# La prosodie pour l'Identification Automatique des Langues

Jérôme FARINAS

23 juin 1998

# Table des matières

**1 Cadre de l'étude** **5**
1.1 Importance de l'Identification Automatique des Langues . . . . . 5
1.2 Le projet de recherche . . . . . 5
1.3 Le sujet de DEA . . . . . 6
1.4 Organisation du rapport . . . . . 7

**2 La prosodie** **8**
2.1 Définition . . . . . 8
2.2 Fonctions . . . . . 9
2.3 Paramètres . . . . . 11
2.3.1 Énergie . . . . . 11
2.3.2 Durée . . . . . 11
2.3.3 Fréquence fondamentale . . . . . 12
2.4 Problèmes . . . . . 13

**3 Méthodes d'extraction de la fréquence fondamentale** **14**
3.1 Principe général . . . . . 14
3.2 Algorithmes de type corrélation . . . . . 15
3.2.1 Autocorrélation . . . . . 16
3.2.2 Fonction de distance (AMDF) . . . . . 16
3.2.3 Super résolution (SRPD) . . . . . 17
3.3 Peigne spectral . . . . . 17
3.4 Programmation dynamique . . . . . 18
3.5 Méthodes combinatoires . . . . . 18
3.6 Conclusion . . . . . 19

**4 Identification Automatique des langues et prosodie** **20**
4.1 Description d'un système d'IAL . . . . . 20
4.1.1 Problématique . . . . . 20
4.1.2 Les sources d'informations . . . . . 21
4.1.3 Principe général d'un système . . . . . 21
4.2 Systèmes d'IAL utilisant la prosodie . . . . . 22
4.2.1 Système de Hazen et Zue . . . . . 22

4.2.2 Système de Muthusamy . . . . . 26
4.2.3 Système de Li . . . . . 27
4.2.4 Système de Itahashi . . . . . 28

**5 Réalisation d'une modélisation automatique de la prosodie pour l'IAL 29**
5.1 Système général . . . . . 29
5.2 Corpus utilisés . . . . . 30
5.3 Paramètres utilisés . . . . . 31
5.4 Modèles prosodiques . . . . . 32
5.5 Modèle de langage . . . . . 35
5.5.1 Description du modèle multigramme [Deligne 96] . . . . . 35
5.5.2 Application au modèle prosodique . . . . . 37
5.6 Résultats . . . . . 37
5.6.1 Résultats de la modélisation de la prosodie . . . . . 37
5.6.2 Résultats en Identification des Langues . . . . . 39

**6 Conclusions et perspectives 40**

# Remerciements

Je tiens tout d'abord à remercier Régine Andre-Obrecht de m'avoir acceuilli dans son équipe, et qui par son dynamise, sa bonne humeur, ses précieux conseils, m'a soutenu tout au long de l'année.

Un grand merci à François Pellegrino : sa patience, ses idées, ses capacités d'analyse m'ont été d'un très grand service.

Merci à Arnaud Galinier, Reshad Khoyratty, Pierre Larrey, Pascal Ortega, Nathalie Parlangeau, Christine Senac, et tous ceux qui m'ont aidé de près ou de loin, de m'avoir fait passer une bonne année.

Merci à Robert Espesser pour la patience et la générosité qu'il a manifesté au cours de la correspondance que nous avons échangée à propos de MESSIGNAIX.

# Chapitre 1

# Cadre de l'étude

## 1.1 Importance de l'Identification Automatique des Langues

Depuis dix ans, le développement économique et social au niveau international a provoqué une mondialisation des télécommunications accompagnée d'une explosion du marché des systèmes de synthèse vocale et des serveurs vocaux interactifs. La prise en compte de l'aspect multilingue par ces systèmes devient une priorité dans de nombreux domaines tels que l'interrogation de systèmes d'information (services de transport, spectacle, hôtellerie, météorologie...), la consultation de bases de données documentaires multilingues, l'enseignement ou encore la traduction automatique [Rigoni 96]. La mise en oeuvre de tels services nécessite par conséquent des systèmes interactifs capables, en temps réel, de diriger un appel prononcé dans une langue quelconque vers l'opérateur ou le service automatique correspondant. Cet aspect fondamental des applications multilingues est pris en charge par le module d'identification automatique de la langue [Muthusamy et al. 94].

## 1.2 Le projet de recherche

Le thème de l'Identification Automatique des Langues (IAL) est abordé dans l'équipe IHM-PT de l'IRIT, depuis deux ans sous son aspect fondamental, dans le cadre du projet Discrimination Automatique Multilingue financé par la DGA, réunissant trois Instituts de Phonétique (ICP Grenoble, ILPGA Paris, DDL Lyon) et l'IRIT. L'étude est orientée suivant deux axes :

- la recherche d'une typologie des langues selon des indices discriminants et détectables automatiquement,

- une meilleure connaissance des systèmes vocaliques (nombre de voyelles, représentation fréquentielle, fréquence d'apparition) et la recherche d'un modèle

probabiliste de l'espace acoustique vocalique pour chaque langue.

Dans le cadre de cette étude, nous avons montré qu'avec une simple détection automatique des noyaux vocaliques présents dans le signal et une modélisation stochastique de type Mélanges de Lois Multigaussiennes, des performances en IAL intéressantes pouvaient être atteintes [Pellegrino et Obrecht 97a] sur le corpus de la parole téléphonique OGI-MLTS.

Notre groupe de travail se propose actuellement de fournir un système complet d'identification automatique de la langue, regroupant des aspects acoustiques, phonotactiques et prosodiques, après avoir cerné les paramètres discriminants. Ce système se décomposera en trois modules :

- un module de décodage acoustico-phonétique généralisé. Il s'agit d'élargir l'approche vocalique décrite ci-dessus à l'ensemble des sons de chaque langue en proposant un modèle probabiliste global [Pellegrino et al. 97] ou plusieurs modèles différenciés,

- un modèle de langage pour gèrer les séquences de sons. Il sera basé sur des modèles probabilistes multigrammes [Deligne 96] qui se substitueront aux traditionnels modèles bigrammes ou trigrammes,

- un modèle prosodique. Ce module devra exploiter les critères discriminants du point de vue prosodique. Cet aspect étant principalement dynamique et suprasegmental, il est nécessaire de définir des unités intrinsèquement prosodique, puis de développer une modélisation statistique adaptée. Ce module n'avait pas fait l'objet d'une étude au sein de l'équipe, le travail de DEA s'insère comme un premier pas dans cette recherche.

## 1.3 Le sujet de DEA

La première phase du stage DEA a consisté à recenser les principales méthodes d'extraction de la fréquence fondamentale F0 (un des principaux paramètres prosodiques) et à en comparer les performances tant sur des bases de données prosodiques classiques [Bagshaw 94] que sur les corpus de référence en IAL (OGI-MLTS). Cette étude a pour but de déterminer la méthode la plus robuste dans le cadre multilingue et en environnement difficile.

La deuxième phase a consisté en l'étude multilingue de la structure prosodique suprasegmentale. Cette étude a pour but de faire émerger des unités ou des motifs prosodiques élémentaires et d'étudier leur modélisation. En effet, les études existantes confirment la pertinence de la prosodie en IAL (cf. système de Itahashi) mais aussi le manque d'efficacité d'une modélisation prosodique à un niveau segmental ; en particulier, une modélisation centiseconde du F0 se révèle assez logiquement inefficace (cf. système de Hazen et Zue). Une fois l'échelle temporelle des motifs prosodiques et leur caractérisation acoustique déterminées, il s'agit d'en proposer une modélisation inspirée des modèles phonétiques classiques mais adaptée aux nouvelles unités. Une modélisation de type Modèles Stochastiques de

Trajectoires (MST) semble particulièrement appropriée car elle autorise la définition de trajectoires assez conforme à la réalité prosodique. Nous n'avons pu avoir accès aux implémentations des MST, nous avons donc employé une modélisation des unités prosodiques par Modèles de Markov Cachés (MMC).

Le modèle prosodique d'une langue se décompose alors en deux niveaux :

- une modélisation des unités de base par MMC,
- un réseau stochastique traitant les enchaînements de ces unités (type bigramme ou multigramme).

Dans le cadre du DEA, l'étude a porté sur deux langues seulement, le français et le japonais. Mais les méthodes et les modules développés on été conçus de telle façon qu'ils restent facilement adaptables au traitement d'une nouvelle langue et que l'ajout d'une nouvelle langue ne nécessite aucun nouveau corpus étiqueté manuellement.

## 1.4 Organisation du rapport

Le chapitre 2 est consacré à la présentation de la prosodie : quelques définitions, ses fonctions, les principaux paramètres utilisés et les problèmes rencontrés.

Le chapitre 3 présente le principe général des méthodes d'extraction de la fréquence fondamentale ainsi que quelques algorithmes robustes qui sont couramment utilisés dans les systèmes prosodiques.

Au cours du chapitre 4, nous abordons le principe de fonctionnement général des systèmes d'IAL ainsi que des systèmes qui utilisent des paramètres prosodiques pour l'identification de la langue.

Le chapitre 5 est dédié aux expérimentations effectuées dans le cadre de la réalisation du module prosodique. Dans un premier temps nous décrivons la structure générale du module, puis le corpus utilisés et les différents paramètres qui en ont été extraits. Ensuite nous abordons le modèle prosodique composé de la modélisation des unités de base et du réseau stochastique multigramme.

Au chapitre 6 un bilan de ce qui a été effectué pendant le stage est effectué et nous présentons les perspectives de ces travaux.

L'annexe A est consacré à MES SIGNAIX, la boîte à outils pour le traitement du signal développée à l'université d'Aix-en-provence. Les algorithmes d'extractions du F0 ont été particulièrement utiles tout au long du stage.

L'annexe B contient les détails du calcul d'une dérivée d'un signal discret utilisée lors de l'extraction de paramètres prosodiques.

L'annexe C présente les bases de données développées pour l'évaluation des différentes méthodes d'extraction du F0, ainsi que quelques résultats obtenus avec les méthodes testées dans le cadre du DEA.

L'annexe D présente le modèle prosodique INTSINT qui a inspiré la modélisation des unités de base de notre modèle prosodique.

# Chapitre 2

# La prosodie

Dans ce chapitre nous abordons quelques définitions sur la prosodie, puis nous présenterons quelques fonctions [Geoffrois 95]. Ensuite nous examinons les différents paramètres utilisés dans les sytèmes prosodiques ainsi que les difficultés qui peuvent surgir lors de leur extraction. Enfin nous nous intéressons à l'apport de la prosodie au domaine de l'identifation des langues.

## 2.1 Définition

Du point de vue acoustique, la prosodie désigne les phénomènes liés à la variation dans le temps des paramètres de hauteur, d'intensité et de durée. La perception de **hauteur** est essentiellement liée à la **fréquence fondamentale** (notée **F0**) qui correspond, au niveau physiologique de la production de la parole, à la fréquence de vibration des cordes vocales. La perception d'**intensité** est essentiellement liée à l'amplitude et à l'**énergie** du son, mais dépend aussi partiellement de sa durée. La perception de **durée** correspond à son temps d'émission, sa durée acoustique.

Si nous nous plaçons sur le plan perceptuel, la variation dans le temps des paramètres ci-dessus correspond à la perception du rythme des phrases, de leur mélodie et de l'accentuation. Le **rythme** des phrases est perçu grâce à l'enchaînement des durées des segments. La **mélodie** de la phrase correspond à l'évolution dans le temps de la hauteur. L'**accentuation** est un phénomène de plus haut niveau, qui consiste à mettre en relief une syllabe ou une more[1] par rapport à son environnement immédiat. On distingue généralement deux catégories de langues :

- les langues à **accent fixe** pour lesquelles la position accentuable du mot (accent d'insistance excepté) est toujours placée sur une syllabe déterminée (en français la dernière syllabe du mot).

- les langues à **accent libre** pour lesquelles la position accentuable dépend de la fonction lexicale et de la structure morphologique du mot.

1. une more est une unité inférieure à la syllabe et qui peut porter l'accent.

## 2.2 Fonctions

La prosodie intervient à tous les niveaux dans les phénomènes de parole et remplit des rôles variés dans la communication parlée. En voici quelques unes extraites de [Geoffrois 95] :

– **Distinction entre homonymes**

Dans les langues à accent libre et dans les langues à ton, la prosodie est associée aux choix lexicaux. Elle permet en particulier de discriminer des homonymes. En anglais, la position de l'accent tonique peut distinguer un nom d'un verbe :

**seg**ment (un segment)
seg**ment** (segmenter)

En espagnol l'accent tonique permet de différencier certains mots :

**pla**tano (une banane)
pla**ta**no (un platane)

– **Structuration de l'énoncé**

Dans les langues dites à accent fixe, la position de l'accent lexical est la même pour tous les mots. Par exemple, en français, il tombe sur la dernière syllabe pleine[2] des mots. En tchèque il tombe sur la première. Il permet ainsi de déterminer les frontières de mots. C'est ce qui est appelé sa fonction **démarcative**.

Plus généralement, dans toutes les langues, la prosodie participe à la segmentation de l'énoncé en groupes plus petits. Elle indique aussi le type de relation que ces groupes entretiennent. Elle permet aussi la hiérarchisation au sein de la phrase. Cette structure n'est pas identique à la structure syntaxique, mais elle est en relation avec elle.

Voici un exemple qui illustre comment la prosodie participe au regroupement de mots :

L'instituteur, dit le directeur, est un incapable.
L'instituteur dit : "le directeur est un incapable".

Les situations où il y a risque de confusion entre deux phrases ou expressions, appelées oronymes, sont aussi courantes en anglais :

That reflects the secretariat's fear of competence.
That reflects the secretariat's sphere of competence.

2. c'est-à-dire excepté un éventuel "e" final.

Dans cette situation la prosodie peut jouer un rôle, en conjonction avec d'autres sources d'information comme le contexte sémantique, pour déterminer le sens exact de la phrase.

Au delà de la phrase, la prosodie joue un rôle dans la structuration du discours. Elle peut par exemple marquer un changement de sujet dans la conversation. L'indice prosodique le plus souvent étudié est la dynamique locale ("pitch range"). Le rythme et l'intensité sont deux autres indices importants.

– **Emphase et Focalisation**

L'accentuation est un moyen d'insister sur tel ou tel mot :

> **Je** vais terminer (par opposition à quelqu'un d'autre)
> Je **vais** terminer (par opposition à une action déjà accomplie)
> Je vais **terminer** (par opposition à une autre action)

L'accentuation peut aussi renforcer une opposition :

> Non, pas le six... le **dix** août.

– **Modalité**

La prosodie, et plus particulièrement la mélodie, est liée au mode de la phrase : affirmatif, interrogatif, impératif, ou exclamatif. En français, il est très courant de donner une intonation interrogative ou exclamative à une phrase tout en conservant sa structure grammaticale d'affirmation :

> Il va venir ?
> Il va venir.
> Il va venir !

– **Attitude et Interaction**

La prosodie véhicule l'attitude du locuteur vis à vis de l'énoncé ou ses intentions vis à vis de l'interlocuteur. En indiquant son adhésion plus ou moins forte envers l'énoncé, le locuteur exprime selon le contexte la conviction ou le doute, l'accord ou le désaccord, l'approbation ou la désaprobation, ou encore une invitation, une incitation. La dynamique de F0 est une des manifestations acoustiques.

La prosodie, surtout si l'on y inclut les phénomènes d'hésitation, permet aussi de gérer les tours de parole dans une conversation. Un ton final, un silence laissent le champ libre à l'interlocuteur pour intervenir. Un ton continuatif, une pause verbale ("filled pause"), servent à conserver son tour de parole.

– **Fonctions non linguistiques**

La prosodie au sens large traduit l'état psychologique du locuteur : calme ou énervé, triste ou gai, enthousiaste, surpris, etc. Elle caractérise aussi le locuteur en tant qu'individu ou membre d'un groupe (exemple : l'accent régional).

## 2.3 Paramètres

Les trois paramètres prosodiques utilisés dans la plupart des systèmes sont l'énergie, la durée et la fréquence fondamentale. Comme pour les paramètres utilisés en reconnaissance (LPC, cepstre, etc.), la première étape de l'extraction de paramètres prosodiques est généralement une analyse à court terme du signal, en faisant l'hypothèse que pour une durée d'analyse assez courte le signal est quasi-stationnaire. En général une fenêtre d'analyse de 30 ms est utilisée et cette analyse est répétée à intervalles réguliers, ou trames, typiquement toutes les 10 ms. Mais pour traduire le caractère suprasegmental[3] de la prosodie, il est nécessaire de considérer une unité au dessus de la trame. D'autres paramètres pourront être dérivés en tenant compte de plusieurs trames successives.

### 2.3.1 Énergie

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

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

Étant donné sa dynamique et pour respecter l'échelle perceptive, elle est généralement exprimée en décibels :

$$E_{db} = 10 \times \log_{10}(\sum_{t=1}^{T} s_t^2)$$

Pour un signal échantillonné quelconque, on calcule l'énergie à court terme en prenant des portions de signal convoluées avec une fenêtre glissante (généralement assez étroite, de l'ordre de 5 à 10 ms).

Pour éliminer la variabilité du gain (dû à des conditions d'enregistrements différentes, par exemple à la distance entre le locuteur et le microphone), l'énergie peut être normalisée par rapport au maximum sur la phrase.

Pour tenir compte de résultats d'études psychoacoustiques selon lesquelles la perception d'intensité dépend de la longueur du son, un paramètre agrégé peut être formé en sommant l'énergie sur l'ensemble de la syllabe.

### 2.3.2 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

3. se dit d'une propriété portée par une unité plus grande que le segment, ou liée aux relations entre segments distants.

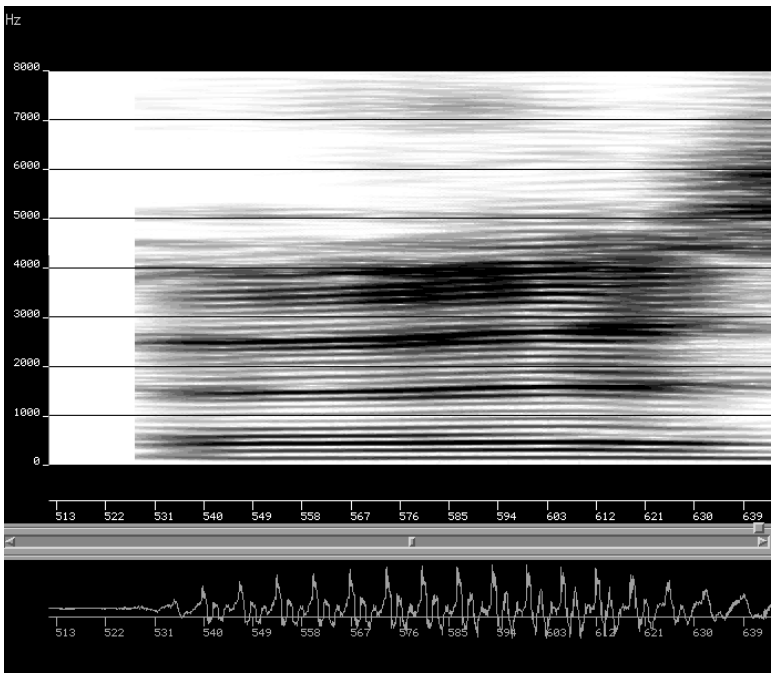



FIG. 2.1 - *un signal de parole et son spectrogramme à bande étroite*

temps. Les indices de durée classiques supposent généralement la donnée d'une segmentation, i.e. des frontières des unités dont on désire mesurer la durée: la durée d'une unité 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.

### 2.3.3 Fréquence fondamentale

La fréquence fondamentale correspond à la fréquence de vibration des cordes vocales. Les algorithmes d'extraction du F0 utilisent généralement une représentation temporelle ou spectrale du signal. Les méthodes temporelles utilisent la similarité du signal d'une période à l'autre pour identifier la période fondamentale. Il est parfois possible de repérer manuellement la periode fondamentale P0 (telle que $F0 = 1/P0$) directement sur le signal (cf. figure 2.1). Dans le domaine fréquentiel, les algorithmes utilisent généralement les harmoniques de la fréquence fondamentale pour trouver le F0. Cette propriété du signal peut être visualisée sur une FFT[4] à bande étroite (cf. figure 2.1).

Pour une description des algorithmes d'extraction de la fréquence fondamentale, voir le chapitre suivant.

4. Transformée de Fourier Rapide (Fast Fourier Transform)

## 2.4 Problèmes

La prosodie possède de nombreuses fonctions, et peut se révéler très utile au niveau de la reconnaissance, de la compréhension ou du dialogue. Or son usage ne s'est pas répandu en reconnaissance de la parole. Deux explications peuvent être avancées pour justifier cet état de fait :

- il manque une théorie unifiée sur la prosodie,
- les paramètres prosodiques manquent de robustesse.

En effet il existe un grand nombre de modèles prosodiques (Hirst, Fujisaki, Delattre, Rossi, Vaissière, Martin, Roach, Taylor [Geoffrois 95, pp. 27-46]), ce qui traduit la difficulté à décrire les phénomènes prosodiques. De plus, la prosodie est apprise et utilisée inconsciemment : elle n'est pas enseignée aux enfants comme on le fait pour les sons de la langue lors de l'apprentissage de la lecture. Elle n'est que très rarement enseignée explicitement dans l'apprentissage des langues étrangères, même si cela serait en fait souhaitable [Bagshaw 94].

Le manque de robustesse des paramètres prosodiques concerne en particulier la fréquence fondamentale. En effet, pour la durée, on dispose déjà de segmentations automatiques suffisament fiables pour estimer les paramètres de durée, et elles le seront de plus en plus, bien que cela passe par une étape de reconnaissance longue et sujette aux fautes. Il existe de nombreuses techniques de post-traitement pour la fréquence fondamentale, mais quoi que l'on fasse, il reste toujours des situations où selectionner une seule valeur de F0 conduit à des valeurs erronnées : l'apparition de valeurs aberrantes, même si c'est en proportion faible, compromet l'utilisation automatique de l'extraction du F0. Pour contourner ce problème, il est possible de développer des représentations intrinsèquement robustes.

# Chapitre 3

# Méthodes d'extraction de la fréquence fondamentale

De nombreuses recherches ont été menées dans le domaine de l'extraction de la fréquence fondamentale des signaux de parole. On peut citer l'ouvrage de référence [Hess 83], "Pitch Determination of Speech Signals - algorithm and devices" de W. Hess qui date de 1983, où sont proposés un grand nombre d'algorithmes.

Dans ce chapitre sont rassemblées quelques méthodes d'extraction de la fréquence fondamentale. Les méthodes présentées sont relativement robustes et il est assez aisé d'en trouver des implémentations.

## 3.1 Principe général

Un algorithme d'extraction de la fréquence fondamentale peut en général se décomposer en trois phases successives (cf. figure 3.1) :

1. un prétraitement et un changement de représentation,
2. l'extraction du fondamental,
3. un post-traitement visant à corriger les erreurs.

Le prétraitement vise à optimiser les caractéristiques du signal en vue de l'extraction en utilisant, par exemple, un filtrage passe-bas, une pré-accentuation ou un filtrage non linéaire. Ensuite des transformations sont appliquées pour adapter la représentation du signal au domaine (temporel, temporel à court terme, fréquentiel....) du traitement fondamental.

La deuxième phase consiste à extraire la fréquence fondamentale et dépend donc de l'algorithme utilisé. Généralement cela revient à optimiser une fonction de la fréquence fondamentale (fonction de coût, résultat d'une transformation, densité de probabilité).

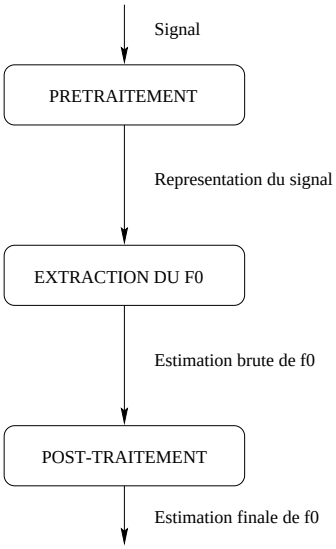



FIG. 3.1 - *schéma global d'un algorithme*

La phase de post-traitement a pour but de diminuer les erreurs qui sont de plusieurs types :

- les erreurs de voisement : lorsqu'une valeur de F0 a été trouvée sur une zone non-voisée, ou lorsqu'aucune n'a été trouvée sur une zone voisée

- les erreurs grossières ("gross-errors") : la fréquence fondamentale correspond a une harmonique ou une sous-harmonique. Ce type d'erreur peut facilement être corrigé en tenant compte du voisinage ou en effectuant un lissage.

- les erreurs fines : la valeur trouvée est située à plus ou moins 10% de la valeur réelle.

Nous allons maintenant présenter plusieurs familles d'algorithmes : l'autocorrélation, l'AMDF et le SRPD sont des méthodes qui travaillent sur une représentation temporelle à court terme alors que la méthode du peigne spectral travaille sur une représentation fréquentielle à court terme. Ensuite nous aborderons une méthode qui privilégie la phase de post-traitement et enfin nous décrirons le principe de fonctionnement des méthodes combinatoires.

## 3.2 Algorithmes de type corrélation

Les algorithmes de type corrélation travaillent dans le domaine temporel à court terme : le signal est extrait trame par trame et aucune transformation n'a été appliquée sur ces trames. La taille des trames est un paramètre important : lorsque elle a une valeur fixe elle contient généralement deux à trois périodes de signal. Pour la plupart des algorithmes il s'agit de trouver un extremum d'une fonction de la période appelée fonction de périodicité (FP).

Les méthodes de type corrélation se basent sur la similarité du signal entre

deux périodes : il est possible de corréler le signal de départ et une version décalée de ce signal, décalage correspondant à la période cherchée. Cette corrélation peut aussi être remplacée par une fonction de dissemblance.

### 3.2.1 Autocorrélation

Dans le cas de l'autocorrélation [Hess 83, pp. 351-356], les deux séquences en entrée sont dérivées du même signal. On introduit un décalage ("lag" en anglais) qui constitue le paramètre $\tau$ de la fonction d'autocorrélation :

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

avec $(s_i)_{i=1,n}$ la suite finie d'échantillons du signal.

Les maxima de la fonction de périodicité correspondent à des multiples de la période fondamentale. Le premier pic (le décalage donnant la meilleure corrélation) indique la fréquence fondamentale.

Cette méthode suppose que le signal soit stationnaire, tout au moins dans la trame utilisée, or cette hypothèse est rarement valide sur les signaux étudiés. Avec un signal de parole, l'extraction de la période sera donc moins simple car la moindre irrégularité du signal peut provoquer l´apparition de pics dont le décalage est inférieur à la période fondamentale. Dans ces conditions, pour améliorer la recherche du pic correspondant à la période, il est possible de choisir d'autres heuristiques, par exemple en ne retenant que les pics d'amplitude supérieure à un certain seuil.

### 3.2.2 Fonction de distance (AMDF)

Le critère de variation d'amplitude à court terme (Average Magnitude Difference Function) [Miller et Weibel 56], au lieu de calculer la corrélation entre deux signaux, 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 $T_0$, $2 * T_0$, $3 * T_0$, ... 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 $2 * T_0$ que pour un décalage de $T_0$.

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 (cf. SRPD).

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

### 3.2.3 Super résolution (SRPD)

Le SRPD (Super Resolution Pitch Determination) est un algorithme proposé par Medan, Yair et Chazan [Medan et al. 91] avec comme objectif initial de réduire le plus possible les "erreurs fines" d'estimation de la fréquence fondamentale. L'idée de l'algorithme est de comparer selon une mesure de ressemblance deux fenêtres de signal décalées de la valeur de la période test. Cela ressemble fort aux algorithmes du type AMDF, la différence essentielle étant que la taille des fenêtres est ici variable, et plus exactement égale à la période de test. Ainsi l'algorithme vise à positionner au mieux deux fenêtres successives représentant deux périodes successives du signal. La fonction de périodicité de l'algorithme SRPD s'écrit alors :

$$FP_{SRPD}(\tau) = \frac{\sum_{i=1}^{\tau} s_i s_{i+\tau}}{\sum_{i=1}^{\tau} s_i^2 \sum_{i=1}^{\tau} s_{i+\tau}^2}$$

Soquet [Soquet 94] a trouvé que le SRPD, malgré sa grande simplicité donne des taux d'erreur très acceptables, seules les erreurs de sous-harmoniques présentaient un score relativement élevé.

## 3.3 Peigne spectral

La méthode d'intercorrélation avec une fonction peigne proposée par Martin [Martin 81] utilise une représentation fréquentielle à court terme. Les algorithmes utilisant cette représentation exploitent généralement la structure harmonique des signaux périodiques.

L'algorithme est fondé sur l'intercorrélation entre le spectre du signal et une série d'harmoniques d'impulsions de Dirac d'amplitude normalisée ("peigne"). Cela revient à cumuler toutes les valeurs des amplitudes du spectre à des positions multiples de la fréquence de test $f$. Le spectre étant noté $\tilde{s}$ et ses coefficients $(\alpha_i)_{i=1,n}$ , la fonction de périodicité s'écrit alors :

$$FP_{peigne}(f) = \sum_{i=1}^{n(f)} \alpha_i \, |\tilde{s}(i * f)|$$

où $n(f)$ désigne le nombre d'harmoniques : il s'agit d'une fonction dépendante de la fréquence fondamentale de test $f$ et qui est à valeurs inférieures à $FMax/f$ où $FMax$ représente la plus haute fréquence prise en compte dans le spectre.

Pour résoudre le problème des sous-harmoniques, Martin applique une fonction de décroissance exponentielle sur les amplitudes des pics du peigne (par exemple $\alpha_i = e^{-\beta i}$). De cette façon les sous-harmoniques ont une intercorrélation inférieure à celle du F0.

Le peigne est une méthode particulièrement robuste qui parvient à retrouver la fréquence fondamentale même quand elle n'est plus contenue dans le signal (cas du signal téléphonique où la bande passante est restreinte à 300-4000Hz). En effet, l'algorithme fonctionnera même s'il manque des harmoniques dans le signal.

## 3.4 Programmation dynamique

Secrest et Doddington [Secrest et Doddington 82] ont appliqué lors de la phase de post-traitement un algorithme de programmation dynamique. La programmation dynamique est une technique puissante pour chercher un chemin optimal dans un réseau de transitions ou pour mettre en correspondance deux séquences. Elle a été initialement utilisée pour la reconnaissance de mots isolés [Sakoe et Chiba 71].

L'utilisation de la programmation dynamique permet de prendre en compte non seulement la première estimation de la fréquence fondamentale mais aussi la deuxième et la troisième. Il est donc nécessaire que l'extracteur délivre plusieurs estimations du F0 ainsi qu'une mesure de robustesse : les analyseurs à court terme peuvent généralement le faire ; de nombreux extracteurs dans le domaine temporel, comme ceux décrits par Gold et Rabiner [Gold et Rabiner 69], peuvent être facilement modifiés dans ce sens. La routine de programmation dynamique vérifie alors les chemins possibles des estimations du F0. Des pénalités sont assignées à toutes les solutions sous-optimales. A la fin le contour du F0 sélectionné correspond au chemin optimal.

## 3.5 Méthodes combinatoires

Il existe un très grand nombre de méthodes pour extraire la fréquence fondamentale, chacune présentant des avantages et des inconvénients, mais aucune ne permet d'évaluer la fréquence fondamentale avec une précision absolue. Ces observations ont conduit Hess [Hess 83] à suggérer de combiner différentes approches pour augmenter les performances globales du système d'extraction.

L'idée est d'appliquer différents analyseurs simultanément sur le signal et de combiner les différentes estimations ainsi obtenues. On peut citer par exemple

le système de combinaison de MES SIGNAIX [Espesser 96] qui utilise une méthode basée sur la stabilité des résultats, faisant intervenir plusieurs passes de traitement.

Soquet [Soquet 94] utilise une approche basée sur la corrélation des différents candidats avec une fenêtre triangulaire. Cette fenêtre étroite se déplace dans le domaine de variation de la fréquence fondamentale (comprise entre une valeur minimum autorisée du F0 et une valeur maximum). La valeur choisie correspond à la position de la fenêtre pour laquelle la corrélation est maximale. Cette méthode permet de diminuer de manière sensible le taux d'erreurs.

## 3.6 Conclusion

La tâche d'extraction de la fréquence fondamentale sur des signaux de parole a provoqué le développement de nombreux algorithmes. Le signal de parole étant très variable, une approche par combinaison de méthodes présente l'intérêt de pouvoir s'adapter à un grand nombre de signaux différents.

Le peigne, l'autocorrélation, l'AMDF et une combinaison de ces trois méthodes ont été implémentés dans MES SIGNAIX (cf. Annexe A). La méthode combinatoire a été utilisée pour extraire le fréquence fondamentale sur les signaux téléphoniques d'OGI MLTS (cf. section 5.2) et s'est révélée très performante compte tenu de la qualité du signal.

# Chapitre 4

# Identification Automatique des langues et prosodie

Ce chapitre aborde l'Identification Automatique des Langues, les problèmes soulevés ainsi que les différentes sources d'information utilisables dans les systèmes. Dans un deuxième temps, nous décrivons les principaux systèmes d'identification automatique qui intègrent des paramètres prosodiques.

## 4.1 Description d'un système d'IAL

Cette section présente l'Identication Automatique des Langues (IAL) ainsi que les différentes sources d'informations qui peuvent être utilisées ainsi qu'un explemple de système d'IAL.

### 4.1.1 Problématique

L'Identification Automatique des Langues consiste à identifier le langage parlé dans un échantillon de parole prononcé par un locuteur inconnu. Dans ce domaine, comme dans celui de la reconaissance de la parole, l'homme demeure le système le plus efficace [Ramus 96] : quelques secondes suffisent pour qu'il détermine si c'est un langage qu'il connaît. Dans des environnements bruités, comme un hall de gare par exemple, il est même possible de porter un jugement sur la langue parlée par des individus sans que le dialogue soit parfaitement intelligible.

La recherche en IAL revêt actuellement deux aspects, à savoir une étude cognitive recherchant les traits perceptuellement discriminants d'une langue, et une étude visant à répertorier des indices présents dans le signal acoustique qui soient robustes au sens de l'extraction et modélisable d'un point de vue statistique pour aboutir à un système d'IAL.

### 4.1.2 Les sources d'informations

Un bon algorithme d'IAL devrait exploiter des informations de toutes les sources suivantes pour arriver à la décision d'identification [Muthusamy et al. 94] :

- **le vocabulaire** : chaque langue possède son propre lexique, même si l'on retrouve certains mots en commun entre les langues, les mots voisins seront sûrement différents.

- **les phonèmes** : chaque langue possède son propre système phonétique. Certains phonèmes sont spécifiques à une langue et même quand les langues possèdent le même jeu de phonèmes, on peut s'intéresser aux fréquences d'occurence des phonèmes.

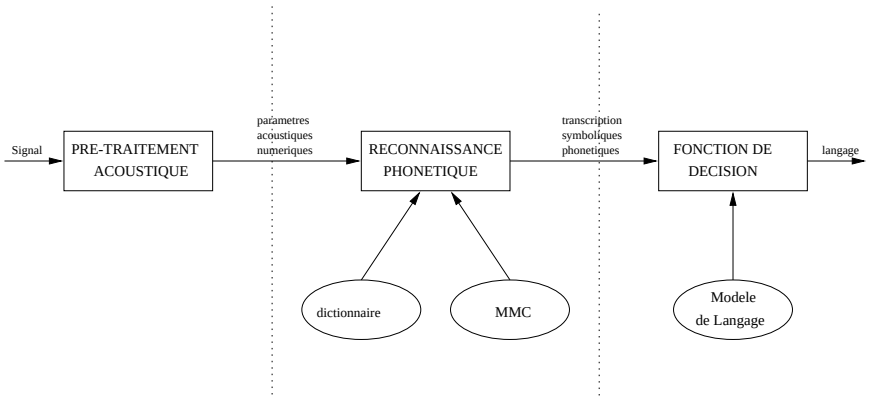
- **la phonotactique** : les règles d'enchaînement des différents sons sont propres à chaque langage. Par exemple la combinaison /sr/ est courante en Tamoul, alors qu'elle n'est pas admise en anglais.

- **la prosodie** : les langues varient en fonction de la durée, du rythme et de l'intonation (cf. section 2.1). Certains langages appartiennent à des catégories prosodiques différentes (langue tonale, accentuée au niveau de la syllabe...) qui ont des propriétés bien distinctes mais il est difficile d'automatiser le traitement des règles d'ordre linguistique en identification des langues.

Les systèmes d'IAL actuels sont principalement basés sur une modélisation statistique du niveau acoustique et phonotactique : un tel système est présenté à la section suivante. La prise en compte d'informations prosodiques reste difficile et peu exploitée car leur connaissance est à ce jour plus qualitative que quantitative, et le problème de fusion de données (acoustiques, phonotactiques, prosodiques) deumeure non résolu.

### 4.1.3 Principe général d'un système

Un système classique d'IAL, utilisant des indices acoustiques et phonotactiques [Yan et Barnard 95], peut se décomposer en trois parties distinctes :




La première partie consiste à extraire du signal des paramètres caractéris-

tiques.

Une transformation numérique/symbolique est ensuite effectuée au cours de l'étape de reconnaissance phonétique. Des techniques de classification de type Quantification Vectorielle [Pellegrino et Obrecht 97b] où une modélisation de type markovienne peuvent être utilisées à cette étape.

La suite des symboles phonétiques est ensuite décodée à l'aide d'un modèle de langage et fournit également un score de vraisemblance. Les séquences de symboles sont alors modélisées. 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...).

Nous allons maintenant présenter des systèmes existants qui utilisent des informations prosodiques.

## 4.2 Systèmes d'IAL utilisant la prosodie

Différentes approchent existent pour l'intégration de paramètres prosodiques dans les systèmes d'IAL : Hazen et Zue appliquent une modélisation par Modèles de Mélanges de Gaussiennes qui ne se révèle pas aussi efficace que pour les paramètres acoustiques ; Muthusamy utilise la fréquence fondamentale et le rythme dans son approche par réseaux de neurones ; Li applique des principes de reconnaissance du locuteur sur des paramètres syllabiques et Itahashi utilise le contour de la fréquence fondamentale comme paramètre principal dans un système de classification.

### 4.2.1 Système de Hazen et Zue

Hazen et Zue [Hazen et Zue 97] ont intégré un modèle prosodique à un système d'IAL dans un cadre probabiliste. Le système est constitué d'un modèle acoustico-phonétique, d'un modèle de langage phonétique et d'un modèle prosodique (cf. figure 4.1 pour un aperçu du système global). Etant donné le vecteur acoustique $\overrightarrow{a}$, la vecteur prosodique $\overrightarrow{f}$ , la séquence phonétique la plus probable $C$ et la segmentation associée $S$, le langage le plus probable $L^\star$ est trouvé en utilisant l'expression suivante :

$$L^\star = \arg\max_i P(L_i/C, S, \overrightarrow{a}, \overrightarrow{f})$$

ce qui revient à écrire :

$$L^\star = \arg\max_i P(\overrightarrow{a}/C, S, \overrightarrow{f}, L_i) P(S, \overrightarrow{f}/C, L_i) P(C/L_i) P(L_i)$$

Les quatre probabilités de cette expression sont organisées de telle sorte que les informations phonétiques et prosodiques soient séparées. Ces termes désignent les différents modèles du système :

1. $P(\overrightarrow{a}/C, S, \overrightarrow{f}, L_i)$ : le modèle acoustico-phonétique
2. $P(S, \overrightarrow{f}/C, L_i)$ : le modèle prosodique
3. $P(C/L_i)$ : le modèle phonétique de langage

et $P(L_i)$ désigne la probabilité a priori du langage. Cette dernière est considérée comme uniforme et a donc été ignorée par la suite.

Le **modèle acoustique** est utilisé pour capturer des informations sur les phonèmes utilisés par chaque langage. Pour simplifier le modèle, Hazen et Zue font l'hypothèse que le vecteur acoustique $\overrightarrow{a}$ est indépendant du vecteur prosodique $\overrightarrow{f}$ :

$$P(\overrightarrow{a}/\overrightarrow{f}, S, C, L_i) = P(\overrightarrow{a}/S, C, L_i) = \prod_{k=1}^{m} P(\overrightarrow{a_k}/c_k, L_i)$$

où $m$ est le nombre de trames dans le signal et $\overrightarrow{a_k}$ est le vecteur de paramètres qui décrit l'acoustique de la $k^{ieme}$ trame. Le vecteur $\overrightarrow{a_k}$ est constitué de 14 MFCC [1] et des 14 dérivées associées.

Le **modèle phonétique de langage** utilisé est un modèle trigramme interpolé, dont l'expression est :

$$\widehat{P}(c_i/c_{i-1}, c_{i-2}) = \lambda_2 P(c_i/c_{i-1}, c_{i-2}) + (1-\lambda_2)(\lambda_1 P(c_i/c_{i-1}) - (1-\lambda_1)P(c_i))$$

où $\lambda_1$ et $\lambda_2$ sont les poids qui dépendent des phonèmes précédant $c_i$ . $\lambda_1$ s'écrit :

$$\lambda_1 = \frac{k_{c_{i-1}}}{k_{c_{i-1}} + K_1}$$

où $k_{c_{i-1}}$est le nombre d'occurences de $c_{i-1}$ dans le corpus d'apprentissage et $K_1$ est une constante. D'une manière similaire nous avons :

$$\lambda_2 = \frac{k_{c_{i-1}, c_{i-2}}}{k_{c_{i-1}, c_{i-2}} + K_2}$$

Le **modèle prosodique** se base sur des informations statistiques simples sur la fréquence fondamentale et la durée des segments. Pour simplifier le modèle, l'expression pour le modèle prosodique peut de développer comme suit :

1. Mel-Frequency Cepstral Coefficients : coefficients caractérisant la forme du conduit vocal dans une échelle fréquentielle perceptive.

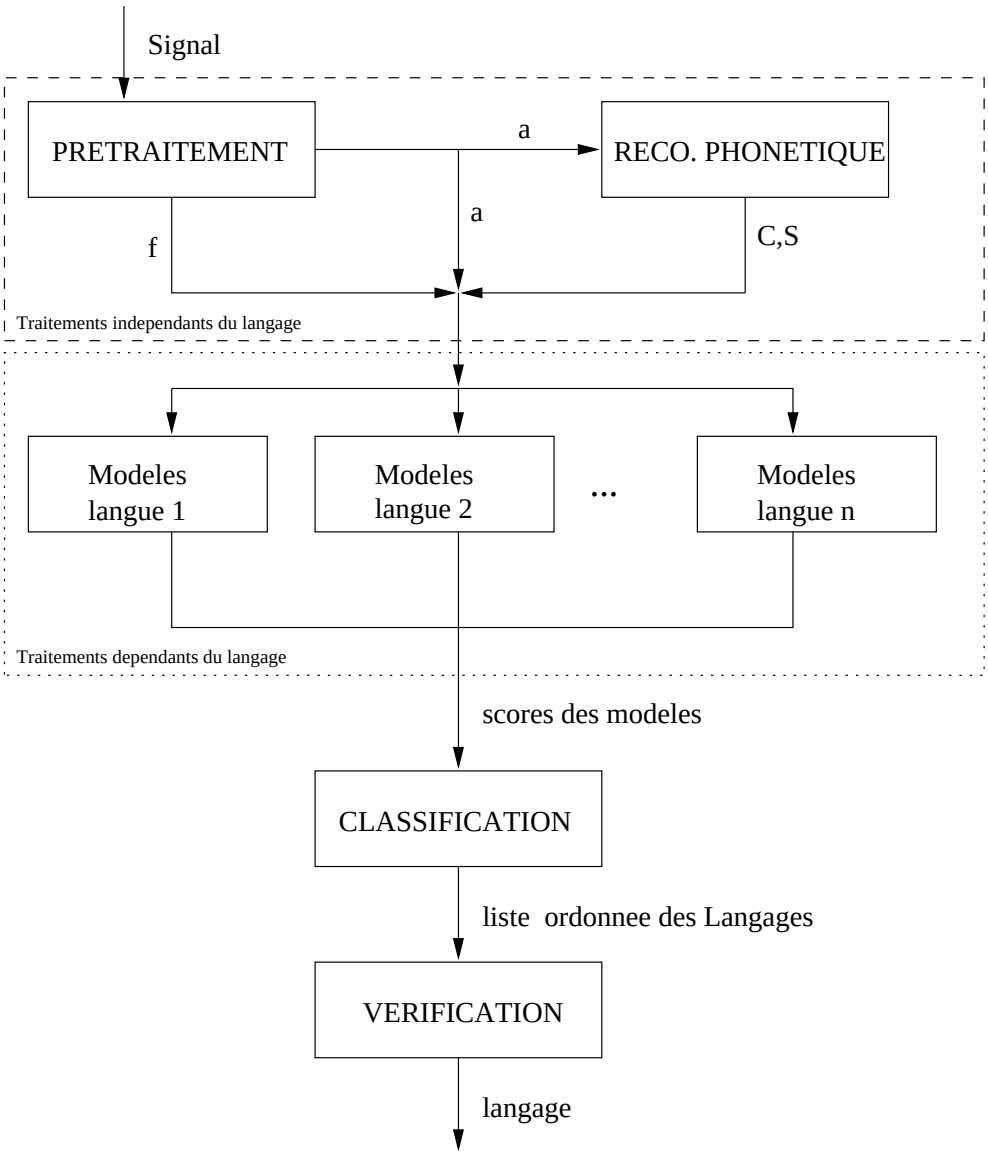



FIG. 4.1 - *vue globale du système de Hazen et Zue*

$$P(S, \overrightarrow{f}/C, L_i) = P(\overrightarrow{f}/S, C, L_i)P(S/C, L_i)$$

Le **modèle de fréquence fondamentale** $P(\overrightarrow{f}/S, C, L_i)$ utilise des informations du contour du F0 d'un signal. Alors qu'il doit y avoir une corrélation entre le contour du F0 et le durée segmentale, cette corrélation est ignorée pour simplifier le modèle. Donc $\overrightarrow{f}$ est considéré comme indépendante de $S$ et $C$ . Avec ces hypothèses, le modèle de fréquence fondamentale donne :

$$P(\overrightarrow{f}/S, C, L_i) = P(\overrightarrow{f}/L_i)$$

Hazen et Zue considèrent maintenant que l'information qui est contenue dans la dynamique du contour du F0 est contenue dans les valeurs de la dérivée du F0 et qu'il n'est donc pas indispensable de modéliser spécifiquement cette dynamique. Pour simplifier le modèle, chaque trame est considérée comme statistiquement indépendante :

$$P(\overrightarrow{f}/L_i) = \prod_{k=1}^{m} P(\overrightarrow{f_k}/L_i)$$

où $m$ est le nombre de trames dans le signal et $\overrightarrow{f_k}$ est un vecteur de paramètres représentant les valeurs du F0 et de la derivée du F0 pour la $m^{ieme}$ trame. Il faut noter que le calcul de cette probabilité n'est effectué que sur les zones voisées, *i.e.* que sur les zones où la fréquence fondamentale est définie. L'expression ci-dessus a été modélisée en utilisant un mélange de Gaussiennes, avec 9 fonctions de densité estimées avec des matrices de covariance pleines.

Le **modèle de la durée segmentale** ne tient pas compte des informations contenues dans l'accentuation des motifs des syllabes, des mots et des phrases du signal. En faisant l'hypothèse d'indépendance des segments, le modèle de durée s'écrit :

$$P(S/C, L_i) = \prod_{k=1}^{n} P(d_k/c_k, L_i)$$

où $n$ est le nombre de segments dans le signal et $d_k$ la durée du $k^{ieme}$ segment. Cette expression est modélisée par mélange de Gaussiennes.

Les résultats ont été calculés à la sortie du système global mais chaque modèle à été évalué séparément, ce qui permet de mesurer l'importance relative de chaque composante. Sur le corpus OGI-MLTS, les résultats pour 11 langues avec des fichiers de parole spontanée de 10 et 45 s. sont les suivants :

| nom | description |
|---|---|
| VOC | voyelle |
| PRVS | pré-vocalique |
| POVS | post-vocalique |
| INVS | intervocalique |
| FRIC | fricative |
| STOP | explosion |
| CLOS | pas de signal de parole |

TAB. 4.1 - *Grandes catégories utilisées par le système de Muthusamy*

| **Modèles** | **10 s.** | **45 s.** |
|---|---|---|
| *Système complet* | 65% | 78% |
| Modèle de langage | 62% | 77% |
| Modèle acoustique | 49% | 53% |
| Modèle de durée | 31% | 44% |
| Modèle de F0 | 12% | 20% |

Ces résultats sont compétitifs avec les autres systèmes d'IAL. Le modèle de langage est la modélisation la plus utile pour le système probabiliste, les informations des modèles acoustico-phonétique et prosodique pouvant tout de même se révéler utiles pour améliorer les performances.

### 4.2.2 Système de Muthusamy

Muthusamy [Muthusamy 93] a développé un système basé sur une classification en larges catégories phonétiques du signal de parole. Ce système a été initialement testé sur un corpus de quatre langues (anglais, japonais, mandarin et tamoul) contitué d'enregistrement de parole de haute qualité (OGI-HQ). Muthusamy a ensuite amélioré ce système, en évaluant différentes combinaisons de paramètres en utilisant le corpus téléphonique OGI-MLTS avec 10 langues.

Le premier système se décompose de deux phases :

1. une segmentation en grandes catégories (cf. tableau 4.1) du signal de parole,
2. une classification de langage utilisant des paramètres extraits des grandes catégories.

La segmentation est basée sur un réseau de neurones entièrement connecté à trois couches qui délivre des scores sur 7 grandes catégories de phonèmes. Une recherche par algorithme de Viterbi qui prend en compte les probabilités sur la durée et les modèles bigrammes est alors utilisée pour trouver les meilleures séquences de labels.

La classification utilise un autre réseau de neurones à trois couches. 80 paramètres prosodiques et phonétiques dérivés de la segmentation en larges catégories sont utilisés en entrée du réseau. Parmi les paramètres prosodiques, notons l'utilisation de la moyenne de la variance de la fréquence fondamentale, prise à l'intérieur de chaque catégorie voisée et sur toutes les catégories voisées, la moyenne et la variance de la durée de chaque catégorie, la moyenne et la variance de la différence de durée entre des segments successifs.

Les résultats en identification des langues sont de 79% avec un corpus de test composé de 6 locuteurs par langue.

La deuxième série d'expérimentations, effectuées sur le corpus téléphonique avec dix langues, utilise le même principe de segmentation/classification mais avec différents jeux de paramètres. Les meilleurs résultats obtenus avec les paramètres exploitant la fréquence fondamentale (variance inter et intra segmentale) ne dépassent pas 19%. Les paramètres prosodiques utilisés (fréquence fondamentale, rythme de parole) se sont révélés moins performants que les paramètres mettant en jeu des règles phonotactiques ou la durée segmentale.

### 4.2.3 Système de Li

Li [Li 94] a adopté des méthodes issues de 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 sylllabiques sont extraits de la même façon et les coefficients spectraux sont comparés à tous ceux sauvés 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 $n$ 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.

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 82% et 91% pour des fichiers de 45 s. et 10 s. respectivement.

### 4.2.4 Système de Itahashi

Itahashi [Itahashi et Du 95] utilise un système d'IAL basé sur la fréquence fondamentale. Son système est constitué de trois parties :

1. Extraction de F0 et segmentation : la fréquence fondamentale est extraite à l'aide de l'algorithme AMDF (cf. section 3.2.2) et les zones voisées sont détectées à partir de l'énergie du signal.

2. Approximation polynômiale du contour F0 : chaque intervalle voisé est approximé par un nombre minimal de lignes (dont la valeur est obtenue par un algorithme de programmation dynamique), ce qui permet de lisser le contour et réduire le taux d'erreur.

3. Analyse discriminante : 21 paramètres statistiques sont extraits de la courbe mélodique et de son approximation polynômiale. La variance, le skewness et le kurtosis sont calculés à la fois à partir de F0, de l'energie E, des lignes polynômiales ayant une pente positive et de celles ayant une pente négative. A ces 12 paramètres sont ajoutés un coefficient de corrélation entre F0 et E, le nombre, la moyenne, la durée et les fréquences relatives de départ et la moyenne de la pente des lignes positives et négatives de l'approximation polynômiale. Une Analyse en Composantes Principales est ensuite appliquée puis la classification est réalisée avec une analyse discriminante

Les résultats sont de 78% environ pour cinq langues (japonais, chinois, coréen, anglais, français et allemand) sur le corpus téléphonique OGI-MLTS, avec les fichiers de dialogue non contraint de 20 s.

# Chapitre 5

# Réalisation d'une modélisation automatique de la prosodie pour l'IAL

La recherche en IAL menée au sein de l'équipe IHM-PT de l'IRIT a déjà mis en évidence la pertinence de la recherche dans le domaine acoustique et phonotactique. Le cadre de l'étude au niveau du DEA consiste à apporter au système d'IAL des connaissances prosodiques. Les systèmes existants évoqués dans le chapitre précédent ont confirmé la pertinence de la prosodie en IAL (Itahashi) mais aussi le manque d'efficacité d'une approche classique (système de Hazen et Zue, Muthusamy) adaptée à la prosodie. Il est nécessaire de proposer une nouvelle approche qui tienne compte du caractère suprasegmental de la prosodie. Après un aperçu de la conception générale du module prosodique, nous présentons le corpus utilisé ainsi que les paramètres qui en ont été extraits. Nous décrivons ensuite les deux composantes principales du système, le modèle prosodique et le modèle de language multigramme. Nous terminons avec la présentation et les commentaires des résultats obtenus.

## 5.1 Système général

Un système classique d'IAL basé sur l'acoustique et la phonotactique est composé de trois parties distinctes (cf. section 4.1.3) :

1. Un prétraitement acoustique,
2. un décodage phonétique,
3. une modélisation de langage.

De façon similaire, le module prosodique que nous proposons se décompose en trois parties :

1. une extraction de paramètres prosodiques,
2. une reconnaissance des unités prosodiques
3. une modélisation du langage régissant ces unités.

L'extraction des paramètres prosodiques est abordée en 5.3. La phase de reconnaissance des unités prosodiques, décrite en 5.4, a pour but d'extraire des informations sur la variation des paramètres prosodiques dans le temps. La dernière partie, décrite en 5.5, présente la modélisation d'unités prosodiques dans chaque langue.

## 5.2 Corpus utilisés

Le corpus utilisé est MLTS (Multi-language Telephone Speech) de l'Oregon Graduate Institute (OGI). Ce corpus a été enregistré pour permettre l'évaluation des systèmes d'Identification Automatique des Langues. Il est composé de 11 langues : l'anglais, le farsi, le français, l'allemand, l'hindou, le japonais, le coréen, le mandarin, l'espagnol, le tamoul et le vietnamien. Les enregistrements ont été effectués sur des lignes téléphoniques réelles et présentent donc une grande variabilité au niveau du microphone, un faible rapport signal/bruit, une faible bande passante et une distortion variable due au canal.

Le corpus est divisé en trois parties homogènes : une partie pour l'apprentissage (APP), une autre pour le développement du système (DEV), et une dernière pour les tests (TST).

Chaque langue contient environ 90 locuteurs et il y a environ 2 minutes d'enregistrement par locuteur. Les locuteurs ont eu à répondre à des questions sur des thèmes précis. Dans le cadre de la réalisation du modèle prosodique, seuls les enregistrements présentant le moins de contraintes ont été sélectionnés. En effet la courbe mélodique varie très peu d'un mot à l'autre lors d'énumérations : fichiers "nums" : nombres de 0 à 9, "dowk" : jours de la semaine (cf. tableau 5.1 pour le détail des fichiers utilisés).

Cette étude s'inscrit dans un contexte de faisabilité, ainsi seuls le français et le japonais ont été utilisés. Ces deux langues possèdent des caractéristiques prosodiques différentes : le français possède un accent lexical fixe (sur la dernière syllabe du mot) alors que le japonais possède un accent lexical libre (comme l'anglais et l'allemand). Il y a des différences tant au niveau des fonctions de la prosodie (cf. section 2.2) que du rythme de la parole.

| suffixe | description |
|---|---|
| htl | certains aspects de la ville qu'il apprécie |
| htc | climat dans sa propre ville |
| meal | son dernier repas |
| room | description de la pièce où il se trouve |
| story-at | texte libre (environ 10s) |
| story-bt | texte libre (environ 45s) |

TAB. 5.1 - *Descriptif des fichiers du corpus utilisé*

## 5.3 Paramètres utilisés

Les paramètres utilisés sont extraits du contour de la fréquence fondamentale et de l'énergie du signal. La courbe de la fréquence fondamentale a été obtenue en utilisant la méthode combinatoire développée dans MES SIGNAIX (cf. Annexe A) du Laboratoire Parole et Langage d'Aix-en-provence. Les méthodes combinées sont l'autocorrélation (voir section 3.2.1), l'AMDF (voir section 3.2.2) et la méthode du peigne spectral (voir section 3.3). Cette méthode fournit des valeurs toutes les 10ms. L'énergie (E) est calculée toutes les 10 ms avec une fenêtre de Hamming de 30 ms.

Ces deux valeurs sont ensuite interpolées pour obtenir environ 10 valeurs par trame de 10 ms de signal. Douze paramètres sont alors extraits à partir de ces 10 valeurs :

**Delta(F0) et Delta(E)** : dénivelé de la fréquence fondamentale et de l'énergie

$$Delta(X) = \sum_{i=1}^{N} X(i+1) - X(i)$$

où $N$ désigne le nombre total d'echantillons et $X = F0$ ou $E$ .

**Dérivée1(F0) et Dérivée1(E)** : dérivée première de la fréquence fondamentale et de l'énergie (cf. annexe B)

**Dérivée2(F0) et Dérivée2(E)** : dérivée seconde de la fréquence fondamentale et de l'énergie (cf. annexe B)

**Ecart type(F0) et Ecart type(E)** : valeurs statistiques classiques sur les valeurs de la fréquence fondamentale.

$$\sigma(X) = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

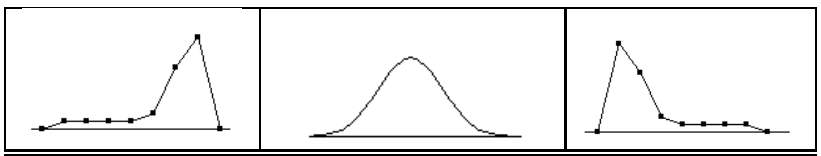
où $N$ désigne le nombre total d'échentillons et $X = F0$ ou $E$.

**Skewness(F0) et Skewness(E)** : valeur statistique basée sur la symétrie des empâtements de la distribution de la fréquence fondamentale et de l'énergie. Il s'agit du troisième moment central divisé par le cube de l'écart type :

$$skewness(X) = \frac{\sum_{i=1}^{N}(X_i - \mu)^3}{N\sigma^3}$$

où $N$ désigne le nombre total d'échantillons et $X = F0$ ou $E$.

Exemple de l'influence de forme de distributions sur la valeur du skewness :

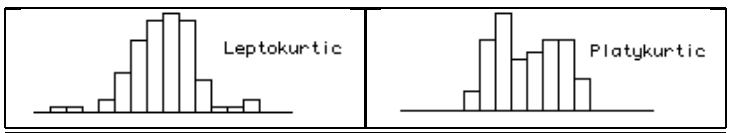


| skewness négatif | skewness nul | skewness positif |
|---|---|---|

**Kurtosis(F0) et Kurtosis(E):** valeur statistique basée sur la taille de l'empattement de la distribution de la fréquence fondamentale et de l'énergie. Il s'agit du quatrième moment central divisé par l'écart-type puissance quatre :

$$kurtosis(X) = \frac{\sum(X - \mu)^4}{\sigma^4}$$

où $N$ désigne le nombre total d'échantillons et $X = F0$ ou $E$.
Exemple de l'influence de la forme de distribution sur la valeur du kurtosis (la distribution à gauche appartient à la famille des "leptokurtic" et celle de droite à celle des "platykurtic") :




| kurtosis positif | kurtosis négatif |
|---|---|

Les paramètres delta et dérivée première permettent de rendre compte de la pente de F0 et de E. La dérivée seconde, calculée à l'aide d'un polynôme du second ordre, permet de rendre compte de variations plus complexes de la courbe. Les paramètres statistiques apportent des informations supplémentaires sur la forme des distributions de probabilités.

## 5.4 Modèles prosodiques

Le but principal est de réaliser une modélisation des différentes unités (formes) prosodiques contenues dans le contour intonatif des zones voisées.

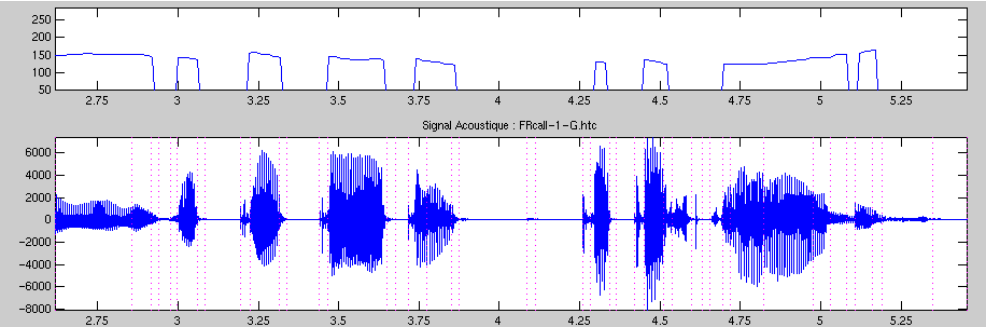
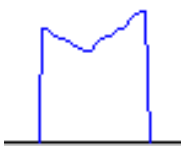



FIG. 5.1 - *Contour mélodique d'un signal parlé*

La première question qui est soulevée par une telle modélisation concerne le choix de l'unité temporelle. Une modélisation basée sur des trames n'est pas concevable, car elle ne reflète pas le caractère suprasegmental de la prosodie. En observant le contour intonatif d'un signal de parole (cf. figure 5.1 : la partie du haut représente l'évolution dans le temps de la valeur de F0, celle du bas le signal acoustique associé), il apparait qu'il existe déjà une segmentation au niveau du voisement : une zone voisée est caractérisée par un contour de la fréquence fondamentale non nulle, dont les valeurs du F0 sont comprises entre 50 et 200 Hz pour les hommes, et 100-400Hz pour les femmes ; une zone non voisée par un contour intonatif nul. La segmentation en unités prosodiques doit-elle être effectuée au niveau de la segmentation voisé/non voisée, ou utiliser un autre niveau, par exemple en choisissant des unités plus petites ? Un motif tel que :



doit-il être caractérisé par une seule unité prosodique ("creux") ou plusieurs (une "descente" suivie d'une "montée") ? Il a été décidé d'essayer une modélisation dont les unités correspondent exactement à une zone voisée du signal dans le cadre d'une première approche.

10 formes élémentaires sont définies :

**ST** (STationary) : forme stationnaire

**BU** (BUmp) : bosse

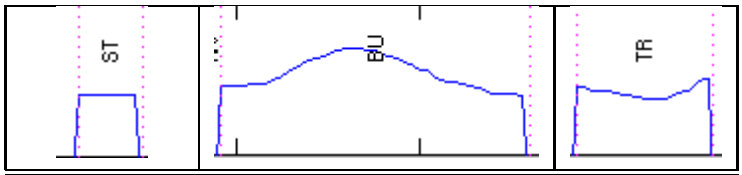
**TR** (TRough) : creux

**UPlo, UPmi, UPhi** (UP low, middle, high) : montées de pentes différentes

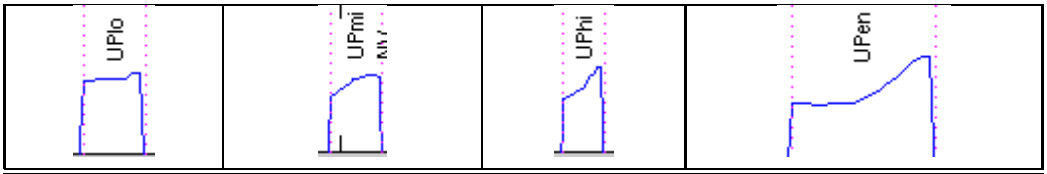
**DWlo, DWmi, DWhi** (DWown low, middle, high) : descentes de pentes différentes

**UPen** (UP end) : montée de fin de phrase

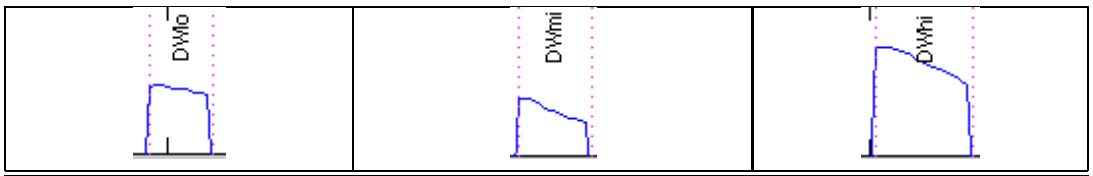
*Exemples de contours mélodiques des modèles :*




| stationnaire | bosse | creux |
|---|---|---|



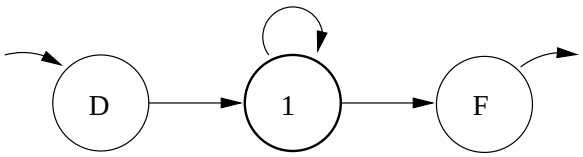

| montée légère | montée moyenne | montée forte | montée de fin de phrase |
|---|---|---|---|



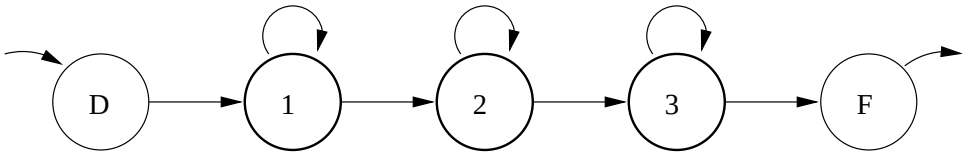

| descente légère (DWlo) | descente moyenne (DWmi) | descente forte (DWhi) |
|---|---|---|

Pour modéliser ces contours, une modélisation par Modèles Stochastiques de Trajectoires [Gong et Haton 94, Illina 97] a été envisagée. Nous n'avons pas pu avoir accès à l'implémentation de ces modèles, et nous avons alors employé une modélisation par Modèles de Markov Cachés (MMC), qui permet de modéliser l'évolution temporelle des formes et qui se révèle intéressante du point de vue de la variabilité des formes.

Parmi les 10 formes considérées, 7 sont modélisées avec un seul état, et les 3 autres à l'aide de trois états (sans compter les états fictifs de début et de fin). Une seule gaussienne est utilisé pour décrire chaque état. Un seul état est utilisé pour modéliser une forme dont l'évolution dans le temps est monotone :




Les modèles à trois états sont assignés aux formes dont l'évolution dans le temps est plus complexe et qui nécessitent donc plus d'un état pour être représenté : il s'agit des formes UPen, TR, BU.




Il faut noter qu'avec une telle modélisation on perd les informations de durée des segments : en effet, un motif très étalé dans le temps et un autre motif du même type de durée plus courte sont modélisés par le même MMC. Ceci est dû au fait que les états bouclent sur eux-même, c'est-à-dire qu'à la sortie d'un état le réseau offre la possibilité de revenir dans cet état ou bien de changer d'état.

Il faut aussi signaler qu'un nombre minimum de vecteurs de paramètres est demandé pour entrer dans un état. Ainsi avec HTK[1], la boîte à outils utilisée pour réaliser cette modélisation, il est nécessaire de disposer d'au moins trois vecteurs prosodiques pour caractériser un état. Ces vecteurs étant calculés toutes les 10 ms. (cf. section 5.3) les formes ne seront étiquetées que si elles sont d'une durée supérieure à 30 ms. Ces formes sont donc étiquetées "**V**" pour "zone Voisée".

L'apprentissage est réalisé de manière empirique par succession d'essais et d'améliorations. Chaque modèle est initialisé par une douzaine de formes prosodiques étiquetées manuellement. Le corpus d'apprentissage est ensuite étiqueté par ce modèle. Lorsqu'une erreur d'étiquetage est décelée, la forme en question est rajoutée dans la liste des formes utilisée pour l'initialisation. L'étiquetage est alors relancé avec le nouveau modèle appris avec les nouvelles listes. Le processus est répété jusqu'à ce que l'étiquetage convienne. Les expérimentations n'ont jamais dépassé 10 cycles de modifications.

La phase de reconnaissance consiste d'une part à découper le signal en segments voisés distincts (à l'aide de la valeur de F0) et d'autre part à étiqueter chaque segment par le modèle prosodique (constitué des 10 MMC prosodiques). La figure 5.2 montre un exemple de segmentation et d'étiquetage.

## 5.5 Modèle de langage

Une fois le modèle prosodique du français et du japonais réalisé, l'objectif est de modéliser les différents enchaînements des formes prosodiques pouvant exister dans chaque langue. Pour cela un modèle de langage multigramme est utilisé à partir de la suite des labels prosodiques obtenus sur les zones voisées.

### 5.5.1 Description du modèle multigramme [Deligne 96]

Un modèle de langage multigramme est un outil statistique qui permet de détecter des motifs récurrents de suites d'observations.

1. Hidden Markov Model Tool Kit édité par Entropic

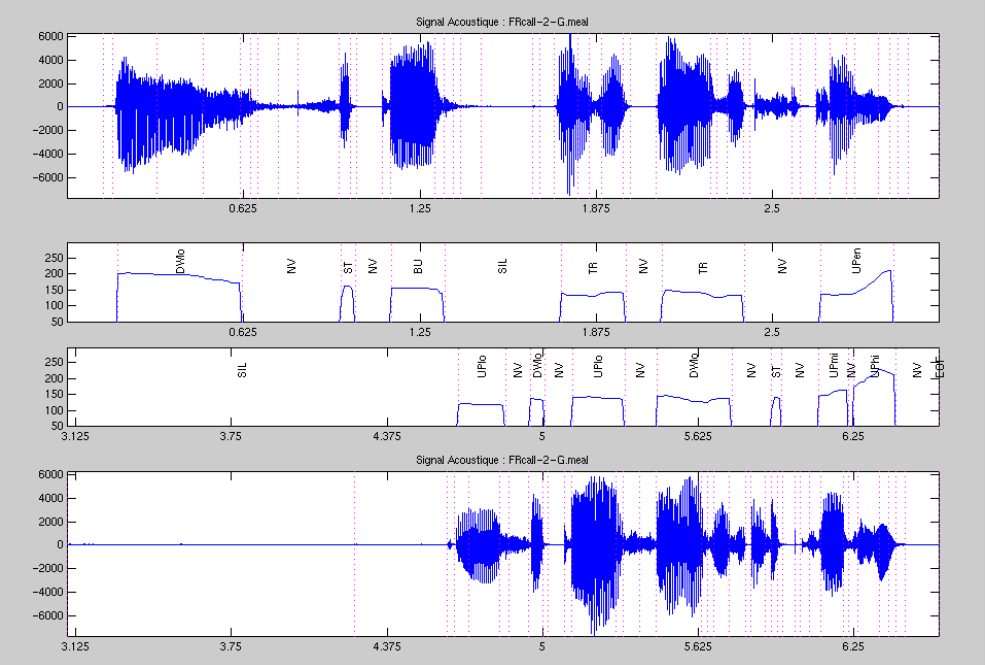



FIG. 5.2 - *Exemple de segmentation et d'étiquetage*

Une source gouvernée par un processus de type multigramme émet des observations (à valeurs continues ou discrètes) par séquences de longueur variable. Seule la concaténation des séquences est accessible à l'observation.

Dans le cas discret (cas de l'étude), le processus observable prend ses valeurs dans un dictionnaire. En phase de reconnaissance il s'agit de trouver à partir d'une suite d'observations $O = (o_1, ..., o_T)$, la segmentation $S = (s_1, s_2, ..., s_{c(s)})$ la plus problable au sens du maximum de vraisemblance :

$$S^\star = \arg\max_S l(O, S)$$

$$l(O, S) = \prod_{i=1}^{c(S)} P(O_{t_{s_i}}, ..., O_{t_{s_{i+1}}-1})$$

$$s_i = O_{t_{s_i}}, ..., O_{t_{s_{i+1}}-1}$$

avec $l(.,.)$ : vraisemblance et $s_i$ : suite d'observations.

L'algorithme d'apprentissage des multigrammes est basé sur un algorithme de type EM itératif et a pour but d'estimer les probabilités de toute séquence d'observations $s_i$. À chaque itération (k+1) à partir du modèle k, la probabilité de la séquence $s_i$ est réestimée après segmentation optimale du corpus d'apprentissage.

$$S^{\star(k)} = \arg\max_{S} l^{(k)}(O, S)$$

$$p^{(k+1)}(s_i) = \frac{c(s_i/S^{\star(k)})}{c(S^{\star(k)})}$$

où $c(s_i/S^{\star(k)})$ est le nombre d'apparitions de $s_i$ dans $S^{\star(k)}$ et $c(S^{\star(k)})$ est le nombre total de séquences dans $S^{\star(k)}$, la segmentation $S^{\star(k)}$ la plus probable étant obtenue par l'algorithme de Viterbi.

### 5.5.2 Application au modèle prosodique

Dans notre cas, la source d'observation est constituée d'une suite de labels de modèles prosodiques. Les fichiers ayant servi à l'apprentissage sont découpés au niveau de pauses longues (pauses superieures à 150 ms), ce qui permet de collecter un ensemble de "propositions". Les étiquettes attribuées par le modèle prosodique sont rassemblées en "propositions" et introduites dans l'algorithme d'apprentissage. Lors de la phase de reconnaissance, l'algorithme de Viterbi fournit pour chaque "proposition" un score basé sur un calcul de vraisemblance. Cette vraisemblance est calculée pour chaque modèle de langage.

## 5.6 Résultats

Nous avons évalué le système prosodique au niveau de la modélisation de la prosodie et au niveau de l'identification des langues.

### 5.6.1 Résultats de la modélisation de la prosodie

Il n'est pas évident de présenter des résultats quantifiables, étant donné que le jeu de formes ne peut pas représenter entièrement toutes les formes disponibles, néanmoins une petite base a été étiquetée manuellement pour évaluer cette modélisation. Les tableaux ci-dessous contiennent les matrices de covariance respectivement du français et du japonais. La première colonne contient les labels étiquetés manuellement et la première ligne les labels qui sont trouvés par la modélisation automatique.

– *Résultats de la modélisation pour le français :*

| | ST | BU | TR | DWlo | DWmi | DWhi | UPlo | UPmi | UPhi | UPen |
|---|---|---|---|---|---|---|---|---|---|---|
| ST | 5 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| BU | 2 | 4 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| TR | 0 | 0 | 12 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| DWlo | 1 | 0 | 0 | 24 | 0 | 0 | 0 | 0 | 0 | 0 |
| DWmi | 0 | 2 | 1 | 0 | 7 | 0 | 0 | 0 | 0 | 0 |
| DWhi | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 1 |
| UPlo | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 0 |
| UPmi | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 | 0 | 0 |
| UPhi | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | 0 |
| UPen | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 6 |

– *Résultats de la modélisation pour le japonais :*

| | ST | BU | TR | DWlo | DWmi | DWhi | UPlo | UPmi | UPhi | UPen |
|---|---|---|---|---|---|---|---|---|---|---|
| ST | 8 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 1 |
| BU | 0 | 10 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| TR | 1 | 1 | 6 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| DWlo | 0 | 0 | 1 | 22 | 0 | 0 | 0 | 0 | 0 | 0 |
| DWmi | 0 | 1 | 0 | 0 | 12 | 0 | 0 | 0 | 0 | 0 |
| DWhi | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 |
| UPlo | 2 | 0 | 1 | 0 | 0 | 0 | 7 | 0 | 0 | 0 |
| UPmi | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 0 | 0 |
| UPhi | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| UPen | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |

Les résultats ne sont pas parfaits, ce qui est attendu : le nombre de formes modélisées est réduit et ne peut pas rendre compte de toutes les formes existantes (en particulier les formes présentant une succession de plus de trois formes montantes et descendantes).

Il faut aussi noter que les erreurs d'extraction de la fréquence fondamentale (en particulier les erreurs d'harmoniques) entrainent une mauvaise modélisation des unités prosodiques. Dans un premier temps nous avons modélisé deux formes supplémentaires pour rendre compte de ces formes ("WAup" et "WAdw" pour waste up et down), mais elles ont été supprimées car elles n'apportaient aucune information sur la prosodie de la phrase, mais une information sur la difficulté d'extraire la fréquence fondamentale sur des signaux très bruités.

Dans l'ensemble la modélisation est de bonne qualité et caractérise bien la courbe prosodique. Il est d'ailleurs surprenant de constater que la modélisation se comporte très bien dans certains cas limites : lorsqu'une forme présente un contour relativement accidenté, le MMC choisit l'unité prosodique la plus proche de la dynamique générale du motif. Certaines erreurs provoqués par l'extraction de la fréquence fondamentale sont alors rectifiées.

### 5.6.2 Résultats en Identification des Langues

Le corpus de test est constitué d'une centaine de fichiers par langue. Les résultats sont calculés à partir de tous les fichiers mais également en se restreignant à ceux de 45 s. de long. Les résultats résultant d'un modèle de langage bigramme sont les suivants :

| type de fichiers | identifications correctes | identifications erronnées | taux |
|---|---|---|---|
| tous | 132 | 72 | 65% |
| 45 s. | 25 | 11 | 69% |

Ces résultats sont un peu décevants. Au cours des différentes manipulations, nous avons constaté que certaines formes étaient spécifiques d'une langue ; l'utilisation de modèles de langage propre à chaque langue, ou bien la définition d'unités prosodiques spécifiques à une langue laisse espérer une augmentation sensible des performances du module prosodique.

# Chapitre 6

# Conclusions et perspectives

Au cours de ce stage, un système complet d'identification automatique des langues (IAL) a été réalisé. Ce système constitue une approche nouvelle quant à l'introduction de la prosodie en IAL. La prosodie constitue un domaine extrêmement riche dont les mécanismes ne sont pas encore clairement définis, ce qui complique d'autant plus des tentatives de modélisation. Le système développé constitue une base de développement sur la prododie et l'IAL. Il a été principalement axé sur l'utilisation de la fréquence fondamentale. Une étude plus approfondie pourra être menée sur la recherche d'unités prosodiques, et en particulier la recherche d'unités spécifiques à chaque langue. Nous espérons également, pour compléter l'exploitation des paramètres prosodiques, pouvoir rapidement adjoindre une modélisation de la durée, du rythme de la parole.

# Annexe A : MESSIGNAIX

MES SIGNAIX(Motif Editor Signal Tool) à été développé pour afficher, étiqueter (MES) et traiter (SIGNAIX) le signal de parole [Espesser 96].

Cette boite à outils composée de commandes de bas niveau executables à la manière de filtres. Elle se veut simple, très modulaire et surtout très ouverte: il existe des outils graphiques distincs des traitements et il n'y a pas d'entête complexe pour les données pour rester accessibles aux autres outils de traitement du signal.

Principaux outils disponibles :

- graphique: affichage de données bidimensionnelles, tridimentionnelles (sonagrammes), affichage d'electropalatogrammes, modélisation interactive du F0 (MOMEL).

- traitement du signal: intensité RMS, densité de passage par zero, analyse spectrale (spectre, spectre LPC, sonagramme, analyse resynthèse par ondelettes), filtrage, detection de F0 (méthodes disponibles : AMDF, peigne spectral, autocorrélation et une combinaison de ces trois méthodes), modélisation de F0, analyse resynthèse (technique PSOLA), variation du débit (technique SOLA).

Le logiciel est librement disponible à :

http://www.lpl.univ-aix.fr/projects/multex/tools.html

# Annexe B : Calcul de la dérivée première et seconde

Pour calculer la dérivée d'un signal discret, il est nécessaire de passer par une approximation du signal à analyser. Le calcul de la dérivée première est basé sur la définition de l'énergie d'une fonction polynômiale du premier ordre.

$$E = \sum_{i=1}^{N} (y_i - ax_i + b)^2$$

en utilisant les conditions initiales :

$$\begin{cases} \frac{\delta E}{\delta a} = 0 \\ \frac{\delta E}{\delta b} = 0 \end{cases}$$

Ce qui donne pour la valeur de $a$ (dérivée première) :

$$a = \frac{\sum_i x_i y_i - \frac{1}{N} \sum_i x_i \sum_i y_i}{\sum_i x_i^2 - \frac{1}{N} (\sum_i x_i)^2}$$

avec $N$ désignant le nombre d'echantillons, $i$ variant de 1 à $N$.

La dérivée seconde est calculée de façon similaire, avec un polynôme du second degré pour caractériser le signal :

$$E = \sum_{i=1}^{N} (y_i - ax_i^2 - bx - c)^2$$

avec les conditions initiales :

$$\begin{cases} \frac{\delta E}{\delta a} = 0 \\ \frac{\delta E}{\delta b} = 0 \\ \frac{\delta E}{\delta c} = 0 \end{cases}$$

On obtient donc pour la valeur de $a$ :

$$a = \frac{\sum_i x_i^2 y_i - \sum_i x_i^3 d - \frac{1}{N} \sum_i x_i^2 \sum_i y_i + \frac{1}{N} \sum_i x_i^2 \sum_i x_i d}{\sum_i x_i^3 c + \frac{1}{N} \sum_i x_i^2 c \sum_i x_i - \frac{1}{N} (\sum_i x_i^2)^2}$$

avec $N$ désignant le nombre d'echantillons, $i$ variant de 1 à $N$ et

$$c = \frac{\sum_i x_i^3 + \frac{1}{N}\sum_i x_i \sum_i x_i^2}{\sum_i x_i^2 - \frac{1}{N}(\sum_i x_i)^2}$$

$$d = \frac{\sum_i x_i y_i - \frac{1}{N}\sum_i x_i \sum_i y_i}{\sum_i x_i^2 - \frac{1}{N}(\sum_i x_i)^2}$$

Remarque : l'expression $d$ correspond à la dérivée première vu plus haut.

# Annexe C : Une base de donnée pour tester les extracteurs de F0

Fundamental Frequency Determination Algorithm (FDA) Evaluation Database est une base de donnée constituée par Paul Bagshaw [Bagshaw 94] pour évaluer les méthodes d'extraction de la fréquence fondamentale à partir du contour laryngographique.

Les données laryngographiques, enregistrées en même temps que le signal de parole, constituent une référence précise sur les vibration des cordes vocales. En effet, la vibration glottique peut être directement recueillie au niveau des cartilages thyroidiensmm à l'aide de capteurs electro mécaniques sensibles aux vibrations de la membrane entourant le larynx. L'évaluation de la fréquence fondamentale est plus simple qu'à partir du signal de parole du fait de l'absence de bruit et de fréquences harmoniques.

Bagshaw a ainsi pu tester quelques méthodes d'évaluation de la fréquence fondamentale. Il différencie quatre types d'erreurs :

- les erreurs de non-voisement ("unvoiced in error") qui désignent les régions non voisées incorrectement classé comme voisé
- les erreurs de voisement ("voiced in error") qui concernent les régions voisées qui auraient dû être classées non-voisées
- les erreurs grossières hautes ("high gross errors")
- les erreurs grossières basses ("low gross errors")

Ces deux dernières erreurs concernes les évaluations de F0 qui dépassent en plus ou en moins de 20% de la valeur de référence.

La base est constitué du signal à 20 kHz en 16 bits, des données laryngographiques échantillonées à 20 kHz en 12 bits et du contour de la fréquence du larynx dérivé de ce dernier.

Voici les résultats calculés par Bagshaw pour les méthodes CPD (Cepstral based F0 Determinator), FBPT (Feature based F0 tracker), HPS (Harmonic product spectrum based method), IPTA (Integrated pitch tracking algorithm), PP (Parallel Processing method), SRPD (Super Resolution) et eSRPD (enhanced SRPD) qui est une méthode développée par Bagshaw. Les résulats sur les locuteurs masculins pour la méthode de combinaison de MESSIGNAIX ont été rajoutés. La partie du haut du tableau représente l'évaluation pour un locuteur

masculin et celle du bas un locuteur féminin.

| Algorithme | Unvoiced in error (%) | Voiced in error (%) | Gross errors Low (%) | Gross errors High (%) | déviations absolues moy. (Hz) | déviations absolues ecart-type |
|---|---|---|---|---|---|---|
| CPD | 18,11 | 19,89 | 4,09 | 0,64 | 2,94 | 3,60 |
| FBPT | 3,73 | 13,90 | 1,27 | 0,64 | 1,86 | 2,89 |
| HPS | 14,11 | 7,07 | 5,34 | 28,15 | 3,25 | 3,21 |
| IPTA | 9,78 | 17,45 | 1,40 | 0,83 | 2,67 | 3,37 |
| PP | 7,69 | 15,82 | 0,22 | 1,74 | 2,64 | 3,01 |
| SRPD | 4,05 | 15,78 | 0,62 | 2,01 | 1,78 | 2,46 |
| eSRPD | 4,63 | 12,07 | 0,90 | 0,56 | 1,40 | 1,74 |
| combinaison | 18,99 | 3,16 | 1,05 | 1,13 | 2,36 | 3,16 |
| CPD | 31,53 | 22,22 | 0,61 | 3,97 | 6,39 | 7,61 |
| FBPT | 3,61 | 12,16 | 0,60 | 3,55 | 5,40 | 7,03 |
| HPS | 19,10 | 21,06 | 0,46 | 1,61 | 4,59 | 5,31 |
| IPTA | 5,70 | 15,93 | 0,53 | 3,12 | 4,38 | 5,35 |
| PP | 6,15 | 13,01 | 0,26 | 3,20 | 6,11 | 6,45 |
| SRPD | 2,35 | 12,16 | 0,39 | 5,56 | 4,14 | 5,51 |
| eSRPD | 2,73 | 0,13 | 0,43 | 0,23 | 4,17 | 5,13 |

La base est disponible sur le réseau à l'adresse :
http://www.cstr.ed.ac.uk/~pcb/home.html

# Annexe D : INTSINT

INTSINT (Intonation...) est un modèle proposé par Hirst et De Cristo, qui vise à proposer une transcription formelle et inversible de la structure mélodique. Il s'appuie pour pour ceci sur un modèle de trajectoires, MOMEL [Hirst et al. 91], 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") : $T_{i-1} < T_i < T_{i+1}$

**D** ("down") : $T_{i-1} > T_i > T_{i+1}$

**S** ("same") : $T_{i-1} = T_i$

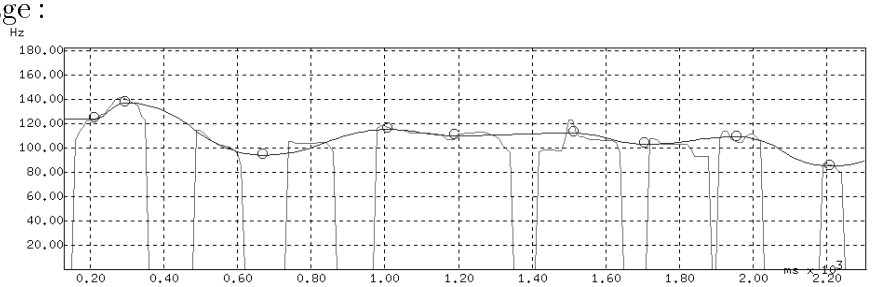
**H** ("higher") : $T_{i-1} < T_i > T_{i+1}$

**L** ("lower") : $T_{i-1} > T_i < T_{i+1}$

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

Ce premier étage d'abstraction est ensuite complété par un niveau d'abstraction phonologique comprenant : unités tonales (TU) et unités intonatives (IU) dont les rapports et la fonction dans le discours sont indiqués par un choix approprié de tons relatifs ou absolus.

Voici un exemple d'interpolation de la courbe mélodique et des points d'ancrage :



# Bibliographie


[Bagshaw 94] BAGSHAW P.C., *Automatic prosodic analysis for computer aided pronunciation teaching*, PhD. of Edinburgh Univ., 1994

[Deligne 96] DELIGNE S., *Modèles de séquences de longueur variable : application au traitement du langage écrit et de la parole*, thèse de l'ENST, Paris, octobre 1996

[Espesser 96] ESPESSER R., *"MES : un environnement de traitement du signal"*, XXIes Journées d'Etude sur la Parole, Avignon, p 447, juin 1996

[Geoffrois 95] GEOFFROIS E., *Extraction robuste de paramètres prosodiques pour la reconnaissance de la parole*, thèse du LIMSI, Paris, decembre 1995

[Gold et Rabiner 69] GOLD B., RABINER L.R., *"Parallel processing techniques for estimating pitch periods of speech in the time domain"*, Journal of the Acoustical Society of America, no. 46, pp. 442-448, 1969

[Gong et Haton 94] GONG Y., HATON J.P., *"Stochastic Trajectory Modeling for Speech Recognition"*, ICASSP'94, Adelaide, avril 1994

[Hazen et Zue 97] HAZEN T.J., ZUE V.W., *"Segmented-based Automatic Language Identification"*, Journal of the Acoustical Society of America, no. 101(4), avril 1997

[Hess 83] HESS W., *"Pitch Determination of Speech Signals - Algorithms and Devices"*, Springer-Verlag, 1983

[Hirst et al. 91] HIRST D., NICOLAS P., ESPESSER R., *"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

[Illina 97] ILLINA I., *Extension du modèle stochastique des mélanges de trajectoires pour la reconnaissance automatique de la parole continue*, Thèse de l'Université Poincaré, Nancy, octobre 1997

[Itahashi et Du 95] ITAHASHI S., DU L., *"Language Identification with phonological and lexical models"*, ICASSP'95, vol. 2, pp. 1359-1362, septembre 1995

[Li 94] LI K.P., *"Automatic Language Identification usiong syllabic spectral features"*, ICASSP'94, vol. 1, pp. 297 à 300, avril 1994

[Martin 81] MARTIN PH., *"Extraction de la fréquence fondamentale par intercorrélation avec une fonction peigne"*, XIIème Journées d'Etude sur la Parole, Montréal,, pp. 223 à 232, mai 1981

[Miller et Weibel 56] MILLER R.L., WEIBEL E.S., *"Mesurements of the fundamental frequency of a speech using a delay line"*, Journal of the Acoustical Society of America, no. 28(A), 1956

[Medan et al. 91] MEDAN Y., YAIR E., CHAZAN D., *"Super Resolution Pitch Determination of Speech Signals"*, IEEE ASSP, Vol. 39, No. 1, pp. 40 à 48, jan 1991

[Morley 97] MORLEY Y., *Génération multiparamétrique de la prosodie du français par apprentissage automatique*, thèse de l'ICP, Grenoble, décembre 1997

[Muthusamy 93] MUTHUSAMY Y.K., *Segmental Approach to Automatic Language Discrimination*, Ph. D. Thesis, Oregon Graduate Institute of Science Ec. Technology, 1993

[Muthusamy et al. 94] MUTHUSAMY Y.K., BARNARD E., COLE R.A., *"Reviewing Automatic Language Identification"*, IEEE Trans. On ASP, octobre 1994

[Pellegrino et al. 97] PELLEGRINO F., GALINIER A., ANDRE-OBRECHT R., *"Alternative Approaches to Phonetic Modeling"*, Workshop COST 254 on Emerging Techniques for Communication Terminals, Toulouse, juillet 1997

[Pellegrino et Obrecht 97a] PELLEGRINO F., ANDRE-OBRECHT R., *"From Vocalic Detection to Automatic Emergence of Vowel Systems"*, ICASSP 97, Munich, Avril 1997

[Pellegrino et Obrecht 97b] PELLEGRINO F., ANDRE-OBRECHT R., " *Vocalic system modeling : a VQ approach*", DSP'97, juillet 1997

[Ramus 96] RAMUS F., *Indices acoustiques permettant la discrimination de langues naturelles*, DEA de sciences cognitives, Ecole Polytechnique, 1996

[Rigoni 96] RIGONI J.J., *"A survey of Industrial Applications of Speech Synthesis"*, Conf. Proc. Of Interface to Real and Virtual Worlds, Montpellier, Mai 1996

[Sakoe et Chiba 71] SAKOE H., CHIBA S., *"A dynamic-programming approach to continuous speech recognition"*, IEEE Trans. ASSP-26, pp. 43 à 49, 1971

[Secrest et Doddington 82] SECREST B.G., DODDINGTON G.R., *"Post-processing technique for voice pitch trackers"*, ICASSP'82, pp. 172 à 175, 1982

[Soquet 94] SOQUET A., *"Approche coopérative de l'extraction de la fréquence fondamentale"*, XXèmes JEP, Trégastel, pp. 229 à 234, juin 1994

[Yan et Barnard 95] YAN Y., BARNARD E., *"An approach to Language identification with Enhanced Language Model"*, Eurospeech'95, Madrid, pp. 1351 à 1354, 1995