, à la définition de l'ensemble des unités destinées à la reconnaissance de la parole, et au groupement d'un grand nombre de modèles markoviens en contexte. Le classement automatique des phonèmes dans un cadre multilingue a été incité par le développement de corpus destinés à l'identification automatique de la langue (IAL) : OGI-TS [Muthusamy 92], CALLHOME, CALLFRIEND, SPEECHDAT, IDEAL.

Dans l'approche que nous avons retenue, les paramètres acoustiques utilisés par l'algorithme de classement automatique correspondent à des coefficients cepstraux plutôt qu'à des paramètres obtenus par extraction de formants. Cette dernière en effet nécessite une segmentation voisé/non voisé, la localisation des noyaux vocaliques et l'extraction de la valeur des formants, décisions qui sont sujettes à erreur. Nous proposons un algorithme de classement hiérarchique. Berkling [Berkling 94] a également appliqué cette méthode à l'IAL de 6 langues du corpus OGI, mais elle utilise une mesure de similitude fondée sur des distances entre vecteurs acoustiques. Köhler [Köhler 96] utilise lui une mesure définie sur des vraisemblances acoustiques ; néanmoins, cette mesure a seulement été appliquée à 3 des 11 langues du corpus OGI. La normalisation introduite dans notre mesure de similitude, elle même fondée sur des vraisemblances acoustiques, a permis d'appliquer le classement automatique à un cadre plus étendu : celui des corpus IDEAL [Corredor-Ardoy 97] et SPEECHDAT. La mesure utilisée s'est révélée robuste par rapport aux différentes conditions d'enregistrement et au contenu linguistique des corpus.

Dans cet article, nous décrivons un algorithme de classement hiérarchique visant à regrouper les phonèmes de six langues européennes (français, anglais, allemand, espagnol, italien et portugais), nous donnons une interprétation linguistique des résultats et nous présentons des résultats obtenus avec ces classes en identification automatique de la langue. La section 2 présente l'algorithme de classement automatique. La section 3 est consacrée aux expériences menées en classement multilingue avec une analyse des classes obtenues en fonction de leur nombre. Un exemple d'utilisation de ce classement en identification automatique de la langue est montré dans la section 4.

2. L'algorithme de classement automatique

2.1. Préliminaire

Depuis l'apparition des premiers systèmes de reconnaissance vocale fondés sur une modélisation phonétique markovienne, il s'est avéré intéressant de pouvoir mesurer la distance entre les modèles [Young 93]. Cette mesure peut être utilisée pour réduire le nombre des modèles en contexte, pour déterminer l'ensemble optimal des phonèmes dans une langue ou pour établir un ensemble phonétique commun à plusieurs langues. Parmi les algorithmes de classement automatique, l'algorithme *k*-moyennes [Duda 73] est très dépendant de l'initialisation des classes. Nous avons ici opté pour une arborescence hiérarchique, même si celle-ci n'autorise pas de retour en arrière (même mal calculée, une classe ne peut être reconsidérée).

Les phonèmes des six langues ont ainsi été groupés par classement automatique hiérarchique [Duda 73]. Dans la phase d'initialisation de l'algorithme, chaque phonème est assigné à une classe. Après chaque itération, les deux classes ayant la plus grande similitude sont regroupées. La procédure se répète jusqu'a l'obtention du nombre de classes désiré. Il faut

noter que certaines classes peuvent contenir plusieurs phonèmes d'une même langue, car aucune contrainte sur l'origine linguistique des phonèmes n'a été prise en compte[1].

2.2. La mesure de similitude

La définition d'une mesure de similitude (ou de dissimilitude) entre phonèmes ou allophones est un sujet qui a été largement traité par la communauté scientifique. Young [Young 93] a défini la dissimilitude entre allophones en exprimant la divergence de deux gaussiennes en fonction de leur moyenne et de leur variance. Cependant, cette approche est seulement applicable à des modèles markoviens à un seul état et une seule gaussienne. En remplaçcant le concept de divergence par la distance de Bhattacharyya, Mak [Mak 96] a proposé une expression alternative de la dissimilitude en fonction des paramètres des modèles. Cette approche est applicable à des modèles markoviens avec mélange de gaussiennes, mais à un seul état. La comparaison de deux modèles markoviens à plusieurs états et mélange de gaussiennes requiert une mesure de similitude fondée sur la vraisemblance des données acoustiques (coefficients cepstraux) par rapport aux modèles markoviens. Dans ce type d'approche, Juang [Juan 85] et Köhler [Köhler 96] ont proposé une mesure de dissimilitude définie comme la différence des vraisemblances acoustiques. La mesure ainsi définie est dérivée du concept de divergence entre modèles markoviens.

Dans le cadre du travail présenté dans cet article, nous avons utilisé une mesure de similitude fondée sur le concept d'information mutuelle [Proakis 89]. Cette mesure constitue une approximation de la probabilité a posteriori $\Pr\left(\lambda_i \middle| \vec{\varphi}_j\right)$:

$$S(\varphi_i, \varphi_j) = \frac{f(\vec{\varphi}_j|\lambda_i)^\gamma}{\sum_{k=1}^{n} f(\vec{\varphi}_j|\lambda_k)^\gamma} \qquad (1)$$

où $\vec{\varphi}_j$ correspond aux données acoustiques du phonème φ_j ; λ_i est le modèle markovien du phonème φ_i ; f est la fonction de densité de probabilité des observations, connaissant les données d'apprentissage ; et n est le nombre d'unités phonétiques. Le coefficient γ a été introduit pour compenser l'hypothèse d'indépendance entre modèles (il a été fixé à 0,5 de façon empirique). Cette mesure étant asymétrique, nous avons utilisé une version symétrique, en prenant la moyenne des mesures de similitude :

$$S_s(\varphi_i, \varphi_j) = \frac{S(\varphi_i, \varphi_j) + S(\varphi_j, \varphi_i)}{2} \qquad (2)$$

Dans le cas où les classes contiennent plus d'un phonème, le concept de mesure de similitude phonétique s'étend à celui de mesure de similitude interclasse. Cette mesure est définie comme la moyenne des similitudes entre phonèmes[2] :

---

[1] Notre approche entre donc dans le type de classement maximal (inter-langue + intra-langue) proposé par Berkling [Berkling 96].

[2] La mesure de similitude entre deux classes peut aussi être calculée comme le maximum ou le minimum des similitudes entre phonèmes [Duda 73].

$$S(C_i, C_j) = \frac{1}{n_i n_j} \sum_{\varphi \in C_i} \sum_{\varphi' \in C_j} S_s(\varphi, \varphi') \qquad (3)$$

où n_i et n_j sont les nombres de phonèmes dans les classes C_i et C_j respectivement, et $S_s(\varphi;\varphi')$ est la mesure de similitude phonétique interclasse.

3. Expériences en classement multilingue

3.1. Description des corpus

L'algorithme de classement hiérarchique a été appliqué aux enregistrements du corpus IDEAL [Corredor-Ardoy 97] et du corpus SPEECHDAT. Le corpus IDEAL a été conçu pour le développement de systèmes d'identification automatique de la langue. Il contient plus de 300 appels téléphoniques en français, anglais, allemand et espagnol. Le contenu linguistique d'un appel est très varié : 18 phrases lues (phrases de journal, phrases phonétiquement équilibrées, dates, heures, adresses, etc.), des réponses à 12 questions d'ordre général (code de l'appel, sexe, âge, code postal, etc.) et à 6 questions destinées à recueillir de la parole spontanée.

Dans le cadre de ce travail, nous avons sélectionné environ 7500 phrases par langue, dont 40% sont lues et 60% constituent des réponses aux questions d'ordre général. Le corpus téléphonique SPEECHDAT a lui été construit pour le développement de systèmes de reconnaissance de la parole dans plusieurs langues européennes. En ce qui concerne l'italien et le portugais, nous avons sélectionné environ 9500 phrases phonétiquement équilibrées par langue.

3.2. Définition des alphabets phonétiques

Les unités phonétiques de chaque langue ont été sélectionnées différemment. Pour le français, l'anglais et l'allemand, nous avons repris les jeux de phonèmes utilisés par le LIMSI dans le cadre de la reconnaissance de la parole continue de ces langues (34, 44 et 46 phonèmes pour le français, l'anglais et l'allemand respectivement) [Gauvain 94, Adda-Decker 96]. Les 24 phonèmes de l'espagnol constituent l'ensemble de base traditionnellement retenu [Quilis 92]. Enfin pour l'italien et le portugais les jeux phonétiques sont issus de SAMPA [Gibbon 97] (49 et 38 phonèmes pour l'italien et le portugais respectivement). Au total, le nombre d'unités phonétiques est de 235 ; Le tableau 1 illustre le nombre de phonèmes par langue et par catégorie.

	Fr.	An.	Al.	Es.	It.	Po.
V simples	14	14	19	5	7	14
diphtongues	-	6	3	-	-	2
glides ou C syl.	3	2	4	-	2	2
C simples	17	20	20	17	17	20
affriquées	-	2	-	1	4	-
géminées	-	-	-	1	19	-
Total	34	44	46	24	49	38

Tableau 1 – nombre de phonèmes par langue (français, anglais, allemand, espagnol, italien, portugais) et par catégorie (V=voyelles, C=consonnes, syl.=syllabiques).

3.3. Calcul de la mesure de similitude

Chaque phonème est caractérisé au moyen d'un Modèle de Markov Caché à trois états indépendant du contexte (CI-CDHMM: *Context Independent Continuous Density Hidden Markov Model*). A chaque état, la loi d'émission d'observation est définie comme la somme pondérée de 32 gaussiennes. Pour chacune des langues, l'ensemble des phrases a été phonétiquement aligné. Cette segmentation a été obtenue en utilisant la transcription orthographique des phrases, le dictionnaire des mots sous forme graphémique et phonémique, et les modèles phonétiques.

A partir des vecteurs cepstraux correspondant aux 235 phonèmes, et à partir de leurs modèles markoviens associés, nous avons calculé la matrice des vraisemblances acoustiques. Avant d'effectuer cette opération, nous avons dû limiter le nombre de vecteurs acoustiques par phonème à moins de 20 000 échantillons (avec une valeur moyenne de 7 000 vecteurs cepstraux[3]), pour réduire le temps de calcul de la matrice de vraisemblance. Ceci correspond à environ 4 heures de parole avec en moyenne une minute de parole par phonème $\vec{\varphi}_j$ pour l'estimation de la mesure de similitude $S(\varphi_i,\varphi_j)$. Une fois cette matrice calculée, nous avons appliqué l'algorithme de classement hiérarchique en faisant varier le nombre de classes désirées.

3.4. Classements multilingues et analyse des classes

On peut regrouper de diverses façons les phonèmes d'une langue, a fortiori de plusieurs. Il est classique, en phonétique, d'adopter une représentation sous la forme d'un arbre plus ou moins hiérarchisée, reflétant un degré d'analyse plus ou moins profond. Ce type de représentation est également possible avec notre algorithme, où l'on choisit à l'avance le nombre de classes, cet algorithme faisant décroître de 1 le nombre de classes à chaque itération. Nous allons aborder rapidement le comportement de l'algorithme de classement avec un nombre variable de classes et donner ensuite une interprétation plus détaillée dans le cas de 90 classes. En effet nous nous intéressons surtout aux résultats de classement multilingue dans une configuration avec un minimum d'information phonématique perdue[4]. Nous ne prétendons pas définir formellement ce concept, qui est difficile à évaluer de manière objective. Le phonème et ses allophones sont traditionnellement définis pour une langue donnée, par l'analyse en paires minimales de commutation, liée à la fonction distinctive. Des recommandations et des conventions sont publiées chaque année par des spécialistes, pour décider d'inclure officiellement ou non des symboles et des diacritiques dans l'Alphabet Phonétique International (API). En même temps, ce n'est pas un hasard, intuitivement, si des sons voisins, dans plusieurs langues, sont représentés avec le même symbole API [Pullum 96]. C'est cette proximité que l'algorithme décrit ci-dessus permet d'explorer objectivement.

[3] Certains phonèmes comme les géminées de l'italien, ont très peu de représentants. Pour ces phonèmes rares, la validité des modèles statistiques peut être remise en question.

[4] Ce type de configuration a été retenu pour des expériences en identification de la langue.

Partant des 235 phonèmes de notre ensemble initial (pour le français, l'anglais, l'italien, l'espagnol, l'allemand et le portugais), on a très rapidement des géminées de l'italien qui se regroupent avec les consonnes simples correspondantes, mais ce n'est que vers 90 classes qu'émerge une classe /t/, commune aux 6 langues. Avec 80 classes, les /z/ se retrouvent regroupés, mais le /o/ est rattaché au /u/. Avec une dizaine de classes, on retombe à peu près sur les groupes suivants, qui s'excluent mutuellement : voyelles ouvertes, voyelles fermées, fricatives, occlusives, nasales, liquides, glides. Même si certains phonèmes sont réfractaires au regroupement, et même si d'autres regroupements ne sont pas ceux que nous souhaiterions, il est intéressant d'observer la pertinence linguistique des classes produites par un calcul sur des traits purement acoustiques (les coefficients cepstraux). Avec un classement maximal de 2 classes, il est intéressant de noter qu'on retrouve en gros une bipartition voyelle-consonne.

Interprétation des 90 classes de phonèmes

Avec la part d'arbitraire que cela comporte (mais il faut trouver un compromis), nous allons détailler maintenant le regroupement en 90 classes. On peut noter que 90 est grosso modo le nombre de symboles API différents utilisés pour décrire nos 6 langues.

Fr	*An*	*Al*	*Es*	*It*	*Po*
i	i	i	i	i	i
-	-	-	-	-	ĩ j̃
-	ɪ e	-	-	-	-
e	-	e: e ɪ	-	-	-
-	-	-	-	-	e ẽ
ɛ	-	ɛ: ɛ	e	e	ɛ
-	ɛ ɛə	-	-	-	-
a	æ ɑ: aʊ	a ɑ	a	a	a
ɑ	-	-	-	-	ɐ
-	-	aʊ	-	-	ɬ
-	aɪ	aɪ	-	-	-
ɔ	ɒ	ɔ	o	ɔ o	ɔ
o	ɔ:	o	-	-	o
-	-	-	-	-	ũ õ
u	-	-	u	u	u w
-	-	u ʊ	-	-	-
y	-	y	-	-	-
œ ø	-	-	-	-	-
-	-	ɐ̯ əʁ	-	-	-
ə	-	ə	-	-	-
-	ɜ˞ ə	-	-	-	ɐ
m	m	m	m	m mm	m
n	n	n ən	n	n	-
-	-	-	-	nn	n
ɲ	-	-	ɲ	ɲ ɲɲ	ɲ
-	ŋ	ŋ	-	-	-
-	-	-	ǰ ʎ	ʎʎ	ʎ
w	-	-	-	w	-
j	-	-	-	j	j

Fr	*An*	*Al*	*Es*	*It*	*Po*
f	f	f	f θ	f ff	f
s	s	s	s	s ss	s
ʃ ʒ	ʃ	ʃ	-	ʃ ʃʃ	ʒ
-	h	h	-	-	-
v	-	-	-	v	v
z	-	-	-	z	z
-	z v	-	-	-	-
-	ʤ ʧ	-	-	-	-
-	-	-	-	ʤ ʤʤ	-
-	-	-	-	ʧ ʧʧ	-
-	-	-	-	ts tsts	-
p	p	p	p	p pp	p
t	t	t	t	t tt	t
k	k	k	k	k kk	k
-	b	b	-	-	-
-	-	-	b	b bb	b
-	d	d	-	-	-
-	-	-	d	d dd	d
b d	-	-	-	-	-
g	g	g	-	-	-
-	-	-	g	g	g
l	-	l	l	l ll	-
-	l	-	-	-	l
ʁ	-	ʁ	-	-	-
-	-	-	x	-	ʀ
-	-	-	r r̄	r rr	r

Tableau 2 – regroupements des voyelles, nasales et glides (gauche), de fricatives, affriquées, plosives et liquides (droite) à travers les 6 langues étudiées.

Parmi les 90 classes on compte 35 singletons, correspondant à presque autant de voyelles que de consonnes : des exemples en sont le /ɥ/ français, le /ð/ et le /ʌ/ anglais, le /ç/ allemand, qui sont propres à leurs langues. Parmi les 55 classes restantes, nous avons pu regrouper des classes correspondantes aux voyelles, aux fricatives, affriquées et plosives, liquides et glides dans les différentes parties de la table 2. Certaines rassemblent uniquement 6 phonèmes de symbole API identique à travers les 6 langues (plus un 7ème pour les consonnes géminées de l'italien) : /i/, /s/, /m/, /t/, /p/, /k/. Dans ces chiffres, il faut bien sûr compter avec le bruit introduit initialement par des jeux de phonèmes hétérogènes (notamment en ce qui concerne les traits de durée, les diphtongues, les glides, les affriquées, les géminées), plus ou moins fins en allophones (vocaliques). Les affriquées et les diphtongues se voient souvent isolées – même si les /aɪ/ allemand et anglais sont regroupés. Notre modèle ne permet pas de déterminer s'ils s'apparieraient avec 2 segments (par exemple /t/ et /ʃ/ pour /tʃ/). Les classes dérivées automatiquement correspondent en général à des unités linguistiques identiques ou similaires.

4. Expériences en identification automatique de la langue

Le classement multilingue peut servir à estimer des HMM de phones indépendants de la langue, ou plus exactement des modèles multilingues. En effet on peut considérer que tous les phonèmes φ_{n_i} d'une même classe nommée C_n seront renommés C_n . Ceci entraîne que tous les vecteurs acoustiques du corpus d'apprentissage correspondants à ces phonèmes φ_{n_i} sont utilisées pour estimer un HMM de phone multilingue (représentant la classe C_n).

Ainsi, tous les vecteurs acoustiques étiquetés ɲ (français, espagnol, italien et portugais) participent à l'estimation d'un seul modèle de phone multilingue, les ɳ (anglais et allemand) à un autre modèle de phone multilingue. Ces nouveaux modèles acoustiques peuvent servir à l'identification de la langue avec une approche phonotactique. L'idée de cette approche est simple : lors de l'apprentissage du système d'identification de la langue, les modèles acoustiques multilingues sont utilisés dans un système de reconnaissance phonétique dans le seul but de transformer le corpus de parole d'une langue donnée en un corpus de suites de phones pour cette langue. A partir d'une telle suite de phones, des modèles phonotactiques sont estimés pour la langue en question. Lors du processus d'identification, une séquence de parole de langue inconnue sera transcrite comme suite de phones de manière automatique par le système de reconnaissance phonétique multilingue. Les modèles phonotactiques dépendants de la langue sont ensuite utilisés pour scorer la séquence de phones ainsi produite. La séquence de phones sera identifiée comme appartenant à la langue dont le modèle phonotactique donne la probabilité la plus élevée.

Cette approche est décrite plus en détail dans [Matrouf 2000]. Nous donnons ici qu'une description succincte des conditions expérimentales et quelques résultats. Les modèles acoustiques de phones multilingues sont estimés à partir du corpus IDEAL uniquement. Les modèles phonotactiques proviennent des données d'apprentissage de OGI (parole spontanée). Lors du test seuls les énoncés spontanés de la partie de test de OGI sont utilisés.

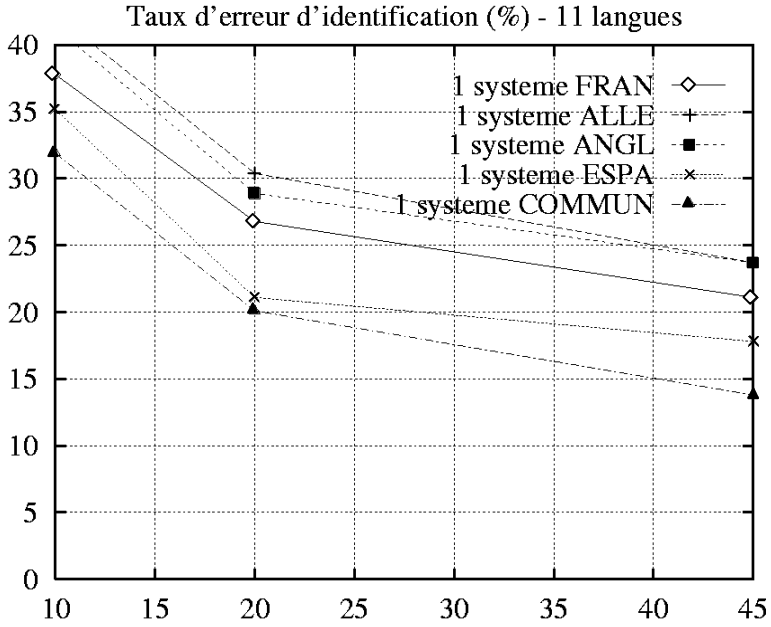

Figure 1 – Taux d'erreur d'identification en fonction de la durée avec un ensemble de test comprenant 11 langues. Les résultats sont obtenus par un système d'identification de la langue à approche phonotactique faisant appel à un système de reconnaissance phonétique dont les modèles acoustiques sont soit dépendants d'une seule langue (FRAN, ALLE, ANGL ou ESPA) soit multilingues (COMMUN).

Nous donnons dans la figure 1 un aperçu des résultats en identification de la langue (11 langues du corpus OGI-TS) sur des énoncés de parole spontanée.

Alors que pour une durée de 45 secondes le meilleur système d'identification utilisant des modèles acoustiques dépendants de la langue (espagnol) a un taux d'erreur autour de 18%, l'utilisation d'un ensemble de modèles acoustiques communs permet de descendre le taux d'erreur autour de 14%. Les résultats montrent donc l'intérêt d'un ensemble de HMMs de phones multilingues, au moins dans cette configuration expérimentale où la quantité de données pour l'apprentissage est limitée. En effet dans le corpus OGI on ne dispose que d'environ 2 heures par langue pour l'apprentissage des modèles phonotactiques (excepté pour l'anglais où plus de 4 heures sont disponibles).

5. Conclusion

Nous avons appliqué une approche de classement automatique aux phonèmes de six langues (français, anglais, allemand, espagnol, italien et portugais). Cette méthode est fondée sur un algorithme de classement hiérarchique, qui permet de suivre facilement l'évolution d'un groupement agglomératif. A partir d'une initialisation avec 235 phonèmes multilingues (donc 235 classes) nous avons itéré le classement tout en analysant la pertinence linguistique des groupes obtenus à chaque étape. Pour une analyse plus détaillée, nous avons retenu le nombre de 90 classes, afin de ne pas perdre trop d'information phonématique. Ces classes correspondent en majeure partie à une réalité linguistique.

Nous avons aussi analysé les résultats du classement dans les niveaux les plus bas de l'agglomération hiérarchique – où l'on retrouve la dichotomie voyelle-consonne. Ceci est un autre résultat qui montre l'intérêt de cette méthode.

Nous devons aussi noter la capacité de cette approche à traiter des données de différentes origines. De l'analyse des résultats obtenus à chaque itération, nous pouvons observer qu'il n'y a pas eu un classement des phonèmes par corpus, ni un bon classement pour les phonèmes de certaines langues et mauvais pour les phonèmes des autres. Ceci semble démontrer la robustesse de la mesure de similitude, face aux variations acoustiques des conditions d'enregistrement, et au contenu linguistique des différents corpus d'apprentissage.

Cette classification a été utilisé dans le cadre de l'identification automatique de 11 langues (corpus OGI-TS, français, anglais, allemand, espagnol, japonais, coréen, mandarin, tamil, farsi, vietnamien, hindi) pour l'estimation de modèles acoustiques multilingues. Les résultats obtenus montrent l'intérêt d'un tel ensemble de modèles acoustiques multilingues pour l'identification de la langue avec une approche phonotactique.

Le classement itératif de phonèmes peut également servir à quantifier la proximité entre les phonèmes et les allophones d'une même langue afin de définir un ensemble économique de phonèmes. Regrouper les phonèmes de plusieurs langues trouve enfin des applications en synthèse vocale multilingue, et en phonétique descriptive, didactique ou corrective.

6. Références

[Adda-Decker 96] M. Adda-Decker, G. Adda, L. Lamel, J.L. Gauvain, "Developments in large Vocabulary Continuous Speech Recognition of German», *in Proc. of IEEE-ICASSP*, Atlanta 1996.

[Andersen 97] O. Andersen, P. Dalsgaard, «Language-identification based on Cross-Language Acoustic models and Optimised Information Combination», *in Proc. of European Conference on Speech Technology, EuroSpeech*, pp. 67-70, Rhodes, septembre 1997.

[Berkling 94] K.M. Berkling, E. Barnard, «Language Identification of six Languages Based on a Common Set of Broad Phonemes», *in Proc. of International Conference on Speech and Language Processing*, Yokohama, Japon 1994.

[Berkling 96] K.M. Berkling, *Automatic Language Identification with Sequences of Language-Independent Phoneme Clusters*, thèse PhD, OGI, 1996.

[Corredor-Ardoy 97] C. Corredor-Ardoy, J.-L. Gauvain, M. Adda-Decker, L. Lamel, «Language Identification With Language-Independent Acoustic Models», *in Proc. of European Conference on Speech Technology, EuroSpeech*, 3, pp. 1423-1426, Rhodes, septembre 1997.

[Corredor-Ardoy 98] C. Corredor-Ardoy, P. Boula de Mareüil, M. Adda-Decker, L. Lamel, J.L. Gauvain, «Classement automatique de phonèmes dans un cadre multilingue», *actes XXIIièmes Journées d'Etudes sur la Parole*, pages 75-78, Martigny, Suisse, juin 1998.

[Delattre 65] P. Delattre, *Comparing the phonetic features of English, German, Spanish and French*, Julius Gross Verlag, Heidelberg, 1965.

[Duda 73] R.O. Duda, P.E. Hart, *Pattern Classification and Scene Analysis*, Wiley-Interscience, 1973.

[Gauvain 94] J.L. Gauvain, L. Lamel, G. Adda, M. Adda-Decker, «Speaker-independent Continuous Speech Dictation», *Speech Communication* 15, pp. 21-37, septembre 1994.

[Gibbon 97] D. Gibbon, R. Moore, R. Winski (eds.), *Handbook of Standards and Resources for Spoken Language Systems*, Mouton de Gruyter, Berlin, 1997.

[Haudricourt 67] A.G. Haudricourt, J.M.C. Thomas, *La notation des langues. Phonétique et phonologie*, Imprimerie de l'Institut Géographique National, Paris, 1967.

[Hazen 94] T.J. Hazen, V.W. Zue, «Recent Improvements in an Approach to Segment-based Automatic Language Identification», *in Proc. of International Conference on Speech and Language Processing*, Yokohama, Japon 1994.

[Juan 85] B.H. Juang, L.R. Rabiner, «A Probabilistic Measure for Hidden Markov Models», *AT&T Technical Journal* 64(2), 1985.

[Köhler 96] J. Köhler, «Multi-lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds», *in Proc. of International Conference on Speech and Language Processing*, Philadelphie, 1996.

[Lamel 94] L.F. Lamel, J.L. Gauvain, «Language Identification Using Phone-based Acoustic Likelihoods», *in Proc. of IEEE-ICASSP*, Adelaide 1994.

[Mak 96] B. Mak, E. Barnard, «Phone Clustering using the Bhattacharyya Distance», *in Proc. of International Conference on Speech and Language Processing*, Philadelphie, 1996.

[Matrouf 2000] Driss Matrouf, Martine Adda-Decker, Jean-Luc Gauvain, Lori Lamel, «Identification automatique de la langue par téléphone», *ailleurs dans ces actes*.

[Muthusamy 92] Y.K. Muthusamy, R.A. Cole, B.T. Oshika, «The OGI multi-language telephone speech corpus», *in Proc. of International Conference on Speech and Language Processing*, 2, pp. 895-898 Banff, octobre 1992.

[Proakis 89] J.G. Proakis, *Digital Communications*, McGraw-Hill Series in Electrical Engineering, 1989.

[Pullum 96] G.K. Pullum & W.A. Ladusaw, *Phonetic Symbol Guide*, The University of Chicago Press, Chicago and London, 1996.

[Quilis 92] A. Quilis et J.A. Fernández, 1992, «Curso de fonética y fonología españolas», *Consejo Superior de Investigaciones Científicas*, Madrid, 1992.

[Vallée 96] N. Vallée, L.J. Boë, C. Abry, J.L. Schwartz, A. Berrah, «La matérialité des structures sonores du langage», *actes XXIièmes Journées d'Etudes sur la Parole*, Avignon 1996.

[Yan 96] Y. Yan, E. Barnard, R.A. Cole, «Development of an approach to automatic language identification based on phone recognition», *Computer Speech and Language*, 10(1), pp. 37-54, janvier 1996.

[Young 93] S.J. Young, P.C. Woodland, «The use of state tying in continuous speech recognition», *in Proc. of European Conference on Speech Technology, EuroSpeech*, pp. 2203-2206, Berlin 1993.

3ème Partie

Expertise phonologique et identification

Use Of 'Rare' Segments For Language Identification

Jean-Marie HOMBERT & Ian MADDIESON

Phonology Laboratory, University of California, Berkeley
Dynamique Du Langage, Université Lumière Lyon 2

hombert@univ-lyon2.fr – ianm@socrates.berkeley.edu

Abstract

Knowledge of the distribution of rare segments across the languages of the world might be used in identifying languages within an open set. Segments which are both discriminatory (i.e. rare) and robust (i.e. easy to identify) are the best targets for efficient language identification. Considering several properties at the same time allows to use more common segments and/or features in a still very discriminatory way.

Résumé

L'étude de la distribution de segments phonologiques rares dans les langues du monde peut se révéler utile dans une tâche d'identification de langues en ensemble ouvert. Les segments à la fois discriminants (c'est-à-dire rares) et robustes (c'est-à-dire faciles à détecter et à identifier) sont évidemment les plus pertinents dans cette tâche. Le fait de prendre en compte de manière conjointe plusieurs propriétés simultanément permet cependant d'obtenir des informations discriminantes à partir de segments et/ou de traits moins rares.

1. Introduction

In recent years automatic language identification (ALI) has been a rapidly growing field of research. However, most of the results are based on a limited number of languages, at most 20, and often much less. In this paper we propose to show how the knowledge accumulated in traditional phonetic and phonological studies across the whole range of spoken languages might be applied to the task of improving our abilities to identify languages correctly. It is important to appreciate that what is being envisaged is the task of identifying languages within an open-ended set, rather than within a closed set, as is usual in current approaches to ALI [Muthusamy 94] and [Pellegrino, this volume].

Typological data on the world's languages has been accumulating at an accelerating rate in recent decades, so that we now have a good basic knowledge of the phonological patterns of almost all the extant languages. These descriptions are often based on limited familiarity with the language, but nevertheless allow a reasonable approximation of the segmental inventory to be obtained. This knowledge permits the construction of large typological surveys of phonological systems. These surveys permit a well-founded appreciation of which segments are common and which are rare, and how these rare segments are distributed within geographical areas and language families. The original purpose of these surveys was to highlight universal sound patterns [Maddieson 84] but they also necessarily show which segment types are most

restricted in their distribution. It is this aspect that can be exploited to extend the possibilities of identifying languages, especially in the context of an open-ended set.

The two databases used in these preliminary studies are those tabulated by Ruhlen [Ruhlen 75] and an extended version of the UPSID database originally described in Maddieson [Maddieson 84], which cover respectively 693 and 451 languages. The criteria used to select languages and interpret their phonologies in these databases differ, but their results are highly convergent with respect to the questions considered in this paper. It should be emphasized that these databases are based on phonological analyses only, abstracted from the phonetic richness of the acoustic signal. Consequently a real implementation of the method we are proposing will require a prior stage of automatic or semi-automatic transformation from the phonetic facts to the phonological level. Some approaches to this stage have been addressed [Ohala 2000]. However, even when segments are described by very general phonological categories, these terms do imply some typical acoustic properties that can be expected to be present in the signal. Such properties must be categorized along two dimensions — their discriminatory power and their robustness of identification.

2. Discrimination and Robustness

In a previous paper [Hombert 98] we emphasized the importance of separating these two dimensions. Discriminatory power refers to how useful a segment is in narrowing the choice of possible languages; the rarest segments will have the greatest discriminatory power. Robustness of identification refers to how easily a segment's presence can be detected; the more salient and less variable a segment's acoustic pattern is, the easier it is to detect. For example, the presence of /s/-like sibilant fricatives is easy to detect automatically, but since the vast majority of languages have such a sound, this property has little value in discriminating between languages. By contrast, very few languages have voiceless non-sibilant dental fricatives (θ), so this is a strongly discriminatory property, but the acoustic signal for this segment is weak and easily confounded with other weak fricatives, such as /f/. Hence, this property also has little value, as it is weakly detectable. The best targets are thus those properties which are both strongly discriminatory and strongly detectable, as indicated in the grid below. The most obvious cases would include clicks, labial-velar stops, and other rare and salient segments. Further suggestions are given in [Hombert 98]. Other segments will fall into an intermediate category along both dimensions.

	weakly detectable	strongly detectable
weakly discriminatory	e.g. /f/	e.g. /s/
moderately discriminatory	e.g. /h/	e.g. /ʃ/
strongly discriminatory	e.g. /θ/	**e.g. clicks**

By definition, the rarest segments occur in few languages. In order to be able to identify more than these few languages, it is necessary to make use of less powerfully discriminatory features. Less discriminatory features become more powerful when used jointly. For example,

consider 3 distinct features which are relatively common — let's say they are each found in 20% of the world's languages. If these features are randomly distributed, then less than 1% would possess all three. In the following sections we will present a detailed illustration of the discriminatory power of features taken in combination in this way.

3. Distribution of Features

The distribution of certain selected features of the consonant systems in 693 languages as given by Ruhlen [Ruhlen 75] is shown in Table 1. Similar data for selected properties of vowel systems is given in Table 2. The family grouping are those provided by Ruhlen. Two of the ten consonantal traits illustrated (a voicing contrast in stops and the presence of glottals) are very widespread, one is truly rare (clicks), but seven are found in between 6% and 19% of the languages. Of the vocalic characteristics in Table 2, one — the presence of a vowel length contrast — is found in almost half the languages but the others are found in less than a quarter. Any value between about 25% and 5% might be considered to indicate a moderately discriminatory property.

The particular features shown in Tables 1 and 2 are features which we think are likely to be usable in practical recognition tasks, but it is not critical that this be so. We are more concerned with demonstrating the concept of how such distributions can be of value in principle. A further practical issue concerns the length of sample that would be required to detect the presence of given feature in the phonological system. Given that it is desirable that samples used in automatic language identification be as short as possible, it is obvious that only positive detection of features can be taken into consideration. The absence of a given feature in a sample does not necessarily imply its absence from the phonology of the language.

	Voice Contrast in Stops	Aspiration Contrast in Stops	Ejectives	Implosives	Prenasalized Stops	Length Contrast in Cons.	Clicks	Labial-Velars	Retroflex Cons.	Glottals
Afro-Asiatic	29 /29	0 /29	6 /29	12 /29	4 /29	16 /29	0 /29	0 /29	3 /29	24 /29
Niger-Kordofanian	50 /51	4 /51	1 /51	15 /51	13 /51	4 /51	1 /51	36 /51	4 /51	28 /51
Nilo-Saharan	24 /25	0 /25	2 /25	11 /25	11 /25	9 /25	0 /25	4 /25	8 /25	17 /25
Khoisan	3 /4	2 /4	2 /4	0 /4	0 /4	0 /4	4 /4	0 /4	1 /4	4 /4
Indo-European	71 /73	13 /73	1 /73	1 /73	1 /73	8 /73	0 /73	1 /73	19 /73	41 /73
Caucasian	37 /37	3 /37	37 /37	1 /37	0 /37	17 /37	0 /37	0 /37	0 /37	37 /37
Uralic	15 /23	0 /23	0 /23	0 /23	0 /23	12 /23	0 /23	0 /23	2 /23	12 /23
Altaic	36 /39	3 /39	1 /39	0 /39	0 /39	5 /39	0 /39	0 /39	0 /39	21 /39

Table 1 Proportion of languages in different families with the consonantal traits identified in the column headings.

	Voice Contrast in Stops		Aspiration Contrast in Stops		Ejectives		Implosives		Prenasalized Stops		Length Contrast in Cons.		Clicks		Labial-Velars		Retroflex Cons.		Glottals	
Paleo-Siberian	3	/8	1	/8	1	/8	0	/8	0	/8	5	/8	0	/8	0	/8	0	/8	7	/8
Dravidian	10	/10	1	/10	0	/10	0	/10	0	/10	3	/10	0	/10	0	/10	9	/10	5	/10
Sino-Tibetan	12	/18	15	/18	0	/18	0	/18	1	/18	0	/18	0	/18	0	/18	3	/18	16	/18
Austro-Asiatic	16	/17	10	/17	0	/17	4	/17	1	/17	0	/17	0	/17	0	/17	8	/17	16	/17
Indo-Pacific	37	/50	2	/50	2	/50	1	/50	18	/50	0	/50	0	/50	1	/50	3	/50	26	/50
Australian	2	/24	1	/24	0	/24	0	/24	1	/24	1	/24	0	/24	0	/24	19	/24	2	/24
Austro-Tai	50	/67	8	/67	1	/67	4	/67	12	/67	6	/67	0	/67	0	/67	10	/67	57	/67
Eskimo-Aleut	1	/5	0	/5	1	/5	0	/5	0	/5	2	/5	0	/5	0	/5	0	/5	4	/5
Na-Dene	7	/12	7	/12	12	/12	0	/12	2	/12	0	/12	0	/12	0	/12	1	/12	12	/12
Macro-Algonguian	3	/13	1	/13	0	/13	0	/13	0	/13	3	/13	0	/13	0	/13	2	/13	13	/13
Salish	3	/10	0	/10	10	/10	0	/10	0	/10	0	/10	0	/10	0	/10	0	/10	10	/10
Wakashan	0	/2	1	/2	2	/2	0	/2	0	/2	0	/2	0	/2	0	/2	0	/2	2	/2
Macro-Siouan	4	/12	1	/12	3	/12	0	/12	0	/12	0	/12	0	/12	0	/12	0	/12	12	/12
Penutian	26	/43	2	/43	31	/43	14	/43	1	/43	3	/43	0	/43	0	/43	9	/43	41	/43
Hokan	5	/19	3	/19	7	/19	0	/19	0	/19	0	/19	0	/19	0	/19	4	/19	19	/19
Aztec-Tanoan	7	/15	2	/15	2	/15	0	/15	0	/15	2	/15	0	/15	0	/15	4	/15	15	/15
Oto-Manguean	11	/14	1	/14	1	/14	1	/14	4	/14	0	/14	0	/14	0	/14	3	/14	14	/14
Macro-Chibchan	7	/10	3	/10	2	/10	1	/10	0	/10	0	/10	0	/10	0	/10	2	/10	10	/10
Ge-Pano-Carib	13	/24	1	/24	1	/24	1	/24	0	/24	0	/24	0	/24	0	/24	8	/24	23	/24
Andean-Equatorial	19	/39	8	/39	3	/39	2	/39	1	/39	1	/39	0	/39	0	/39	9	/39	35	/39
TOTAL/693	501		93		129		68		70		97		5		42		131		523	
Percent	72.3		13.4		18.6		9.8		10.1		14.0		0.7		6.1		18.9		75.5	

Table 1 (continued) Proportion of languages in different families with the consonantal traits identified in the column headings.

In addition to the overall frequency of these features, the tables show how some of them have quite marked variations in their frequency distribution in the different language families listed. These variations also correlate strongly with a geographical pattern of distribution (for the geographical location of language families see, for example, www.sil.org/ethnologue).

	Nasalized V's	Front Rounded V's	Back Unrounded V's	Long V's
Afro-Asiatic	0/29	0/29	0/29	17/29
Niger-Kordofanian	27/51	3/51	1/51	24/51
Nilo-Saharan	1/25	0/25	1/25	12/25
Khoisan	4/4	0/4	0/4	2/4
Indo-European	19/73	15/73	0/73	33/73
Caucasian	14/37	12/37	1/37	15/37
Uralic	0/23	13/23	5/23	13/23
Altaic	1/39	25/39	21/39	23/39
Paleo-siberian	0/8	0/8	0/8	4/8
Dravidian	2/10	0/10	0/10	10/10
Sino-Tibetan	3/18	7/18	5/18	5/18
Austro-Asiatic	6/17	0/17	2/17	7/17
Indo-Pacific	2/50	1/50	0/50	14/50
Australian	0/24	0/24	0/24	9/24
Austro-Tai	2/67	3/67	5/67	38/67
Eskimo-Aleut	0/5	0/5	0/5	3/5
Na-Dene	8/12	0/12	0/12	9/12
Macro-Algonguian	2/13	0/13	1/13	11/13
Salish	0/10	0/10	0/10	3/10
Wakashan	0/2	0/2	0/2	1/2
Macro-Siouan	8/12	0/12	0/12	8/12
Penutian	0/43	0/43	3/43	26/43
Hokan	1/19	0/19	0/19	15/19
Aztec-Tanoan	3/15	1/15	3/15	11/15
Oto-Manguean	13/14	0/14	3/14	3/14
Macro-Chibchan	6/10	0/10	0/10	0/10
Ge-Pano-Carib	9/24	0/24	10/24	6/24
Andean-Equatorial	23/39	0/39	5/39	13/39
TOTAL/693	154	80	66	335
Percent	22.2	11.5	9.5	48.3

Table 2. Proportion of languages in different families with the vocalic traits identified in the column headings.

The labial-velar stops /k͡p, g͡b/ are found almost exclusively in two African families of languages, Niger-Congo and Nilo-Saharan. Implosives are present especially in all three large language families of Africa — Niger-Congo, Afro-Asiatic and Nilo-Saharan — and in the Penutian family of North America. Front rounded vowels occur almost exclusively in five language families mainly present on the Eurasian landmass, Indo-European, Caucasian, Uralic, Altaic, and Sino-Tibetan. Such patterns of distribution mean that, given a knowledge of its segmental features it is often possible to focus in on a likely area in which a language is spoken, or to say which family it belongs to even if it is not possible to identify the specific language.

4. Joint Discrimination

Three of the features in Tables 1 and 2 were utilized in a test of their joint discriminatory power. These three features are: the occurrence of nasalized vowels, of labial-velar stops, and of retroflex consonants. In the expanded UPSID database of 451 languages, 22.6% (102) have nasalized vowels in their inventories. There are marked regional disparities; none (0%) of the 72 languages in the major families of Oceania — Australian and Papuan — have nasalized vowels, but 20 of 49 (or 40.8%) of the North American languages (not including Eskimo-Eyak) have this feature.

Labial-velar stops are found in a substantially smaller number of languages than nasalized vowels, 41 or 9.1% of the languages in the database. If we search for those languages with both these traits, there are only only 19 (4.2%). All of these are African. Just these two features eliminate languages from other areas, such as North America or Europe. When a third feature, the occurrence of retroflex consonants is added, only three languages are retained from the original set of 451, which is less than 0.7% of the original search area. To select between the remaining three languages, a large number of traits, some of which are in themselves not at all rare, can be used. For example, Lelemi has a simple velar nasal /ŋ/ which is not found in Dan or Sango.

Another way to take advantage from the phonological databases is to reduce the list of candidates, if the sample to identify is long enough to be phonetically representative of the language. Figure 1 displays histograms corresponding with the proportion of languages in each family that different kinds of vowels. For example, this chart illustrates that a sample with both nasal vowels and front rounded vowels may only be language from one of the following families: Aztec, Niger-Kordofanian, Indo-European, Caucasian, Altaic, Sino-Tibetan, Austro-Asiatic or Indo-Pacific.

In the same way, the inventory of consonantal segments may lead to very strong cues to eliminate candidate languages (see. Figure 2). It illustrates the previous considerations about the strong discriminative power of some types of segments, as clicks or labial-velar consonants, that are present only in Africa.

5. Conclusion

Existing databases of phonological systems can be used to provide a geographical distribution of segments found in the world languages. Segments which are both rare and easy to identify are extremely valuable in an automatic language identification task. But it is also important to point out that even less restricted (found in 5 to 25% of the sample can be very discriminatory when used jointly (e.g. nasalized consonants, labial velar stops and retroflex consonants are found in less than 0.7% of the samples).

6. Bibliography

[Hombert 98] Hombert, J.-M. & Maddieson, I. (1998) Automatic language identification : a linguistic point of view. UCLA Working Papers in Phonetics 97: 119-124.

[Maddieson 84] Maddieson, I. (1984) Patterns of Sounds. Cambridge: CUP.

[Muthusamy 94] Muthusamy, Y. K., E. Barnard & R. A. Cole. (1994) Automatic language

recognition: A review/tutorial. *in IEEE Signal Processing Magazine* 10/94: 33-41.

[Ohala 2000] Ohala, J. J. & Marsico, E. (1999) Differentiating phonetic from phonological events in speech. In [].

[Pellegrino 2000] Pellegrino, F (1999) Ed. Actes de la Première Journée d'Etude sur l'Identification Automatique des Langues : de la caractérisation à l'identification des Langues".

[Ruhlen 75] Ruhlen, M. (1975) Guide to the World's Languages. Stanford University.

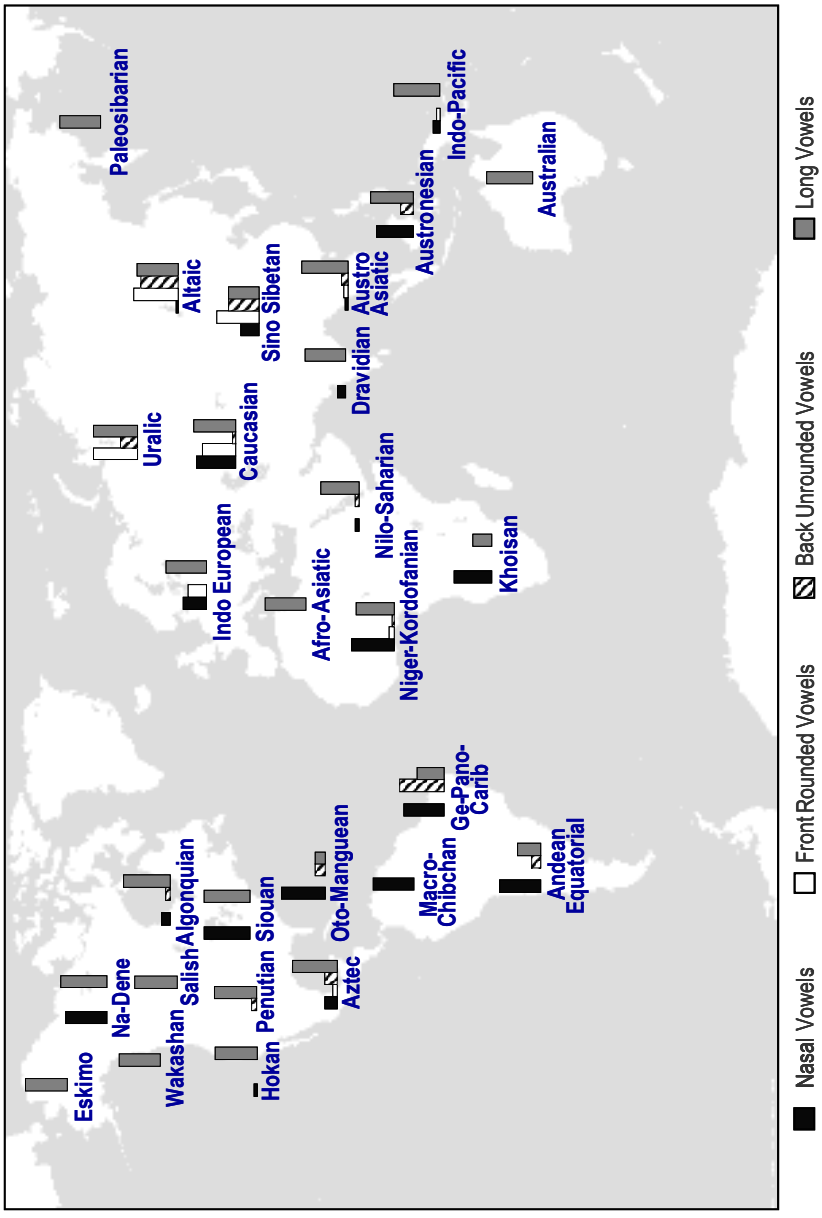

Figure 1 – Proportions of languages in each family exhibiting some vocalic features (from UPSID).

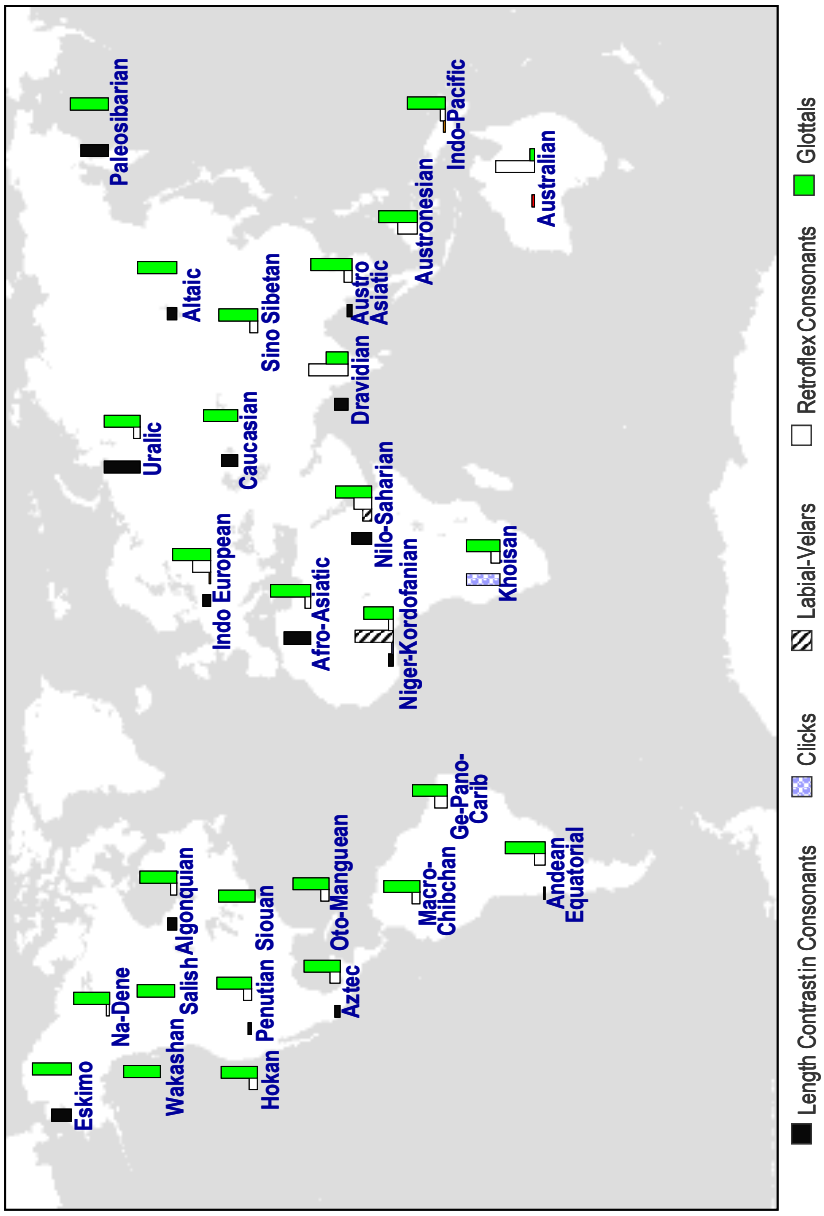

Figure 2 – Proportions of languages in each family exhibiting some consonantic features (from UPSID).

Détermination d'Indices Acoustiques Robustes pour l'Identification Automatique des Parlers Arabes

Melissa Barkat

Laboratoire Dynamique Du Langage
14, avenue Berthelot, 69363 Lyon Cedex 07

Melissa.Barkat@ish-lyon.cnrs.fr

Abstract

This work aims at determining some reliable acoustic cues for the Automatic Identification of Arabic dialects by geographical zone. We used perceptual experiments to determine a set of efficient acoustic cues for the discrimination of Western vs. Eastern Arabic dialects. Among the entire set of discriminating criteria that emerged from perceptual studies, two were, *a priori*, relevant for our goal, since they occur exclusively in one area or the other. Specifically, these two criteria are vocalic distribution and the realization of distinctive vowel length contrast in spontaneous speech. Our acoustic studies revealed that Western dialects develop central vocalic positions whereas their Eastern counterparts tend to prefer peripheral ones. Besides, the ratios long vowel/short vowel are comparable for dialects in the same area. On the contrary, there are significant differences between Western and Eastern dialects, the latter attesting a higher long/short ratio. In order to evaluate the robustness of those two criteria Automatic Language Identification, we performed a set of experiments using a model based on an automatic vowel detection algorithm and statistical vowel system modeling. Results show that Western and Middle-Eastern dialects may be discriminated using spectral features combined with vowel duration cues since the rate of correct identification ranges from 70 % to 90 % depending on the number of cues retained.

Résumé

Ce travail a pour objectif la recherche d'indices acoustiques robustes en vue de l'identification automatique des parlers arabes par zones géographiques. Par le biais d'expériences perceptuelles, nous avons déterminé un faisceau d'indices discriminants permettant la distinction des parlers maghrébins vs. orientaux. Parmi l'ensemble des traits dégagés, deux sont apparus comme étant *a priori* pertinents pour notre propos du fait de leur réalisation exclusive sur l'une ou l'autre des deux aires dialectales concernées. Il s'agit de la distribution des segments vocaliques et de la réalisation de l'opposition de durée vocalique. Différentes analyses acoustiques ont permis de caractériser la distribution des voyelles dans l'espace acoustique et d'établir une opposition pertinente entre les parlers maghrébins privilégiant la génération de voyelles centrales et les parlers orientaux préférant les positions périphériques. Du point de vue quantitatif, notre étude révèle que l'opposition de durée est réalisée, en parole spontanée, dans des rapports comparables pour les parlers appartenant à une même zone géographique. En revanche, d'une zone dialectale à l'autre, les rapports mis en œuvre sont significativement différents, la tendance étant que les rapports voyelle longue/voyelle brève croissent d'Ouest en Est. La co-occurrence de ces deux critères de

discrimination sur chacune des deux aires dialectales étudiées les désigne comme potentiellement pertinents pour la discrimination automatique des parlers arabes par zone géographique. Nous avons ainsi mis en place des expériences d'identification automatique en utilisant un modèle de reconnaissance basé sur la détection automatique des voyelles et la modélisation statistique des systèmes vocaliques. Les résultats obtenus à l'issue de ces expériences valident la robustesse et la pertinence des deux critères de discrimination définis dans ce travail. En effet, nous montrons qu'il est possible de discriminer les parlers maghrébins des parlers orientaux sur la base des caractéristiques spectrales et quantitatives des segments vocaliques et nous obtenons entre 70 % et 90 % d'identification correcte en fonction du nombre de paramètres de modélisation retenus.

1. Introduction

La plupart des systèmes de reconnaissance automatique se basent sur les formes standardisées des langues. Or, un grand nombre de langues naturelles se déclinent sous des formes dialectales plus ou moins ressemblantes. Certains parlers arabes, par exemple, présentent si peu de caractéristiques communes qu'il est parfois difficile d'établir des rapprochements d'une forme linguistique à l'autre. Il serait donc intéressant, dans le cadre des recherches sur l'Identification Automatique des Langues, d'élargir le champ des investigations aux parlers non-standards afin d'optimiser les résultats dans ce domaine. Cette étude vise spécifiquement à déterminer la validité d'un indice acoustique « robuste » [Hombert 97] pour l'identification automatique des parlers arabes. Elle se base, d'une part, sur les résultats d'une expérience perceptuelle à partir de parole naturelle menée auprès de sujets originaires de différents pays du Maghreb et du Moyen-Orient. Cette première étape nous a permis de repérer de manière expérimentale les indices acoustiques utilisés par les arabophones eux-mêmes lors d'une tâche d'identification dialectale. Afin de vérifier la réalité physique des impressions auditives de nos sujets, nous avons procédé, dans un second temps, à l'analyse acoustique d'un des critères évoqués en étudiant plus particulièrement les variations qualitatives et quantitatives discriminantes des segments vocaliques. Les résultats obtenus à l'issue de cet examen nous ont permis de caractériser les parlers arabes en fonction des caractéristiques qualitatives et quantitatives des segments vocaliques et de retenir la distribution vocalique ainsi que les rapports de durée vocalique (i.e. voyelle longue / voyelle brève) comme des indices *a priori* pertinents pour l'identification automatique des parlers arabes. Enfin, afin de valider le pouvoir discriminant de ces critères, des tests de reconnaissance automatique ont été effectués sur de la parole spontanée en arabe dialectal, à l'aide d'un système différencié basé sur la modélisation acoustique des systèmes vocaliques [Pellegrino 98].

2. Identification dialectale et détermination expérimentale d'indices discriminants

2.1. Géographie dialectale du Monde Arabe

Chaque parler arabe présente des caractéristiques qui lui sont propres. Celles-ci peuvent être d'ordre syntaxique, lexical, segmental et/ou prosodique. Certains parlers ont innové à l'intérieur de chacun de ces domaines et ne présentent aujourd'hui que très peu de caractéristiques communes avec l'arabe standard moderne. En effet, bien que les parlers

contigus sur le plan géographique attestent de nombreux traits communs et que l'ensemble des parlers arabes présente un air général de ressemblance sensible aux sujets parlants, l'intercompréhension devient toute relative — voire inexistante — lorsque l'on prend en considération des points situés aux antipodes du domaine. Une vue d'ensemble du domaine linguistique arabe (figure 1) permet de distinguer deux zones principales à l'intérieur du continuum dialectal : à l'Ouest, la zone occidentale regroupant les parlers du Maghreb ; à l'Est, la zone orientale constituée des parlers Moyen-Orientaux [Cohen 70] [Fleish 75].

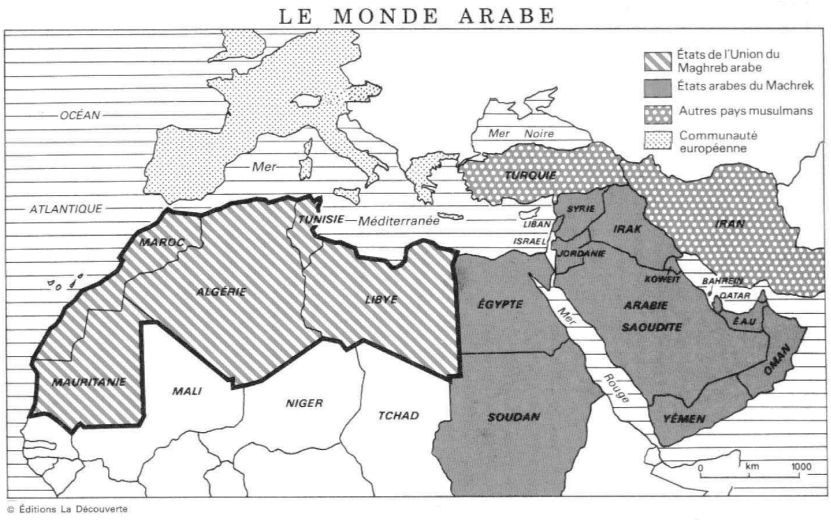

Figure 1 – Géographie dialectale du monde arabe

Il convient, par ailleurs, de souligner la présence d'une zone intermédiaire aux frontières linguistiques relativement vagues dont les parlers — définis comme transitoires — se caractérisent par des caractéristiques mixtes. Toutefois, les parlers arabes constituent de manière générale des formes linguistiques aisément identifiables par le système perceptuel des locuteurs/auditeurs arabophones.

2.2. Corpus et méthode

Les travaux relevant de l'identification dialectale arabe sont peu nombreux. En effet, seules deux études ont abordé le problème de la variabilité interdialectale arabe au niveau phonétique [Abu-Haidar, 91] [Rjaibi-Sabhi, 93]. L'idée maîtresse sous-tendant ces travaux étant que le système phonologique du dialecte maternel des sujets arabophones transparaît au niveau de leurs productions en arabe standard sous la forme de transferts linguistiques, lesquels fournissent une indication pertinente quant à l'origine dialectale du sujet parlant. Notre travail vient compléter les travaux précédemment cités. Il est basé non pas sur la production mais sur la *perception* de ces marqueurs linguistiques — étant entendu que de la même manière qu'un individu est capable de reconnaître une voix familière à partir d'un échantillon de parole très bref — les individus qui possèdent une connaissance moyenne de différentes langues et/ou dialectes, devraient être capables de l'identifier dans un délai relativement court (de l'ordre de quelques secondes) sur la base de certaines propriétés linguistiques discriminantes. D'un point de vue méthodologique, nous avons choisi de travailler directement sur les formes vernaculaires

de l'arabe afin de contrôler toute normalisation perceptuelle vers la variété standard liée à un effet de filtre laquelle pourrait conduire à la non-perception des traits dialectaux les plus typiques. L'hypothèse que nous avançons ici est que ces caractéristiques linguistiques — si tant est qu'elles soient perçues des sujets — pourraient constituer un faisceau d'indices de discrimination pertinents dans le cadre d'une tâche d'identification dialectale. Afin de repérer les indices discriminants utilisés par les auditeurs arabophones pour identifier l'origine dialectale de locuteurs s'exprimant en arabe, nous avons mis au point une expérience perceptuelle d'identification dialectale ayant pour objectif de déterminer de manière expérimentale les indices jugés comme discriminants par une population de locuteurs/auditeurs arabophones [Barkat 98]. Dans ce cadre, six variétés dialectales représentatives des deux pricipales zones géographiques du Monde Arabe (i.e. Maghreb vs Moyen-Orient) ont été sélectionnées (tableau 1) et une base de données acoustiques a été élaborée après enregistrement de 12 locuteurs natifs (6 maghrébins et 6 orientaux) arabophones s'exprimant de manière spontanée dans leur dialecte maternel sur la base d'un livre d'images sans texte[1].

Maghreb	**Moyen-Orient**
Maroc (MA)	Syrie (SY)
Algérie (AL)	Liban (LI)
Tunisie (TU)	Jordanie (JO)

Tableau 1 – Origines dialectales représentées à l'intérieur des stimuli et chez les sujets testés.

Les enregistrements — effectués en chambre insonorisée sur magnétophone DAT — ont été digitalisés à 22 kHz, 16 bits, monophonique sous Sound Forge©. Quatre-vingt-seize échantillons de parole (soit 2×2 énoncés[2] complets/locuteur × 24) d'une durée variable allant de sept à trente secondes[3] ont été extraits du corpus. Ils ont ensuite été réorganisés en ordre aléatoire et présentés comme matériel de stimulation lors d'un test d'identification dialectale,

[1] Mayer, M., 1980, *Frog, where are you ?*, Dial Books for Young Readers, New-York.

[2] Plusieurs études d'identification à partir d'échantillons de langues inconnues ont établi l'importance de la phase d'entraînement (Ohala & Gilbert, 1979 ; Maidment, 1983). Dans cette première expérience, nous sommes partis du postulat que les locuteurs arabophones disposent tous d'une connaissance relative des variations de production en fonction de l'origine géographique laquelle est favorisée par le contact entre ces populations dans la communauté étudiante qui nous a servi de "vivier à sujets". Ce degré de connaissance a été considéré comme suffisant pour ce travail. Par conséquent, nous n'avons pas retenu de période d'apprentissage dans ce protocole. En revanche, nous avons accordé du temps à l'explication de la troisième tâche (i.e. définition des indices permettant la discrimination). Celle-ci a été présentée en termes « communs » (i.e. « sons » typiques pour « indices phonétiques », « mélodie, musique de la langue » pour « indices prosodiques » etc). Par ailleurs, nous avons établi sur la base d'exemples *ad hoc* un code de transcription exploitable pour l'analyse. Pour ce qui est de la mise en valeur de la prononciation d'un segment spécifique ayant contribué à l'identification dialectale, il a été convenu que les sujets (sachant tous écrire en arabe) devaient transcrire l'item en arabe classique et cercler le phonème (i.e. « la lettre ») caractérisée par une prononciation particulière.

[3] Des recherches traitant de l'identification de l'identité sociale véhiculée par les usages dialectaux ont montré que 10 à 15 secondes sont suffisantes pour reconnaître l'origine sociale et/ou dialectale (Lee, 1971).

chaque stimulus étant précédé d'un numéro de passage. Les 96 échantillons de parole ont été présentés à 18 sujets arabophones « naïfs » (i.e. étudiants et étudiantes non spécialistes de linguistique et/ou de dialectologie), soient : 9 sujets maghrébins et 9 sujets moyen-orientaux originaires des six mêmes pays que ceux représentés dans les stimuli mais n'ayant jamais eu de contact avec les personnes dont les voix ont été utilisées pour le matériel expérimental, et ce, afin d'éviter toute identification dialectale par reconnaissance du locuteur. Les sujets disposaient de quinze secondes pour exécuter les trois tâches suivantes : (i) identifier l'origine dialectale du locuteur entendu en fonction de la zone géographique (i.e. Maghreb vs Moyen-Orient) ; (ii) identifier le pays dont pouvait être originaire le locuteur parmi les 6 choix proposés sur la grille de réponses (tableau 1) ; (iii) définir — dans la mesure du possible — les indices prosodiques, segmentaux et/ou lexicaux ayant permis l'identification. Les réponses apportées aux trois tâches précédentes devant être reportées sur une grille de réponses pré-formatée distribuée aux sujets en début d'expérience. Outre la détermination expérimentale d'indices discriminants *a priori* pertinents pour l'identification automatique des parlers arabes, nous avons voulu vérifier auprès de nos sujets les hypothèses conceptuelles suivantes : (i) l'identification par zone géographique principale est aisée pour l'ensemble des sujets. Elle constitue de ce fait une réalité linguistique perceptible par des sujets arabophones même non entraînés[4] ; (ii) les meilleurs scores d'identification sont obtenus pour la distinction du dialecte maternel par rapport aux autres parlers ; (iii) Les résultats les moins probants concernent les parlers proches à l'intérieur d'une même zone, ce qui impliquerait — pour ces parlers spécifiquement — une analyse plus fine des traits distinctifs, couvrant plusieurs niveaux de la langue, voire à plus long terme dans le cadre de l'IAL, l'intégration d'un modèle d'apprentissage particulier (basé, par exemple, sur la détection d'items lexicaux spécifiques).

2.3. Résultats des tests d'identification dialectale par zone

Les résultats de l'identification dialectale par zone révèlent que 97 % des stimuli maghrébins et 99 % des stimuli moyen-orientaux ont été identifiés correctement. Les erreurs de classification, soit respectivement 3 % et 1 %, résultant — pour le cas des stimuli maghrébins — du fait de l'emploi, pourtant impropre, par l'une des locutrices originaire de la zone occidentale, d'items morpho-syntaxiques appartenant à l'arabe classique. Les parlers moyen-orientaux présentant plus de caractéristiques communes avec l'arabe classique, la perception de certains indices morpho-syntaxiques a conduit certains sujets maghrébins à juger ces productions comme plutôt orientales. Au niveau lexical, l'emploi — pourtant syntaxiquement erroné— du numéral classique [iθnætæjn] « deux » au lieu de la forme invariable [zuːdʒ] typique des parlers occidentaux semble constituer pour la plupart des sujets maghrébins un premier élément de perturbation pour la catégorisation dialectale. Du point de vue morpho-syntaxique, la présence — inattendue en arabe dialectal — de la marque du « duel » en [-tæjn] sur l'adjectif numéral [iθnætæjːn] « deux » et sur le mot [dʒrana] « grenouille » (i.e. [dʒranæ-

[4] Les études citées ci-dessus études ont parfaitement établi le rôle de la période d'entraînement pour l'amélioration des performances d'identification linguistiques et/ou dialectales. Nous postulons ici que les sujets arabophones possèdent tous une connaissance relative des différentes variétés linguistiques arabes du fait de leur co-habitation sur le territoire Français et des émissions radiophoniques diffusées aujourd'hui dans la plupart des dialectes arabes.

tæjn] « deux grenouilles) ») a conduit au même type d'interprétation erronée. La forme duel est en effet remplacée dans la plupart des dialectes modernes par une marque de pluriel simple. Elle conduirait, dans notre exemple, soit à l'adjonction d'une marque de pluriel en [j] à l'intérieur de l'item lexical, soit [dʒra-jən] "des grenouilles", soit à des formes plurielles de type [ʒranæːt] (pluriel ordinaire dénombrable) et/ou [ʒraːn] « des grenouilles » (pluriel collectif, indénombrable). Le choix de l'item lexical [sˁɑqatˁo] « il est tombé » semble être connoté par nos sujets maghrébins comme [+ oriental]. Ceci peut s'expliquer par le fait qu'il possède un équivalent plus fréquent en arabe dialectal du Maghreb (i.e. [tˁɑħ] « il est tombé »). Son usage, relevé dans un stimulus maghrébin emprunt d'emphase narrative, a ainsi conduit certains sujets à classifier l'item maghrébin parmi le groupe oriental. Pour ce qui est des stimuli orientaux, l'erreur de classification par zone a été provoquée par la perception du phénomène d'*imala*[5] *finale* (i.e. antériorisation de la voyelle ouverte /a/ > [e] voire [i]). Celle-ci — bien que peu fréquente dans la plupart des parlers maghrébins[6] — est caractéristique de certains parlers tunisiens comme celui de Bizerte (ville côtière du Nord de la Tunisie dont est originaire l'un des locuteurs ayant participé à l'étude), et plus particulièrement en position finale des items monosyllabiques. L'*imala* est en effet définie comme un fait plutôt oriental [Fleish 75]. Sa perception dans un énoncé très court (moins de 2 secondes) et par un sujet arabophone « naïf » (i.e. n'ayant pas connaissance de cette particularité dialectale) a conduit à une catégorisation incorrecte en termes de zone géographique. On imagine néanmoins, qu'un stimulus plus long comportant *ipso facto* plus d'indices de discrimination aurait permis au sujet en question de s'appuyer sur d'autres critères avant de prendre sa décision.

Néanmoins, malgré la présence de ces quelques erreurs induites par la présence d'emprunts à l'arabe classique et le caractère bref de certains stimuli, les scores d'identification par zone restent très élevés. Ceci nous permet donc de confirmer l'idée selon laquelle la bi-partition dialectale du domaine linguistique arabophone en termes de zones géographiques (parlers maghrébins vs parlers orientaux) transparaît à travers des éléments perceptuels identifiables par l'ensemble des locuteurs arabophones originaires de l'une ou l'autre de ces deux régions, constituant ainsi une réalité linguistique bien établie au plan perceptuelle qu'il

---

[5] Le phénomène de 'l'imala' déjà décrit par les Grammairiens Arabes du 8ème siècle consiste en une antériorisation de la voyelle ouverte [a] en position *interne* (i.e. médiane) et/ou *finale*. Elle se manifeste sur le plan acoustique par une baisse de F1 et une montée de F2 [Benkirane 82] et permet de distinguer entre parlers à imala interne et/ou finale *"forte"* (i.e. [a] > [e] ; [ɛ] ou [i]) comme en Syrie, au Liban et dans certains villages côtiers de Tunisie, vs. parlers à imala interne et/ou finale *"moyenne"* (i.e. [a] > [æ]) vs. parlers connaissant une imala interne moyenne mais pas d'imala finale (i.e. (i.e. [-a#] = [-a#]) comme au Maroc ou en Algérie (voir Barkat & al., 1997). Selon [Kaye 97] : *"Imala (lit. inclination) refers to /a-/raising, often due to the umlauting influence of /i/. A classical word such as /ʔibaːd/ "slaves" could have a dialectal pronunciation [ʔibeːd] or [ʔibiːd]. Imala has produced the very distinctive high vowel pronunciation of /a/ in many Syro-Lebanese dialects, i.e. [beːb] or [biːb], equivalent to classical Arabic /baːb/ "door and [bæːb] as uttered, say, by a Cairene*". [Kaye 97].

[6] Dans les parlers du Maghreb, on atteste en général un degré d'*imala* intermédiaire (i.e. moyenne) correspondant au passage de /a/ à [-æ-] en position médiane uniquement et en l'absence de consonnes d'arrière (i.e. pharyngalisées en particulier).

serait intéressant d'analyser plus en détail du point de vue linguistique afin de constituer une liste exhaustive des traits dialectaux (lexicaux, phonético-phonologiques, prosodiques et/ ou rythmiques, etc…), propres à chacune de ces deux variétés dialectales.

2.4. Identification dialectale par pays

Les résultats obtenus par nos sujets pour la seconde tâche (i.e. identification par pays) confirment, dans la grande majorité des cas, les hypothèses avancées ci-dessus selon lesquelles les meilleurs taux d'identification devraient concerner (i) les dialectes proches des parlers maternels, et (ii) les parlers appartenant à la même zone géographique que ceux-ci. Il semble donc que les sujets perçoivent — à l'intérieur des stimuli les plus représentatifs d'un parler — certains indices acoustiques propres à chaque variété dialectale qui leur permettent d'effectuer une discrimination fine en termes de « pays ». Le tableau 2 présente de manière synthétique les scores d'identification par pays obtenus par l'ensemble des sujets testés sur la totalité des stimuli présentés. Il nous permet d'ores et déjà d'observer que tous les locuteurs/auditeurs arabophones — à l'exception des sujets syriens — attestent un haut taux de reconnaissance pour, au moins, le parler le plus proche de leur dialecte d'origine.

	Sujet MA	Sujet AL	Sujet TU	Sujet SY	Sujet LI	Sujet JO	Tx moyen
Stim MA	**94**	63	67	33	54	75	64
Stim AL	88	**92**	48	69	48	65	68
Stim TU	73	85	**96**	56	83	46	73
Stim SY	31	35	19	**90**	75	88	56
Stim LI	27	33	38	96	**94**	100	65
Stim JO	23	46	35	83	85	**100**	62

Tableau 2 – Matrice de confusion des scores d'identification par pays en fonction de l'origine dialectale des sujets (%)

• Identification perceptuelle « intra-zone »

Bien que les sujets syriens aient été plus performants pour la reconnaissance du libanais (i.e. dialecte proche mais non-maternel), le score moyen d'identification du parler d'origine (tous sujets, toutes variétés et toutes zones géographiques confondus) frôle les 94 %, et ne présente pas d'écart significatif d'une origine dialectale à l'autre comme l'a confirmée l'analyse de variance effectuée sur la base des résultats obtenus pour chacun des sujets lors de l'identification de son dialecte d'origine.

Le tableau 2 permet d'une part d'observer que la reconnaissance des parlers les plus proches du dialecte d'origine s'effectue sans encombre pour la quasi-totalité des sujets. A ce niveau de l'analyse, ce résultat autorise à nuancer l'approche — souvent tranchée dans les travaux de dialectologie traditionnelle — selon laquelle la définition d'entités linguistiques appelées *« arabe maghrébin »* vs. *« arabe oriental »* ne constitue pas une réalité aussi nettement contrastée. Il semble en effet que les sujets se soient appuyés sur des macro-critères discriminants efficaces pour d'une part, distinguer des parlers maghrébins par rapport à leurs pendants moyen-orientaux et, d'autre part, arriver à une classification plus fine en termes de dialecte national (i.e. parler jordanien, par exemple).

• Identification perceptuelle « inter-zones »

La figure 2 révèle une asymétrie importante des scores associés à la distinction des parlers inter-zones selon l'origine des sujets. Les taux d'identification correcte obtenus par les sujets maghrébins pour la discrimination des parlers moyen-orientaux (32%) présentent une différence significative en comparaison des résultats obtenus pour l'identification des parlers maghrébins plus ou moins proches de leur dialecte d'origine (78%) [$t_{(8)} = 149$, $p \leq .0001$]. De la même manière, les taux d'identification des sujets orientaux attestent des valeurs nettement différenciées selon qu'il s'agisse d'un stimuli maghrébin (59%) ou moyen-oriental (90%) [$t_{(8)} = 224$, $p \leq .0001$]. Par ailleurs, on observe que pour la même tâche (i.e. discrimination par pays des stimuli n'appartenant pas à la zone dialectale d'origine), les deux populations font preuve d'un comportement différent : alors que les sujets moyen-orientaux parviennent à identifier correctement 59% des stimuli maghrébins, les sujets maghrébins présentent quant à eux un taux de discrimination des stimuli orientaux de l'ordre de 32% seulement. Cette remarquable dissymétrie dans les proportions de reconnaissance des parlers inter-zones se révèle statistiquement significative [$t_{(8)} = 4$, $p = .0023$].

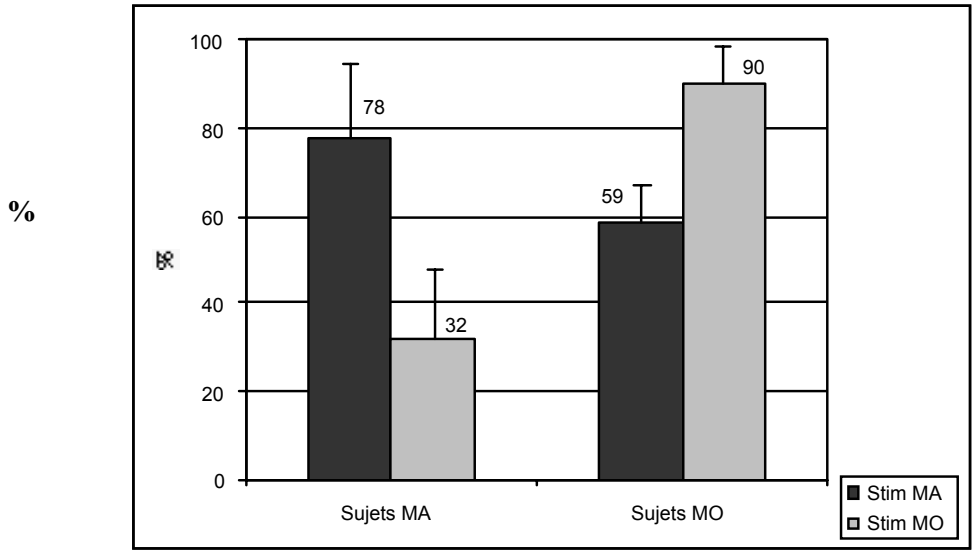

Figure 2 – Scores d'identification intra et inter zones observés chez les 2 populations de sujets (MA = maghrébins / MO = orientaux).

Notons enfin que les scores obtenus par les deux populations pour l'identification des 48 stimuli maghrébins — soient 78% pour les sujets maghrébins et 59% pour les sujets orientaux — présentent une différence plus significative encore [$t_{(8)} = 184$, $p \leq .0001$]. On retrouve la même opposition pour la discrimination des stimuli moyen-orientaux, puisque les sujets maghrébins présentent un taux d'identification correcte de 32% largement inférieur à celui obtenu par les sujets originaires de l'Est (90%). Cette différence s'avère elle aussi largement significative [$t_{(8)} = 170$, $p = \leq .0001$].

Ces résultats corroborent deux des trois hypothèses de départ, à savoir : (i) que l'on reconnaît significativement mieux les parlers proches du dialecte maternel, (ii) que la discrimination des parlers appartenant à la même zone géographique pose — en règle générale

— moins de problèmes que l'identification des parlers représentatifs de la zone dialectale opposée. De plus, ils montrent, que certains sujets — en l'occurrence les moyen-orientaux — semblent plus performants que les maghrébins pour les tâches d'identification dialectales en termes de pays. Cette différence pouvant être attribuée au fait que la population maghrébine est — en France — supérieure à celle des moyen-orientaux, ces derniers sont plus exposés aux différentes variétés dialectales arabes occidentales et donc plus sensibles aux particularismes dialectaux des parlers maghrébins.

2.5. Détermination expérimentale d'indices discriminants pour les parlers arabes

Parmi l'ensemble des critères discriminants dégagés à l'issue de l'expérience perceptuelle, la distribution des voyelles dans l'espace acoustique ainsi que la réalisation de l'opposition de durée vocalique (i.e. rapports voyelles longues / voyelles brèves) peuvent a priori être considérés comme pertinentes pour la reconnaissance automatique des parlers arabes par zones géographiques principales. En effet, les sujets ont majoritairement constaté qu'il existait bien des différences inter dialectales typiques de l'une et/ou l'autre des deux zones dialectales au niveau des réalisations vocaliques tant au niveau qualitatif qu'au niveau qualitatif. Notre objectif a donc été de caractériser au plan acoustique les différences pressenties au niveau perceptuel.

3. La distribution des segments vocaliques en arabe maghrébin vs oriental

Le système vocalique de l'arabe standard est généralement décrit comme étant composé de 3 phonèmes brefs [i ; u ; a] auxquels s'opposent 3 voyelles longues de mêmes timbres. Toutefois, des constatations concernant l'introduction secondaire de certaines nouvelles voyelles définies comme centrales et/ou d'aperture moyenne ont été faites pour la majorité des dialectes [Ghazali 79].

3.1. Analyse acoustique

Nous avons procédé à l'analyse acoustique de la dispersion vocalique dans les parlers présentés dans le tableau 1. Pour chacune des deux zones dialectales concernées, nous nous intéressons plus particulièrement à la caractérisation qualitative des segments vocaliques tels qu'ils sont effectivement réalisés en parole spontanée (i.e. non-lue). Ceci nous permettra de caractériser au niveau acoustique (i) la distribution des segments vocaliques dans les parlers maghrébins et orientaux et (ii) la réalisation de l'opposition de durée. Les analyses ont été effectuées à partir d'enregistrements effectués en chambre sourde sur la base d'un corpus unique consistant en la traduction spontanée en arabe dialectal d'un texte court (i.e. « La Bise et le Soleil »). Chacune des 12 phrases qui le compose ont été transmises oralement aux locuteurs par le biais d'une langue commune (i.e. français ou anglais). Pour chaque phrase, nous avons acquis trois répétitions par locuteur. Pour les analyses nous avons utilisé à chaque fois la troisième occurrence (correspondant de manière générale à la meilleure répétition). Nous avons alors procédé à l'étiquetage manuel par timbre des 1500 segments vocaliques (monophtongues) présents dans les échantillons de parole. Une analyse LPC a ensuite été effectuée au centre de chacune de ces voyelles, de manière à obtenir les valeurs formantiques associées à leur état stable. Ainsi, bien que conscients que le nombre de syllabes dans le mot, la nature de ces

syllabes, la position de l'accent, l'articulation spécifique des segments vocaliques, la nature du contexte consonantique environnant et le débit d'élocution jouent un rôle important sur la qualité et/ou la quantité vocaliques, nous n'avons pas tenu compte de ces paramètres variationnels afin d'établir une représentation statistique globale de la distribution vocalique propre à chacun des parlers et de dégager les informations prises en compte implicitement par le modèle de reconnaissance automatique utilisé pour les expériences d'identification automatique.

3.2. Résultats

• Aspect qualitatif

Les tableaux 3 et 4 récapitulent les valeurs formantiques moyennes obtenues pour chacun des timbres vocaliques brefs rencontrés dans nos données en arabe maghrébin et oriental.

	F1	écart-type	F2	écart-type
a	647	44	1392	89
i	429	56	1840	104
æ	560	55	1682	115
u	448	66	1208	109
ə	471	42	1516	74

Tableau 3 – valeurs formantiques moyennes des segments vocaliques brefs de l'arabe maghrébin (en Hz)

	F1	écart-type	F2	écart-type
a	668	48	1355	128
i	360	48	2140	125
æ	593	62	1768	102
u	362	53	994	125
ə	512	48	1522	100
o	491	22	1082	47
e	498	46	2084	80

Tableau 4 – valeurs formantiques moyennes des segments vocaliques brefs de l'arabe oriental (en Hz)

La comparaison des réalisations de [a] montre que la variante orientale est significativement plus ouverte et plus postérieure que la réalisation maghrébine ($p = .01$ sur F1 et $p = .05$ sur F2). La voyelle d'aperture moyenne [æ] présente dans ses réalisations orientales un degré d'aperture significativement supérieur à celui caractéristique des dialectes maghrébins ($p = .0002$), de plus, tout comme pour la voyelle ouverte, [æ] est au Maghreb plus centralisé (i.e. valeur de F2 inférieure par rapport à la réalisation orientale ($p = .0001$). Pour ce qui concerne les voyelles fermées d'avant et d'arrière, soient [i] et [u], on constate à nouveau qu'elles sont (i) plus fermées dans les dialectes moyen-orientaux qu'au Maghreb ($p = .0002$ pour [i] et $p = .0001$ pour [u], et (ii) plus centralisées en arabe maghrébin, puisque [i] est significativement plus antérieur dans ses réalisations orientales ($p = .0001$) et [u]

significativement plus postérieur (p = .0001). Les résultats statistiques que nous venons de présenter permettent donc de définir la distribution vocalique des segments brefs des parlers maghrébins comme étant significativement plus centrale que celle des parlers orientaux. Cette tendance à la « périphérisation » typique des parlers orientaux, s'observe également pour la voyelle centrale elle-même. La comparaison des valeurs formantiques du [ə] montre que cette voyelle ne connaît aucune différence interdialectale au niveau de sa position sur l'axe avant ~ arrière, ceci nous permet de poser une valeur moyenne pour $F2_{[ə]}$ = 1519 Hz. En revanche, la réalisation orientale est significativement plus ouverte que la réalisation maghrébine (p = .0001). Ce phénomène nous semble particulièrement important à souligner, car la mise en valeur d'une différence significative au niveau statistique quant au degré d'aperture pour la réalisation maghrébine vs orientale d'une voyelle, par définition, « centrale », peut *ipso facto* être interprétée comme la preuve d'une distribution différenciée : les parlers maghrébins privilégiant la génération de timbres vocaliques [+ centralisés] alors que les dialectes orientaux préfèrent les positions périphériques.

	F1	écart-type	F2	écart-type
aː	653	42	1373	101
iː	394	76	1918	166
æː	564	51	1660	103
uː	433	64	1069	121
əː	476	40	1522	83

Tableau 5 – valeurs formantiques moyennes des segments vocaliques longs de l'arabe maghrébin (en Hz)

	F1	écart-type	F2	écart-type
aː	696	45	1302	108
iː	341	76	1918	166
æː	605	64	1703	90
uː	386	53	958	121
eː	522	38	1981	120
oː	507	24	1013	83

Tableau 6 – valeurs formantiques moyennes des segments vocaliques longs de l'arabe oriental (en Hz)

Les tableaux 5 et 6 récapitulent les valeurs formantiques moyennes pour chacun des timbres vocaliques longs rencontrés dans nos données en arabe maghrébin et oriental. Statistiquement, on observe que les voyelles longues de l'arabe oriental présentent des caractéristiques formantiques différentes de celles de l'arabe maghrébin. La voyelle ouverte [aː] est ainsi significativement plus ouverte (p = .0001) et plus postérieure (p = .0001) que son pendant maghrébin. Les voyelles fermées [iː] et [uː] sont significativement plus fermées dans les parlers orientaux (p = .0001), [iː] et [uː] étant respectivement plus antérieur (p = .0001) et plus postérieur (p = .0001) en arabe oriental. Enfin, pour la voyelle d'aperture moyenne [æː], on observe que la réalisation orientale présente des différences articulatoires significatives tant sur

l'axe F1, où l'on constate que la réalisation orientale est plus ouverte qu'en arabe maghrébin (p = .0001), que sur l'axe F2, la réalisation orientale étant plus antérieure (p = .006). Les résultats statistiques présentés ici tendent ainsi à montrer que l'on retrouve, pour les voyelles longues, les mêmes schémas de dispersion que ceux observés pour les voyelles brèves. L'espace acoustique préférentiel de l'arabe maghrébin pouvant ainsi être défini comme [+ central] par rapport à celui des parlers orientaux.

Les analyses acoustiques menées à partir de parole spontanée en arabe maghrébin et oriental nous ont permis de caractériser au niveau phonétique la distribution des voyelles dans l'espace acoustique et d'établir une distinction entre parlers occidentaux privilégiant la génération de voyelles intérieures (i.e. centrales) résultant d'un processus de réduction vocalique et parlers orientaux préférant les positions périphériques. Nous avons vu que les voyelles brèves se distribuent, à l'Ouest vs à l'Est du domaine, suivant deux schémas de dispersion vocalique distincts (figure 3).

les parlers maghrébins présentent, en effet, une distribution vocalique plus condensée et nettement plus centralisée que celle des parlers orientaux ce qui vient préciser la remarque de Ph. Marçais [Marçais 75] selon laquelle « *les parlers du Maghreb se caractérisent par la ruine considérable de leur matériel vocalique* ». Ces deux schémas de dispersion vocalique différenciés apparaissent également au niveau du vocalisme long : les voyelles longues du Maghreb apparaissant comme plus centrales en comparaison de leurs correspondantes orientales en arabe oriental (figure 4).

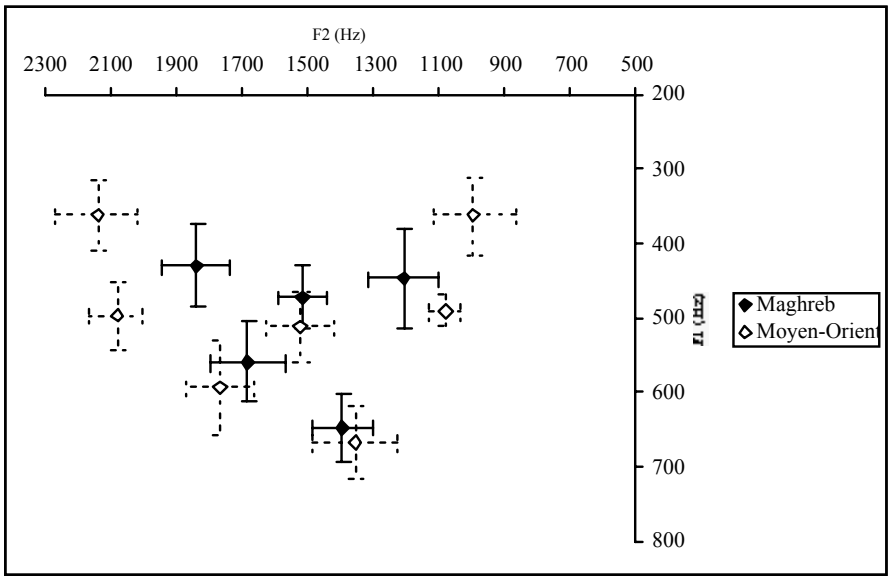

*Figure 3 – Représentation globale de la dispersion acoustique des **voyelles brèves** en arabe maghrébin vs moyen-oriental*

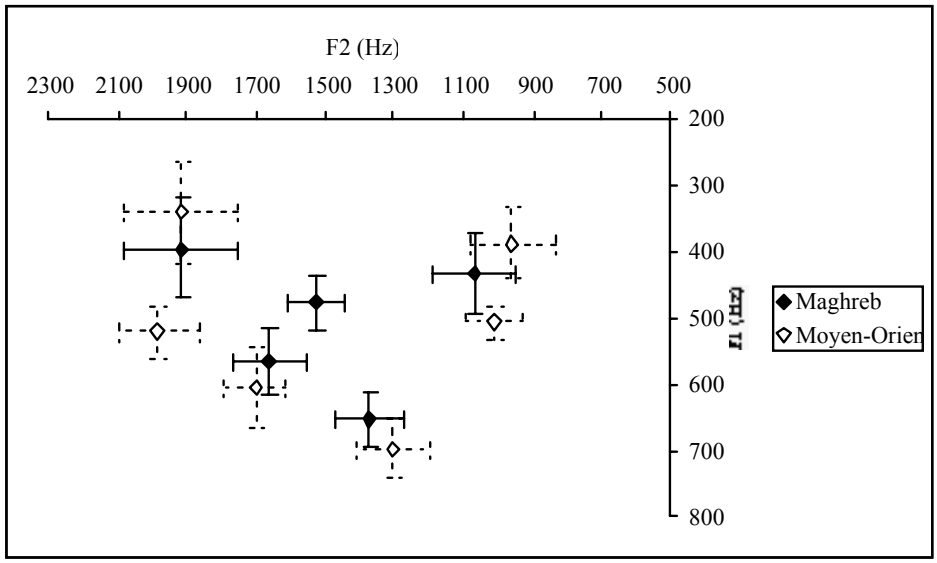

*Figure 4 – Représentation de la dispersion acoustique des **voyelles longues** en arabe maghrébin vs moyen-oriental*

Ainsi, bien que les systèmes phonologiques varient substantiellement d'un dialecte à l'autre à l'intérieur d'une même zone géographique [Barkat 00], l'observation comparée de la distribution des segments vocaliques brefs tels qu'ils sont effectivement réalisés en parole spontanée montre que les parlers maghrébins — contrairement aux parlers orientaux — tendent à favoriser la génération de timbres centralisés, ce qui conduit à une certaine réduction des contrastes vocaliques d'origine. Les parlers orientaux en revanche privilégient les positions périphériques et ont tendance à conserver une dispersion vocalique maximale tout en en développant par ailleurs de nouvelles voyelles d'aperture moyenne de type [e] et [o] et centrale (i.e. [ə]). Afin de mieux comprendre la dispersion des voyelles brèves dans les parlers du Maghreb, il convient de rappeler qu'un certain nombre d'entre elles ont purement et simplement disparu du système : les voyelles courtes finales, par exemple, déjà instables de par leur position, ne subsistent généralement plus. Certaines autres voyelles brèves positionnées à l'intérieur de syllabes ouvertes peuvent, selon le parler, soit s'abréger au point de devenir de simples points vocaliques de timbre plus ou moins neutre selon le contexte articulatoire et d'une durée généralement inférieure à 25 ms, soit disparaître complètement du fait d'une règle de syncope touchant, à de rares exceptions près, toutes les voyelles brèves ainsi positionnées. Enfin, il semblerait que les voyelles brèves qui sont conservées, tout en entretenant les oppositions qualitatives de base (i.e. [ɪ ə ʊ]), soient articulées à l'intérieur d'un espace articulatoire plus réduit que celui des parlers orientaux.

Outre la caractérisation, sur le plan acoustique, de la distribution des segments vocalique en arabe maghrébin vs oriental, cette étude nous autorise à considérer la distribution des voyelles brèves dans l'espace acoustique comme un critère potentiellement pertinent pour l'identification automatique des parlers arabes. Par ailleurs, notre travail confirme l'idée déjà avancée dans [Ghazali 79] que proposer un système vocalique identique à celui généralement postulé pour l'arabe standard n'est pas compatible avec les faits phonétiques observables en arabe dialectal, ce qui constitue un argument supplémentaire quant au développement de

modèles d'identification basés sur un décodage acoustico-phonétique à partir de parole spontanée tenant compte des variations dialectales et individuelles.

• Plan quantitatif

Les dialectes arabes opposent des voyelles longues aux voyelles brèves. Cette opposition de durée est admise par tous, à l'exception du dialecte marocain pour lequel elle est vivement controversée[7]. Elle est inscrite dans le système phonologique sous-jacent de la langue et permet de distinguer des significations par la seule variation de la quantité vocalique (par exemple, /χal/ « *vinaigre* » vs /xa:l/ « *oncle maternel* »). Les travaux traitant de la durée vocalique en arabe dialectal sont rares et jusqu'alors assez dispersés [Mitleb 84] [Jomaa & Abry 88] [Ghazali & Braham 92] [Alioua 91-92] [Agoujard 1979]. Toutefois, bien que ce type de connaissance soit indispensable pour l'élaboration des règles phonétiques significatives liées au phénomène de temporalité, l'étude de l'opposition de la durée vocalique telle qu'elle est réalisée en parole naturelle (et non à l'intérieur de mots isolés comme dans la plupart des études précédemment citées) reste primordiale. En effet, les connaissances ainsi acquises permettraient d'améliorer les performances des modèles de reconnaissance et des systèmes de synthèse de la parole dont l'objectif ultime est, d'une part, de pouvoir fonctionner à partir de parole spontanée dans le cas des systèmes d'identification et d'autre part, d'être le plus proche possible des phénomènes observés en parole naturelle pour ce qui concerne la synthèse de la parole.

Nous avons ainsi retenu les mesures acoustiques des segments vocaliques brefs et longs de 6 variétés d'arabe dialectal différentes (tableau 1) avant d'établir, sur la base de nos données, le rapport moyen V_L/V_B tel qu'il est effectivement réalisé en parole spontanée, c'est-à-dire non-lue. Le choix de cette méthodologie s'explique par le fait qu'il nous semble contestable d'établir les caractéristiques phonétiques de l'arabe dialectal à travers le filtre de l'arabe standard dont les règles sont nécessairement transposées à l'arabe dialectal par l'usage d'un corpus écrit. De ce fait, procéder à des analyses phonétiques en s'appuyant sur un corpus lu alors que l'arabe dialectal est une langue orale — qui plus est sans écriture conventionnelle jusqu'à nos jours — nous semble limité. Il est de ce fait évident que l'analyse des caractéristiques phonétiques de l'arabe dialectal se doit d'être réalisée à partir d'un corpus oral spontané, seul capable de refléter une image moins déformée de l'usage quotidien qu'ont les locuteurs arabophones de leur langue maternelle.

[7] Plusieurs auteurs affirment que le système vocalique de l'arabe marocain oppose des voyelles longues aux brèves. On peut citer entre autres, Benkaddour (1982), Fennan (1986), Harris (1942 et 1946), Hassaoui (1980), Idrissi (1987), Khomsi (1975), Laabi (1975), Rhardiss & al. (1992). Pourtant, certains autres chercheurs adoptant une démarche synchronique, fondée sur l'observation des faits physiques par des outils expérimentaux, aboutissent généralement à l'absence ou à la disparition de l'opposition de quantité en arabe marocain. Benhallam & Dahbi (1990) affirment par exemple que l'arabe marocain a abandonné cette opposition. De même, Amrani (1997) et Embarki (1997) observent que sur le plan acoustique, l'arabe marocain ne connaît pas (c'est-à-dire plus) d'opposition entre voyelles brèves et voyelles longues et que la perception que certains auditeurs ont de cette opposition est due (1) à une distribution différente de la durée entre les différents segments de la syllabe ; (2) à la nature de la syllabe ; (3) à l'influence de la connaissance de la morphologie classique liée diachroniquement à la variété dialectale marocaine.

Le tableau 7 présente les valeurs moyennes de durée vocalique pour chacun des parlers étudiés et permet de remarquer — qu'en moyenne — les voyelles brèves et/ou longues présentent d'une zone dialectale à l'autre des différences de durée remarquables.

	Durée moyenne des V_B (en ms)	Durée moyenne des V_L (en ms)
Maroc	58	106
Algérie	52	103
Tunisie	51	102
Moyenne *Maghreb*	**54**	**104**
Syrie	61	140
Liban	58	145
Jordanie	54	131
Moyenne *Orient*	**58**	**139**

Tableau 7 – Moyenne des durées vocaliques tous timbres confondus par pays et par zone

Rapports $_{VL/VB}$ moyens par pays (tous timbres et tous locuteurs confondus)			
Maghreb		Moyen-Orient	
Maroc	1.8	Syrie	2.3
Algérie	2.0	Liban	2.6
Tunisie	2.0	Jordanie	2.1
Moyenne des Rapports $_{VL/VB}$ par zone			
1.9		**2.3**	

Tableau 8 – Rapports R $_{VL/VB}$ moyens en arabe maghrébin et moyen-oriental

Afin de caractériser au niveau statistique ces écarts, nous avons reporté dans le tableau 8 les rapports V_L/V_B calculés pour chaque parler, sur la base desquels un T-test a été effectué.

L'analyse statistique révèle que les rapports mis en œuvre dans l'une et l'autre des deux zones dialectales sont significativement différents $T_{(3, 2,35)} = 2.50$ ($p = .04$), les parlers orientaux attestant des rapports significativement plus élevés que les parlers du Maghreb. Cela signifie que dans les parlers du Moyen-Orient, l'opposition de durée vocalique s'établit de manière plus contrastée qu'au Maghreb, où l'on constate néanmoins que, même au Maroc, cette opposition subsiste.

4. Identification automatique des parlers arabes

La dernière étape de ce travail a consisté à mettre au point des expériences de reconnaissance automatique à partir des caractéristiques de la distribution et des rapports de durée vocalique en arabe dialectal maghrébin vs oriental, lesquelles sont considérées à l'issue de l'expérience perceptuelle et des analyses acoustiques, comme potentiellement pertinentes pour l'identification des parlers arabes par zones géographiques principales. Le modèle utilisé [Pellegrino 98b] est basé sur la modélisation acoustique des systèmes vocaliques (i.e. système différentié), il permet d'obtenir des modèles d'apprentissage à partir de données non-étiquetées (i.e. approche non-supervisée). Cette pratique présente, entre autres, l'avantage de ne pas être

influencée par les connaissances phonologiques, qui dans le cadre de l'arabe, apparaissent souvent comme biaisées compte tenu des phénomènes d'interférence et d'hypercorrection avec la variété haute de la langue (arabe classique et/ou moderne). En détectant sur le signal les sons possédant une structure formantique, le système est en mesure d'établir un certain nombre de classes vocaliques gaussiennes qu'il attribue dans une phase d'apprentissage à des modèles de langue (i.e. modèle maghrébin vs modèle moyen-oriental). Les décisions prises lors des tests de reconnaissance étant fonction de la vraisemblance des voyelles détectées à partir de l'échantillon de parole non-connue avec les modèles d'apprentissage.

4.1. Corpus et méthode

• Corpus d'apprentissage

Le corpus d'apprentissage est élaboré à partir de la traduction spontanée du texte « La Bise et le Soleil » par 10 locuteurs et locutrices arabophones originaires de différents points du domaine linguistique arabophone. Pour chacun de ces 10 locuteurs, 4 répétitions du texte ont été acquises (correspondant à environ 2 minutes de parole). Les données acoustiques enregistrées auprès des locuteurs et de locutrices natifs ont ensuite été digitalisées à 16 kHz, 16 bits. La zone occidentale est représentée dans le corpus d'apprentissage à travers des stimuli en arabe algérien et marocain[8] La zone orientale par des stimuli en arabe égyptien, syrien, libanais et jordanien. Le modèle maghrébin (i.e. MA) est basé sur les réalisations vocaliques de cinq différents locuteurs : un locuteur algérien (originaire de Oran) et quatre locuteurs marocains (originaires de Rabat et Casablanca) (tableau 9). La modélisation de l'espace acoustique moyen-oriental a été obtenue à partir des réalisations vocaliques détectées dans des corpus de cinq locuteurs orientaux originaires de Syrie (deux locuteurs originaires de Alep), du Liban (un locuteur originaire de Beyrouth) et de Jordanie (deux locuteurs originaires de Irbid) (tableau 10).

Locuteur	Pays	Ville d'origine
B0008	Algérie	Oran
B0009	Maroc	Rabat
B0010	Maroc	Rabat
B0016	Maroc	Casablanca
B0017	Maroc	Casablanca

Tableau 9 – Variétés dialectales occidentales représentées dans le corpus d'apprentissage

[8] Les parlers tunisiens présentant des caractéristiques spectrales et temporelles différentiellement moins marquées que les autres parlers maghrébins ont été volontairement écartés de cette étude afin de ne pas dégrader les résultats obtenus sur la seule base de la distribution et de l'opposition de durée vocaliques. Cette décision a été prise du fait du relativement faible nombre de locuteurs (de manière générale, on parle de modèles indépendants du locuteur au delà de 50 locuteurs d'apprentissage !). L'utilisation de locuteurs tunisiens aurait ainsi conduit à rajouter au modèle de la variabilité alors que nous ne disposons pas, à l'heure actuelle, d'un nombre suffisant de données pour estimer convenablement nos modèles.

Locuteur	Pays	Ville d'origine
B0004	Syrie	Alep
B0005	Jordanie	Irbid
B0007	Jordanie	Irbid
B0020	Liban	Beyrouth
B0021	Syrie	Alep

Tableau 10 – Variétés dialectales orientales représentées dans le corpus d'apprentissage

L'objectif de l'expérience est de confirmer ou d'infirmer la robustesse de l'indice de dispersion vocalique et de l'opposition de durée pour l'identification automatique des parlers arabes par zones. Pour chacune des deux aires dialectales, nous avons élaboré quatre modèles en fixant, à chaque fois, un nombre de classes gaussiennes différent (5, 10, 15, et enfin 20). L'objectif est de déterminer le nombre de classes optimal à partir duquel les meilleurs taux de reconnaissance seront observés[9].

• Corpus de test

Pour la phase de test, nous avons utilisé les réalisations dialectales de dix autres locuteurs originaires de différents pays du Maghreb et du Moyen-Orient. Les parlers maghrébins sont représentés par des parlers marocains et algériens (i.e. locuteurs originaires de Touggourt et Jijel pour l'Algérie et Rabat, Tétouan et Casablanca pour le Maroc). La zone orientale est représentée par des réalisations en arabe égyptien (locuteur d'Assouan), syrien (locuteur originaire de Homs), palestinien (Hébron et Haïfa) et enfin, jordanien (locuteurs de Irbid) (Tableau 11).

Locuteur	Pays	Ville d'origine
B0011	Algérie	Touggourt
B0013	Algérie	Jijel
B0022	Maroc	Tétouan
B0024	Maroc	Rabat
B0002	Maroc	Casablanca
B0001	Egypte	Assouan
B0006	Palestine	Hébron
B0023	Syrie	Homs
B0025	Palestine	Haifa
B0031	Jordanie	Irbid

Tableau 11 – Variétés dialectales représentées dans le corpus de test

Les expériences d'identification automatique ont été réalisées, dans un premier temps, sur la base des quatre répétitions de chacun des dix locuteurs (40 tests)[10]. Dans un second temps,

[9] Des tailles de modèles supérieures à 20 composantes ont été envisagées, mais il s'avère que le nombre de données d'apprentissage est insuffisant pour parvenir à les estimer convenablement.

[10] Notons toutefois que le fait d'utiliser plusieurs répétitions d'un même locuteur a pour conséquence que les quatre tests ne peuvent pas être considérés comme indépendants. Ce cadre expérimental a été retenu afin d'observer l'effet de la durée des énoncés sur les taux d'identification atteints.

nous avons considéré les quatre répétitions de chaque locuteur comme un bloc unique à la fin duquel la décision d'identification dialectale était prise (10 tests). Les taux obtenus à l'issue de cette seconde expérience correspondent ainsi aux scores atteints pour dix tests (i.e. 1 décision par locuteur). Lors de l'utilisation de 40 tests, deux conditions expérimentales ont été testées. La première a consisté à déterminer le pouvoir discriminant de la dispersion vocalique seule. Dans ce premier temps, le modèle d'apprentissage est élaboré à partir des caractéristiques formantiques seules (i.e. 8 MFCC). Dans un second temps, le modèle est appris à partir de l'utilisation conjointe des caractéristiques formantiques et de l'information de durée (i.e. 8 MFCC + D). Notre hypothèse consiste à supposer que l'utilisation conjointe de ces deux critères discriminants conduit à améliorer les taux de reconnaissance de manière significative.

4.2. Résultats

Nos résultats — bien que relativement peu significatifs d'un point de vue statistique compte tenu du faible nombre d'enregistrements disponibles pour l'élaboration des modèles d'apprentissage — révèlent des tendances générales intéressantes (figure 5).

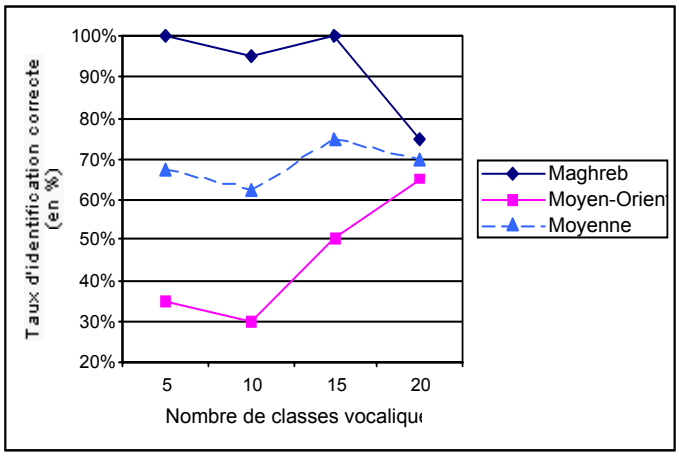

Figure 5 – Taux d'identification correcte pour la discrimination par zones en fonction de la taille du modèle
(caractéristiques spectrales : 8 MFCC).

Les modèles composés d'un faible nombre de classes gaussiennes distinguent assez mal les deux zones dialectales représentées dans les différents stimuli : la plupart des échantillons de parole sont ainsi identifiés comme étant de l'arabe maghrébin (cf. le taux d'identification élevé pour la catégorie Maghreb et le faible score correspondant à l'identification de la zone orientale). Cette tendance sous-entend qu'un modèle plus complexe (i.e. comportant un nombre de classes vocaliques supérieur) est nécessaire pour parvenir à caractériser de manière plus fine l'organisation vocalique des parlers orientaux. Toutefois, quand la taille du modèle augmente cet effet tend à disparaître et l'on obtient, avec le nombre — ici optimal — de 20 classes vocaliques, un score d'identification par zone de 70 %.

L'utilisation du test statistique de Pearson (Khi^2) établit que ce score est significativement supérieur à la chance (P $\chi^2 > 3.84$) = .05. Le t-test effectué sur les taux d'identification moyens obtenus pour chaque zone dialectale ne révèle quant à lui aucune différence significative, ce qui prouve que le modèle est aussi performant pour la discrimination de l'une et l'autre des deux aires dialectales à identifier. Si l'on compare ces résultats avec ceux obtenus à partir d'un apprentissage fondé *conjointement* sur les caractéristiques spectrales et sur la durée des segments vocaliques, on observe que — dans ce dernier cas et pour le même nombre de composantes gaussiennes — les scores d'identification correcte atteignent 78 %, (Figure 6).

Les taux d'identification correcte obtenus avec des modèles de petites tailles sont sensiblement supérieures lorsque l'on prend en considération l'information de durée vocalique. Alors que le taux correspondant à l'identification de la variété occidentale reste élevé, le score d'identification des variétés orientales passe à 50 % avec 5 classes vocaliques (classification résultant du hasard), et à 60 % avec 10 composantes. Ces résultats moyens s'expliquent du fait que le modèle n'est pas assez complexe (en termes de classes vocaliques) pour permettre une bonne discrimination des systèmes vocaliques orientaux. La prise en compte du paramètre de durée vocalique tend à faire disparaître le biais des « petits » modèles (i.e. composés de 5 ou 10 classes) qui ont tendance à classifier la plupart des échantillons de parole dans la catégorie Maghreb. Avec 20 classes gaussiennes, nous obtenons 78 % d'identification correcte. L'analyse statistique de ce score permet d'écarter l'hypothèse qu'il soit dû au hasard (P $\chi^2 > 3.84$) = .05.

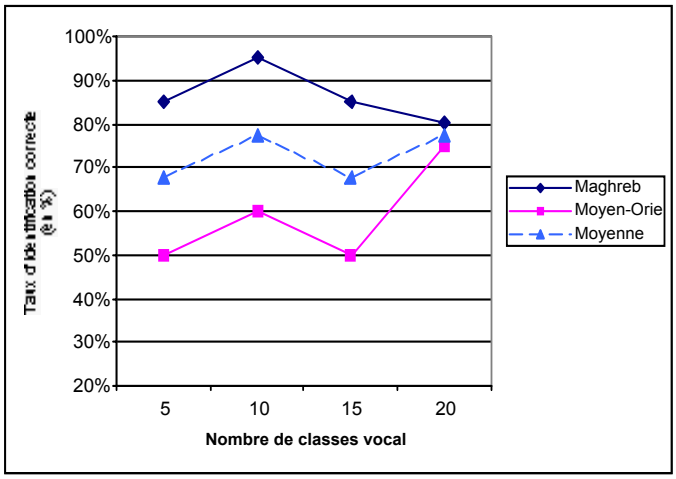

Figure 6 – Taux d'identification correcte pour la discrimination par zones en fonction de la taille du modèle
(caractéristiques spectrales : 8 MFCC + Durée).

Les différences de scores obtenus par le modèle à 20 composantes pour la discrimination du Maghreb vs du Moyen-Orient se révèlent non-significatives sur le plan statistique, ce qui nous autorise à dire que le modèle est aussi performant pour l'identification des stimuli maghrébins que pour la discrimination des parlers orientaux.

L'utilisation *conjointe* de plusieurs indices discriminants (i.e. ici la dispersion et la réalisation de l'opposition de durée vocaliques) conduit à une amélioration sensible des scores de reconnaissance. En effet, pour 40 tests effectués, nous obtenons 70 % d'identification correcte avec un modèle prenant en compte les caractéristiques formantiques des systèmes

vocaliques seules. Lorsque l'on modélise conjointement les caractéristiques spectrales ainsi que l'information de durée des segments détectés, le score passe à 78 %. La différence des taux d'identification observée entre l'une et l'autre de ces deux conditions expérimentales présente un écart significatif au plan statistique ($T_{(39\ 1.68)} = 1.78$; $p = 0.5$). Ces deux critères peuvent donc être définis comme robustes pour la discrimination automatique des parlers arabes par zone géographique principale. Il convient cependant de rappeler que le nombre de classes vocaliques constitue un facteur essentiel quant aux performances des systèmes d'identification automatique. Compte tenu du faible nombre de données dont nous disposons à l'heure actuelle, nos modèles d'apprentissage sont élaborés à partir des réalisations vocaliques de cinq locuteurs par zone uniquement, soit sur la base de la modélisation de quelque 2000 voyelles détectées par zone[11]. De ce fait, nous n'avons pas pu tester les modèles appris avec un nombre supérieur de composantes gaussiennes. Par ailleurs, afin de vérifier l'influence du facteur « durée de test » sur nos résultats, nous avons répété l'expérience d'identification par zone avec 10 tests. Rappelons que dans cette condition nous avons considéré les quatre répétitions de chaque locuteur comme un bloc unique. La décision d'identification porte donc sur deux minutes de parole continue et non pas sur trente secondes comme dans l'expérience précédente. Les propriétés du modèle utilisé ici correspondent au modèle optimal retenu dans l'expérience précédente (i.e. 20 classes vocaliques). Les résultats obtenus à l'issue de cette seconde expérimentation apparaissent dans le tableau 12.

Conditions expérimentales (paramètres de modélisation)	**Durée des tests**	
	30 secondes (40 tests)	2 minutes (10 tests)
MFCC	70%	70%
MFCC + Durée	78%	90%

Tableau 12 – Taux d'identification correcte (en %) en fonction de la durée des échantillons à identifier et des paramètres de modélisation

Avec 10 tests, nous obtenons pour la première condition expérimentale (modélisation des caractéristiques formantiques seules) 70 % d'identification correcte. Dans ce cas, le manque de données ne nous permet pas d'écarter l'hypothèse que ce score puisse être dû au facteur chance (correspondant à 50 %). Toutefois, Lorsque l'on intègre au modèle le paramètre de durée vocalique les scores atteignent 90 % ce qui nous permet de rejeter statistiquement cette éventualité ($P\ \chi^2 > 3.84) = .05$. Enfin, nous avons vu que sur 40 tests, l'intégration du paramètre de durée conduisait à améliorer de manière significative les scores d'identification (de 70 % à 78 %). Sur 10 tests, bien que les pourcentages obtenus à l'issue de la seconde condition

[11] Le modèle a détecté 2000 segments vocaliques en arabe maghrébin et 2300 voyelles en arabe oriental. Cette différence est probablement liée à la présence de voyelles ultra-brèves présentes en arabe maghrébin que le modèle a du mal à détecter de manière automatique. Du point de vue linguistique, il est possible d'imputer *a priori* ces différences au phénomène de chute des voyelles brèves en syllabe ouverte, fait rencontré fréquemment dans les parlers du Maghreb. Il conviendrait d'observer de plus près les signaux afin de déterminer l'origine de cette différence.

(i.e. 8 MFCC + D) semblent indiquer de meilleures performances lorsque l'on prend en compte le paramètre de durée pour la modélisation des systèmes vocaliques (de 70 % à 90 %), la différence observée s'avère non-significative au plan statistique.

5. Conclusion

L'objectif ultime de ce travail était de confirmer la pertinence de la distribution des segments vocaliques ainsi que la réalisation de l'opposition de durée pour l'identification automatique des parlers arabes par zones géographiques. Bien qu'à l'heure actuelle, nos résultats soient peu représentatifs au niveau statistique, ils mettent en valeur des tendances intéressantes. En effet, les résultats obtenus pour l'identification automatique des dialectes arabes à partir de la modélisation acoustique de leurs systèmes vocaliques sont, si ce n'est probants, tout au moins encourageants. En effet, selon le nombre de tests effectués (i.e. 40 et/ou 10) et en fonction du nombre de paramètres retenus pour la modélisation des systèmes vocaliques (caractéristiques formantiques seules et/ou caractéristiques formantiques + paramètre de durée vocalique), les taux d'identification varient entre 70 %, 78 % et 90 %. Cette étude finit ainsi de lever l'inconnue présentée en introduction de ce travail quant à la possibilité de distinguer — à l'intérieur du monde arabophone — deux zones dialectales principales et distinctes caractérisées par des organisations vocaliques suffisamment différenciées pour permettre de les considérer comme des indices robustes de discrimination dialectale pertinents dans le cadre d'une tâche d'identification automatique.

6. Références

[Abu-Haidar 91] F. Abu-Haidar, *Variabilité et invariance du système vocalique de l'arabe standard*, Thèse de Doctorat nouveau régime en Sciences du Langage, Université de Besançon, (1991).

[Barkat 97] M. Barkat, J.M. Hombert & C. Taine-Cheikh, « Détermination d'indices acoustiques robustes pour l'identification automatique des parlers arabes : les variations qualitatives de la voyelles [a] dans les parlers arabes », in *Actes des Journées d'Etude sur la Voyelle*, Université de Nantes, pp 141-46, (1997).

[Barkat 98] M. Barkat, « Identification dialectale et détermination expérimentale d'indices discriminants pour les parlers arabes », in *Actes des XXIIèmes Journées d'Etudes sur la Parole*, Martigny, pp. 71-74, (1998).

[Barkat 00] M. Barkat, *Détermination d'indices acoustiques robustes pour l'identification automatique des parlers arabes*, Thèse de Doctorat nouveau régime en Sciences du Langage, Université Lumière Lyon 2, 300p, (2000)

[Benkirane 82] T. Benkirane, *Etude et fonctions de la syllabe en arabe marocain*, Thèse de 3° cycle en Sciences du Langage, Université d'Aix-Marseille, (1982).

[Cohen 70] D. Cohen, « Koiné, langue commune et dialectes arabes », in *Janua Linguarum*, Série Practica vol 81, Mouton Eds, pp.105-25, (1970).

[Fleish 75] H. Fleish, Article « Arabiyya » in *Encyclopédie de l'Islam*, tome 1, Brill & Maisonneuve Ed., pp. 593, (1975)

[Ghazali 79] S. Ghazali, « Du statut des voyelles en arabe », in *Analyses et Théorie* 2/3, pp. 199-217, (1979).

[Ghazali 96] S. Ghazali, « Fréquence d'occurrence des voyelles en arabe standard », étude non-publiée, IRSIT, Tunis, (1996).

[Hombert 97] J.M Hombert & I. Maddieson, « A linguistic approach to Automatic Language Recognition », in *Actes du Congrès International des Linguistes*, (1997).

[Kaye 97] A.S. Kaye, *Phonologies of Asia and Africa (including the Caucasus)*, A. S. Kaye Ed., Eisenbrauns, Wiona lake, (1997).

[Marçais 75] Ph. Marçais, *Esquisse grammaticale de l'arabe maghrébin*, Maisonneuve, (1977).

[Pellegrino 98a] F. Pellegrino & A. Galinier, « Identification des langues par discrimination automatique des systèmes vocaliques », in *Actes des XXIIèmes Journées d'Etudes sur la Parole*, Martigny, pp. 179-182, (1998).

[Pellegrino 98b] F. Pellegrino, *Une approche phonétique en identification automatique des langues : la modélisation acoustique des espaces vocaliques*, Thèse de doctorat, Université des Sciences de Toulouse, (1998)

[Rjaibi-Sabhi 93] N. Rjaibi-Sabhi, *Approches historique, phonologique et acoustique de la variabilité dialectale arabe : caractérisation de l'origine géographique en arabe standard*, Thèse de Doctorat nouveau régime en Sciences du Langage, Université de Besançon, (1993)

Differentiating phonetic from phonological events in speech

John J. OHALA & Egidio MARSICO

Phonology Laboratory, University of California, Berkeley
Dynamique Du Langage, Université Lumière Lyon 2

ohala@cogsci.Berkeley.edu – Egidio.Marsico@ish-lyon.cnrs.fr

Abstract

This study sheds light on the phenomenon in which, when looking into the speech signal for a specific cue that is supposed to belong to the phonology of a particular language, one encounters the result of a contextual phonetic variation of another language, thus leading to wrong identification. We present some examples where phonetic and phonological events have the same – or almost the same – representation in the signal, and some methodological strategies to deal with phonological contrasts rather than just phonetic variation.

Résumé

L'idée principale de cette étude est d'éclairer le fait qu'en recherchant dans le signal de parole une propriété spécifique supposée appartenir à la phonologie d'une langue, on puisse, en fait, n'être confronté qu'à une variation phonétique contextuelle d'une autre langue, et par là même, être amené à une mauvaise identification. Nous donnons ici quelques exemples où des évènements phonétique et phonologique ont les mêmes – ou presque les mêmes – représentations dans le signal, ainsi que quelques conseils méthodologiques permettant d'être certain d'avoir affaire à un contraste phonologique plutôt qu'à une variation phonétique.

1. Introduction

When embarking on a project of automatic language identification, one no doubts consults the phonological literature to see what some of the phonological properties are of the languages that have to be differentiated. These properties even if they're representing phonological features, must be described as acoustic properties that can be found when examining the signal.

Unfortunately, there could be some "hidden" traps in this approach. Although a given language may have a certain feature (or phoneme) phonologically, other languages may have the same feature phonetically. For example, French, Hindi and Portuguese have distinctively nasalized vowels, e.g. Hindi /sas/ "mother-in-law" vs. /sãs/ "breath", but contextually nasal vowels are found in numerous languages, including English and Japanese.

Also, English, Spanish, and Italian have contrastive affricates, e.g. English /tʃɪl/ vs. /tɪl/ vs. /ʃɪl/ and French does not. However, Quebec French has contextually predictable affricate allophones of /t/ and /d/ before /i/, e.g. [paʁtsi] "parti".

Pharyngealization is a well known feature of Arabic. But in many cases its principal cue is its coloring of adjacent vowels, e.g. /ɑ/ instead of /a/ and /e/ instead of /i/ appears next to

pharyngealized consonants. But the vowels which are allophones in Arabic may be present distinctively in other languages, e.g. French.

Tones are a distinctive feature of Punjabi, having developed from stops that once had a distinctively breathy voiced release. But in present day Hindi (and other modern Indo-Aryan languages) these same breathy voiced stops have a characteristic influence on the F0 of adjacent vowels that is similar to the lexically distinct tones in Punjabi.

To understand how there can be such parallels between distinctive phonological features of some languages and non-distinctive phonetic features of others, we have to understand where phonological features come from.

The answer here is similar to the answer to the question : Where do all the diverse life forms come from ? Answer : they evolve and are shaped by constraints internal to living organisms (e.g., the need for chemical fuel) and constraints external to the organism (e.g., the environment, including other organisms competing for the same resources).

Phonological features also "evolve" (in a sense) via sound changes and the output is shaped by a variety of factors, some of which are the physical phonetic constraints of speech production and perception.

Sound change (a change in a pronunciation norm) arises when the listener misinterprets or misparses a speech signal made ambiguous due to speakers "shortcuts" or due to inherent constraints of the speech production and perception systems.

For example, the release of apical stops before the approximant [ɻ] as in English *draft* /dɻæft/ may become non-distinctively affricated, [dʒɻæft], due to the exiting airflow being channeled through the narrow mid-palatal constriction. If, in addition, the final cluster is simplified, it can become homophonous with giraffe [dʒɻæf] where the affrication is distinctive. The affricate in *actual* arose in a similar way : [ækt] + [juəl] > [æktʃuəl].

This is how distinctive phonetic features in language X can be present non-distinctively in language Y.

Although the task of differentiating between the two is not an easy one, we present some ideas on how to accomplish this. We'll examine several ambiguous cases where our knowledge enables us to give measurable parameters of differentiation.

2. Examples

2.1. Front rounded vowels

Boyd Michailosvky, [Michailosvky 75], gives the following examples of sound changes in Tibeto-Burman: written Tibetan {a, o, u}> Lhasa Tibetan {ɛ, ø, y} / _ d, n, l, s ; e.g., bod > phøø "Tibet", bdud > tyy "demon".

A detailed acoustic study revealed that alveolar consonants raise the F2 value of the preceding vowel, corresponding to a forward movement in the articulation of the vowel. Table 1 demonstrates the variation.

	F2		
	o	**u**	**i**
-t	+ 400	+ 170	- 95
-d	+ 310	+ 370	- 240
-n	+ 220	+ 385	- 300
-s	/	+ 420	- 200

Table 1 – F2 transitions before finals in English CVC utterances. Lehiste & Peterson, 1960.

We thus have a clear explanation of how the misparsing of the acoustic signal may occur: unable to normalize the contextual F2 variation, the listener attributes this cue to the vowel and gives rise to the change. So, how can one be sure to deal with phonological front rounded vowels when looking into the speech signal rather than just a contextual variation ? One way is to first determine the F2 frequency at the beginning and the end of the vowel, seeking an eventual transition, and then determining the place of articulation of the following consonant.

2.2. Effects of /s/ on adjacent vowels [Ohala 93]

The first case is the one of spontaneous nasalization studied by John Ohala, [Ohala 75], [Ohala 80].

We find in the literature the following sound changes, Sanskrit sarpa > Hindi [sãp] "snake", Chinese /p^{h}antʃjo/ "plantain".

There seems to be no plausible articulatory or aerodynamic reason for the emergence of nasality near an /s/, but we can give a possible explanation based on a complex interaction of articulatory and acoustic-auditory facts. The high airflow required for voiceless fricatives implies a wider than normal glottal opening, and the induced abductory gesture of the vocal cords spreads onto the adjacent vowel. In fact, this slightly open glottis during voicing lowers the amplitude and increases the bandwidth of F1, thus, mimicking the effects of nasalization.

The misparsing, then, occurs when transfering these "nasalized-like" properties as real nasalization.

The second case correspond to the phenomenon called s-aspiration.

Colombian Spanish [loh libroh] vs Spanish [los libros], Indo-European septm > Classical Greek [hɛpta].

Widdison [Widdison 91] suggested that the above mentioned changes that /s/ imposes on the margin of adjacent vowels may sound sufficiently [h]-like to imply the loss of the final [s], being replaced by a "latent" [h]. He conducted an experiment in which he cross-spliced vowels between spanish words like pasta and pata as spoken by a Mexican Spanish speaker. (Mexican Spanish does not exhibit s-aspiration). His results showed that subjects misinterpreted pata as pasta 33% of the time. So, like spontaneous nasalization, s-aspiration appears to result from a misparsing of the peculiar transition between vowels and voiceless fricatives.

Last case, the phenomenon of nasal effacement:

Latin institutus > Italian istituto, Latin constringere > Italian costringere

Although this change seems to be the reverse of spontaneous nasalization, and thus, hard to explain by the same process, we hypothesize that, indeed, the same process is responsible. Spontaneous nasalization arises when, due to misparsing, listeners dissociate the acoustic effects of the voiceless fricative and guess there is nasalization. Nasal effacement occurs when an original nasal element is misparsed as a spurious effect due to the voiceless fricative and thus discounted.

A perceptual experiment conducted by Grazia Busá and John Ohala [Busá 92], has shown that a nasal consonant is harder to detect in the vicinity of final [s] and [θ] than of [t], thus supporting the correctness of the hypothesis.

Going back to automatic identification, we can say that when detecting nasality in the signal, the presence of an /s/ in the vicinity might be a good indication of phonetic variation.

2.3. F0 variations due to voice contrast

We know that phonetically, voiced and voiceless consonants induce non-distinctively a lower and higher F0 on following vowels. Some languages have developed phonologically distinctive tone from the consonantally-induced F0 difference.

For example, we have : Southern Kammu /klaaŋ/-/glaaŋ/ vs. Northern Kammu /kláaŋ/-/klàaŋ/ ("eagle"-"stone").

In this case one might differentiate phonetic F0 contours from phonological F0 contours (i.e. tones) by discounting an F0 contour by, say, -10 Hz if the preceding consonant is voiced and by +10 Hz if the preceding consonant is voiceless. The “discounting” would be gradient: more immediately adjacent to the consonant, less further away (up to 100 msec. after the consonant).

The key to the problem, then, is to look at the correlation between the feature, F, and the possible causal factor, C, especially if one can vary the magnitude of C. If the correlation is continuous and “lawful”, F – the dependent variable –, is assumed to be a phonetic event caused by C – the independent variable. If the correlation between F and C is discontinuous and “not lawful”, then F is assumed to be a phonological event not caused by C. Figure 1 illustrates this opposition.

The difficult part in this task is to be able to identify and quantify all the causal (the independent) variables for a given (posited) dependent variable. We give in the next part, some examples that seem to illustrate the above way of differentiating phonetic features from phonological ones.

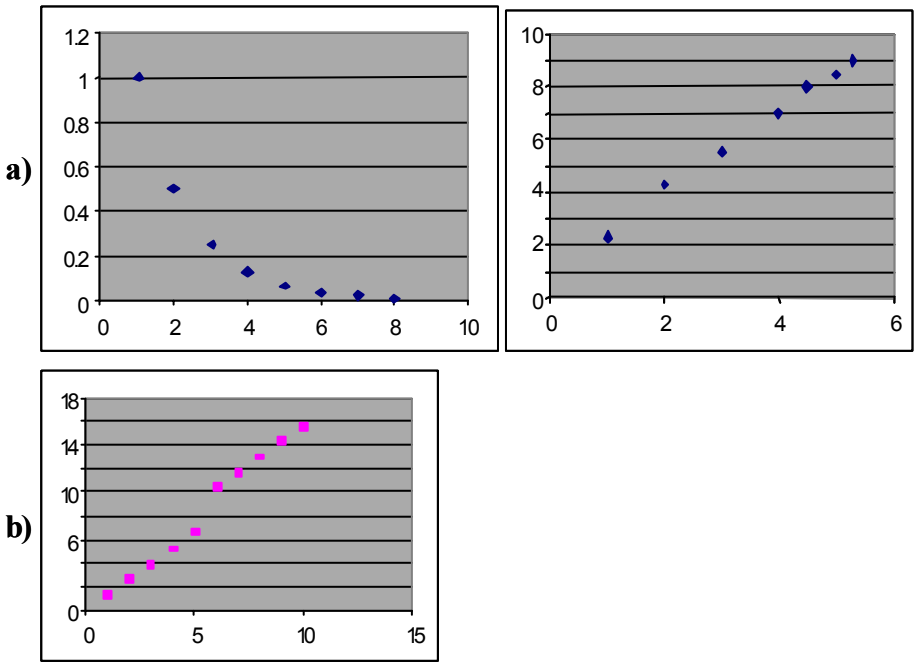

Figure 1 – Illustration of the opposition between lawful (a: continuous) and non lawful (b: discontinuous) correlations.

3. Case studies.

3.1. Vowel reduction as a function of vowel duration.

Lindblom, [Lindblom 63], measured F1 and F2 of several Swedish vowels spoken by a single speaker uttering C1VC2 syllables where the V was one of 10 Swedish vowels and C1 = C2.

He showed (Figure 2) that the shorter the vowel, the more the vowel formant frequencies – especially F2 – approached the formant frequencies of the consonant.

This was interpreted as articulatory "undershoot". In fact, it results for the consonantal articulation being more "overlaid" upon the vowel when the vowel is shorter.

Thus vowels' quality can change as a function of the rate of speech and consonantal context. Figure 2 shows the range of change. F1 and F2 for isolated vowels are given as dashed lines at far right in each figure. Here the variation as a function of duration is continuous and lawful and thus phonetic.

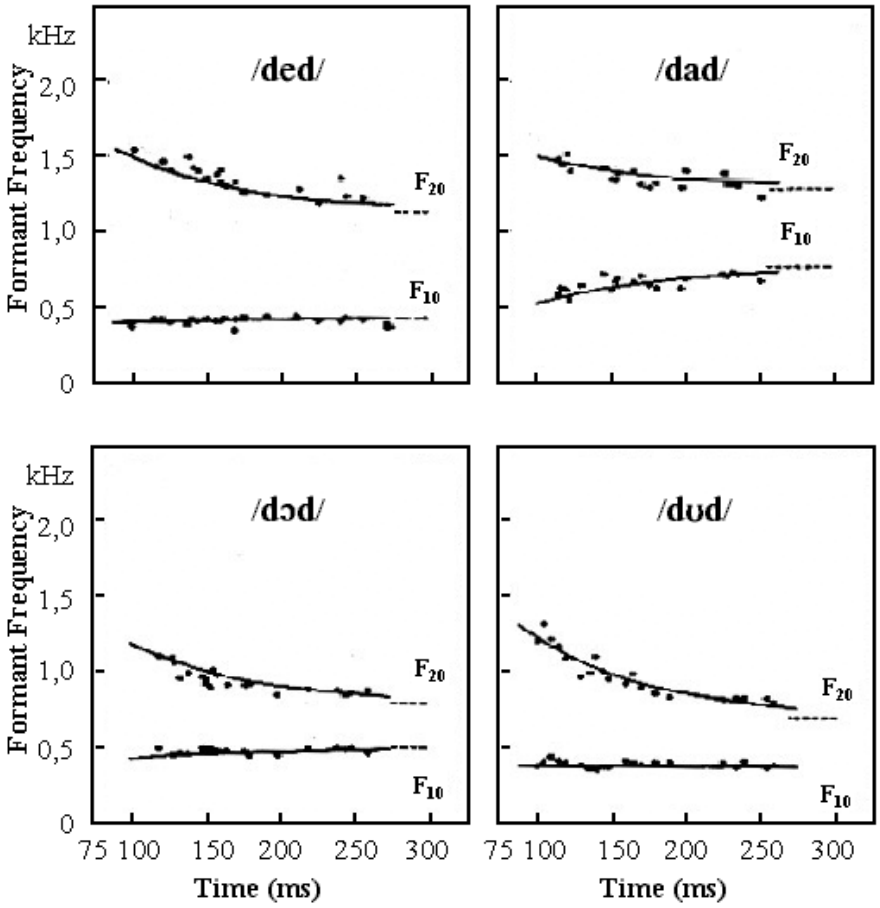

Figure 2 – Changes in F1 and F2 of CVC syllables as a function of vowel duration.

In a later study, Nord [Nord 74] measured the formant frequencies of selected Swedish vowels in final and non-final position and when stressed and unstressed.

Figure 3 shows the F2 value for the vowel in the syllable /sel/ in two degrees of stress and two positions in the utterance. Most F2 values are lawfully related to vowel duration, so that variation can be attributed to phonetic factors. But the reduction in the unstressed final position (Δ) argues for it showing phonological variation.

To determine the likelihood that an [y] is either /y/ or /u/ (or some other vowel), we suggest :

- Determine the vowel duration and compare this with the distribution of such vowels obtained from a larger corpus (cf. Campbell, Fant & Kruckenberg). Estimate rate of speaking from this.
- Determine place of articulation of adjacent consonants, discount F1 and F2 values to allow for their consonantally-caused perturbation.

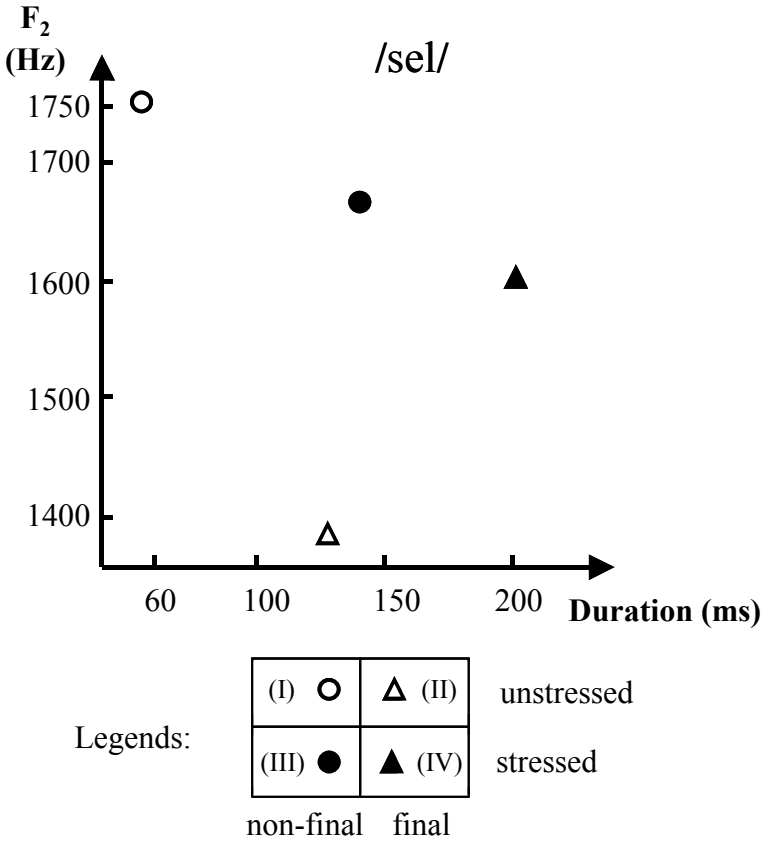

Figure 3 – Vowel quality as a function of stress level and word position

3.2. VOT variation as a function of the closeness of the following vowel.

Chang, [Chang 99], demonstrated that there is a positive correlation between VOT and the fall time of oral pressure (Po) and that VOT was greater in clusters like /tw-/ /tr-/ and greater before high close vowels /ti-/ /tu-/ than before more open vowels, /tæ/ /tɑ/.

Presumably a delay in the Po also delays the achievement of a sufficient transglottal pressure differential required to initiate voicing.

An interesting point is that English has distinctive aspiration word initially, but Japanese does not (Figure 4). The problem, then, is that the patterns are very similar for phonological and phonetic aspiration. Below are some hints for differentiating the two:

- Measure the duration of VOT.
- Measure the F1 of the following vowel or approximant.
- Discount a fraction of the VOT as a function of the inverse of F1.
- If adjusted VOT is greater than, say, 45msec, the stop is aspirated.

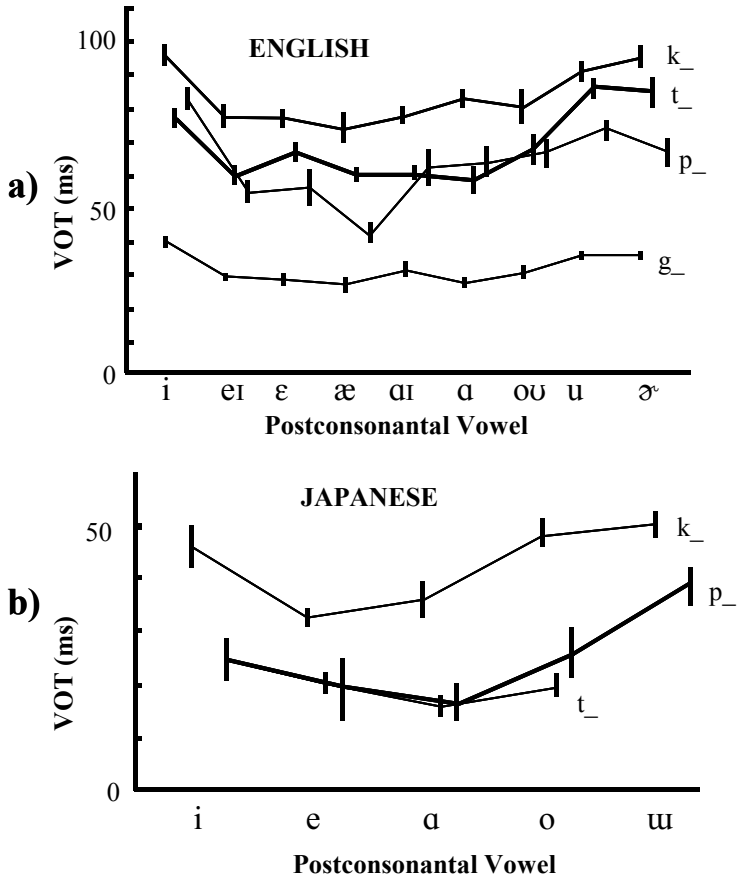

Figure 4 – VOT variation for stops as a function of following vowel in English (a) and Japanese (b).

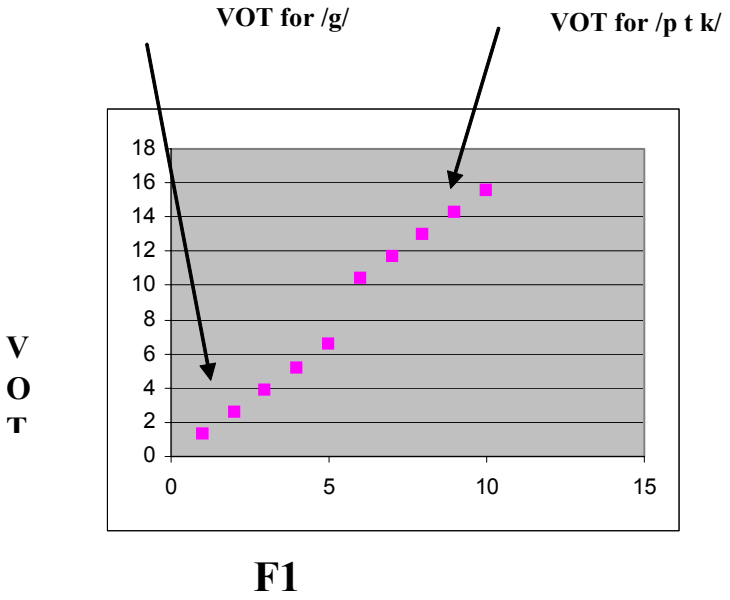

Figure 5, English stops VOT's variations plotted against F1.

However, a caveat is that the degree of aspiration also varies as a function of rate of speech. We have not taken this into consideration.

If the distribution of VOT in English were plotted against F1 (or some other measure of vowel closeness), one would presumably derive a discontinuous function similar to that shown in figure 5, from which one could deduce, correctly, that aspiration was phonological (distinctive) in English.

3.3. Complex clusters, or the problem of the so-called epenthetic stops.

Complex clusters might differentiate languages that have them from others that don't. But when is [mpθ] to be interpreted as /mpθ/ and when as /mθ/ ? We turn toward the problem of epenthetic stops (we prefer the term 'emergent'), well-known in historical phonology. Consider the following examples.

English	*Thompson*	<	*Thom* + *son*
English	*dempster*	<	*deem* + *ster* 'judge'
Sotho	*vontʃa*	<	* *voniʃa* 'see (caus.)'
Cl. Greek	*andros*	<	*anē ros* 'man'
French	*chambre*	<	Latin *kaměra* 'room'
Spanish	*alhambra*	<	Arabic *al hamra* 'the red'
Latin	*templum*	<	**tem* - *lo* 'a section'

The stops are anticipatory denasalization of the latter portion of the nasal under the influence of a following oral segment. But these emergent stops may also appear in words where they may not be phonological; they may be purely phonetic, as must have happened in an intermediate stage between Thompson < Thom + son, e.g.

English	youngster ['jʌŋkstə˞]	<	jʌŋ + stə˞
English	[wɔɻmpθ]	<	warm + θ

How can one distinguish a phonological stop from a phonetic stop ?

A possible answer is to compare durations of the segments adjacent to the stops in those cases where the phonological or phonetic character is clear.

Figure 6 shows audio and oral air pressure recordings sampled just behind the lips of a) the neologism sump+ster, b) the neologism sum+ster, and c) the existing word teamster. The three had stops whose status is : phonological, phonetic, and unknown, respectively. Using neologisms allows one to control whether the stop that appears in these clusters is phonological or phonetic. From a larger corpus using neologisms clamp+ster and clam+ster, the histograms (figure 7) of VN duration were obtained.

These duration differences allow us to differentiate phonological from phonetic stops in such clusters.

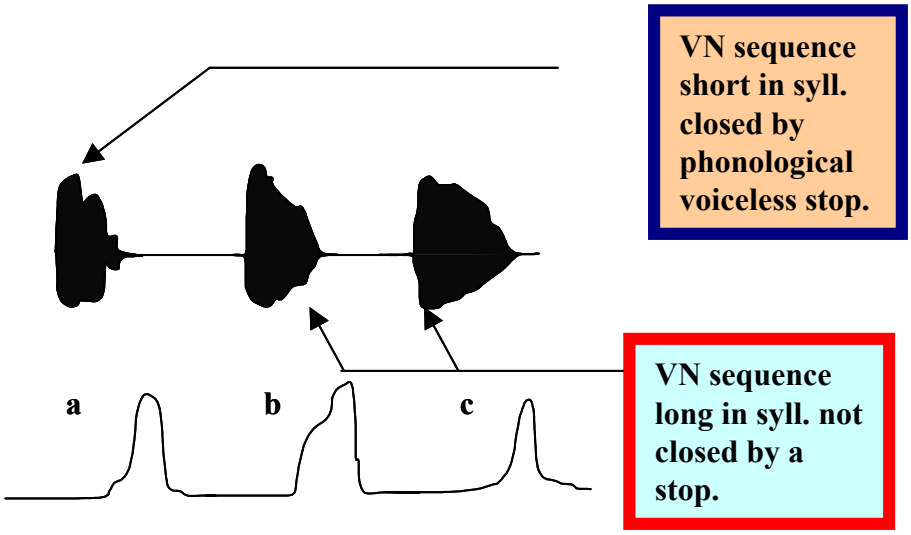

Figure 6, Schematic audio and oral air pressure recordings of VN sequences from different syllable structures.

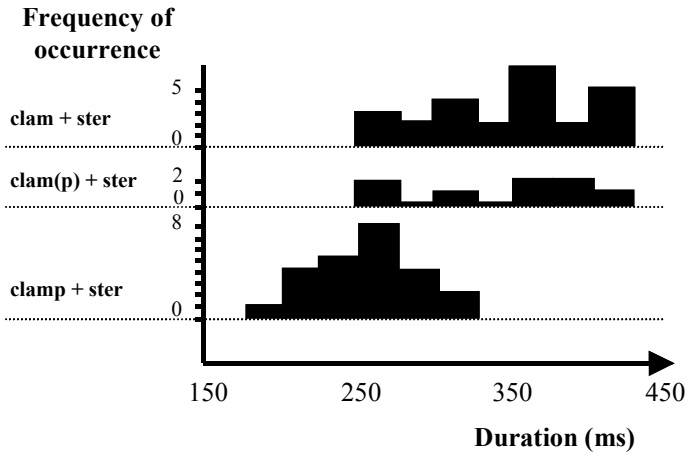

Figure 7 – Duration of VN sequences for neologisms (clamp+ster vs. clam+ster)

3.4. Vowel nasalization.

Both Spanish and American English show nasalization on vowels before nasal consonants, but that in English is temporally more extensive. How can one determine whether it is phonetic or phonological?

Figure 8 shows data taken from a study by Solé, [Solé92], on the duration of vowel nasalization before nasal consonants in /ta'Vn/ sequences as a function of speech rate in Iberian Spanish and American English (nasalization detected using the nasograph).

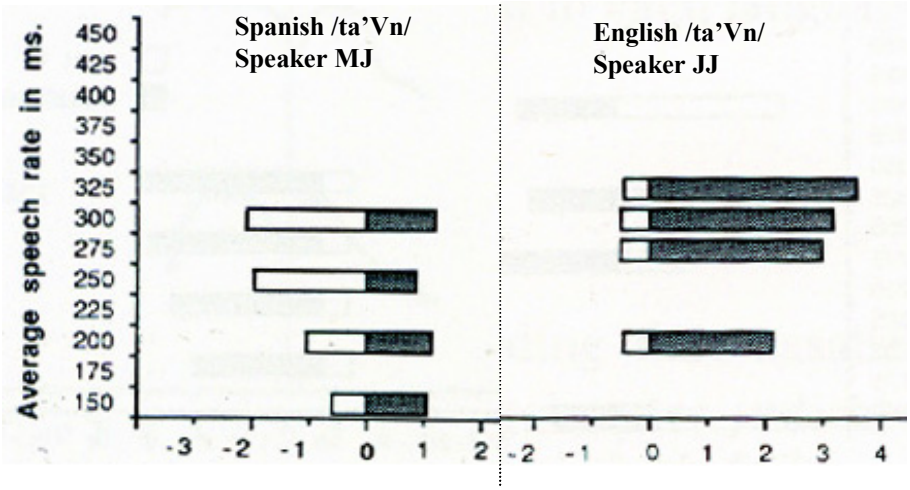

Figure 8 – Data from Solé, 1992, duration of vowel nasalization as a function of speech rate.

The data show that vowel nasalization in Spanish (filled bars) is nearly constant in duration over different rates, whereas in English the initial orality (open bars) is constant but the nasalization occupies nearly the whole vowel, no matter the speech rate. Solé's found a correlation between temporal spread of nasalization and another variable, vowel duration, thus undermining the notion that it would be correlated with degree of velic opening.

The conclusion then, is that vowel nasalization in English is phonological, whereas in Spanish it is phonetic.

3.5. Vowel duration differences before final voiced and voiceless consonants.

Vowels are longer before voiced than voiceless consonants in virtually all languages where this has been studied. The phonetic causes are not known. In English the ratio of this duration difference, estimated as 3:2 in stressed syllables, is larger than its ratio in other languages, estimated as 10:9. Can we establish the phonological, language-specific, character of this ΔV_{dur} ?

Caisse, [Caisse 88], claimed that under changes in rate of speech, duration differences that are phonological tend to vary in proportion to the rate change, and when they interact show a multiplicative effect whereas phonetic differences are constant over rate and show an additive effect when they interact. Thus, for English,

- phonological differences : ΔV_{dur} , "tense" vs. "lax" (/i/ vs. /I/), pre-pausal lengthening,
- phonetic differences : ΔV_{dur} before stops and fricatives.

A related study by Barnwell showed that even in English in minimal pairs such as retool [ɻi'tul] / reduel [ɻi'dul], the ΔV_{dur} of the two /i/ is much smaller than normal for English presumably because the /t/ and /d/ were in different syllables.

4. Conclusion

We have tried in this presentation to give some hints to help differentiate phonetic from phonological features. Though we showed several examples, there's a lot more to be done, and this task requires in particular, a detailed understanding of the factors influencing variation in speech, the need to take into account the phonetic context of the target feature, and accumulating a great deal of data which allows cross-language comparison of phonetic features.

5. Bibliography

[Busá 95] M.G. Busá & J. J. Ohala, "Nasal loss before voiceless fricatives : a perceptually-based sound change", *Rivista di Linguistica*, 7, pp 125-144, (1995)

[Caisse 88] Michelle Caisse, "Modeling english vowel duration", *Thesis*, University of California, Berkeley, (1988)

[Chang 99] S. Chang, "Vowel dependent VOT variation", *Proceedings of the , San Francisco*, (1999).

[Lindblom 63] B. Lindblom, "Spectrographic study of vowel reduction", *Journal of the Acoustic Society of America*, 35, pp1773-1781, (1963)

[Michailovsky 75] B. Michailovsky, "On some Tibeto-Burman sound changes", *Proceedings of the annual meeting of the Berkeley Linguistic Society*, 1, pp 322-332, (1975)

[Nord 74] L. Nord, "Vowel reduction-centralization or contextual assimilation", in *Preprints of the Speech Communication Seminar*, Stockholm Aug. 1-3, 1974, Vol. 2, Speech production and synthesis by rule, Stockholm : Speech transmission lab., KTH. pp. 149-154

[Ohala 75] J. Ohala, "Phonetic explanations for nasal sound patterns", in C.A. Ferguson, L.M. Hyman & J.J. Ohala (eds.), *Nasálfest : Papers from a symposium on nasals and nasalization*, Stanford : Language Universals Project, pp 289-316, (1975)

[Ohala 80] J. Ohala, "The application of phonological universals in speech pathology", in N.J. Lass (ed.), *Speech and Language. Advances in basic research and practice*, Vol. 3, New York : Academic Press, (1980)

[Ohala 93] J. Ohala, "Sound change as nature's speech perception experiment", *Speech Communication*, 13, pp 155-161, (1993b)

[Solé 92] M.J. Solé, "Phonetic and phonological processes: the case of nasalization", *Language and Speech*, 35.29-43, (1992)

[Widdison 91] K. Widdison, "The phonetic basis for s-aspiration in Spanish", Doc. Diss., University of California Berkeley (1991)

4ème Partie

Prosodie et identification

La discrimination des langues par la prosodie : Modélisation linguistique et études comportementales

Franck Ramus[1]

Laboratoire de Sciences Cognitives et Psycholinguistique
(EHESS/CNRS)

ramus@lscp.ehess.fr

Abstract

Spoken languages have been classified by linguists according to their rhythmic properties. Although researchers have measured many speech signal properties, they have failed to identify reliable acoustic characteristics for language classes. This paper presents instrumental measurements based on a consonant/vowel segmentation for eight languages. The measurements suggest that intuitive rhythm types reflect specific phonological properties, which in turn are signaled by the acoustic/phonetic properties of speech. The data support the notion of rhythm classes and also allow the simulation of language discrimination experiments with human subjects. Four such experiments are reported and establish the overall consistency of the model. Consequences for automatic language identification are considered.

Résumé

Les langues du monde ont été classées par les linguistes selon leurs propriétés rythmiques. Bien que de nombreuses mesures aient été effectuées sur le signal de parole, aucune n'a permis de rendre compte correctement des classes rythmiques de langues. Dans cet article nous proposons des mesures basées sur une segmentation de la parole en consonnes/voyelles effectuée en huit langues. Ces mesures suggèrent que les types de rythme reflètent des propriétés phonologiques précises, qui sont elles-mêmes détectables au niveau acoustique/phonétique. Nos données sont compatibles avec la notion de classes de rythme, et permettent la simulation d'expériences de discrimination de langues chez des sujets humains. Quatre expériences sont présentées qui renforcent la cohérence globale du modèle. Des applications à l'identification automatique des langues sont envisagées.

1. Introduction

Le rythme de la parole semble être un bon moyen de caractériser les langues du monde et de les classifier, au moins en un petit nombre de groupes. En effet, les linguistes ont

[1] Je remercie Marina Nespor et Jacques Mehler pour leur collaboration, la Délégation Générale pour l'Armement pour son soutien financier, et les participants de la 1ère journée sur l'identification automatique des langues pour une discussion intéressante de mes résultats.

traditionnellement distingué les langues accentuelles *(stress-timed*), englobant notamment les langues germaniques, slaves, ainsi que l'arabe, et les langues syllabiques *(syllable-timed*), comprenant les langues latines, ou encore le yoruba et le telegu [Abercrombie 1967, Pike 1945]. Un troisième groupe, les langues moraïques *(mora-timed*), comprenant le japonais ou le tamoul, a également été proposé [Ladefoged 1975]. Il était supposé que toutes les langues du monde avaient une organisation rythmique bien déterminée, appartenant à l'une de ces trois classes.

L'intuition derrière cette classification était que la production de la parole repose sur la répétition d'unités semblables, comme le pied, la syllabe ou la more, chaque langue utilisant un seul type d'unité, d'où l'existence de trois classes distinctes. Il était par ailleurs supposé que ces unités se répétaient à intervalles réguliers, les accents toniques étant régulièrement espacés.dans les langues accentuelles, et de même pour les syllabes dans les langues syllabiques et les mores dans les langues moraïques: c'est l'hypothèse *d'isochroni*e. Suivant cette hypothèse, il serait possible, en mesurant les durées séparant les accents, les syllabes ou les mores dans un échantillon d'une langue, de déterminer la classe rythmique de celle-ci. Malgré de nombreuses recherches, cette hypothèse n'a pas été validée empiriquement, les accents n'étant pas plus régulièrement espacés dans les langues accentuelles que dans les langues syllabiques, ni vice versa pour les syllabes [Bolinger 1965, Dauer 1983, Roach 1982]. A ce stade, la caractérisation rythmique des langues est donc assez incertaine. Dans ce qui suit nous proposons une nouvelle approche de cette caractérisation.

2. Corrélats du rythme dans le signal de parole[2]

2.1. Bases phonologiques du rythme

Notre approche repose sur une conception du rythme de parole, non plus comme primitive de l'organisation temporelle des langues, mais comme conséquence de certaines de leurs propriétés phonologiques [Bertinetto 1981, Dasher 1982, Dauer 1983], notamment: la complexité des syllabes, la corrélation entre poids syllabique et accent, la présence ou non de réduction vocalique... Selon cette conception, les langues dites syllabiques sont des langues n'autorisant que des syllabes simples et n'admettant pas de réduction vocalique. Les syllabes sont donc de taille relativement stable, donnant ainsi l'impression d'un rythme syllabique régulier. Dans les langues accentuelles, au contraire, des syllabes complexes sont autorisées, et celles-ci portent en général l'accent tonique. Les syllabes plus simples, en revanche, ne sont pas accentuées, et font au contraire l'objet d'une réduction vocalique, accentuant le contraste entre les syllabes fortes et les syllabes faibles, ce qui induit un rythme syllabique moins régulier, porté par les seules syllabes accentuées.

2.2. Etude empirique du rythme en 8 langues

L'approche phonologique du rythme de parole prédit qu'une analyse de la complexité syllabique d'une langue devrait permettre de déterminer sa classe rythmique. C'est ce que nous nous proposons de tester, en mesurant la complexité syllabique par le biais d'une segmentation de la parole en consonnes/voyelles.

[2] Une partie des données présentées dans cette section sont tirées de [Ramus 1999b].

Le matériel utilisé a été extrait du corpus multilingue du LSCP [Nazzi 1997]. Huit langues (Anglais, Néerlandais, Polonais, Français, Espagnol, Italien, Catalan, Japonais) et quatre locutrices natives de chaque langue ont été choisies. Un corpus de 160 phrases a été constitué, 20 par langue, 5 par locutrice. Les phrases ont été sélectionnées pour avoir un nombre de syllabes et des durées comparables à travers les langues. Les phonèmes de chaque phrase ont été marqués manuellement et alignés avec le signal de parole, puis classifiés en consonne ou voyelle[3]. Afin de mesurer plus directement la complexité syllabique, nous ne nous sommes pas intéressés aux durées des phonèmes individuels, mais aux durées des intervalles vocaliques (du début à la fin d'une séquence de voyelles) et consonantiques (du début à la fin d'une séquence de consonnes). La complexité syllabique est donc capturée notamment par la durée des groupes consonantiques.

De ces durées nous avons dérivé 3 variables prenant une valeur par phrase:

- %V la proportion (en durée) d'intervalles vocaliques dans la phrase[4];
- ΔV l'écart type des durées d'intervalles vocaliques par phrase;
- ΔC l'écart type des durées d'intervalles consonantiques.

La figure 1 donne les valeurs moyennes de %V, ΔC et ΔV par langue, les barres d'erreur représentant l'erreur standard de la moyenne.

Le plan (%V, ΔC) fait ressortir clairement 3 groupes, et il s'avère que ces groupes correspondent aux classes rythmiques décrites dans la littérature : Anglais, Néerlandais et Polonais pour les langues accentuelles, Espagnol, Italien, Français et Catalan pour les langues syllabiques, et Japonais pour les langues moraïques. Une ANOVA introduisant un facteur « classe de rythme » montre un effet de classe significatif à la fois pour %V et ΔC ($p<0.001$). Des comparaisons multiples (Tukey test) montrent par ailleurs que chaque classe est significativement différente des deux autres, pour %V ($p<0.001$) et ΔC ($p<0.005$).

Il semble donc que les variables %V et ΔC capturent l'essentiel des propriétés rythmiques de ces langues, et cela peut s'interpréter par les propriétés phonologiques mentionnées ci-dessus. En effet, plus une langue admet des syllabes complexes, plus elle admet des groupes de consonnes de taille importante, donc plus le rapport consonne/voyelle augmente. En conséquence les langues accentuelles ont un faible %V. De plus, les langues admettant des syllabes complexes admettent aussi les syllabes plus simples, ce qui implique que les groupes consonantiques peuvent être de taille plus variable que dans les langues n'admettant que des syllabes simples. Les langues accentuelles ont donc un ΔC plus élevé que les langues syllabiques. Ainsi que le montre le plan (%V, ΔV), la variable ΔV semble moins directement liée aux classes de rythme (pas d'effet de classe significatif). Par rapport à ΔC, ΔV apporte un élément supplémentaire d'information, suggérant que le Polonais a des différences importantes avec les autres langues accentuelles. De fait, le statut rythmique du Polonais ne fait pas l'unanimité, certains auteurs le considérant comme une langue syllabique ou encore intermédiaire [Nespor 1990, Rubach 1985]. Ce désaccord est dû à l'absence de réduction

[3] Les semi-voyelles pré- ou inter-vocaliques ont été considérées comme des consonnes, les semi-voyelles post-vocaliques comme des voyelles.

[4] La proportion d'intervalles consonantiques est égale à 1-%V et n'apporte donc pas d'information supplémentaire.

vocalique en Polonais, alors que cette propriété est généralement associée aux langues accentuelles. Cela se traduit dans nos données par un plus faible ΔV pour le Polonais (moins de variations des durées de voyelles).

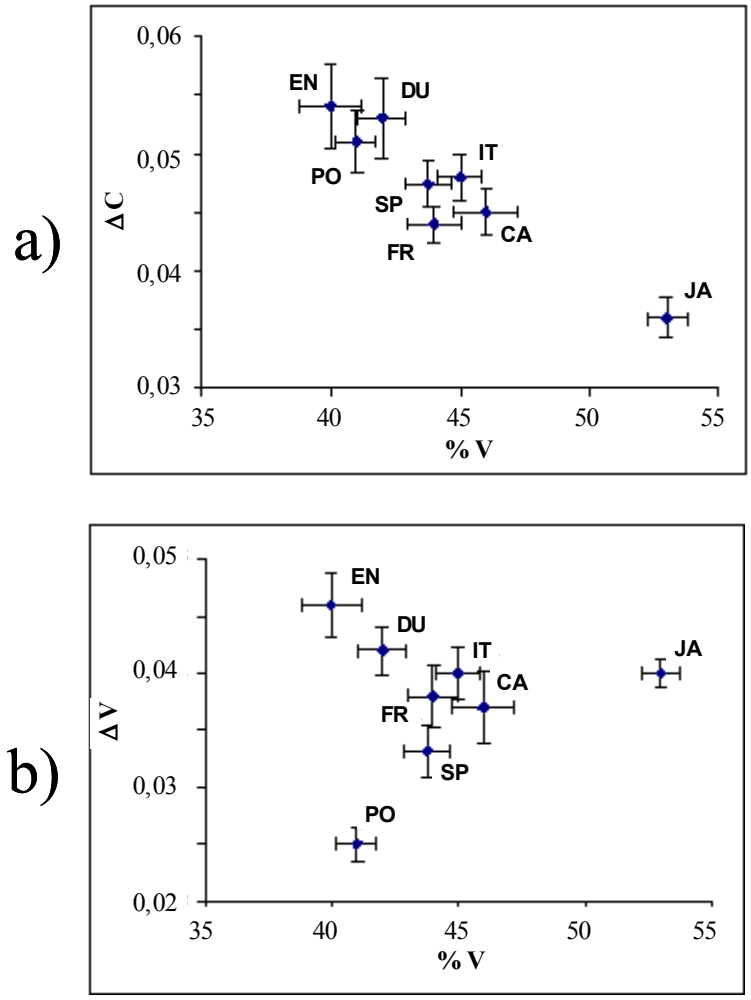

Figure 1 – Projection des huit langues dans les plans (%V, ΔC – Figure 1a.) et (%V, ΔV – Figure 1b)

2.3. Analyse discriminante

Jusqu'ici nous avons montré que, malgré l'échec des recherches sur l'isochronie, il est possible de trouver dans le signal de parole des corrélats phonétiques des classes rythmiques. On peut maintenant se demander si, étant donné une phrase, il est possible de déterminer de manière fiable la classe rythmique de la langue en question. Dans ce but nous avons effectué une analyse discriminante avec cross-validation. Dans cette procédure, chaque phrase est classifiée à l'aide de fonctions discriminantes calculées sur l'ensemble des données sauf la phrase en question. Ici nous utilisons exclusivement les variables %V et ΔC, qui sont les mieux

reliées aux classes de rythme. Les pourcentages de classifications sont présentés dans le tableau 1, le niveau du hasard étant 33%.

	Classe prédite		
classe réelle	moraïque	syllabique	accentuelle
moraïque	95 %	5 %	0 %
syllabique	13,8 %	53,8 %	32,5 %
accentuelle	5 %	31,7 %	63,3 %

Tableau 1 – Classification par fonctions discriminantes appliquées aux variables %V et ΔC.

Ainsi, il est possible, étant donnée une seule phrase (environ 3 s.) d'une langue indéterminée, de prédire avec une bonne fiabilité la classe rythmique à laquelle elle appartient. Afin de mieux apprécier la valeur absolue des scores rapportés ici, notons que :

- Les modèles de classes de langue sont calculés à partir de 20 phrases (environ 60 secondes) par langue, ce qui est peu. Un corpus plus important permettrait l'élaboration de meilleurs modèles et donc une meilleure classification.
- Les échantillons de parole testés sont relativement courts (une phrase, en moyenne 3 secondes). L'utilisation d'échantillons plus longs permettrait une évaluation plus précise de %V et ΔC pour chaque échantillon et améliorerait donc également la classification. Le même type d'analyse peut être restreint à des paires de langues et ainsi prédire si la discrimination est possible pour une paire donnée. Le tableau 2 donne les résultats de classification correcte prédits par l'analyse discriminante pour les paires de langues testées dans la section suivante (hasard = 50%).

Paire de langues	**Variables**	**Pourcentage de phrases correctement classifiées**
Anglais/Japonais	%V, ΔC	95%
Anglais/Espagnol	%V, ΔC	62,5%
Polonais/Espagnol	%V, ΔC	62,5%
Anglais/Polonais	%V, ΔC	40%
Anglais/Polonais	%V, ΔV	90%

Tableau 2 – Classification correcte obtenue pour des paires de langues par analyse discriminante.

Comme la figure 1a le suggère, les variables %V et ΔC prédisent une discrimination lorsque les langues appartiennent à 2 classes différentes, mais pas pour 2 langues de la même classe (Anglais/Polonais). Néanmoins, si l'on incorpore ΔV dans le modèle, l'Anglais et le Polonais deviennent alors discriminables. Dans un but de discrimination automatique, ΔV apporte donc une information supplémentaire. Dans un but de modélisation de résultats comportementaux, l'intérêt de ΔV reste une question empirique, que l'on peut aborder en testant la discrimination Anglais/Polonais.

3. Résultats comportementaux de discrimination de langues

Les données décrites ci-dessus prédisent qu'il devrait être facile de distinguer des langues appartenant à des classes rythmiques différentes, et difficile de distinguer des langues appartenant à la même classe, si toutefois l'on ne dispose que du rythme. C'est également l'hypothèse faite par [Mehler 1996] pour le nouveau-né, étant supposé que ce dernier est particulièrement sensible à la prosodie. Cette hypothèse a été étayée par de nombreuses expériences chez le nouveau-né, montrant que celui-ci pouvait distinguer le Français du Russe, l'Anglais de l'Italien, de l'Espagnol ou du Japonais, mais pas l'Anglais du Néerlandais [Mehler 1988, Moon 1993, Nazzi 1998] Il a même été montré que la discrimination était possible au niveau de la classe, des nouveau-nés discriminant un mélange d'Anglais et Néerlandais contre un mélange d'Espagnol et d'Italien, mais pas Anglais + Italien contre Néerlandais + Espagnol [Nazzi 1998]. La plupart des expériences sus-citées ayant été effectuées ou répliquées avec de la parole filtrée, il semble que le nouveau-né discrimine bien à l'aide des propriétés suprasegmentales de la parole, et donc probablement du rythme. Cependant, le filtrage de parole n'est pas idéal pour étudier le rythme dans la mesure où il préserve également certaines informations phonétiques, de même que la fréquence fondamentale, donc l'essentiel de l'intonation.

3.1. Stimuli

La resynthèse est une technique plus souple pour délexicaliser la parole, car elle permet d'en manipuler les propriétés phonétiques, phonotactiques ou prosodiques [voir Ramus 1999a pour de plus amples détails]. Ici les stimuli ont été resynthétisés de manière à n'en préserver que le rythme. Les phrases d'origine sont celles décrites au 2.2. Leur segmentation en consonne/voyelle a été utilisée pour les resynthétiser en remplaçant toutes les consonnes par /s/ et toutes les voyelles par /a/ [cf. Dutoit 1996 pour des informations sur le logiciel de synthèse MBROLA]. Les stimuli étaient donc des phrases de type /sasasasas…/ ayant précisément les mêmes intervalles vocaliques et consonantiques que les phrases d'origine, et une fréquence fondamentale constante à 230 Hz. Ainsi, seul le rythme des phrases d'origine était préservé[5].

3.2. Procédure

La discrimination des paires de langues suivantes a été testée : Anglais/Japonais, Anglais/Espagnol, Polonais/Espagnol, ainsi que Anglais/Polonais. Chaque paire de langue a fait l'objet d'une expérience sur 16 sujets pour la plupart de langue française. Pour la paire Anglais/Japonais, il s'agissait d'une tâche de catégorisation de langue avec entraînement et feedback [voir Ramus 1999]. Pour les autres paires, il s'agissait d'une tâche de détection d'intrus AAX: un essai consistait en 2 phrases de familiarisation dans la même langue, puis une phrase dans une langue indéterminée, dont le sujet devait dire s'il elle était ou non de la même langue que les 2 premières. Du feedback était fourni après chaque réponse. Les essais étaient regroupés en 2 blocs pendant lesquels la langue de familiarisation était maintenue constante. Les expériences ont été programmées à l'aide du langage EXPE [Pallier 1997].

[5] Des échantillons des types de stimuli utilisés peuvent être écoutés à l'adresse Internet suivante : http://www.ehess.fr/centres/lscp/persons/ramus/resynth/ecoute.htm

L'analyse du paragraphe 2.3 prédit que les trois premières paires peuvent être distinguées sur la base du rythme. Le résultat de la quatrième dépend des variables que l'on incorpore dans le modèle : si l'on suppose que seules %V et ΔC modélisent la perception du rythme, alors l'Anglais et le Polonais devraient être confondus ; si l'on suppose que ΔV joue également un rôle, alors les 2 langues devraient être discriminées.

3.3. Résultats

Les indices de discrimination A' [Snodgrass 1988] pour les phases de test ainsi que la significativité des comparaisons (test T) par rapport au niveau de chance (50%) sont présentés dans le tableau 3.

Paire de langues	A'	Significativité
Anglais/Japonais	0.72	p<0.001
Anglais/Espagnol	0.65	p<0.001
Polonais/Espagnol	0.74	p<0.001
Anglais/Polonais	0.61	p=0.006

Tableau 3 – Résultats de significativité des tests de discrimination.

Conformément aux prédictions, le rythme de l'Anglais est discriminé de celui de l'Espagnol et du Japonais, de même que pour l'Espagnol et le Polonais. De plus l'Anglais et le Polonais sont également discriminés, ce qui suggère que ΔV aussi reflète des propriétés pertinentes pour les sujets.

4. Discussion et application à l'identification automatique

Partant de l'hypothèse que le rythme de parole a pour source des propriétés phonologiques, nous avons montré que ces propriétés ont des corrélats phonétiques qui peuvent être mesurés directement sur le signal de parole. Ainsi, partant du signal, il est possible de retrouver la typologie rythmique des langues et de prédire quelles paires peuvent être distinguées sur la base du rythme.

On peut maintenant envisager un certain nombre d'améliorations et d'extensions au modèle actuel :

- Augmenter la taille du corpus de base, et ce dans plusieurs directions :
  - Augmenter la taille du corpus pour chaque langue ;
  - Augmenter le nombre de langues ;
  - Etendre le corpus à de la parole spontanée ;
    Ces trois extensions, qui permettraient d'asseoir la validité empirique et l'intérêt théorique du modèle, ne peuvent s'envisager raisonnablement dans le cadre d'une segmentation manuelle du signal, ce qui nous amène au point suivant.
- Etendre le corpus à de la parole segmentée automatiquement. En principe il devrait s'agir d'une segmentation indépendante de la langue et donnant des frontières de phonèmes robustes, ce qui peut être problématique pour une segmentation phonétique détaillée.

Néanmoins le présent modèle ne requiert qu'une segmentation en consonnes/voyelles, ce qui rend le problème sensiblement plus simple [voir Corredor-Ardoy 1997 et Adda-Decker dans ce volume pour une approche de la segmentation automatique multilingue].

Ainsi, l'extension du corpus par segmentation automatique permettrait 1) de tester à plus grande échelle la validité empirique du présent modèle, 2) d'augmenter sa plausibilité psychologique, étant entendu qu'il ne peut contribuer à modéliser la perception du rythme par l'humain que si ce dernier peut disposer effectivement d'un algorithme effectuant la segmentation en consonnes/voyelles, 3) d'améliorer les modèles de langues déterminés ici par analyse discriminante, 4) d'explorer les propriétés rythmiques de langues jusqu'à présent peu étudiées.

On peut à partir de là envisager l'élaboration d'un système automatique pour reconnaître le rythme des langues. Un tel système devrait comporter plusieurs niveaux :

- Extraction des silences de la parole spontanée et normalisation pour le débit du locuteur (important pour ΔV et ΔC) ;
- Segmentation automatique de la parole en consonnes/voyelles.
- Exploitation des durées des intervalles vocaliques et consonantiques mesurés pour déterminer la classe rythmique de l'échantillon par rapport à des modèles de classes de.langues prédéterminés [voir Dominey 2000 et Dominey dans ce volume pour une implémentation de ce niveau par des réseaux de neurones].

Il va sans dire que cette reconnaissance du rythme ne permettra pas d'identifier des langues au sein des classes de rythme, mais pourrait néanmoins réduire l'espace de recherche pour des systèmes d'identification des langues se basant sur d'autres approches (acoustiques, phonétiques ou phonotactiques).

Il reste à déterminer si la caractérisation rythmique d'un échantillon de parole peut réellement apporter des informations qui ne sont pas déjà obtenues par les autres approches. Par exemple, on peut penser que les informations de type phonotactique, notamment les probabilités de transition entre consonnes et voyelles, incluent de fait les informations rythmiques liées à la complexité des syllabes et aux groupes consonantiques. Néanmoins, le rythme de parole n'est peut-être pas réductible aux probabilités de transition entre consonnes et voyelles, notamment parce que les informations de durée ne sont pas prises en compte actuellement dans les modèles phonotactiques. Les données ci-dessus en donnent un exemple: l'Anglais et le Polonais ont la même complexité syllabique, mais diffèrent par la présence (en anglais) ou non (en polonais) de réduction vocalique, ce qui se traduit dans nos données par un ΔV plus élevé pour l'Anglais. Il s'agit donc là d'une différence rythmique reposant strictement sur la durée des voyelles et donc non accessible aux modèles phonotactiques.

Dans le cas particulier de l'Anglais et du Polonais, on pourrait bien sûr argumenter que la discrimination est néanmoins possible sur la base de différences purement phonétiques, mais cela n'est pas nécessairement vrai de toutes les paires comportant des différences rythmiques.

De plus, les modèles phonétiques peuvent manquer de robustesse sur de courts échantillons de parole, car un échantillon court ne comporte qu'un ensemble réduit de phonèmes et donc sous-détermine grandement le répertoire phonétique de la langue. La robustesse des caractéristiques rythmiques que nous avons montrée sur de courts échantillons nous laisse donc penser qu'elles peuvent apporter de l'information significative au moins dans ce cas précis.

5. Bibliographie

[Abercrombie 1967] D. Abercrombie, *Elements of general phonetic*s. Chicago: Aldine, (1967).

[Bertinetto 1981] P. M. Bertinetto, *Strutture prosodiche dell' italiano. Accento, quantità, sillaba, giuntura, fondamenti metric*i. Firenze: Accademia della Crusca, (1981).

[Bolinger 1965] D. Bolinger, "Pitch accent and sentence rhythm," in *Forms of English: Accent, morpheme, orde*r. Cambridge, MA: Harvard University Press, (1965).

[Corredor-Ardoy 1997] C. Corredor-Ardoy, J. L. Gauvain, M. Adda-Decker, & L. Lamel, "Language identification with language-independent acoustic models," presented at Eurospeech, Rhodes, (1997).

[Dasher 1982] R. Dasher & D. Bolinger, "On pre-accentual lengthening," *Journal of the International Phonetic Associatio*n, vol. 12, pp. 58-69, (1982).

[Dauer 1983] R. M. Dauer, "Stress-timing and syllable-timing reanalyzed," *Journal of Phonetic*s, vol. 11, pp. 51-62, (1983).

[Dominey 2000] P. Dominey & F. Ramus, "Neural network processing of natural language: I. Sensitivity to serial, temporal and abstract structure in the infant," *Language and Cognitive Processes*, vol. 15, pp. 87-127, (2000).

[Dutoit 1996] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, & O. van der Vrecken, "The MBROLA Project: Towards a set of high-quality speech synthesizers free of use for non-commercial purposes," presented at ICSLP'96, Philadelphia, (1996).

[Ladefoged 1975] P. Ladefoged, *A course in phonetic*s. New York: Harcourt Brace Jovanovich, (1975).

[Mehler 1996] J. Mehler, E. Dupoux, T. Nazzi, & G. Dehaene-Lambertz, "Coping with linguistic diversity: The infant's viewpoint," in *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisitio*n, J. L. Morgan & K. Demuth, Eds. Mahwah, NJ: Lawrence Erlbaum Associates, (1996), pp. 101-116.

[Mehler 1988] J. Mehler, P. Jusczyk, G. Lambertz, N. Halsted, J. Bertoncini, & C. Amiel-Tison, "A precursor of language acquisition in young infants," *Cognitio*n, vol. 29, pp. 143-178, (1988).

[Moon 1993] C. Moon, R. P. Cooper, & W. P. Fifer, "Two-day-olds prefer their native language," *Infant Behavior and Developmen*t, vol. 16, pp. 495-500, (1993).

[Nazzi 1997] T. Nazzi, "Du rythme dans l'acquisition et le traitement de la parole.", Paris: Ecole des Hautes Etudes en Sciences Sociales, (1997).

[Nazzi 1998] T. Nazzi, J. Bertoncini, & J. Mehler, "Language discrimination by newborns: towards an understanding of the role of rhythm," *Journal of Experimental Psychology: Human Perception and Performanc*e, vol. 24, pp. 756-766, (1998).

[Nespor 1990] M. Nespor, “On the rhythm parameter in phonology,” in *Logical issues in language acquisitio*n, I. M. Roca, Ed. Dordrecht: Foris, (1990), pp. 157-175.

[Pallier 1997] C. Pallier, E. Dupoux, & X. Jeannin, “EXPE: An expandable programming language for on-line psychological experiments,” *Behavior Research Methods, Instruments, & Computer*s, vol. 29, pp. 322-327, (1997).

[Pike 1945] K. L. Pike, *The intonation of American Englis*h. Ann Arbor, Michigan: University of Michigan Press, (1945).

[Ramus 1999a] F. Ramus & J. Mehler, “Language identification with suprasegmental cues: A study based on speech resynthesis,” *Journal of the Acoustical Society of Americ*a, vol. 105, pp. 512-521, (1999).

[Ramus 1999b] F. Ramus, M. Nespor, & J. Mehler, “Correlates of linguistic rhythm in the speech sign*al,” *Cognition*, vol. 73, pp. 265-292, (1999).

[Roach 1982] P. Roach, “On the distinction between "stress-timed" and "syllable-timed" languages,” in *Linguistic controversie*s, D. Crystal, Ed. London: Edward Arnold, (1982).

[Rubach 1985] J. Rubach & G. E. Booij, “A grid theory of stress in Polish,” *Lingu*a, vol. 66, pp. 281-319, (1985).

[Snodgrass 1988] J. G. Snodgrass & J. Corwin, “Pragmatics of measuring recognition memory: Applications to dementia and amnesia,” *Journal of Experimental Psychology: Genera*l, vol. 117, pp. 34-50, (1988).

A Neural Network Model of Language Classification Based on Prosodic Structure

Peter Ford Dominey & Franck Ramus

Institut des Sciences Cognitives, CNRS – UPR 9075
67, Blvd Pinel, 69675 BRON Cedex, France
Laboratoire de Sciences Cognitives et Psycholinguistique, CNRS/EHESS, 54 Blvd Raspail, 75006 Paris, France

dominey@lyon151.inserm.fr – ramus@lscp.ehess.fr

Abstract

Human infants are sensitive at birth to the contrasting rhythms or prosodic structures of languages, that can serve to bootstrap acquisition of grammatical structure. We present a novel recurrent network architecture that simulates this sensitivity to different temporal structures. Recurrent connections in the network are non-modifiable, while forward connections from the recurrent network to the output layer are modified by a simple reinforcement rule. This avoids recurrent credit assignment complexity, and provides a flexible system for exploring the effects of temporal structure. The network is exposed to human speech that has been processed to preserve only the temporal component of prosodic structure. The network is trained to categorize individual sentences by their rhythm class, and can then generalize this learning to new sentences. These results demonstrate 1) a recurrent sequence learning architecture that is capable of learning and generalization of temporal structure, and 2) a neurophysiologically plausible mechanism by which human infants could extract the prosodic structure from natural language.

1. Introduction

In behavioral sequences including language, skilled motor control, dance, music etc. the temporal organization of the sequence is of nearly the same importance as the serial order of events, and the two are quite often well correlated. In human language the correlation between the temporal structure and phonological, morphological and syntactic structure is likely exploited by the infant to help bootstrap language acquisition [Mehler 96, Nespor 96]. Indeed, different languages have different global rhythmic or temporal structure (a major aspect of prosody), and the early ability to detect these differences provides a substantial advantage in reducing the possible degrees of freedom for the syntactic structure of the language [Nespor 96]. Already within the first days after birth, human infants are capable of discriminating between unfamiliar languages from different rhythm classes (eg. stress-timed, syllable-timed, mora-timed) based on prosodic (temporal) structure [Nazzi 98]. In these experiments sentences that have been filtered to preserve only the temporal structure are presented out loud to the infants, and the behavioral measure is the rate of sucking on a pacifier. After habituation to sentences in one rhythm class, discrimination is observed as an increase in sucking rate when sentences in a different rhythm class are presented in the test phase. In contrast, no change is observed in the control groups when sentences from the same rhythm class (but different speakers) are

presented in the test phase. In [Nazzi 98] infants discriminate between stress-timed English and mora-timed Japanese (Exp 1), but fail to discriminate between stress-timed English and stress-timed Dutch (Experiment 2). In Exp 3, infants heard different combinations of sentences from English, Dutch (stress-timed), Spanish and Italian (syllable-timed). Discrimination was observed only when English and Dutch sentences were contrasted with Spanish and Italian sentences. These results demonstrate a general capability to classify language based on prosodic structure.

The question remains, what is the mechanism that permits this sensitivity to temporal prosodic structure? Answering this question will provide an important step in establishing how prosodic structure is used to bootstrap the acquisition of morphological and syntactic structure [Nespor 96]. We have previously demonstrated that a recurrent reinforcement network based on the neuroanatomy of the primate frontostriatal system [Dominey 95B] can learn both the serial and the temporal structure of sensorimotor sequences [Dominey 98a, Dominey 98b]. In this paper we attempt to determine if this model can be trained to classify sentences from languages of different rhythm classes, and to then generalize this trained ability to new sentences, thus demonstrating a functional capability equivalent to that of human infants.

2. Recurrent Associative Network

The sequence learning capacity of recurrent networks has been demonstrated in a variety of settings where the encoding of previous states or events allows prediction of future events [e.g. Elman 90, Jordan 90, Cleeremans 91]. In a number of recurrent network studies, learning modifies the recurrent connections. This poses the interesting technical challenge of how to keep track of the effect of a given connection over several cycles that precede the final result [Almeida 87, Pineda 89, Pearlmutter 95]. The current paper presents a recurrent model that avoids this problem by maintaining all connections into and within the recurrent network fixed. The recurrent network thus provides a form of state transition mechanism, and learning allows the association of states with appropriate responses via a simple reinforcement learning rule [Dominey 95B, Dominey 95]. This approach is novel in that is quite simple, and it also provides a very convenient tool for investigating the encoding of serial as well as temporal structure.

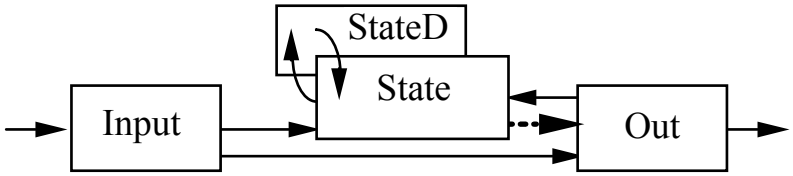

Figure 1. Sequence learning model based on recurrent state representation and associative reinforcement learning. Each of the structures are 5x5 arrays of leaky-integrator neurons. Sequence elements are presented as activation of single units in the Input array. Responses are generated in Out. Out units are influenced by Input, and also by modifiable, non-topographic projections from State that form the associative memory. State is a recurrent network that encodes sequence state as a function of visual input from Input, response copy from Out, and recurrent self input from $State_D$. This network, State, generates a time varying sequence of internal states. These states become associated with Out activity for the successive responses by a reinforcement learning mechanism that modifies State-Out connections. Implemented in Neural Simulation Language [Weitzenfeld 91].

The recurrent network architecture described in Figure 1 is similar to previous recurrent models [Almeida 87, Elman 90, Jordan 90, Pearlmutter 95, Pineda 89] with three important differences. First, there is no learning in the recurrent connections (i.e. those that project from $State_D$ to State), only between the State units and the Out units. Second, learning is based on a simple reinforcement mechanism rather than back-propagation of error, or related error-gradient calculation methods [Pearlmutter 95]. Third, in the temporal domain, a) the computing elements are leaky integrators, and b) simulation time steps are not tightly coupled to input, output and learning processing. Indeed, the experimenter's capability to specify the time delays between external events is an integral part of this model.

2.1. Recurrent State Representation

Equations (1.1) and (1.2) describe how State is influenced by external inputs from Input, responses from Out, and a recurrent inputs from $State_D$. In (1.1) the leaky integrator, s(), corresponding to the membrane potential or internal activation of State is described. In (1.2) the output activity level of State is generated as a sigmoid function, f(), of s(t). The term t is the time, Δt is the simulation time step, τ is the leaky integrator time constant. As τ increases with respect to Δt, the charge and discharge times for the leaky integrator increase. The unit of time in the simulations is referred to as a simulation time step or sts, and corresponds to a single update cycle of the simulation.

$$s_i(t+\Delta t) = (1-\frac{\Delta t}{\tau}) + s_i(t) + \frac{\Delta t}{\tau}(\sum_{j=1}^{n} w_{ij}^{IS} Input_j(t) + \sum_{j=1}^{n} w_{ij}^{SS} StateD_j(t) + \sum_{j=1}^{n} w_{ij}^{OS} Out_j(t)) \quad (1.1)$$

$$State(t) = f(s(t)) \quad (1.2)$$

The connections w^{IS}, w^{SS} and w^{OS} define the projections from units in Input, $State_D$, and Out to State. These connections are one-to-all, and are mixed excitatory and inhibitory, and do not change with learning. This mix of excitatory and inhibitory connections ensures that the State network does not become saturated by excitatory inputs, and also provides a source of diversity in coding the conjunctions and disjunctions of input, output and previous state information.

Recurrent input to State originates from the layer $State_D$. $State_D$ (Equation 2.1 and 2.2) receives input from State, and its 25 leaky integrator neurons have a distribution of time constants from 20 to 400 simulation time steps, while State units have time constants of 2 simulation time steps. This distribution of time constants in $State_D$ yields a range of temporal sensitivity similar to that provided by using a distribution of temporal delays [Kühn 92].

$$sd_i(t+\Delta t) = (1-\frac{\Delta t}{\tau})sd_i(t) + \frac{\Delta t}{\tau}(State_i(t)) \quad (2.1)$$

$$State_D = f(sd(t)) \quad (2.2)$$

2.2. Associative Memory

In the current study after a sequence is presented it must then be “classified” by a response generated in Out corresponding either to language L1 or L2. During learning, for each correct response, the pattern of activity in State at the time of the response becomes linked, via

reinforcement learning in a simple associative memory, to the responding element in Out. The required associative memory is implemented in a set of modifiable connections (w^{SO}) between State and Out, described in equation (3). When a response is evaluated, the connections between units encoding the current state in State, and unit encoding the current response in Out are strengthened as a function of their rate of activation and learning rate R. R is positive for correct responses and negative for incorrect responses. Weights are normalized to preserve the total synaptic output weight of each State unit.

$$w^{SO}_{ij}(t+1) = w^{SO}_{ij}(t) + R^* \text{State}_i * \text{Out}_j \qquad (3)$$

The network output is thus directly influenced by the Input, and also by State, via learning in the w^{SO} synapses, as described in Equations (4.1) and (4.2). The model explains primate cortical sequence encoding [Dominey 95B) and learns complex sequences in reproduction [Dominey 95] and serial reaction time tasks [Dominey 98a,b].

$$o_i(t+\Delta t) = (1 - \frac{\Delta t}{\tau})\, o_i(t) + \frac{\Delta t}{\tau}\,(\text{Input}_i(t) + \sum_{j=1}^{n} w^{SS}_{ij}\text{State}_i(t)) \qquad (4.1)$$

$$\text{Out} = f(o(t)) \qquad (4.2)$$

3. Temporal Discrimination

The current study is based on the hypothesis that speech rhythm processing relies on extracting regularities from the durations of vocalic and consonantal intervals. This hypothesis is supported both by measurements on the speech signal in various languages [Ramus 2000b] and by perceptual experiments using resynthesized reiterant speech [Ramus 2000a]. Thus the simulations took as input speech that was pre-segmented into consonant (C) and vowel (V) durations, sampled and coded at 5 ms intervals. The main idea is that presentation of C-V sequences with different temporal structures should result in different vectors of activity in the 25 State neurons, that can be associated, by learning, with the correct classification responses. We first exposed the network to 10 3-second sentences from each of 5 languages, and recorded the State vector resulting from each of these sentences. We then tested whether these State vectors could be correlated with the languages from which they originated. Table 1 (Correlation) indicates that for the pairs in which the discrimination required was coherent (i.e. languages from different rhythm classes), there was a significant correlation between State vector activity and language, and the opposite for incoherent discriminations (i.e. languages from the same rhythm class).

	Languages	Coherence	[NAZZI 98]	Correlation	Performance
Expe 1	Eng vs Jap	Coherent	p < .01	R^2 = .75 p < .001	78% p < .001
Expe 2	Eng vs Dut	InCoherent	p = .16	R^2 = .17 p = 0.87	52% p = .86
Expe 3a	E+D vs S+I	Coherent	p < .01	R^2 = .49 p < .001	73% p < .001
Expe 3b	E+S vs D+I	InCoherent	p = .20	R^2 = .1 p = 0.34	55% p = .42

Table 1. Human and simulated performance on language discrimination Experiments 1 – 3. English, Dutch (stress-timed), Spanish, Italian (syllable-timed), Japanese (Mora-timed).

Given this indication that the recurrent State network is sensitive to rhythm class differences, we then simulated the 3 experiments of [Nazzi 98]. In order to study the stable population behavior of the network, we report results from a population of 10 model "subjects" created by using different random number generator seed values in initializing the non-topographic weights. Simulations assessed in three conditions the ability to discriminate between sentences from two languages, in teRamus 2000a of the percentage of correct classifications, where 50% represents chance performance.

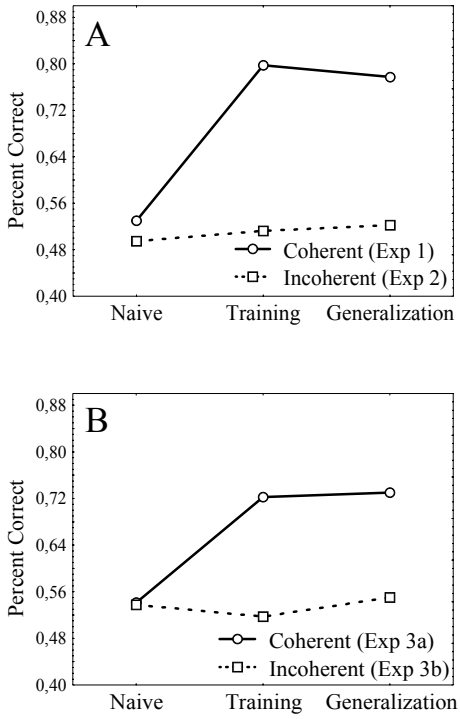

Figure 2. (A) Simulation performance in Experiments 1,2 and (B) Experiments 3a, 3b.

The *Naive* condition tested discrimination on 10 sentences (20 in Exp 3) from each of the two languages with no learning. The *Training* condition tested discrimination on 10 new sentences (20 in Exp 3) from each language, using the reinforcement rule. The *Generalization* condition then tested the trained model (with learning now inactivated) on the original sentences from the Naive condition, to assess the generalization of learning acquired during the Training condition (Fig 2). Four speakers per language were used to ensure that learning did not depend on a particular speaker's characteristics. Simulation performance in the *Generalization* condition is compared with that of infants from [Nazzi 98] in Table 1 (Performance).

As seen in Figure 2, naive model performance is near chance level. For coherent discriminations that contrast sentences from the same rhythm class with sentences from a different rhythm class (Exp 1, Exp 3a), the model learns, and can then generalize this learning to

new sentences uttered by new speakers. Learning and generalization fail for incoherent discriminations (Exp 2, Exp 3b) that contrast sentences from one (or 2) rhythm class(es) with sentences from the same rhythm class(es). That is, between-class contrasts succeed, while within-class contrasts fail.

4. Conclusion

We previously demonstrated that our recurrent network based on the primate frontostriatal system [Dominey 95B] is capable of learning and discriminating between specific temporal structures in sequences that share the same serial structure [Dominey 98a,Dominey 98b]. The current study extends these results by demonstrating that the model can learn the temporal component of prosodic structure in natural language. This learning can generalize to permit classification of new sequences into their respective rhythm classes, with performance that corresponds to that of a newborn human. Aside from the technical interest of demonstrating temporal generalization in recurrent networks, these results are of particular interest from the perspective of human language acquisition. It is becoming increasingly apparent that the acquisition of knowledge of lexical and syntactic structure in language relies on "bootstrapping" from prosodic structure [Christophe 98, Mehler 96, Morgan 96]. Until now, a functional characterization of this bootstrapping mechanism has been lacking. The current results with our sequence learning model now provide a preliminary characterization of this mechanism that can be used both in predicting and explaining psycholinguistic results, and also as a basis for the further elaboration of a model of language acquisition.

5. REFERENCES

[Almeida 87] Almeida LB (1987) A learning rule for asynchronous perceptrons with feedback in a combinatorial environment. *in Proc IEEE 1t Intl Conf. Neural Networks*, San Diego, CA, June 21-24, p.609-618.

[Christophe 98] Christophe A, Guasti T, Nespor M, Dupoux E, Van Ooyen B (1998) *Reflections on phonological bootstrapping : its role for lexical and syntactic acquisition. Language and Cognitive Processes*, In press

[Cleeremans 91] Cleeremans A, McClelland JL (1991) Learning the Structure of Event Sequences, *J Exp Psychol: General*, 120,3:235-253

[Dominey 95] Dominey PF (1995) Complex Sensory-Motor Sequence Learning Based on Recurrent State-Representation and Reinforcement Learning. *Biological Cybernetics*. 73, 265-274

[Dominey 98a] Dominey PF (1998a) Influences of Temporal Organization on Transfer in Sequence Learning: Comments on Stadler (1995) and Curran and Keele (1993), *J Exp Psychology: Learning, Mem and Cog*, 24, 1, 234-248

[Dominey 98b] Dominey PF (1998b) A shared system for learning serial and temporal structure of sensorimotor sequences? Evidence from simulation and human experiments. *Cog Br Res*, 6, 163-172.

[Dominey 95B] Dominey PF, Arbib MA, Joseph JP (1995) A Model of Cortico-Striatal Plasticity for Learning Oculomotor Associations and Sequences. *J Cog Neuroscience*, 7:3, 311-336

[Elman 90] Elman JL (1990) Finding Structure in Time. *Cognitive Science*, 14:179-211.

[Jordan 90] Jordan MI (1990) Learning to articulate: Sequential networks and distal constraints, in M Jeannerod (Ed), *Attention and Performance XIII*, Hillsdale, NY: Lawrence Erlbaum.

[Kühn 92] Kühn R, van Hemmen JL (1992) Temporal Association, in *Physics of Neural Networks*, (Eds E. Domanay, JL van Hemmen and K Schulten), Springer-Verlag, Berlin, 213-280

[Mehler 96] Mehler, J., Dupoux, E., Nazzi, T., & Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: The infant's viewpoint. In J. L. Morgan & K. Demuth (Eds.), *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition* (pp. 101-116). Mahwah, NJ: Lawrence Erlbaum Associates.

[Morgan 96] Morgan JL, Demuth K (1996) Signal to Syntax: Bootstrapping from speech to grammar in early acquisition. Mahwah NJ: Lawrence Erlbaum Associates.

[Nazzi 98] Nazzi T, Bertoncini J, Mehler J (1998) Language discrimination by newborns: Towards an understanding of the role of rhythm. *Journal of Exp Psych. Human Percept & Perform*, In press

[Nespor 96] Nespor M, Guasti T, Christophe A (1996) "Selecting word order: The rhythmic activation principle," in *Interfaces in Phonology*, U. Kleinhens, Ed. Akademic, Verlag, Berlin. 1 – 26.

[Pearlmutter 95] Pearlmutter BA (1995) Gradient calculation for dynamic recurrent neural networks: A survey. *IEEE Trans Neural Networks*, 6(5) 1212-1228.

[Pineda 89] Pineda FJ (1989) Recurrent Backpropagation and the Dynamical Approach to Adaptive Neural Computation. *Neural Computation*, 1, 161-172.

[Ramus 99a] Ramus F, Mehler J, (1999), Language identification with suprasegmental cues: A study based on speech resynthesis, *Journal of the Acoustical Society of America*, 105(1)

[Ramus 99b] Ramus F, Nespor M, Mehler J (1999) Correlates of linguistic rhythm in the speech signal, *Cognition*, 73(3)

[Wang 96] Wang DL, Liu X, Ahalt SC (1996) On temporal generalization in simple recurrent networks. In Press, *Neural Networks*

[Weitzenfeld 91] Weitzenfeld A (1991) NSL Neural Simulation Language, v 2.1, *Center for Neural Engineering, University of Southern California Technical Report TR91-5*

Conclusion

Cet ouvrage rassemble dix articles issus de communications données en janvier 1999 à Lyon au cours du premier colloque organisé par le Groupe Francophone de la Communication Parlée sur l'Identification automatique des langues. Les articles de la première partie constituent un support introductif à ce thème. Le lecteur a ainsi pu découvrir les enjeux liés à l'identification ainsi que la réalité linguistique sous-jacente. Il s'avère que, suivant les applications envisagées (identification de la langue, recherche de l'origine géographique d'un locuteur, indexation de documents…), les impératifs, en terme de vitesse d'identification et de précision varient considérablement. La nécessité d'adapter les solutions techniques aux demandes et dès lors évidente (page 9).

L'état de l'art proposé présente un panorama des systèmes actuels mettant en évidence le rôle prépondérant joué par les approches phonotactiques (page 17). Ce type d'information est, en effet, celui ayant conduit aux meilleurs résultats jusqu'à présent. Cependant, peu d'améliorations ont été apportées depuis quelques années, tandis que le déploiement des premières applications opérationnelles nécessitera probablement une amélioration significative des performances, en particulier en terme de diminution de la durée des enregistrements utilisés.

Pour s'affranchir des limites de la modélisation phonotactique, plusieurs approches sont proposées. De manière générale, elles visent à mieux intégrer des informations caractéristiques des langues jusqu'alors sous exploitées. Le système proposé par Matrouf et al. (page 65) prend en compte des informations lexicales, particulièrement discriminantes. Sur le plan acoustico-phonétique, plusieurs pistes s'offrent aux chercheurs. En effet, l'étape de projection du signal acoustique dans un espace phonémique discret semble être le point faible de la plupart des systèmes actuels, principalement par manque de données étiquetées. Une partie importante de l'information phonétique est alors perdue, bien qu'elle puisse se révéler cruciale.

Parmi les approches visant à s'affranchir de ce problème, celle proposée par Boula de Mareüil et al. (page 74) tire profit d'un décodage acoustico-phonétique réalisé dans un espace multilingue obtenu par agglomération de phonèmes dépendants de la langue. Cette technique permet de limiter le biais existant entre les modèles phonétiques des différentes langues et de disposer d'une quantité de données d'apprentissage des modèles plus importante. Cependant, l'exploitation de l'information phonétique reste implicite puisque le décodage acoustico-phonétique n'est là encore qu'un pré-traitement appliqué à la modélisation phonotactique.

Une autre approche, décrite dans Barkat (page 95) et appliquée à l'identification dialectale de l'arabe, exploite plus spécifiquement les informations phonétiques en les modélisant dans un cadre phonologique : le système vocalique de chaque zone dialectale est modélisé globalement et non sous forme de phonèmes distincts. Cette méthode, s'inspirant des travaux réalisés en linguistique sur les typologies de langues et présentés dans la première partie (cf. Boë et Vallée, page 35), peut se révéler complémentaire de la classique modélisation phonotactique, tout en requérant moins de données.

Bien que les informations phonétiques, phonotactiques et lexicales contribuent fortement à l'identité de chaque langue, il reste fondamental, si l'on cherche à optimiser la procédure d'identification, de ne pas négliger les niveaux de description phonologique et prosodique.

Malheureusement, ce type de traits se révèlent particulièrement difficile à modéliser de manière robuste, et beaucoup de progrès restent à venir. Cependant, l'article proposé par Hombert (page 87) fournit un cadre global basé sur la détection de traits phonologiques dans lequel des travaux de modélisation peuvent s'inscrire. Cette approche est particulièrement prometteuse dès lors que l'on cherche à étendre de manière importante le nombre de langues à identifier. La réflexion menée par Ohala et Marsico (page 117) permet quant à elle d'estimer les difficultés qu'il faudra surmonter, et elle fournit ainsi un intéressant éclairage sur ce problème qui ne sera probablement résolu que par une collaboration efficace entre linguistes et informaticiens.

Si la caractérisation phonologique automatique des langues reste un problème ouvert, il est intéressant de constater que des progrès considérables ont été obtenu par des approches pluridisciplinaires dans le domaine connexe de la modélisation prosodique. L'importance du rythme et de la prosodie des phrases dans des domaines tels que la synthèse ou la compréhension de la parole n'est plus à démontrer. Il s'agit cependant là encore d'un domaine où la modélisation est ardue : l'absence d'unités discrètes clairement identifiées, la nécessité de prendre en compte des paramètres hétérogènes (voisement, fréquence fondamentale, intensité mais surtout durée) résulte en un problème complexe. Depuis quelques années, on voit cependant apparaître des systèmes de modélisation, en particulier dans le domaine de la synthèse. Il est dès lors raisonnable d'envisager une exploitation efficace de la prosodie en identification des langues à moyen terme. A ce titre, les expériences réalisées par Ramus (page 131) et Dominey et Ramus (page 141) fournissent des approches, statistiques ou neuro-mimétiques, qui s'engagent dans cette voie.

En conclusion, il semble que l'identification automatique des langues est devenu un domaine du traitement du langage parlé à part entière, et non plus seulement une simple émanation des travaux entrepris en reconnaissance de la parole comme à l'origine. Les chercheurs en technologie de la parole ont atteint une excellente maîtrise de certaines techniques de modélisation (en particulier phonotactique), mais il s'avère maintenant nécessaire de se tourner vers d'autres types d'information. L'exploitation de *modèles multiniveaux*, c'est-à-dire exploitant en parallèle la phonétique, la phonologie, la phonotactique, le lexique et/ou la prosodie devient dès lors un défi pour les années à venir. De tels systèmes ne pourront s'appuyer que sur une dynamique associant des psycholinguistes, des linguistes et des cogniticiens aux chercheurs en traitement de la parole. En effet, l'identification automatique des langues constitue un creuset fortement pluridisciplinaire duquel peuvent émerger des progrès importants, tant en ingénierie qu'en théorie linguistique.