Mémoire de DEA

Jean Luc Rouas

6 juillet 2001

Table des matières

Chapitre 1

Introduction

1.1 Introduction à l'Identification Automatique des Langues (IAL)

L'identification automatique des langues est un domaine récent dans le cadre du traitement de la parole. L'objectif est de déterminer la langue employée, à partir d'un énoncé prononcé par un locuteur. Cette définition, un peu simpliste, ne prend pas en compte la diversité des applications possibles. Les conditions peuvent varier suivant le domaine applicatif (interrogation de systèmes d'information, consultation de bases de données documentaires multilingues, enseignement ou traduction automatique), notamment en nombre de locuteurs, nombre de langues à identifier, conditions d'enregistrement (signal bruité ou non...).

La recherche en identification des langues revêt actuellement deux aspects :
- Une étude cognitive recherchant les traits perceptuellement discriminants pour chaque langue.
- Une étude de type ingénierie pour répertorier des indices pouvant être extraits du signal acoustique d'une façon robuste et modélisables d'un point de vue statistique, le but étant d'aboutir à un système d'IAL.

La signature acoustique d'une langue peut être recherchée à plusieurs niveaux :
- Acoustico-phonétique (nature des sons et fréquences d'occurrences relatives)
- Phonotactique (règles d'enchaînement des unités phonétiques)
- Phonologique (organisation des unités phonétiques en tant que système)
- Prosodique (motifs rythmiques ou mélodiques)
- Morphosyntaxique ou lexical.

1.2 Sujet de stage

L'objectif du stage est d'appronfondir l'étude de paramètres prosodiques (c'est à dire liés à la perception auditive) en identification des langues.

Le travail effectué pendant le stage a été principalement d'ordre bibliographique, ce qui explique l'importance donnée à la partie état de l'art. D'un point de vue pratique, des expérimentations ont étés menées sur la validité de différents paramètres tels que la prise en compte de la durée.

1.3 Organisation du rapport

Dans une première partie, nous ferons un état de l'art pour ce qui concerne l'identification automatique des langues, et nous décrirons quelques systèmes récents utilisant des paramètres prosodiques. Dans le troisième chapitre, nous nous intéresserons plus en détail au système d'IAL tel qu'il est envisagé à l'IRIT, nous présenterons l'existant et nous discuterons des possibilités d'amélioration. Le travail d'expérimentation effectué au cours du stage de DEA est décrit dans le chapitre quatre.

Chapitre 2

État de l'art

Les paramètres pris en compte dans un système d'identification automatique des langues sont issus de la reconnaissance automatique de la parole et ils permettent de caractériser des informations acoustico-phonétiques et phonotactiques. Les systèmes les plus performants sont basés sur une modélisation statistique de ces indices. La méthode employée consiste, pour chacune des langues, à estimer un - ou plusieurs - modèles à partir d'enregistrements acoustiques dits d'apprentissage, puis à comparer ces modèles avec l'énoncé à identifier et à déterminer lequel est le plus probable [Zis96].

De plus en plus, des systèmes prenant en compte les aspects prosodiques s'imposent malgré les difficultés de modélisation ([Ita99] ou [Li94]). Les résultats obtenus par ces systèmes sont comparables à ceux obtenus en utilisant uniquement les paramètres acoustico-phonétiques ou phonotactiques. Mais la fusion des différents résultats obtenus suivant les modèles reste un problème majeur dans la décision d'identification.

2.1 Description d'un système général d'Identification Automatique des Langues

Un bon algorithme d'IAL devrait exploiter des informations de différentes sources pour arriver à la décision d'identification [Mut94] :

- LE VOCABULAIRE
- LES PHONÈMES
- LA PHONOTACTIQUE
- LA PROSODIE

Un système classique d'IAL, utilisant les indices acoustiques, phonotactiques, et prosodiques, peut se décomposer en trois parties distinctes :

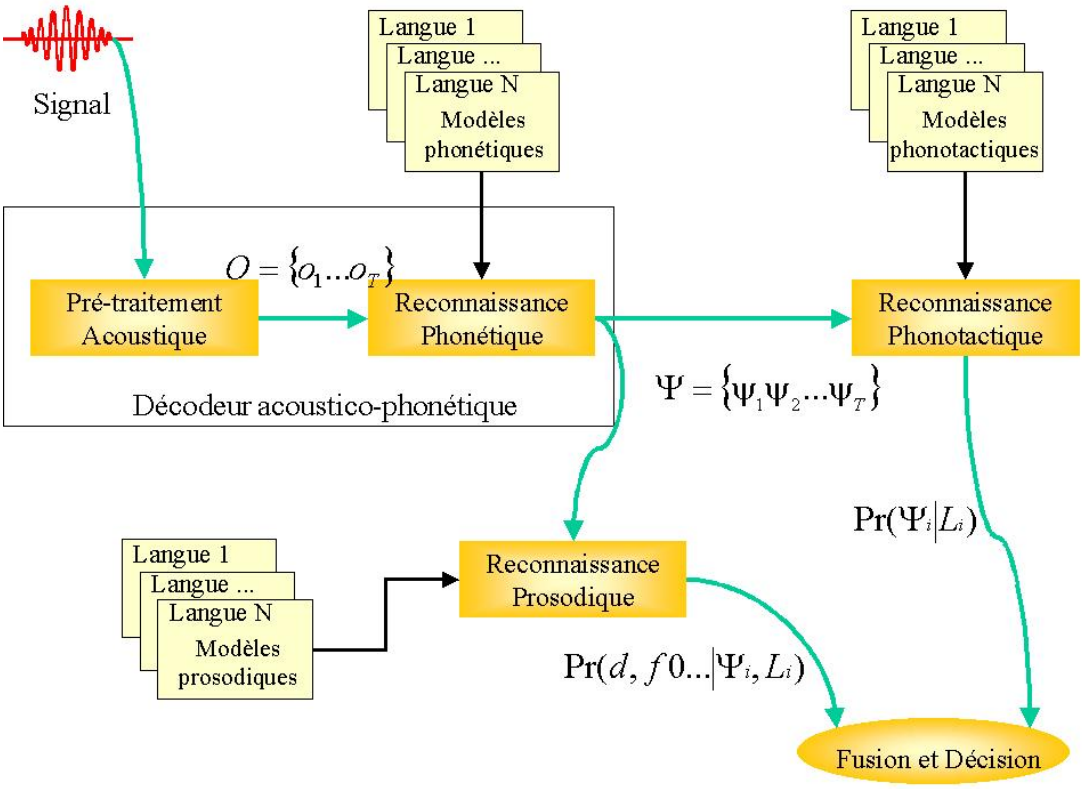

Figure 1 - Description d'un système général d'identification des langues

La première partie (décodeur acoustico-phonétique) consiste à extraire du signal des paramètres caractéristiques. La première étape est un pré-traitement du signal acoustique (O). Ensuite une reconnaissance de symboles phonétique permet de générer la suite de phonèmes décrivant le signal (Ψ).

Le module de reconnaissance phonotactique se propose d'identifier la langue en se basant sur des règles phonotactiques. La suite de symboles phonétiques issue du décodage peut être obtenue de différentes façons : des techniques de classification de type Quantification Vectorielle [Pel97] ou une modélisation de type markovienne peuvent être utilisées pour modéliser le langage. La suite des symboles phonétiques est ensuite décodée à l'aide d'un modèle de langage proposant les combinaisons phonétiques les plus fréquentes et fournissant un score de vraisemblance. Dans un système d'IAL, il y a en général autant de modèles de langage que de langues à identifier, alors qu'au niveau du décodage phonétique, les approches peuvent être très variées (emploi d'un décodeur unique, d'un décodeur par langue...).

En parallèle, à partir d'un modèle prosodique pour chaque langue un score de vraisemblance prosodique est également fourni. Les paramètres utilisés par ce modèle varient selon les systèmes. Nous allons présenter rapidement ce que l'on entend par prosodie, ainsi que les paramètres les plus employés.

2.2 Paramètres prosodiques utilisés en Identification Automatique des Langues

Acoustiquement parlant, la prosodie désigne les phénomènes liés à la variation dans le temps des paramètres de hauteur, d'intensité et de durée. La notion de hauteur est essentiellement liée à la FRÉQUENCE FONDAMENTALE (notée F0), qui correspond à la fréquence de vibration des cordes vocales. La perception de l'intensité est principalement liée à l'amplitude et à l'ÉNERGIE du son, mais dépend aussi partiellement de la durée. La DURÉE correspond au temps d'émission du son.

D'un point de vue perceptuel, la variation dans le temps des paramètres suscités correspond à la variation de rythme des phrases, de leur mélodie et de leur accentuation. Nous percevons le RYTHME grâce à l'enchaînement des durées des phonèmes. La mélodie est la perception des variations de la FRÉQUENCE FONDAMENTALE. L'ACCENTUATION est un phénomène de plus haut niveau, qui consiste à mettre en relief une syllabe par rapport à son environnement immédiat. Les linguistes classent en général les langues suivant deux catégories :

- les langues à accent fixe (français, anglais...), où l'accentuation se place toujours sur la même syllabe (en français la dernière syllabe du mot)
- les langues à accent libre (espagnol, ...), où la position de l'accentuation détermine le sens du mot.

Trois paramètres prosodiques sont principalement utilisés dans les systèmes d'identification automatique des langues :

- l'énergie,
- la durée,
- et la fréquence fondamentale.

La première étape de l'extraction des paramètres prosodiques est généralement une analyse à court terme du signal, en faisant l'hypothèse que pour une durée d'analyse suffisamment courte, le signal est quasi-stationnaire.

2.2.1 Energie

L'énergie est un paramètre couramment utilisé dans les systèmes de reconnaissance (par exemple le premier coefficient cepstral). C'est le paramètre prosodique le plus facile à calculer. L'énergie d'un signal échantillonné $(s_t)_{t=1,T}$ à support borné est définie par :

$$E = \frac{1}{T}\sum_{t=1}^{T} s_t^2$$

En général, l'énergie est exprimée en décibels :

$$E_{dB} = 10 * log_{10}(\frac{1}{T}\sum_{t=1}^{T} s_t^2)$$

L'énergie est calculée sur des portions de signal convoluées avec une fenêtre glissante étroite. Pour éliminer la variabilité du gain, l'énergie peut être normalisée par rapport au maximum sur la phrase.

2.2.2 Fréquence fondamentale

La fréquence fondamentale correspond à la fréquence de vibration des cordes vocales. Les principales techniques d'extraction de la fréquence fondamentale utilisent une représentation temporelle ou spectrale du signal. Les méthodes spectrales se servent des harmoniques de la fréquence fondamentale (méthode du peigne spectral). Les méthodes basées sur les représentations temporelles utilisent la similarité du signal d'une période à l'autre, afin de repérer la période fondamentale. Plusieurs de ces méthodes sont couramment employées [Hes83], comme la méthode d'autocorrélation, la méthode SRPD (super résolution) ou la méthode AMDF (Avergage Magnitude Différence Function) décrite en Annexe A.

2.2.3 Durée

La durée est le paramètre le plus difficile à préciser car rien n'indique comment le système de contrôle, de production ou de perception de parole mesure le temps. Les indices de durée classiques supposent généralement la donnée d'une segmentation, i.e. des frontières des unités dont on souhaite mesurer la durée. La durée est alors mesurée par le nombre de trames qui séparent ses frontières de début et de fin. La plupart des systèmes prosodiques utilisent une segmentation basée sur le phonème. Récemment des expérimentations ont été réalisées sur une segmentation basée sur d'autres unités que le phonème, notamment en pseudosyllabes.

Nous allons décrire quelques systèmes employant des paramètres prosodiques.

2.3 Quelques systèmes d'IAL récents utilisant la prosodie

Le premier système présenté est celui de S. Itahashi [Ita99]. Il utilise la fréquence fondamentale (paramètre prosodique) et les coefficients cepstraux (paramètres segmentaux).

Le deuxième système est celui de F. Cummins [Cum99]. Les paramètres utilisés sont dérivés de la fréquence fondamentale et de l'enveloppe d'amplitude.

Le troisième système est donné par [Sav91]. Il emploie les contours de la fréquence fondamentale et des modèles de Markov cachés.

Le dernier système évoqué dans ce rapport est le système développé par Li [Li94]. Il est basé sur la reconnaissance du locuteur, et emploie comme paramètres des noyaux syllabiques.

2.3.1 Le système d'Itahashi

Le système décrit dans [Ita99] utilise des paramètres extraits de la fréquence fondamentale et des coefficients cepstraux. Les tests sont réalisés en prenant soit uniquement les paramètres extraits de la fréquence fondamentale, soit les coefficients cepstraux, soit l'ensemble des paramètres.

2.3.1.1 Fréquence fondamentale

Les contours de la fréquence fondamentale sont extraits grâce à la méthode AMDF [Hes83] (voir Annexe A), avec :
- une fréquence d'échantillonnage de 8 kHz,
- une quantification sur 16 bits,
- une fenêtre d'analyse de 30 ms, intervalles de 10 ms.

Des erreurs dites d'harmoniques peuvent se produire lors de l'extraction de la fréquence fondamentale par la méthode AMDF. La fréquence estimée peut être le double ou la moitié de la fréquence réelle.

Une méthode de correction de ces erreurs grossières est proposée par [Bag94]. Le principe est de calculer l'autocorrélation de toutes les valeurs candidates de la période fondamentale à chaque instant. La fréquence fondamentale estimée correspond au candidat qui possède la plus forte valeur d'autocorrélation.

Le motif décrit par la fréquence fondamentale est alors modélisé soit par des lignes polygonales soit par des fonctions exponentielles. Au total, on extrait 7 paramètres du contour de la fréquence fondamentale, et 9 paramètres des lignes approximées ou des fonctions exponentielles. Les 7 paramètres extraits du contour de F0 sont :
- 1,2,3 : écart-type, skewness et kurtosis de F0,
- 4,5,6 : écart-type, skewness et kurtosis de l'énergie,
- 7 : coefficient de corrélation entre F0 et l'énergie.

Pour les deux modélisations, on a :

Lignes polygonales : $y_k(t) = a_k(t_k - t_{k-1}) + b_k,\ k = 1, 2, ..., K$. On détermine a_k et b_k de façon à minimiser l'erreur des moindres carrés. Le nombre de lignes est déterminé de sorte que l'erreur d'approximation soit inférieure à un seuil.

Les paramètres extraits sont :
- 1 : rapport de durée entre la pente positive et la pente négative,
- 2,3 : nombre de lignes par unité de durée (pour les pentes positives et négatives),
- 4,5 : pente moyenne (positive et négative),

- 6,7 : écart-type moyen des pentes (positives et negatives),
- 8,9 : fréquence de départ relative des pentes positives et négatives.

Fonctions exponentielles : $y(t) = a(\frac{e}{\tau})e^{(-\frac{t}{\tau})} + bt + c$

Les paramètres extraits sont :

- 1 : durée moyenne de l'intervalle d'approximation,
- 2 : nombre de fonctions par unité de durée,
- 3,4 : moyenne et écart-type de l'amplitude a,
- 5,6 : moyenne et écart-type de la pente b,
- 7,8 : moyenne et écart-type de la constante de temps τ
- 9 : fréquence de départ relative.

2.3.1.2 Modèles de Markov Cachés (HMM) pour les coefficients cepstraux

Les coefficients cepstraux (suivant l'échelle de Mel) sont extraits du signal acoustique en utilisant :

- une fréquence d'échantillonnage de 8 kHz,
- une fenêtre d'analyse (Hamming) de 5 ms, à intervalles de 10 ms.

On extrait alors 12 coefficients cepstraux, on calcule 12 dérivées de ces coefficients, et la dérivée de l'énergie. On utilise un modèle de Markov par langue, en faisant l'apprentissage sur 30 locuteurs. Le nombre d'états de chaque modèle est variable, des tests sont effectués pour 4, 8, 16, 32 et 64 états. Cette modélisation est de nature acoustico-phonétique.

2.3.1.3 Expérimentations

Les expériences sont réalisées sur le corpus Multilingual Telephone Speech Corpus de l'Oregon Graduate Institute (OGI-MLTS) (Annexe B) [Mut92]. Itahashi utilise des données de 45 secondes de parole spontanée par locuteur, pour 50 locuteurs dans 10 langues (anglais, français, espagnol, farsi, chinois, coréen, japonais, tamoul et vietnamien). L'apprentissage se fait sur 30 locuteurs, les 20 locuteurs restants (différents de ceux employés à l'apprentissage) sont utilisés pour les tests.

2.3.1.4 Résultats

Aux résultats obtenus par chacune des approches viennent s'ajouter ceux obtenus en combinant les deux approches initiales.

Modèles de markov cachés : Les taux d'identifications ont étés calculés en utilisant uniquement les modèles de Markov, pour cinq nombres différents d'états (4, 8, 16, 32, 64). Les meilleurs résultats sont de 56% d'identification correcte. Ils sont obtenus à la fois pour 32 et 64 états.

Fréquence fondamentale : Une analyse discriminante des paramètres dérivés de la fréquence fondamentale donne les résultats suivants :
- modèle lignes polygonales : 25,5% d'identification correcte,
- modèle fonctions exponentielles : 28,0% d'identification correcte.

Méthode combinée : Afin de fusionner les résultats obtenus par les deux méthodes, on normalise les scores en vraisemblance de façon à avoir une moyenne nulle et une variance de 1. On pondère le résultat obtenu par les HMM avec un poids w. En combinant les deux méthodes, le meilleur taux de reconnaissance (60%) est obtenu pour des HMM avec 32 états, le modèle de lignes polygonales, et un poids w=2,4.

Les résultats montrent que l'apport de la fréquence fondamentale augmente de 5% le score obtenu par les modèles de Markov seuls dans le cas d'une modélisation par lignes polygonales, ou 2% pour les fonctions exponentielles. La prise en compte de la fréquence fondamentale peut apporter une information sur la langue à identifier, mais l'apport réalisé par comparaison aux résultats donnés par les modèles de Markov cachés seuls est faible.

2.3.2 Le système de Cummins

Le système [Cum99] utilise une décision par réseaux de neurones, en fusionnant les résultats obtenus par estimation de la fréquence fondamentale et de l'enveloppe d'amplitude.

2.3.2.1 Estimation de la fréquence fondamentale

La fréquence fondamentale est estimée à intervalles de 1 ms. On prend alors la dérivée puis on sous-échantillonne à 100 Hz (fenêtre glissante rectangulaire), ensuite on effectue un lissage (avec un fenêtre rectangulaire de 15 points), et enfin on fait un changement d'échelle pour avoir des valeurs comprises entre -1 et 1. Le paramètre obtenu ainsi est noté ΔF0.

2.3.2.2 Estimation de l'enveloppe d'amplitude

Le signal est filtré au moyen d'un filtre de Butterworth de faible ordre, passe-bande centré sur 1000 Hz, de largeur de bande 500 Hz. La valeur absolue de ce signal est alors calculée, puis fournie en entrée d'un autre filtre Butterworth passe-bas (fréquence de coupure 10 Hz) qui effectue alors un lissage. La dérivée est évaluée, on la sous-échantillonne à 100 Hz, on fait un lissage et un changement d'échelle. Le paramètre ainsi obtenu est noté ΔEnv.

2.3.2.3 Réseau de neurones

Les neurones utilisés sont dits à Long Short Term Memory (Hochreiter & Schmidthuber 1997). Dans un réseau LSTM, les unités cachées conventionnelles sont remplacées par des *blocs* mémoire contenant une ou plusieurs cellules (voir Figure 2).

Une cellule est une unité linéaire avec une connection récurrente de poids 1. Ce poids permet à la cellule de rester activée en l'absence d'entrée. Le flux d'activation entrant (net_c) passe par une grille d'entrée (*input gating*) et une fonction signoïde. L'entrée (net_{in}) pour chaque cellule est multipliée par l'activité de la grille d'entrée, autorisant la grille d'entrée à décider à quelle information la cellule est exposée. La sortie se comporte de la même façon. L'apprentissage est une combinaison de Back Propagation Through Time et de Real Time Recurrent Learning.

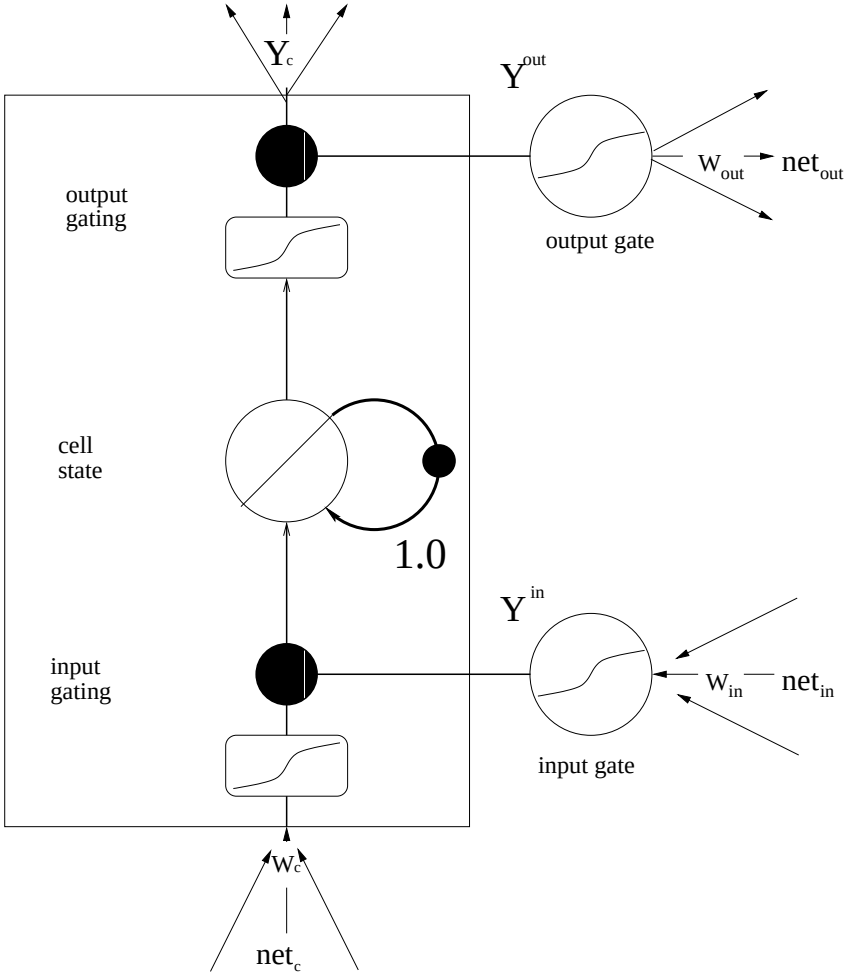

Figure 2 - Bloc LTSM contenant une seule cellule [Cum99]

2.3.2.4 Expérimentations

Le corpus utilisé est le corpus OGI-MLTS. 5 langues ont été considérées (anglais, japonais, espagnol, mandarin et allemand). L'apprentissage se fait sur 50 locuteurs par langue sur des fichiers de parole spontanée courts. Les tests sont effectués sur 20 locuteurs par langue (différents de ceux employés à l'apprentissage) sur des enregistrements de plus grande durée de parole spontanée.

2.3.2.5 Résultats

Les expérimentations sont faites en utilisant un seul critère à la fois : ΔF0 ou ΔEnv. Lors des tests, les tâches d'identification consistent à choisir parmi deux langues. Les tests sont donc effectués sur chaque paire de langues (anglais/allemand, anglais/espagnol,...).

- En utilisant uniquement ΔEnv on obtient un taux d'identification correcte variant selon les paires de langues entre 50 et 63 %.
- Avec ΔF0, on obtient des taux d'identification correcte légèrement supérieurs, entre 50 et 69 %.

Ces résultats concordent avec la littérature [Thy96] où le paramètre F0 est une variable plus discriminante que la modulation de l'amplitude. Cependant, les expériences suggèrent que la modulation de l'enveloppe soit un paramètre plus exploité.

2.3.3 Le système de Savic

Le système de Savic [Sav91] est basé sur l'association de deux méthodes : une modélisation phonotactique des langues par modèles de Markov cachés et une estimation des contours de la fréquence fondamentale.

2.3.3.1 Méthodologie

Modèles de Markov cachés : La probabilité de transition entre un état et lui-même représente la fréquence d'occurrence d'une classe de phonèmes dans une langue. Les probabilités de transition entre différents états indiquent la structure phonétique de la langue. Les résultats expérimentaux montrent qu'un modèle de Markov à cinq états est optimal d'un point de vue performance et coût de calcul.

Analyse des contours de la fréquence fondamentale : Les contours de la fréquence fondamentale sont déterminés grâce à un algorithme basé sur l'autocorrélation.

2.3.3.2 Expérimentations

Les données utilisées proviennent d'enregistrements de locuteurs natifs ou ayant un accent peu détectable. Les locuteurs effectuent la lecture d'un texte dans leur langue natale. Les données ont été enregistrées dans un environnement non bruité pendant approximativement dix minutes. Les données sont ensuite converties au format numérique au moyen d'un convertiseur 12 bits avec une fréquence d'échantillonnage de 10 kHz. Le signal numérique est découpé en segments de 15 secondes et passé dans un filtre passe-bas avec une fréquence de coupure de 4.5 kHz.

2.3.3.3 Résultats

Les résultats expérimentaux décrits sont incomplets, nous n'avons qu'une information sur la validité des paramètres choisis :

- Modèles de Markov cachés : lorsque l'on observe les matrices de transition, on s'aperçoit que quel que soit le locuteur dans une même langue, les probabilités de transition sont similaires. Certaines langues semblent avoir des similarités de probabilités de transition pour quelques états. Cependant, ces langues ont des probabilités de transition très différentes pour les autres états, ce qui permet une discrimination facile.
- Les contours de la fréquence fondamentale apportent également une information intéressante. On voit qu'en espagnol, par exemple, les segments voisés sont en général longs et varient lentement, contrairement au chinois.

Les résultats montrent que les structures phonétiques (probabilités de transition dans les modèles de Markov) et les contours de la fréquence fondamentale décrivent les différences entre les langues. Toutefois, chacun des critères ne donnera pas dans tous les cas une réponse définitive, mais il indiquera la bonne direction. Un classificateur devra être utilisé pour combiner les résultats obtenus par chaque critère.

2.3.4 Le système de Li

Li [Li94] a développé un système d'identification automatique des langues basé sur la reconnaissance du locuteur. Son idée est de classer un signal en mesurant la similarité entre son locuteur et les locuteurs les plus proches dans chaque langue.

Durant l'apprentissage, un réseau de neurones est utilisé pour extraire tous les noyaux syllabiques. Des coefficients spectraux sont extraits à différents endroits des noyaux et sauvegardés.

Durant la phase de reconnaissance, les noyaux syllabiques sont extraits de la même façon et les coefficients spectraux sont comparés à tous ceux mémorisés pour chaque locuteur. La plus petite différence entre chaque noyau du fichier à examiner et les noyaux des autres locuteurs est alors calculée. La somme des

différences est considérée comme la différence entre le locuteur et chacune des références. La différence moyenne des locuteurs les plus semblables dans chaque langage constitue la différence entre le fichier de test et les langages cibles : le langage ayant la plus petite différence est sélectionné.

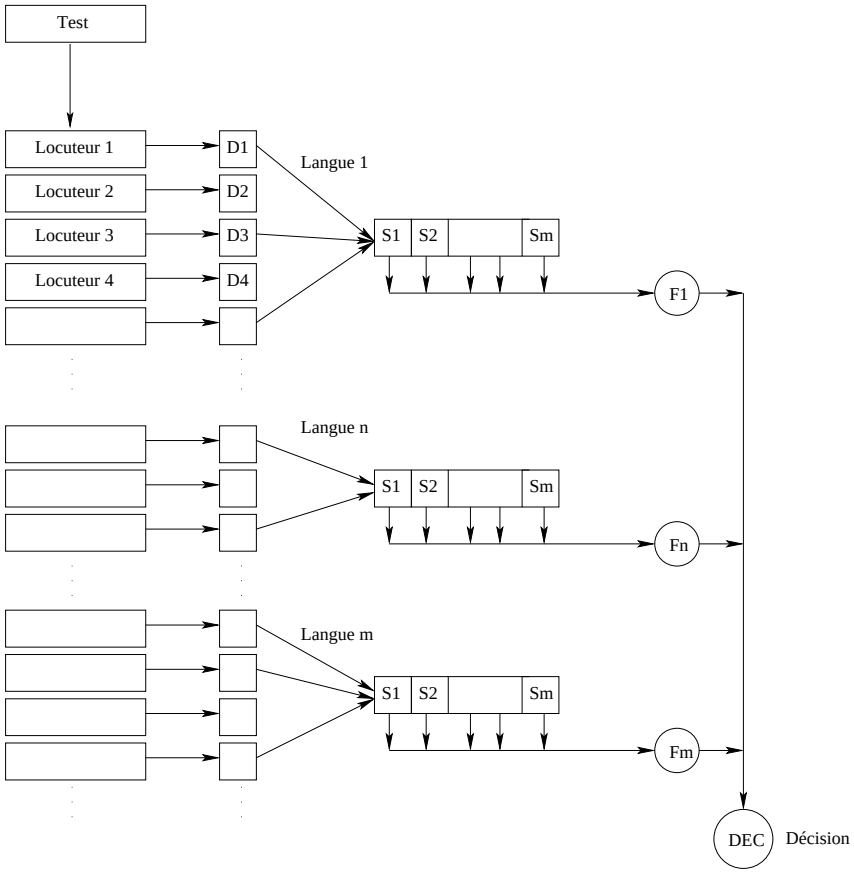

Figure 3 - Schéma descriptif du système de K. P. Li [Li94]

Le système a été évalué sur le corpus téléphonique OGI-MLTS, en utilisant 449 locuteurs répartis sur 10 langues. Les résultats sont d'environ 78% et 58% pour 10 langues en utilisant des séquences de respectivement 45 s. et 10 s. Si l'on ne considère que les identifications par paire de langues, les taux moyens sont de 91% et 82% pour des fichiers de 45 s. et 10 s. respectivement.

2.4 Conclusion

Les différents systèmes étudiés dans ce chapitre démontrent la validité de l'emploi de différents types de paramètres :

- les modèles de Markov cachés, utilisés dans deux des systèmes présentés, [Ita99] et [Sav91], permettent d'obtenir des résultats classiques de l'ordre de 60% d'identification correcte [Ita99],

- la fréquence fondamentale est employée dans trois des quatres systèmes présentés, [Ita99] [Cum99] [Sav91]. Elle n'a cependant qu'une influence limitée sur l'amélioration des résultats (5% d'amélioration lorsque l'on fusionne les résultats obtenus par modèles de Markov cachés et ceux obtenus avec les paramètres dérivés des contours de la fréquence fondamentale [Ita99]),
- la prise en compte de l'information apportée par l'enveloppe d'amplitude [Cum99] pourrait apporter un plus à un futur système d'IAL,
- la technique employée par Li [Li94], basée sur la reconnaissance du locuteur, est intéressante puisqu'elle permet de s'affranchir du biais occasionné par le changement de locuteur.

La plupart de ces systèmes d'IAL obtiennent de bons résultats mais ils ne permettent pas une exploitation industrielle. Cependant, la prise en compte d'informations prosodiques améliore toujours les performances, même si l'amélioration est parfois faible.

Chapitre 3

L'identification automatique des langues à l'IRIT

Le point de départ des travaux réalisés à l'IRIT a été la thèse de F. Pellegrino [Pel98]. Le laboratoire de Dynamique du Langage de Lyon (DDL), l'Institut de Phonétique de Grenoble (ICP) et l'Institut de Recherche en Informatique de Toulouse (IRIT) ont collaboré sur ce sujet (projet soutenu par la DGA), ce qui a permis la validation de la différenciation des langues au travers de modèles basés sur leurs systèmes vocaliques.

Les expériences d'identification automatique menées à partir de ces seuls modèles ont montré que les performances obtenues étaient comparables à celles des systèmes traditionnels, mais sans qu'aucune information phonotactique ne soit prise en compte.

Depuis, les expériences menées dans [Far00] ont montré la pertinence d'une approche multigramme [Del96] pour le modèle phonotactique. Les modèles de langage "classiques" emploient des modèles N-grammes, avec N fixe, c'est-à-dire qu'ils calculent les fréquences d'apparition de séquences de N élements. Les modèles multigrammes fonctionnent de la même manière mais prennent en compte des séquences de taille variant de 1 à N.

L'importance de la prise en compte d'informations telles que la fréquence fondamentale [Ita99] et la durée des syllabes [Far01] dans le module prosodique a été mis en évidence.

3.1 Le système aujourd'hui

Le système existant à l'IRIT peut se représenter comme suit :

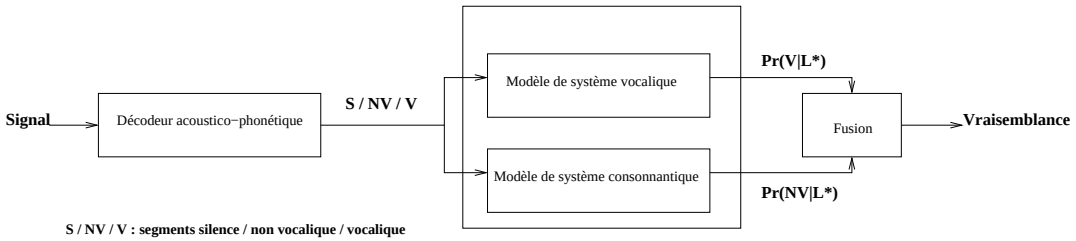

Figure 4 - Système d'IAL existant à l'IRIT

3.1.1 Le décodeur acoustico-phonétique

Le décodeur acoustico-phonétique permet actuellement une segmentation en segments vocaliques / non vocaliques / silences (S/NV/V). Le taux d'identification correcte des segments vocaliques atteint 93% en moyenne, les tests étant réalisés sur des enregistrements d'une durée de 4 à 5 minutes pour chacune des cinq langues traitées (français, japonais, coréen, espagnol et vietnamien).

Une différenciation entre les grandes classes de consonnes (plosives, fricatives, sonantes, occlusives) est à l'étude mais ne donne pas des résultats satisfaisants pour le moment (le taux de reconnaissance est de 60%, mais avec une *accuracy* (fiabilité) de 40% seulement, qui ne permet pas encore aux modèles de langage d'effectuer la discrimination de façon efficace). Ce DAP multilingue devra être amélioré dans le prochain système avant de permettre des modélisations phonétiques des espaces acoustiques et des modélisations phonotactiques efficaces.

3.1.2 Modélisation différenciée des systèmes vocaliques et consonantiques

Cette modélisation se base sur la segmentation effectuée par le décodeur acoustico-phonétique. Elle a été mise au point par F. Pellegrino [Pel98]. La modélisation est faite par des modèles de mélanges de gaussiennes, le nombre de lois est déterminé automatiquement par l'algorithme LBG-Rissanen [And93].

La validation de cette approche a été faite sur le corpus OGI-MLTS [Mut92]. L'apprentissage s'est fait sur des fichiers de parole spontanée d'une durée totale d'environ 50 minutes. Les tests sont effectués avec des locuteurs différents de ceux employés à l'apprentissage et sur des fichiers de parole spontanée d'environ 45 secondes chacun.

La modélisation des systèmes vocaliques seule fournit un taux de reconnaissance de 78%. Les mêmes résultats sont obtenus en employant uniquement la modélisation des systèmes consonantiques. La fusion des deux modèles permet d'obtenir un taux d'identification correcte d'environ 85%.

Les résultats obtenus grâce à ce système sont comparables à ceux obtenus par des systèmes employant des modèles phonotactiques. Afin d'améliorer ses performances, le système évoluera suivant deux axes :

- L'approche par modélisation différenciée des systèmes vocaliques et consonantiques mérite d'être élargie afin de modéliser les différents systèmes consonantiques (fricatif, occlusif, ...).
- La prise en compte d'une information prosodique peut se révéler efficace (cf. chapitre 2). Le futur système devra intégrer un module prosodique pour compléter les résultats donnés par le modèle de langage.

Nous allons maintenant décrire le système tel qu'il est envisagé à long terme.

3.2 Description générale du système ultérieur

Le système proposé regroupe des aspects acoustiques, phonotactiques et prosodiques. Ce système se compose en trois modules :

- un module de décodage acoustico-phonétique généralisé, pour prendre en compte la diversité des langues au niveau des espaces acoustiques,
- un module phonotactique pour gérer les occurrences et les séquences de sons, avec un modèle de langage par langue,
- un module prosodique pour prendre en compte le rythme et l'intonation des langues. Ce module comprendra également un modèle par langue.

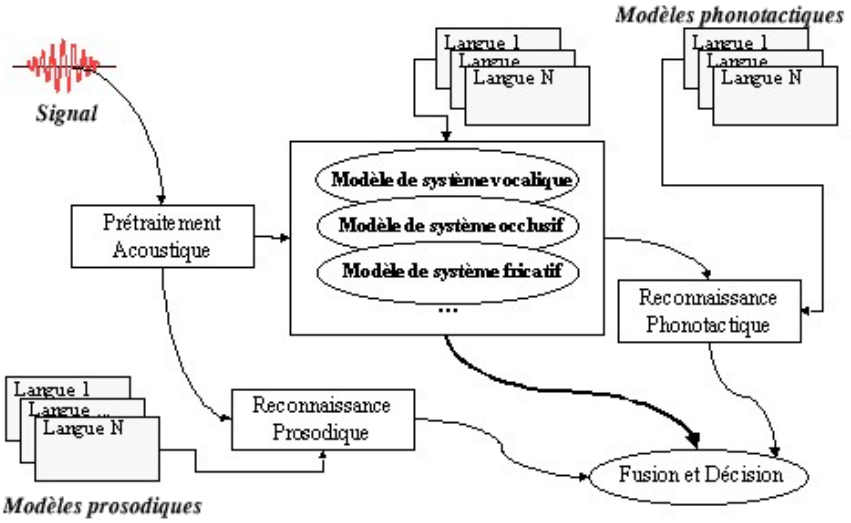

Figure 5 - Description du système d'identification envisagé à l'IRIT.

3.2.1 Le module de décodage acoustico-phonétique généralisé

L'approche proposée ici est la Modélisation Phonétique Différenciée. Elle est basée sur les constats suivants :

- les typologies des langues distinguent les systèmes vocaliques des systèmes consonantiques, et à l'intérieur de ces systèmes, certaines consonnes sont encore représentées séparement (fricatives, plosives,...),
- ces typologies peuvent être utilisées pour identifier une langue (ou un groupe de langues) à partir de sa description phonologique,
- les paramètres les plus pertinents pour caractériser un son peuvent dépendre de sa classe phonétique,
- lorsqu'on modélise ensemble des sons de nature homogène (par exemple des voyelles), on peut plus facilement prendre en compte certaines contraintes spécifiques (limites de l'espace acoustique dans cet exemple).

Ces différents constats ont conduit à envisager une modélisation différenciée de chaque système phonologique (système vocalique, système des consonnes fricatives...) redéfini en tenant compte de contraintes liées à la représentation acoustico-phonétique de la parole spontanée.

3.2.2 Le module phonotactique

De très bonnes performances sont atteintes par des systèmes utilisant la discrimination des modèles phonotactiques. L'un des principaux intérêts de ces systèmes est qu'ils peuvent être appris automatiquement à partir de données brutes non étiquetées (contrairements aux apprentissages classiques). Le module phonotactique du système se situe dans le prolongement direct du module de décodage acoustico-phonétique : la modélisation est effectuée à partir des unités acoustico-phonétiques dégagées par la Modélisation Phonétique Différenciée. Le modèle de langage employé est un modèle multigramme qui permet de différentier les langues par des fréquences d'occurence de suites de phonèmes.

3.2.3 Le module prosodique

Nous avons vu dans le chapitre 2 que la prosodie désigne des phénomènes liés à des variations de hauteur (fréquence fondamentale), d'intensité (énergie) et de durée d'un son. Dans le cadre de l'identification automatique des langues, une modélisation prosodique se doit de prendre en compte à la fois les particularités des langues au niveau de l'intonation, mais aussi au niveau du rythme. Le module prosodique comprendra donc deux grandes composantes : intonation et rythme.

L'amélioration du système existant se fera en plusieurs étapes. Les prochaines mises en œuvre sont décrites dans la partie suivante.

3.3 Extensions restant à mettre en œuvre

3.3.1 Modélisation des systèmes phonétiques

Les systèmes d'identification récents ne permettent pas d'exploiter convenablement les informations phonétiques présentes dans le signal. Une approche par modélisation différenciée des systèmes vocaliques et consonantiques [Pel98] s'est révélée efficace, grâce à une détection automatique Voyelle/Non Voyelle.

Notre but est d'étendre cette modélisation en différenciant les consonnes entre elles, en les regroupant en grandes classes (plosives, fricatives, sonantes). L'architecture différenciée du système se justifie par des considérations acoustico-phonétiques et phonologiques :

- Les caractéristiques des consonnes et des voyelles ne sont pas communes,
- L'espace acoustique consonantique est discontinu, contrairement à l'espace acoustique vocalique,
- Les paramètres pertinents, pour caractériser un son, peuvent dépendre de sa classe phonétique.

3.3.2 Discrimination des motifs rythmiques

L'utilisation de motifs rythmiques en identification automatique des langues est prometteuse mais difficile. Nous envisageons de mettre en place des modèles d'unités rythmiques, basés sur des séquences Consonnes / Voyelles issues d'une segmentation automatique.

Les travaux effectués à ce jour montrent la pertinence d'une modélisation syllabique mais la détermination automatique des frontières syllabiques est difficile car ces unités ne sont pas de nature acoustique [Li94]. Les études menées dans [Far01] montrent l'efficacité d'une segmentation en pseudo-syllabes, c'est-à-dire de séquences constituées d'une série de consonnes interrompues par une voyelle.

3.3.3 Fusion des informations

Les différents modèles employés permettent de traiter des observations extraites selon des échelles temporelles différentes : l'information acoustique relève de l'échelle centiseconde alors que les motifs rythmiques sont observés à l'échelle de la syllabe. Il faudra donc se poser le problème de la fusion des informations issues des modèles acoustiques, phonétiques et rythmiques, fournies de manière asynchrone, afin de disposer d'un système d'identification automatique des langues complet.

Dans le chapitre suivant, nous discuterons de la validité de certains traits prosodiques pouvant être pris en compte dans un système d'identification automatique des langues.

Chapitre 4

Caractérisation des langues par la prosodie

Pendant le stage de DEA, des tests prospectifs ont étés menés afin de déterminer la petinence de différents paramètres peuvent être utilisés en identification automatique des langues.

Le premier paramètre examiné est la fréquence fondamentale. Un étiquetage décrivant les variations de la fréquence fondamentale permet de modéliser celle-ci en employant des multigrammes.

Ensuite, nous avons étudié la modélisation de la durée des consonnes sur le corpus OGI-MLTS (Annexe B). Enfin, la dernière partie du travail de test effectué est l'étude d'une segmentation en pseudosyllabes dans le but d'une modélisation de la durée.

4.1 Tests prospectifs par rapport à la prise en compte de la prosodie

4.1.1 Étude de la fréquence fondamentale

4.1.1.1 INTSINT

INTSINT est un outil proposé par Hirst et De Cristo, qui permet d'obtenir une transcription formelle et inversible de la structure mélodique. Il utilise un modèle de trajectoires, MOMEL [Hir91], permettant d'extraire automatiquement un certains nombre de points cibles connectés par des fonctions spline quadratiques. Chaque point cible est codé par un symbole :

- T ("top") : valeur la plus haute
- B ("bottom") : valeur la plus basse
- M ("mid") : premier point cible (à moins qu'il ne soit déja codé T ou B)
- U ("up") : $F_{i-1} < F_i < F_{i+1}$

- D ("down") : $F_{i-1} > F_i > F_{i+1}$
- S ("same") : $F_{i-1+} = F_i$
- H ("higher") : $F_{i-1} < F_i > F_{i+1}$
- L ("lower") : $F_{i-1} > F_i < F_{i+1}$

Les symboles T, B et M codent les tons absolus et U, D, S, H et L des tons relatifs.

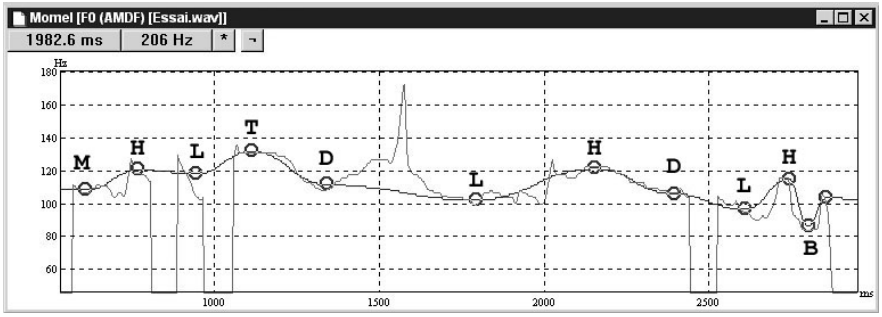

Figure 6 - Exemple de labelisation des contours de la fréquence fondamentale par l'algorithme MOMEL

4.1.1.2 Modélisation

Nous avons modélisé la suite de symboles par des multigrammes de longueurs variables [Del96]. Les multigrammes permettent de détecter des motifs récurrents dans des suites d'observations. Ces motifs peuvent avoir une longueur variable. Modéliser un signal par des multigrammes revient à trouver la segmentation $S = (s_1, ..., s_{n(s)})$ la plus probable d'une séquence d'observations $O = (O_1...O_T)$:

$$S^* = \arg\max \lambda(O, S)$$

avec la vraisemblance

$$\lambda(O, S) = \prod_{i=1}^{n(S)} P(Z_i)$$

$$Z_i = (O_{s_i}...O_{s_{(i+1)-1}})$$

L'algorithme d'apprentissage est un algorithme itératif de type Expectation Maximisation. La segmentation la plus probable est calculée avec l'algorithme de Viterbi.

Lorsque l'apprentissage est terminé, on possède un dictionnaire contenant les séquences Z_i les plus probables et leurs vraisemblances. La qualité du modèle est estimée d'après la perplexité d'une séquence d'observations O en utilisant la segmentation la plus vraisemblable :

$$PP_{V_i}(O) = 2^{-\frac{1}{T} log \lambda^*(O)}$$

avec $T =$ nombre d'observations et $\lambda^*(O) = \arg\max \lambda(O, S)$

4.1.2 Expérimentations

Nous avons fait varier les paramètres suivants :
- nombre maximal de mots dans une séquence multigramme,
- nombre d'occurrences au dessus desquelles une séquence de mots est incluse dans l'inventaire initial des séquences,
- nombre d'occurrences au dessus desquelles une séquence de mots est incluse dans l'inventaire des séquences durant les itérations,
- nombre d'itérations.

Les tests sont effectués sur le corpus MULTEXT [Cam98] contenant des enregistrements non bruités de parole, et ce pour 5 langues européennes (anglais, français, allemand, espagnol et italien). Ce corpus est constitué d'enregistrements de parole de bonne qualité, échantillonnée à 16 kHz, codée sur 16 bits, enregistrée dans une chambre anéchoïque. Les données sont divisées en deux parties (test et apprentissage) pour chaque langue. On se limite volontairement au même nombre de fichiers par langue :
- pour l'apprentissage, nous utilisons les 80 premiers fichiers pour chaque langue, soit 8 locuteurs (environ 27 minutes par langue),
- pour les tests, nous employons les 20 fichiers suivants, pour 2 locuteurs différents de ceux d'apprentissage (environ 20 secondes par fichier).

Les résultats sont présentés ici et plus loin sous la forme de matrices de confusion :
- sur les lignes du tableau sont répertoriés l'appartenance réelle des fichiers de tests,
- en colonnes, nous avons repertorié les langues identifiées,
- à l'intérieur des tableaux, les chiffres correspondent au nombre de fichiers identifiés comme appartenant à la langue de sa colonne.

Les meilleurs résultats sont obtenus pour :
- nombre max de mots dans une séquence multigramme : 5,
- nombre d'occurrences au dessus desquelles une séquence de mots est incluse dans l'inventaire initial des séquences : 3,
- nombre d'occurrences au dessus desquelles une séquence de mots est incluse dans l'inventaire des séquences durant les itérations : 3,
- nombre d'itérations : 10.

Pour ces valeurs de paramètres nous obtenons les résultats suivantes :

	English	German	Italian	French	Spanish
English	10	5	0	2	3
German	2	6	4	6	2
Italian	5	9	1	4	1
French	5	4	2	7	2
Spanish	0	7	4	1	8

Taux moyen d'identification correcte : 32%

Tableau 1 - Résultats obtenus par modélisation par multigrammes des étiquettes représentant les variations de la fréquence fondamentale

4.1.2.1 Interprétation

Les résultats obtenus sont juste au-dessus des 30%. Dans la littérature, Itahashi [Ita99] obtient des scores du même ordre mais sur un corpus nettement plus bruité. On peut donc craindre une baisse des performances si on emploie cette méthode sur un corpus bruité. Les symboles représentant la fréquence fondamentale ne semblent pas suffisament précis pour être employés dans une tâche d'identification des langues.

4.1.3 Travail sur la durée des systèmes consonantiques

En travaillant sur le corpus OGI-MLTS [Mut92], nous avons voulu examiner la validité d'une différenciation possible entre consonnes et voyelles au travers du paramètre "durée". Le corpus OGI-MLTS a été étiqueté manuellement par des linguistes (voir Annexe B). Les étiquettes fournies contiennent la classe du phonème reconnu (voyelle, consonne fricative, ...) ainsi que sa durée dans le temps. Nous avons visualisé la séparation entre les langues au travers des durées des classes phonétiques pour l'ensemble des individus du corpus OGI-MLTS :

- un individu est un fichier de 45 secondes de parole spontanée,
- il est caractérisé par deux paramètres : moyenne de la durée des voyelles sur le fichier et moyenne de la durée d'une sous classe de consonnes (occlusives, plosives, fricatives voisées / non voisées, et sonantes).

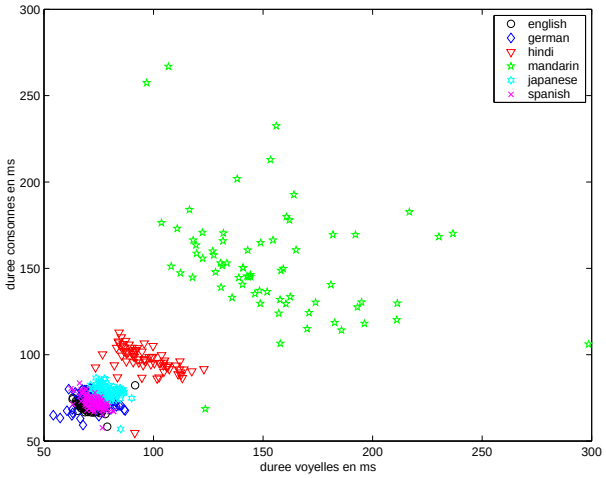

Figure 7 - Discrimination entre les langues d'après les durées des voyelles et des consonnes sans distinction de classe.

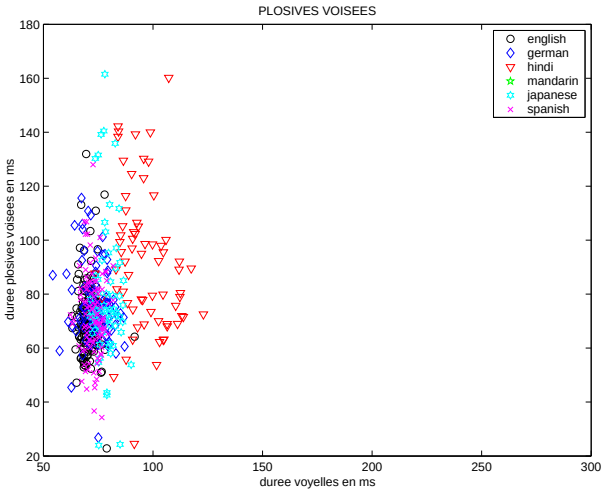

Figure 8 - Discrimination entre les langues d'après les durées des voyelles et des consonnes plosives voisées.

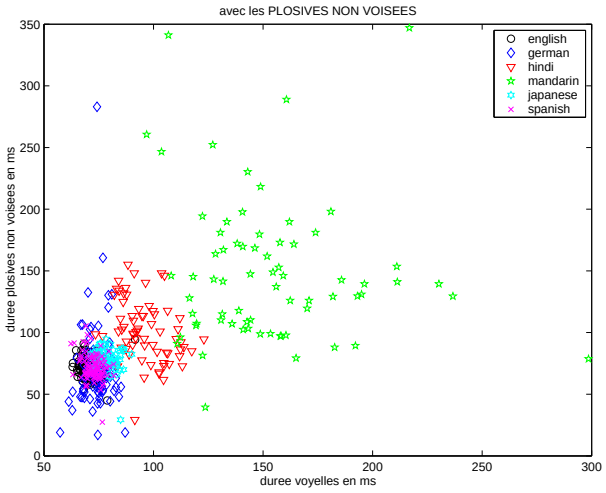

Figure 9 - Discrimination entre les langues d'après les durées des voyelles et des consonnes plosives non voisées.

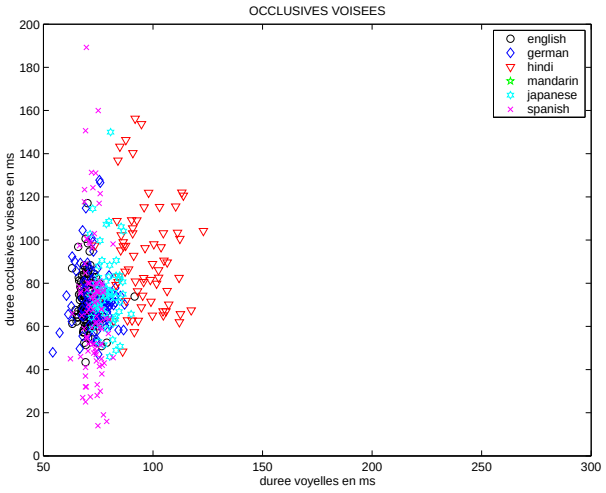

Figure 10 - Discrimination entre les langues d'après les durées des voyelles et des consonnes occlusives voisées.

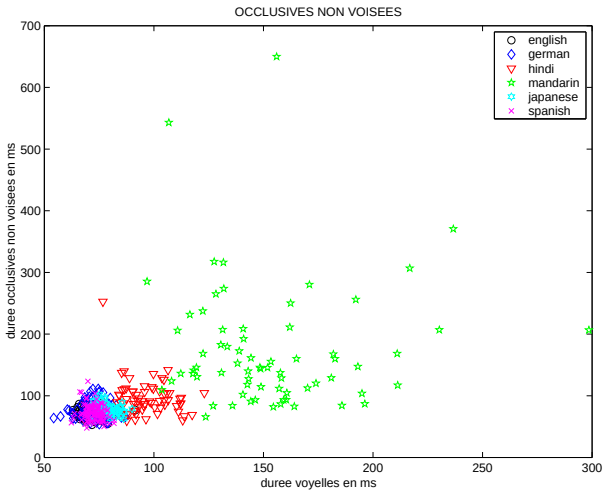

Figure 11 - Discrimination entre les langues d'après les durées des voyelles et des consonnes occlusives non voisées.

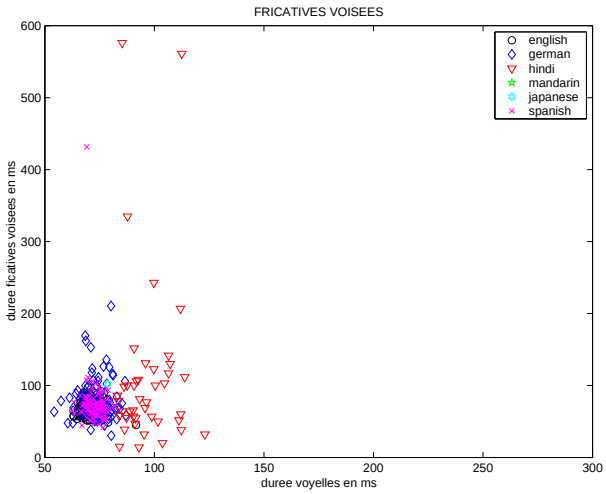

Figure 12 - Discrimination entre les langues d'après les durées des voyelles et des consonnes fricatives voisées.

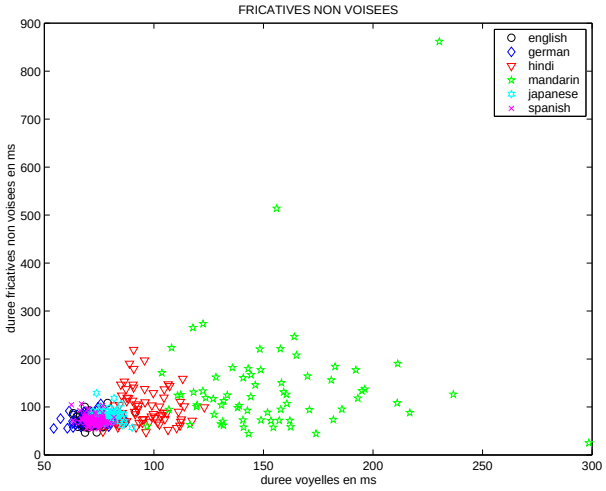

Figure 13 - Discrimination entre les langues d'après les durées des voyelles et des consonnes fricatives non voisées.

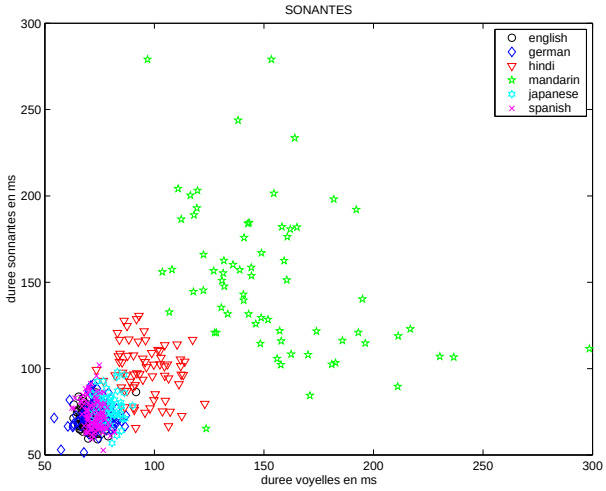

Figure 14 - Discrimination entre les langues d'après les durées des voyelles et des consonnes sonantes.

On observe sur les figures 7, 9, 11 et 14 un détachement net des points correspondant au Mandarin par rapport aux autres langues. Une autre langue se détache également sur les figures 7 et 14, il s'agit de l'hindi. Alors que les voyelles et les consonnes du mandarin sont plus longues que celles des autres langues, on note l'absence de consonnes voisées en Mandarin. Les langues européennes restent regroupées dans les bas des figures (durées des consonnes et voyelles courtes).

D'après les observations, on peut penser qu'un critère basé sur la durée des systèmes consonantiques peut être employé en identification automatique des

langues. Il pourra certainement aider à la décision, et permettra sûrement une bonne discrimination entre langues européennes et asiatiques. Il reste cependant à améliorer le pré-traitement automatique nécessaire en amont.

Toutefois, une réserve est à faire quant aux résultats obtenus : plusieurs linguistes ont été sollicités pour l'étiquetage manuel, même sur une seule langue, (voir Annexe B) ce qui peut indroduire un biais au niveau des étiquettes.

4.2 Modélisation du rythme

4.2.1 Segmentation en pseudosyllabes

Pour modéliser le rythme, nous nous sommes basés sur une segmentation en pseudo-syllabes, décrite dans [Far01], à partir d'une segmentation en segments vocaliques.

4.2.1.1 Segmentation en segments vocaliques

La détection des segments vocaliques décrite dans [Pel98] est basée sur la recherche d'évenements caractéristiques d'une voyelle par analyse spectrale. La principale difficulté est d'obtenir des règles robustes par rapport aux conditions d'enregistrement. au changements de locuteurs et aux changements de langues.

Deux prétraitements sont appliqués au signal : une segmentation statistique *a priori* et une détection d'activité vocale. La segmentation statistique *a priori* emploie l'algorithme de divergence forward-backward [And88]. La détection d'activité vocale se fait par seuillage. Le seuil d'activité vocale est défini par :

$$T_a = \alpha . \min_{s_i}(\sigma_{s_i(z)})$$

avec

- z : signal acoustique,
- $S = (s_1, s_2, ..., s_N) = \{s_i\}_{i=1...N}$ la suite de N segments,
- $\sigma_{s_i}(z)$: écart-type du signal z pour le segment s_i
- $\alpha = 2,5$

La détection des segments vocaliques se fait par analyse spectrale, au moyen d'une fenêtre glissante de 32 ms, avec un recouvrement de 16 ms. La Transformée de Fourier Rapide (TFR) de ce signal est ensuite calculée selon l'échelle de Mel. A partir d'un vecteur d'énergie dans 24 canaux cepstraux, la fonction $Rec(t)$ (Reduced Energy Cumulating) est alors calculée :

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

avec :

- $E_i(t)$: énergie du signal sur la bande de fréquence i à l'instant t.

- $E_{BF}(t) = \sum_{i=1}^{10} \alpha_i E_i(t)$
- $\underline{E(t)} = \sum_{i=1}^{24} \alpha_i E_i(t)$
- $\overline{E_(t)}$: moyenne spectrale de l'énergie.
- $\forall\, i \begin{cases} 350\,Hz \leq F_i \leq 3500\,Hz,\ \alpha_i = 1 \\ F_i \leq 350\,Hz\, ou\, F_i \geq 3500\,Hz,\ \alpha_i = 0 \end{cases}$
  pour se restreindre à la bande passante du canal téléphonique.

Les pics de la fonction $Rec(t)$ indiquent les positions des segments vocaliques. Un seuillage est alors effectué afin de limiter les erreurs de fausse détection. Le seuil S_e est fixé arbitrairement à un dixième de la valeur moyenne de la fonction $Rec(t)$ sur T trames : $S_e = \frac{1}{10T} \sum_{i=1}^{T} Rec(t)$

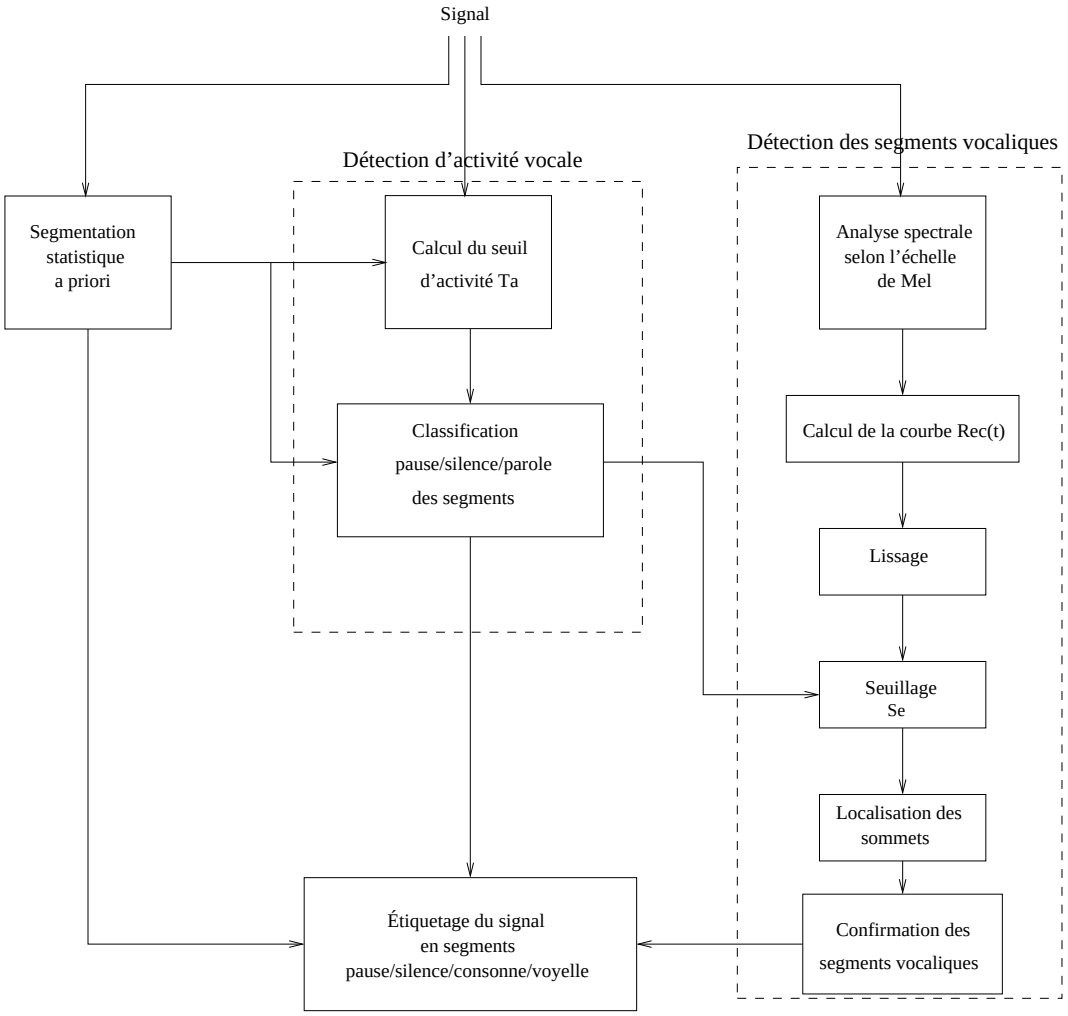

Figure 15 - Système de détection des segments vocaliques [Pel98]

4.2.1.2 Pseudosyllabes

Chaque pseudo-syllabe est composée d'une suite de segments non vocaliques se terminant par un segment vocalique. Le nombre de segments non vocaliques par pseudosyllabe est limité, pour éviter les erreurs de non-détection des voyelles. L'information contenue dans chaque pseudosyllabe est limitée à un vecteur contenant :

- la durée totale des segments non vocaliques contenus dans la pseudosyllabe,
- la durée du segment vocalique terminant la pseudosyllabe,
- le nombre de segments non vocaliques dans la pseudosyllabe.

4.2.2 Expérimentations

L'ensemble des observations (pseudosyllabes) d'une langue est modélisé par un mélange de gaussiennes. Le nombre Q de composantes gaussiennes dans le mélange est déterminé *a priori*. L'initialisation conjointe des paramètres des lois gaussiennes se fait par le biais d'une phase préliminaire de Quantification Vectorielle (QV). Cette étape vise à trouver une partition optimale de l'ensemble d'apprentissage en Q cellulles et à initialiser chaque composante gaussienne avec le centroïde et la distortion de chacune d'entre elles.

Le modèle initial une fois fixé, on applique un algorithme itératif d'optimisation de loi multigaussienne. Il s'agit d'un algorithme de type Expectation-Maximisation (EM). Une reformulation dans un cadre statistique de la modélisation obtenue précedemment par QV permet d'utiliser le modèle multigaussien comme classificateur.

Les expérimentations sont effectuées sur le corpus MULTEXT [Cam98] contenant 5 langues (anglais, français, allemand, italien et espagnol) (voir pg.25).

Nous avons fait varier différents paramètres, tels que le nombre maximal de consonnes par pseudo-syllabe, le nombre de gausiennes du mélange, la taille des fichiers de test. Puisque les données sont très limitées, en particulier en nombre de locuteurs, on utilise neuf des dix locuteurs pour l'apprentissage pendant que le dixième sert pour les tests. Ensuite, les tests portent sur le neuvième locuteur pendant que l'apprentissage est effectué sur tous les autres locuteurs (y compris le dixième). Cette procédure est itérée pour tous les locuteurs.

Nous avons effectuer les tests pour des longueurs de fichiers différentes :

- dans le premier test, on regroupe tous les fichiers où le locuteur de test apparaît en un seul, au total on a 10 fichiers de test par langue,
- dans le deuxième test, les dix fichiers prononcés par le locuteur de test sont testés un par un, ce qui fait 100 fichiers de test par langue.

Nous obtenons donc le tableau suivant :

Langue	Test 1	Test 2	Apprentissage
Anglais	4.4 min.	17.5 s.	40 min.
Français	3.6 min.	21.6 s.	35 min.
Allemand	7.4 min.	22.1 s.	65 min.
Italien	5.4 min.	21.7 s.	50 min.
Espagnol	5.4 min.	21.5 s.	50 min.

Tableau 2 - Durée des fichiers de test et d'apprentissage utilisés

4.2.3 Résultats

Les résultats obtenus varient bien évidemment en fonction des variables de fonctionnement. Les meilleurs résultats sont obtenus pour :

- le nombre maximal de segments non vocaliques par pseudosyllabe : 15,
- le nombre de lois gaussiennes utilisées pour le mélange : 16.

	English	French	German	Italian	Spanish
English	7	0	3	0	0
French	0	7	0	0	3
German	3	0	7	0	0
Italian	1	0	0	7	2
Spanish	0	0	0	0	10

Taux moyen d'identification correcte : 76%.

Tableau 3 - Test 1 - Durée moyenne des fichiers de test : 5.2 min.

	English	French	German	Italian	Spanish
English	53	2	28	7	10
French	4	78	0	0	18
German	33	0	57	8	2
Italian	11	0	6	61	22
Spanish	3	18	0	11	68

Taux moyen d'identification correcte : 63.4%.

Tableau 4 - Test 2 - Durée moyenne des fichiers de test : 20.8 s.

4.2.4 Interprétation

Les résultats obtenus en employant uniquement la durée des pseudosyllabes sont légèrement supérieurs à ceux obtenus grâce à une modélisation phonotactique. La taille des fichiers employés est inférieure à celle de ceux employés traditionnellement (45 s.). Cependant, le corpus utilisé pour les tests est un corpus de parole "propre", enregistré dans de bonnes conditions. Il reste donc à expérimenter sur un corpus de parole bruitée (téléphone) tel que le corpus OGI-MLTS, et à observer dans quelle mesure le taux de reconnaissance se dégrade.

Conclusion et perspectives

Durant ce stage, j'ai exploré les pistes permettant d'aller vers l'amélioration du système existant pour obtenir le système final désiré. L'approche par modélisation de la fréquence fondamentale n'apporte qu'une information limitée d'un point de vue identification des langues. La modélisation de la durée des systèmes consonantiques paraît prometteuse, mais la mise au point d'une segmentation automatique en grandes classes de consonnes est difficile. Le point le plus encourageant est la modélisation de la durée au travers de pseudosyllabes. Mais cette méthode mérite d'être validée sur des données de moins bonne qualité (corpus OGI-MLTS par exemple) avant de pouvoir être intégrée au système global. Une fois cette étape réalisée, il faudra résoudre le problème posé par la fusion des informations issues des modèles prosodiques et phonotactiques.

Bibliographie

[And88] R. André-Obrecht, *A new statistical approach for automatic speech segmentation*, IEEE Trans. on ASSP, vol. 36, n.1, pp.29-40, 1988.

[And93] R. André-Obrecht, *Segmentation et parole ?*, Habilitation à diriger des recherches, Rennes, 1993.

[Bag94] P. C. Bagshaw, *Automatic prosodic analysis for computer aided pronunciation teaching*, PhD. of Edinburgh Univ., 1994

[Cam98] E. Campione & J. Véronis, *A multilingual prosodic database*, Proc. of ICSLP 98, Sidney, 1998.

[Cum99] F. Cummins, F. Gers & J. Schmidhuber, *Langage identification from prosody without explicit features*,in Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech 99), Budapest, Hungary, September 1999.

[Del96] S. Deligne, *Modèles de séquences de longueurs variables, application au traitement du langage naturel et de la parole*, Thèse de l'ENST, Paris, 1996.

[Far98] J. Farinas, *La prosodie pour l'identification automatique des langues*, Mémoire de DEA, Université Paul Sabatier, Toulouse, juin 1998.

[Far00] J. Farinas, R. André-Obrecht, *Identification automatique des langues : variations sur les multigrammes*, Actes XXIIIèmes Journées d'Etude sur la Parole, JEP'2000, Aussois, 19-23 juin 2000.

[Far01] J. Farinas & F. Pellegrino, *Automatic rhythm modeling for language identification*, accepté à Eurospeech, 2001.

[Hes83] W. Hess, *Pitch determination of speech signals - Algorithms and devices*, Springer-Verlag, 1983.

[Hir91] D. Hirst, P. Nicolas, R. Espesser, *Coding the F0 of a continuous text in french : an experimental approach*, Proceeding of the XIIth International Conference on Phonetic Sciences, Vol. 5, Aix-en-Provence, pp. 234-237, août 1991.

[Ita99] S. Itahashi, T.Kiuchi, M. Yamamoto, *Spoken language identification using speech fundamental frequency and cepstra*, Eurospeech 99, Budapest, Hungary, September 1999.

[Li94] K. P. Li, *Automatic language identification using syllabic spectral features*, Proc. of ICASSP'94, Adelaide, pp.297-300, 1994.

[Mut92] Y. K. Muthusamy, R. A. Cole and B. T. Oshika *The OGI multi-language telephone speech corpus*, in Proceedings of the 1992 International Conference on Spoken Language Processing (ICSLP 92), Alberta, October 1992.

[Mut94] Y. K. Muthusamy, E. Barnard and R. A. Cole, *Reviewing Automatic Language Identification*, IEEE Signal Processing Magazine, October 1994.

[Pel97] F. Pellegrino, R.Andre-Obrecht, *Vocalic system modeling : a VQ approach*, IEEE Digital Signal Processing'97, July 1997, Santorini.

[Pel98] F. Pellegrino, *Une approche phonétique en identification automatique des langues : la modélisation des systèmes vocaliques*, Thèse de 3^{eme} cycle, Université Paul Sabatier, Toulouse, décembre 1998.

[Thy96] Thymé-Gobbel & Hutchins, *On using prosodic cues in automatic language identification*, Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP 96), Philadelphia, USA, October 1996.

[Sav91] M. Savic, E. Acosta & S. K. Gupta, *An automatic language identification system*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 91, Toronto, Canada, May 1991.

[Zis96] M. A. Zissman, *Comparison of four approaches to automatic language identification*, IEEE Transactions on speech and audio processing, vol.4, n.1, pp.31-44, janvier 1996.

Annexes

Annexe A

Méthode d'extraction de la fréquence fondamentale : AMDF [Hes83]

Le critère de variation d'amplitude à court terme (Average Magnitude Difference Function) [Miller et Weibel 56] utilise la valeur absolue des différences point par point. La fonction de périodicité de l'AMDF est :

$$FP_{AMDF}(\tau) = \sum_{i=1}^{n} |s_i - s_{i+\tau}|$$

Cette fonction est ensuite normalisée par n ou par $\sum_{i=1}^{n} |s_i|$ pour que la valeur de l'AMDF puisse être comparée à un seuil absolu dans le but de décider si le signal est périodique ou non.

La fonction de périodicité présentera un minimum au niveau des multiples de la période. Cette méthode n'utilise pas l'hypothèse de stationnarité du signal. D'ailleurs, l'ambiguïté entre les pics T_o, $2 * T_o$, $3 * T_o$, ... est souvent atténuée par la non-stationnarité du signal analysé : plus le décalage est grand, plus le signal, de part sa non-stationnarité, intègre de différences par rapport à la trame de départ ; le signal présentera alors plus de différences pour un décalage de $2 * T_o$ que pour un décalage de T_o.

Une conséquence de la non utilisation de l'hypothèse de stationnarité est que le choix de la taille des fenêtres de signal et de signal décalé est libre : l'AMDF utilise deux fenêtres de taille fixe, mais il est possible de concevoir des algorithmes avec des tailles de fenêtre variable, par exemple égale au décalage testé (méthode SRPD).

Cette résistance au problème des erreurs grossières et la rapidité de calcul font de l'AMDF une méthode couramment employée.

Annexe B

Le corpus OGI-MLTS [Mut92]

Cette base de données est la référence dans le monde de l'IAL. Sa réalisation est due au NIST (organisme de normalisation américain), qui l'a élaborée en vue d'effectuer des campagnes d'évaluation en identification automatique des langues. Elle a été utilisée successivement avec 6, 9, 10, puis 11 langues. Il s'agit de parole téléphonique échantillonnée à 8 kHz dans une ambiance le plus souvent bruitée. Les locuteurs doivent répondre à un certain nombre de 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. Énumerez 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. ...

A cela s'ajoute l'enregistrement d'une minute de parole spontanée pour chaque locuteur, enregistrée sous le nom de "story". Chaque réponse est enregistrée avec un temps imparti pour la réponse limité (10 secondes pour la question 5 par exemple). On obtient finalement pour chaque locuteur environ 2 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 génants ; en particulier, pour le corpus dit "français", plusieurs locuteurs ont un fort accent de français canadien, ou tout au moins de français non métropolitains. Dans ce cas précis, il s'agit plus d'un corpus francophone que français. Il est vraisemblable que cet aspect se retrouve sur d'autres langues (anglophone, hispanophone...).

D'autre part, 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, ...) soit sous forme phonétique classique.

L'étiquetage phonétique classique est disponible sur 6 langues 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. L'étiquetage obtenu se révèle parfois inexact ou ambigu (cas de bruits non produits par le locuteur et étiquetés comme des occlusives).

Ces quelques remarques étant formulées, on peut considérer que OGI est la base de données sur laquelle les performances de la plupart des systèmes sont évaluées à l'heure actuelle, et qu'elle représente à ce titre une contribution majeure au domaine de l'IAL. La plupart des tests sont réalisés avec le signal nommé "story", limité aux 45 premières secondes.