Synthèse de la parole à partir du texte

Le Beux Sylvain

CPE Lyon
Option Télécoms

Projet de Fin d'études réalisé à

IRCAM
1,place Igor Stravinsky
75004 Paris
France

Responsable IRCAM : Xavier Rodet
Responsable CPE Lyon : Nicole Gache

Février 2004 - Juillet 2004

Chapitre 1

Introduction

La synthèse de la parole à partir du texte s'entend comme la programmation d une machine capable de lire un texte et de transcrire celui-ci en un texte à voix haute. Si de prime abord un système TTS peut paraître simple du point de vue analytique, il en est tout autre quant à sa réalisation. Pour s'en convaincre il suffit de se rendre compte du nombre d'organes mis en jeu lors de la production de la parole, non seulement les organes de production (cordes vocales, souffle), et d'articulation (pharynx, cavité buccale, lévres) mais également le cortex qui permet la coordination de ces organes en vue d'une production cohérente de la parole. Cependant la construction d'une machine reproduisant synthétiquement ces différents organes est irréalisable car bien trop complexe à modéliser fidèlement. C'est pourquoi les sytèmes TTS actuels reposent sur des techniques bien différentes de sa correspondance humaine. Il est toutefois nécessaire de bien comprendre le fonctionnement humain afin de pouvoir réaliser un système synthétique.
Dans notre cas, la synthèse sera comprise comme une synthèse par concaténation, c'est-à-dire une juxtaposition de phonèmes les uns à la suite des autres permettant de reproduire le texte typographié, se posant alors le problème de la production la plus naturelle possible du texte donné. Pour cela une étude prosodique du texte apparaît indispensable, en ce sens qu'elle permet de lisser la parole concaténée. Typiquement, par exemple, une extraction de la fréquence fondamentale est nécessaire afin de diminuer les discontinuités dues à la suite de phonèmes provenant de structures (mots, phrases) parfois complètement différentes. Nous étudierons donc successivement les organes de production de la parole ainsi que des notions de phonétique (Chapitre 1), puis nous aborderons les outils indispensables pour l'analyse du signal de parole (Chapitre 2). Ensuite nous dresserons l'état de l'art des différentes méthodes utilisées en synthèse de la parole (Chapitre 3). Puis, nous traiterons plus précisément la synthèse de la parole par concaténation d'unités de taille variable (Chapitre 4). Enfin, nous présenterons les résultats obtenus lors de l'élaboration de notre système de synthèse(Chapitre 5).

Chapitre 2

Production et perception de la parole

2.1 Production de la parole

La parole est le résultat de l'activité des appareils respiratoires et articulatoires. Ces phénomènes reposent sur la modélisation source-filtre ; avec par exemple les cordes vocales dans le rôle de la source, produisant une oscillation quasi-périodique, et les cavités supraglottiques dans le rôle de filtres, dont la modification de formes générera des sons différents (ces dernières sont appelés les articulateurs).
Ainsi la production de la parole peut se résumer en :

- La génération d'un flux d'air qui va être utilisé pour faire naître une source sonore.
- La génération d'une source sonore sous la forme d'une onde quasi-périodique résultant de la vibration des cordes vocales et/ou sous la forme d'un bruit résultant d'une constriction du conduit vocal : c'est le rôle de la source vocale.
- la mise en place des cavités supraglottiques (conduits nasal et vocal) pour obtenir le son désiré.

2.2 Les sons de la parole

La parole est constituée d'un nombre finis d'unités élémentaires appelés phonèmes. On peut définir un phonème de la façon suivante : "Les phonèmes sont les éléments sonores les plus brefs qui permettent de distinguer différents mots".
Ainsi, l'étude des sons du langage est souvent divisée en deux approches :

- La phonétique qui s'intéresse à la manière dont les sons de parole sont produits, transmis et perçus.
- La phonologie qui s'intéresse à découvrir comment ces sons participent au fonctionnement de la langue dans l'acte de parole et à son codage.

On pourra distinguer ces deux approches grâce à un exemple. Lorsque le mot "rocailleux" est prononcé, il peut l'être soit avec un [r] roulé (produit avec le bout de la langue) soit avec un [r] grasseyé (produit avec le dos de la langue dans la gorge). Ainsi, on dira qu'ils sont phonétiquement distincts et phonologiquement semblables. Dans la suite, nous allons

uniquement nous attacher à décrire lesz différences phonétiques de la langue française.

2.2.1 Notions de phonétique

La phonétique s'intéresse à la manière dont les sons sont produits. Ainsi, il est possible de classer les phonèmes selon des caractères distinctifs.
Pour les voyelles nous avons donc :

- La nasalité : la voyelle a été prononcée à l'aide du conduit vocal et du conduit nasal suite à l'ouverture du velum
- L'ouverture du conduit vocal
- La position de la constriction principale réalisée entre la langue et le palais.
- La protusion des lèvres.

De même, les consonnes seront classées à l'aide de 3 traits distinctifs :

- Le voisement : la consonne a été prononcée avec une vibration des cordes vocales ou non.
- Le mode d'articulation.
- Le lieu d'articulation qui contrairement aux voyelles n'est pas nécessairement réalisé avec le corps de la langue.

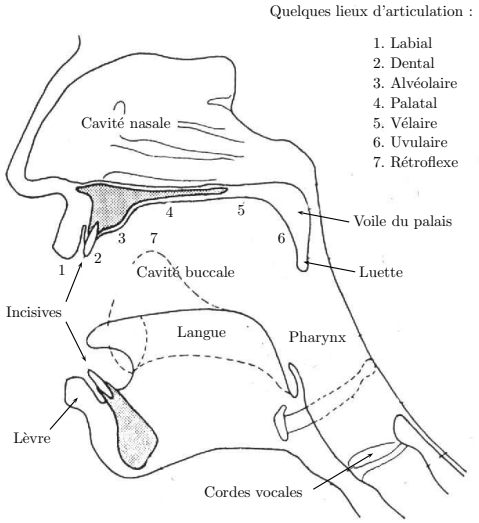

FIG. 2.1 – Description anatomique du conduit vocal et de ses differents constituants

Nous allons voir ci-dessous de manière un peu plus précise, les caractéristiques de chaque classe de sons.

2.2.2 Les voyelles

Les voyelles sont produites grâce à la vibration des cordes vocales, la distinction entre elles résultant du changement de forme du conduit vocal dû aux articulateurs.

- **Les voyelles antérieures/postérieures**
  Le lieu de la constriction du conduit vocal définit des voyelles antérieures, centrales et postérieures. Ainsi, pour une voyelle postérieure (comme /u/ dans "houx"), le corps de la langue sera placé très en arrière du conduit vocal, alors que pour une voyelle antérieure (comme /i/ dans "lit"), le corps de la langue sera ramené vers les dents.
- **Les voyelles ouvertes et fermées**
  L'ouverture du conduit vocal définit des voyelles ouvertes ou fermées. Ainsi, pour une voyelle fermée (comme /i/ dans "lit"), on aura un conduit vocal avec une importante constriction. Cette forme du conduit vocal correspond à une position haute de la langue. Pour une voyelle ouverte, à l'inverse, on aura une position de la langue plus basse et ainsi une constriction moins importante (comme /a/ dans "patte")
- **Les voyelles arrondies**
  La protrusion des lèvres définit des voyelles arrondies (ou labialisées) lorsqu'elles sont prononcées en avançant les lèvres vers l'avant (comme pour le son /u/ dans "houx"). A l'opposée, on trouve des voyelles non-arrondies (telles que le /i/ dans "lit") qui sont prononcées en étirant les lèvres.
- **Les voyelles nasales**
  Certaines voyelles mettent également en jeu le conduit nasal dont l'excitation est rendue possible grâce à l'abaissement du voile du palais. C'est notamment le cas de /an/ dans "pente".

2.2.3 Les consonnes

Comme pour les voyelles, les consonnes vont pouvoir être regroupées en traits distinctifs. Contrairement aux voyelles par contre, elles ne sont pas exclusivement voisées et ne sont pas nécessairement réalisées avec une configuration stable du conduit vocal.

- **Les consonnes voisées**
  On parlera de consonnes voisées lorsqu'elles auront été produites avec une vibration des cordes vocales (comme par exemple /b/ dans "bol" où les cordes vocales vibrent avant le relâchement de la constriction). Lorsqu'en plus du voisement, une source de bruit est présente due à une constriction du conduit vocal, on pourra parler de consonnes à excitation mixte (c'est le cas par exemple du /v/ dans "vent").
- **Les fricatives**
  Elles sont produites par un flux d'air turbulent prenant naissance au niveau d'une constriction du conduit vocal. On distingue plusieurs fricatives suivant le lieu de cette constriction principale :

  - Les labio-dentales, pour une constriction réalisée entre les dents et les lèvres (comme pour le /f/ dans "foin")
  - Les dentales, pour une constriction au niveau des dents (comme pour le /th/ anglais dans "thin")
  - Les alvéolaires, pour une constriction juste derrière les dents (comme pour le /s/ dans "son")
  - Les palatales, pour une constriction au niveau du palais dur (comme pour le /s/ dans chat).
  - Les laryngales, pour une excitation au niveau de la glotte (comme pour le /h/ anglais dans "he")
- **Les plosives**
  Elles sont réalisées en fermant le conduit vocal en un endroit. De même, que pour les fricatives, l'un des traits distinctifs entre les plosives est le lieu d'articulation. Pour les plosives, on aura ainsi :
  - Les labiales, pour une occlusion réalisée au niveau des lèvres (comme pour le /p/ dans "par")
  - Les dentales, pour une occlusion au niveau des dents (comme pour le /t/ dans "tarte").
  - Les vélo-palatales, pour une occlusion au niveau du palais (comme pour le /k/ dans "cake").
- **Les consonnes nasales**
  Elles sont en général voisées et sont produites en effectuant une occlusion complète du conduit vocal et en ouvrant le vélum permettant au conduit nasal d'être l'unique résonateur. Comme pour les autres consonnes, on aura, suivant le lieu d'articulation :
  - Les labiales, pour une occlusion du conduit vocal réalisée au niveau des lèvres (comme pour le /m/ dans "main")
  - Les dentales, pour une occlusion du conduit vocal au niveau des dents (comme pour le /n/ dans "non").
  - Les vélo-palatales, pour une occlusion du conduit vocal au niveau du palais (comme pour le /n/ dans "parking").
- **Les glissantes et les liquides**
  Cette classe de consonnes regroupe des sons qui ressemblent aux voyelles. Les liquides sont d'ailleurs parfois appelées semi-consonnes ou semi-voyelles. Les glissantes et les liquides,sont en général, voisées et non nasales. Les glissantes, comme leur nom l'indique, sont des sons en mouvement et précédent toujours une voyelle (ou un son vocalique). On aura :
  - la glissante vélo-palatale /R/ comme dans "rat"
  - la dentale /l/ comme dans "lit".

  Les liquides (ou semi-voyelles) sont des sons tenus, très similaires aux voyelles mais en général avec une constriction plus conséquente et avec l'apex de la langue plus relevé. On aura :
  - la labiale "Wé", notée /w/ que l'on trouve dans "loi" pour former le son s'intercalant entre le /l/ et le /a/.

– la dentale "Ué", notée /y/, que l'on trouve dans "nuit" pour former le son s'intercalant entre le /u/ et le /i/.En français, ce son est toujours suivi du phonème /i/.
– la vélo-palatale ("yod") comme /j/ pour former le son "ill" entre le /i/ et le /e/ dans "piller".

2.3 Notions de perception des sons de la parole

2.3.1 Description du signal de parole

Description temporelle

Sur des périodes de temps assez courtes (inférieures à 100 ms), le signal de parole peut être assimilé à un signal quasi-stationnnaire. Cependant sur des périodes plus longues, la variation temporelle du signal peut être très différente. On peut toutefois distinguer les parties voisées (enveloppe temporelle quasi-périodique) des parties non voisées (signal aléatoire de faible amplitude). Malgré tout, il sera difficile de distinguer une partie voisée de faible amplitude d'une partie non voisée. De plus, une telle représentation ne permet pas d'identifier/repérer les voyelles entres elles.

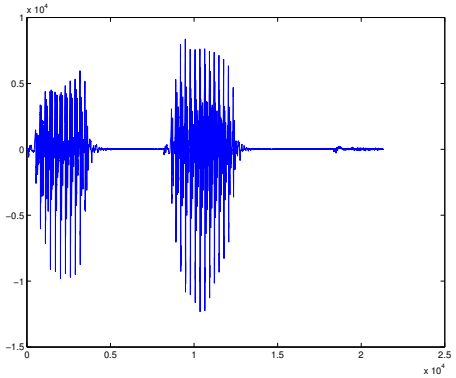

FIG. 2.2 – Signal temporel de la phrase "papap"

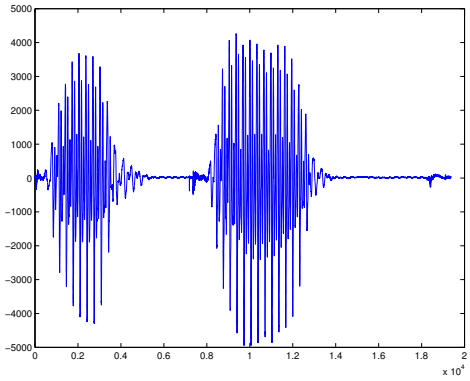

FIG. 2.3 – Signal temporel de la phrase "popop"

Description fréquentielle

La représentation la plus couramment utilisée du signal de parole est la représenataion spectrale ou spectrogrammme. Un spectrogramme représente la répartition fréquentielle du signal en fonction du temps. Plus précisément, le spectrogramme représente le module de la transformée de Fourier discrète calculé sur une fenêtre temporelle plus ou moins longue. La TFD étant donnée par :

$$X_i(k) = \sum_{n=0}^{N-1} x(n) e^{-\frac{2jkn}{N}}$$

Le spectrogramme est ensuite donné par une matrice dont chaque vecteur représente le module de la TFD d'une trame du signal de parole :

$$SPEC = \|X_0\|\|X_1\| \cdots \|X_L\|$$

où L est le nombre de fenêtres du signal de parole.
La taille de la fenêtre d'analyse est un paramètre important pour cette représentation. Pour de petites fenêtres (typiquement de l'ordre de 3 à 10 ms), on obtiendra une représentation avec une très bonne localisation temporelle mais avec une précision fréquentielle moins précise. On aura dans ce cas un spectrogramme à bande large. Dans le cas contraire où l'on choisit des fenêtres d'analyse de plus grande taille (typiquement supérieures à 20 ms), on obtient une plus grande précision fréquentielle au prix d'une localisation temporelle plus approximative. On parlera dans ce cas de spectrogramme à bande étroite. Pour la parole, les deux types de représentations sont utilisées suivant que l'on souhaite observer la structure fine du contenu fréquentiel (qui est clairement visible sur le spectrogramme à bande étroite) ou que l'on souhaite observer l'enveloppe spectrale ou les formants (qui sont plus clairement visible sur un spectrogramme à bande large). Les harmoniques sont alors très clairement identifiées sur le spectrogramme à bande étroite. Les formants sont plus particulièrement visibles sur les spectrogrammes à large bande : ils sont matérialisés par des zones plus sombres indiquant des zones fréquentielles de plus forte énergie. Cette représentation donne une "section" du spectrogramme et permet également de voir la structure fine (les harmoniques) et les formants à travers l'enveloppe spectrale. Il est ainsi possible de représenter les voyelles en fonction de la position de leurs deux (ou trois) premiers formants F1 et F2. Bien sûr, en pratique, une voyelle suivant les locuteurs et suivant leur prononciation ne possédera pas une position des formants rigoureusement stable. On donne dans les figures suivantes un certain nombre de spectrogrammes permettant de mettre en évidence certaines caractéristiques. Nous ne rentrerons pas ici dans le détail.

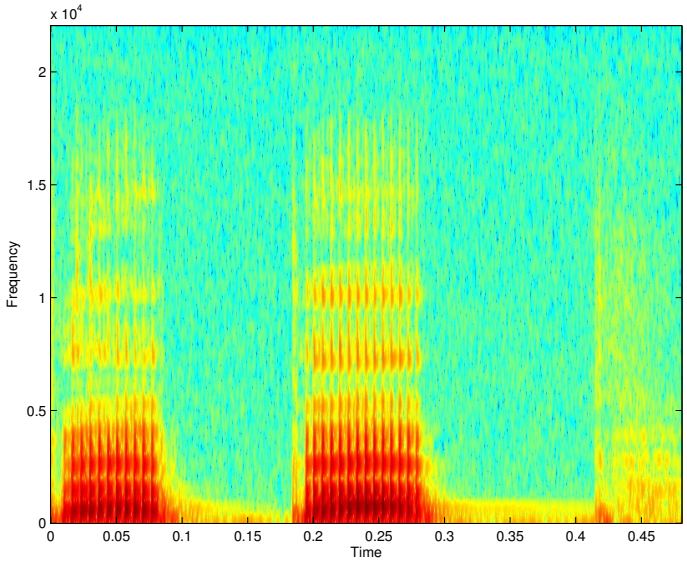

FIG. 2.4 – Spectrogramme de la phrase “papap” avec une fenêtre de 128

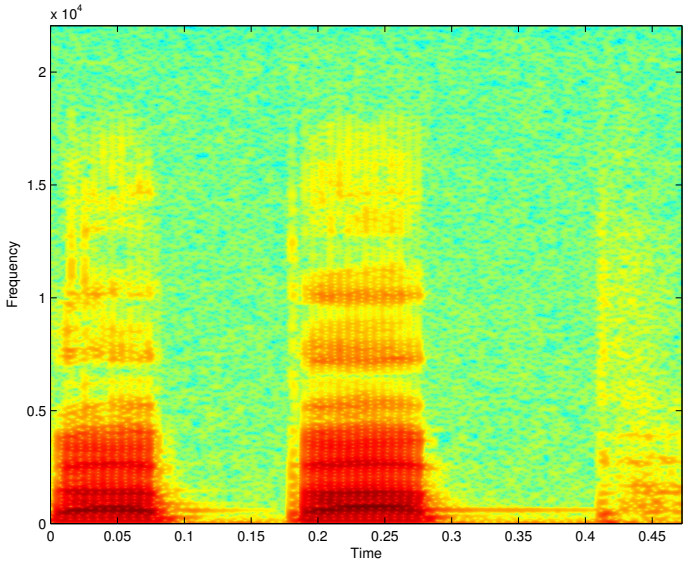

FIG. 2.5 – Spectrogramme de la phrase “papap” avec une fenêtre de 512

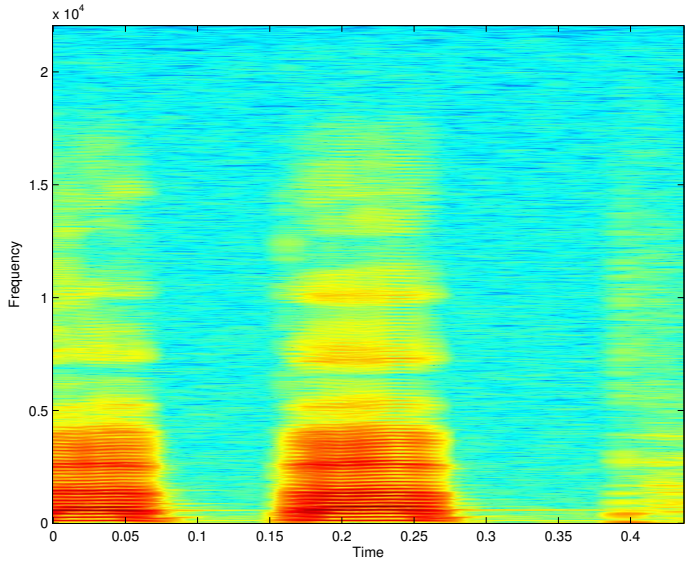

FIG. 2.6 – Spectrogramme de la phrase “papap” avec une fenêtre de 2048

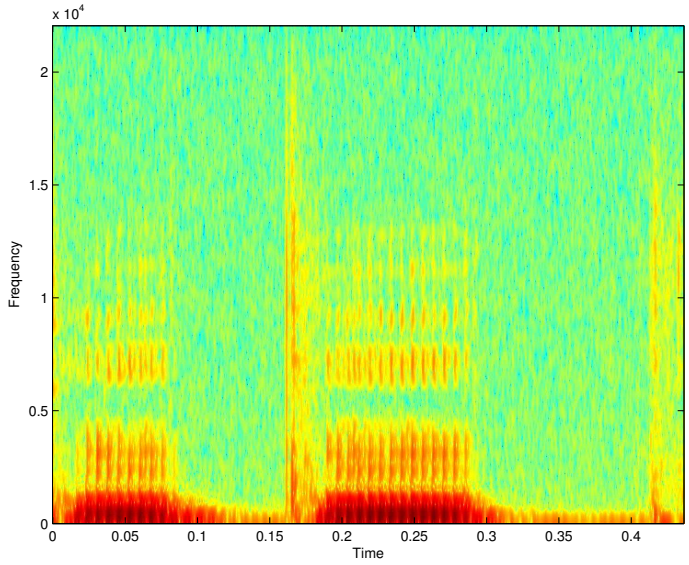

FIG. 2.7 – Spectrogramme de la phrase “popop” avec une fenêtre de 128

Chapitre 3

Outils d'analyse de la parole

3.1 L'échelle Mel

L'échelle Mel correspond à une approximation de la sensation psychologique de hauteur d'un son. De même que pour les formules analytiques de l'échelle Bark, il n'existe pas d'échelle Mel unique. Une relation couramment utilisée reliant la fréquence f et l'échelle Mel, mel(f), est donnée par :

$$melf = 1000 \times log_2(1 + \frac{f}{1000})$$

Notons que la fréquence 1000 Hz correspond à la valeur 1000 mel. L'utilisation de l'échelle Mel conduit à l'une des paramétrisations les plus utilisées en reconnaissance de la parole : les coefficients MFCC (pour Mel Frequency Cepstral Coefficients) utilisent évidemment cette échelle.

3.2 Représentation cepstrale

Comme nous l'avons vu précédemment, la parole peut être représentée sous la forme d'un modèle source-filtre. Cette représentation permet ainsi de représenter le signal de parole s(t) sous la forme du convolution du signal source g(t) par la réponse impulsionnelle du filtre h(t) représentant le conduit vocal :

$$s(t) = g(t) * h(t)$$

L'étude de ce signal à l'aide de la FFT présente un défaut particulier liée à cette convolution qui rend difficile l'observation de la seule contribution du conduit vocal. Le cepstre (parfois appelé lissage cepstral) permet de séparer les contributions respectives de la source et du conduit vocal. En effet, l'équation précédente se réécrit dans le domaine spectral sous la forme :

$$S(\nu) = G(\nu)H(\nu)$$

où $S(\nu), G(\nu)$ et $H(\nu)$ représentent respectivement les transformées de Fourier de s(t),g(t) et h(t). Le cepstre qui est défini par le logarithme de la transformée de Fourier inverse du module de $S(\nu)$ s'écrit donc sous la forme :

$$c(\tau) = FFT^{-1}log|S(\nu)| = FFT^{-1}log|G(\nu)| + FFT^{-1}log|H(\nu)|$$

On peut alors noter que le spectre s'exprime comme la somme de deux termes. Le premier terme $FFT^{-1}log|G(\nu)|$ est caractéristique de la source et représente ainsi la structure fine, tandis que le second terme est caractéristique de l'enveloppe spectrale et représente la contribution du conduit vocal. Le paramètre homogène à un temps est appelé quéfrence. A l'aide de cette représentation, il est possible d'isoler soit le pic (qui correspond au pitch) qui se trouve dans la région des hautes quéfrences (on a ici une méthode d'estimation de la fréquence fondamentale) soit d'isoler la partie correspondant aux basses quéfrences qui représente une version lissée de l'enveloppe spectrale. Ce procédé de séparation des éléments cepstraux est appelé un liftrage (par dérivation de l'appellation filtrage). Lorsque le cepstre est obtenu en calculant la transformée de Fourier discrète, on obtient la forme suivante :

$$c_n = \frac{1}{N}\sum_{k=0}^{N-1} log|X(k)|e^{2j\Pi\frac{kn}{N}} \; pour \; 0 \leq n \leq N-1$$

3.3 La paramétrisation MFCC

La paramétrisation MFCC (Mel-Frequency Cepstral Coefficients) est probablement la paramétrisation la plus répandue dans les systèmes de reconnaissance actuels. De même que pour les coefficients LPCC (Linear Predictive Cepstrum Coefficients), un certain nombre d'étapes sont nécessaires pour cette paramétrisation. Nous ne développerons ci-dessous que les étapes qui ne se retrouvent pas dans la paramétrisation LPCC :

- Fenêtrage du signal similairement à la paramétrisation LPCC
- Calcul de la transformée de Fourier rapide (FFT)pour chaque trame du signal de parole
- Filtrage par un banc de filtre MEL. Cette opération permet d'obtenir à partir du spectre $S(k)$ de chaque trame, un spectre modifié qui est en fait une suite de coefficients, noté $\tilde{S}(k)$, représentant l'énergie dans chaque bande fréquentielle k (définies sur l'échelle Mel), pour k = 1...K. En pratique, on utilise des filtres triangulaires de largeur de bande constante et régulièrement espacées sur l'échelle Mel (On peut par exemple choisir un espacement entre filtres de 150 mels et une largeur des filtres triangulaire prise à leur base de 300 mels).
- Calcul des coefficients MFCC : Les coefficients MFCC sont alors obtenus en effectuant une transformée en cosinus discrète inverse (de type II) du logarithme des coefficients $\tilde{S}(k)$:

$$\tilde{c}_n = \sum_{k=1}^{K}(log\tilde{S}k)\cos[n(k-\frac{1}{2})\frac{\Pi}{K}]pourn = 1, 2, \ldots, L$$

où L est le nombre de coefficients cepstraux désirés.
Dans notre implémentation,la paramétrisation MFCC consiste à prendre les 19 premiers coefficients cepstraux (en omettant l'énergie représentée par c_0) et à construire des vecteurs acoustiques de 57 éléments incluant les dérivées première δ et seconde $\delta\delta$ de ces coefficients.

3.4 Alignement Temporel et Programmation dynamique

Nous avons vu dans les sections précédentes plusieurs approches pour la comparaison de spectres de parole sur la base d'un segment (une trame) de parole. Bien évidemment, cette comparaison doit être menée pour l'ensemble du mot ou de la phrase prononcée. Hors, cette comparaison est confrontée au fait que deux mots ou phrases sont très rarement prononcées avec la même vitesse d'élocution et ainsi les deux séquences X (entrée que l'on cherche à reconnaître) et Y (référence apprise) n'auront pas en général la même durée. La solution la plus simple sera alors d'effectuer une déformation temporelle linéaire, c'est à dire associer plusieurs vecteurs de référence à un vecteur d'entrée (ou vice-versa si le vecteur d'entrée est plus long que le vecteur de référence). Ainsi une déformation temporelle linéaire pourra s'écrire :

$$d(\chi, \xi) = \sum_{i_x=1}^{T_x} d(i_x, i_y)$$

où i_x et i_y vérifient la relation

$$i_y = \frac{T_y}{T_x} i_x$$

Cependant, cet alignement n'est pas optimal car il suppose que le mot d'entrée est prononcé entièrement plus rapidement (resp. plus lentement) et toujours dans la même proportion. En pratique, il est possible que certaines parties (phonèmes) soient prononcées plus rapidement sur le mot test que pour le mot de référence alors que d'autres sections seraient prononcées plus lentement. On peut ainsi définir un alignement temporel plus général qui est couramment appelé Déformation Temporelle dynamique (ou DTW pour Dynamic Time Warping). Cette déformation utilise deux fonctions de déformation x et phiy qui relient les indices des deux segments de parole (ix et iy respectivement) à un axe temporel commun k :

$$i_x = \Phi_x(k) pour k = 1, 2, \ldots, T$$

$$i_y = \Phi_y(k) pour k = 1, 2, \ldots, T$$

Il est ensuite possible de définir une mesure de similarité $d_\Phi(\chi, \xi)$ à partir des fonctions de déformations sous la forme :

$$d(\chi, \xi) = \sum_{k=1}^{T} d(\Phi_x(k), \Phi_y(k))) \frac{m(k)}{M_\phi}$$

où $d(\Phi_x(k), \Phi_y(k)))$ mesure la distorsion spectrale pour les vecteurs $x_{\Phi_x(k)}$ et $y_{\Phi_y(k)}$, m(k) est un coefficient (non-négatif) de pondération le long du chemin et M_Φ est un facteur de normalisation. Pour compléter la définition d'une mesure de similarité pour la paire (χ, ξ), il est nécessaire de spécifier un chemin Φ . Ainsi, le problème est ramené à choisir un chemin de telle sorte que la mesure de similarité soit consistante. Un choix naturel (et populaire) est de définir $d(\Phi_x(k), \Phi_y(k))$ comme étant le minimum de $d_\Phi(\Phi_x(k), \Phi_y(k))$ sur tous les chemins possibles, soit :

$$d(\chi, \xi) = min_\Phi d_\Phi(\chi, \xi)$$

3.4.1 Programmation dynamique

La programmation dynamique (ou Dynamic programmming) est une approche qui permet, sous certaines conditions, d'obtenir la solution optimale à un problème de minimisation d'un critère d'erreur sans devoir considérer toutes les solutions possibles. Pour chercher la meilleure distance $D(T_x, T_y)$ entre deux séquences x et y, il suffit alors de chercher le chemin dans cette matrice D de façon à minimiser la somme des distances locales rencontrées pour aller d'un point initial (généralement (1,1) correspondant au début des mots test et référence) au point final (T_x, T_y) (correspondant à la fin des deux séquences). La mise en oeuvre de cet algorithme se fait alors de manière très simple. La distance optimale est obtenue en calculant, pour chaque entrée (i_x, i_y), la distance cumulée $D(i_x, i_y)$ correspondant à la distance optimale que l'on obtient en comparant les deux sous-séquences (sous-politiques) correspondant aux ix premiers vecteurs de test et aux iy premiers vecteurs référence. La distance accumulée minimale sur le chemin entre (1,1) et (i_x, i_y) sera ainsi donnée par :

$$D(i_x, i_y) = min_{\phi_x, \phi_y, T'} \sum_{k=1}^{T'} d(\phi_x(k), \phi_y(k)) m(k)$$

où

$$\phi_x(T') = i_x; \phi_y(T') = i_y$$

Notons que le coefficient de pondération M a été ici omis puisqu'il ne dépend pas du chemin suivi et qu'il peut être déduit des contraintes. Il sera ainsi ré-injecté une fois que le point final aura été atteint. Ce facteur de normalisation est couramment pris comme la somme des poids le long du chemin choisi soit :

$$M\phi = \sum_{k=1}^{T} m(k)$$

L'algorithme de programmation dynamique avec contraintes devient alors :

$$D(i_x, i_y) = min_{(i'_x, i'_y)}[D(i'_x, i'_y) + \zeta((i'_x, i'_y), (i_x, i_y))]$$

où ζ est la distance pondérée entre le point (i'_x, i'_y) et le point (i_x, i_y) :

$$\zeta((i'_x, i'_y), (i_x, i_y)) = \sum_{l=0}^{L_s} d(\phi_x(T' - l), \phi_y(T' - l)m(T' - l)$$

où L_s est le nombre de déplacements dans le chemin pour aller de (i'_x, i'_y) à (i_x, i_y). Notons que :

$$\phi_x(T' - L_s) = i'_x et \phi_y(T' - L_s) = i'_y$$

Notons cependant que la contrainte la plus utilisée est aussi la plus simple. Si la programmation dynamique est une technique utilisée dans de très nombreux domaines, son utilisation en reconnaissance vocale permet de définir des contraintes supplémentaires telles que :

- des contraintes de monotonie du chemin : le chemin commence au début des deux mots (point (1,1)) et se termine à la fin des deux mots (point (Tx,Ty)).
- des contraintes globales : par exemple certaines contraintes permettant de réduire l'espace de recherche (en imposant que le chemin optimal reste dans une zone déterminée proche de la diagonale).
- des contraintes locales : les prédécesseurs sont limités à quelques éléments proches et garantissent un chemin strictement gauche droite (les phonèmes sont prononcés dans le même ordre dans le mot "test" et le mot "référence". On ajoutera des pénalités de transition ou poids suivant les chemins pris.

En pratique on peut résumer l'implémentation de la programmation dynamique sous la forme :

- Initialiser la matrice D_A des distances cumulées avec la distance locale entre le premier vecteur de test et le premier vecteur de référence $D_A(1, 1)$ = d(1,1)m(1) où m(1) = 1
- Calculer les distance locales pour tous les autres éléments de la première colonne de D (soit d(1,i) c'est à dire les distances entre le premier vecteur de test et tous les vecteurs de référence)
- Si la transition verticale est autorisée, calculer les distances accumulées $D_A(1, i)$ correspondant à la première colonne. Si la transition n'est pas autorisée, les distances accumulées de la première colonne est égale à l'infini (sauf bien entendu pour le point (1,1)).
- Passer à la colonne suivante, calculer les distances locales d(2,i) et ensuite calculer les distances accumulées D(2,i) associées. Itérer sur toutes les colonnes.
- Lorsque le dernier point est atteint, réinjecter le coefficient de normalisation $d(\chi, \xi) = \frac{D_A(T_x, T_y)}{M\phi}$

Notons qu'après chaque itération, il n'est nécessaire de ne garder en mémoire que la dernière colonne de distances accumulées.

3.4.2 Reconnaissance de mots enchaînés à l'aide de la programmation dynamique

La reconnaissance de mots enchaînés est un problème plus complexe puisqu'il existe ici une co-articulation entre les mots et que les mots ne sont plus séparés par des silences. Comme il n'est pas envisageable de mettre en mémoire toutes les séquences de mots possibles, il va être nécessaire de segmenter (de façon automatique) la séquence d'entrée en terme des unités (mots) de référence. Plusieurs approches ont été proposées pour adapter l'algorithme de programmation dynamique. Nous ne décrirons ici que l'une d'entre elles, l'approche de programmation dynamique en une passe (one-pass dynamic time warping) en raison de sa faible complexité mais aussi parce qu'elle est à la base du décodage de Viterbi utilisé dans les systèmes HMM. L'algorithme en une passe est très semblable à l'algorithme DTW pour les mots isolés. Cet algorithme, comme pour la reconnaissance de mots isolés, commence par construire une grande matrice de distances locales entre tous les vecteurs constituant les mots de références (les mots du vocabulaire) et tous les vecteurs de la phrase test. On fait alors la programmation dynamique à travers toute la matrice, avec les conditions suivantes :

- au départ, le chemin peut commencer à partir de n'importe quel début de mot (en d'autres termes, le chemin ne commence pas nécessairement au point (1,1,1) correspondant au point (1,1) pour le mot de référence 1, mais peut commencer à l'un des points correspondant au début d'une référence soit (1,1,k) où k représente la kieme référence)
- à chaque instant n, l'ensemble des successeurs possibles associés au début de chaque mot (n,1,k) contient également la coordonnée (n - 1,J(k),k) correspondant au dernier indice de tous les mots k pouvant précéder k.
- à l'intérieur des références, les prédécesseurs possibles sont identiques au cas des mots isolés et dépendent des contraintes locales retenues.

3.4.3 Discussion

La DTW a été utilisée dès les années 1970. C'est cependant dans les années 1980 qu'elle est devenu un standard pour la reconnaissance vocale. L'intégration de distances locales dans le temps est devenue une notion essentielle qui est à la base de tous les systèmes modernes de reconnaissance et notamment ceux basés sur les modèles de Markov cachés. De nombreuses variantes et améliorations ont été apportées à ces approches. Notons, que nous avons toujours supposé que chaque mot de vocabulaire n'était représenté que par une seule prononciation. Il est clair qu'utiliser plusieurs prononciations du même mot permet d'envisager de meilleurs taux puisqu'une certaine variabilité sera alors prise en compte. La solution la plus simple avec l'approche par DTW est de prendre plusieurs références par mot à reconnaitre et d'effectuer plusieurs reconnaissance DTW. Cette solution peut être suffisante pour des systèmes mono-locuteurs mais est vite impraticable pour des systèmes multilocuteurs. L'une des améliorations consiste à utiliser la quantification vectorielle permettant de regrouper soit plusieurs références d'un mot en une seule soit de regrouper

les vecteurs acoustiques représentant ces références. On peut, par exemple, utiliser l'algorithme des K-means pour définir des vecteurs de mots prototypes à partir de l'ensemble des vecteurs acoustiques des mots de référence. Notons qu'il n'est pas ici nécessaire de savoir à quel mot appartiennent les vecteurs acoustiques. Les vecteurs acoustiques constituant les mots de référence sont ensuite remplacés par l'étiquette du vecteur prototype le plus proche. Cette quantification vectorielle engendre un certain lissage des références et représente un pas vers les modèles HMM. Les améliorations majeures apportées à cette approche de base DTW concernent principalement les notions de distances statistiques et les procédures d'entraînements qui y sont liées.

3.5 L'analyse TD-PSOLA

3.5.1 Introduction

La méthode PSOLA (Pitch Synchronous OverLap Add) a été proposée par Moulines, Charpentier et Hammond à l'ICASSP en 1986. Elle a été modifiée et améliorée par Geoffroy Peeters dans sa thèse à l'IRCAM en 1996. C'est cette dernière version que nous allons décrire.
"La méthode de superposition/addition synchrone à la période fondamentale, PSOLA (Pitch Synchronuous Overlap-Add), repose sur une décomposition d'un signal en une série de formes d'onde élémentaires. Ces formes d´onde élémentaires sont obtenues par un fenêtrage exactement centré sur les périodes fondamentales du signal. Le signal de synthèse est alors reconstitué par superposition/addition (Overlap-Add) de ces formes d´onde élémentaires. La modification de la distance relative entre deux formes d´onde élémentaires, ainsi que la modification du nombre de formes d´onde élémentaires, permet de modifier la hauteur et l´axe temporel du signal."

3.5.2 La synthèse PSOLA

La synthèse est différente selon le type de forme d'onde rencontré, nous en définissons trois :

1. Les zones périodiques ou voisées,
2. Les régions non-périodiques absentes de singularités, ou zone bruitée ou non-voisée.
3. Les singularités non-périodiques ou transitoires,

Une région du signal peut contenir dans son spectre une partie voisée dans les basses fréquences et une partie bruitée dans les hautes fréquences. Ainsi nous définissons une fréquence de coupure entre ces deux zones. Une régions sera dite non-voisée lorsque $f_c < f_{min}$.

3.5.3 Positionnement des marques d'écriture

La position des marques d´écriture $\tilde{m}_j$ dépend de la fréquence fondamentale $f(t)$ désirée pour le signal de synthèse. Pour les région périodiques, elles sont calculées comme suit :

$$\tilde{m}_{j+1} = \tilde{m}_j + \frac{1}{f(\hat{co}_j)}$$

où $\hat{co}_j$ est le temps dit de "correspondance" c´est à dire le temps correspondant à $\tilde{m}_j$ sur le signal original. Il dépend de la fréquence fondamentale désirée $f(t)$ et de la dilatation de l´axe temporel $D(t)$:

$$\hat{co}_{j+1} = \hat{co}_j + \frac{1}{f(\hat{co}_j).D(\hat{co}_j)}$$

Dans le cas des régions bruitées, le calcul est le même en gardant la fréquence $\bar{f}_0(t)$ (inter-distance entre formes d'onde élémentaires) identique à celle du signal original [1].

Enfin le traitement des singularités non-périodiques doit prendre en compte leur spécificité en empêchant la réutilisation ou l'omission d'une forme élémentaire renfermant un transitoire. L'algorithme recopie donc tous les transitoires sans les modifier et en concervant les périodes qui les séparent de leur prédécesseur et successeur. Nous obtenons donc dans ce cas :

$$\hat{co}_{j+1} = t_s$$
$$\hat{co}_{j+2} = t_{ss}$$

où t_s est le temps où intervient la singularité dans le signal original et t_{ss} est son successeur, et

$$\tilde{m}_{j+1} = \tilde{m}_j + \hat{co}_{j+1} - \hat{co}_j$$
$$\tilde{m}_{j+2} = \tilde{m}_{j+1} + \hat{co}_{j+2} - \hat{co}_{j+1}$$

3.5.4 Sélection et interpolation des formes d'onde élémentaires

La forme d´onde élémentaire $\tilde{s}_j$ associée à la marque $\tilde{m}_j$ est l´interpolation des formes d´onde voisines de $\hat{co}_j$.

$$\begin{cases} l = \arg\min_i |\hat{co}_j - m_i| & tel\ que\ \hat{co}_j > m_i \\ l' = \arg\min_i |\hat{co}_j - m_i| & tel\ que\ \hat{co}_j < m_i \end{cases}$$

$$\begin{cases} \alpha = \frac{\hat{co}_j - m_l}{m_{l'} - m_l} \\ \tilde{s}_j = F(s_l(t + m_l), s_{l'}(t + m_{l'}), \alpha) \end{cases}$$

où m_l et $m_{l'}$ sont les marques d'analyse entourant $\hat{co}_j$, α est le coefficient d'interpolation et F est la fonction d'interpolation.

[1] Lors de l'analyse, dans les régions non-voisées les marques sont positionnées avec une inter-distance égale à la moyenne de la période fondamentale des régions voisées voisines.

3.5.5 Addition/Recouvrement

Le signal de synthèse $\tilde{s}(t)$ est obtenu par addition/recouvrement des formes d´onde :

$$\tilde{s}(t) = \sum_j \tilde{s}_j(t - \tilde{m}_j)$$

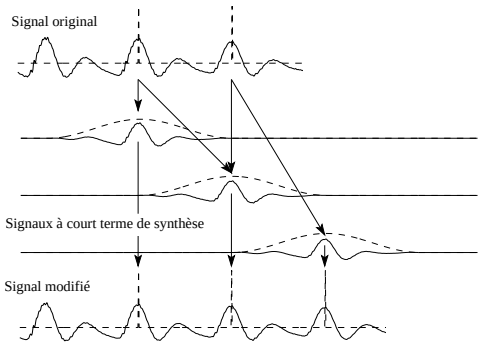

FIG. V.11 – *Modification de la durée du signal par la méthode TD-PSOLA. En haut, le signal original, au milieu trois signaux à court-terme générés à partir des deux signaux à court-terme centrés autour des deux premières marques d'analyse. En bas, signal modifié.*

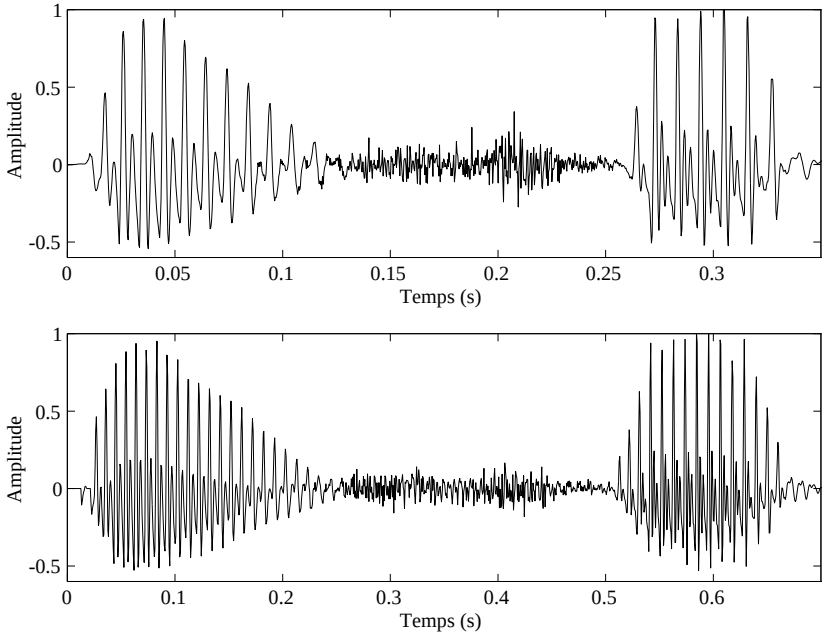

FIG. V.12 – *Original: "il s'est"(d'après [17])*

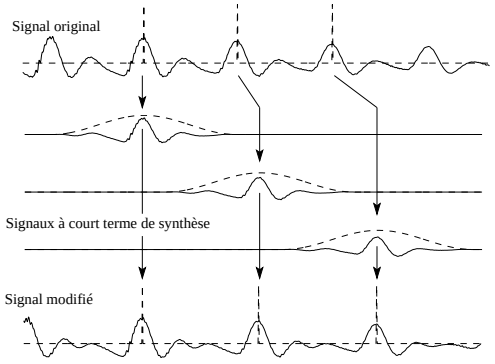

FIG. V.13 – *Modification de la hauteur du signal par la méthode TDPSOLA. En haut, le signal original, au milieu trois signaux à court-terme générés à partir des trois premières marques d'analyse. En bas, signal modifié. L'écartement des marques de synthèse n'est pas identique à celui des marques d'analyse.*

Chapitre 4

Etat de l'art de la synthèse de la parole

4.1 Techniques de synthèse de parole

La synthèse "par règles" et la synthèse "par concaténation d'unités acoustiques" sont des méthodes de synthèse : elles concernent la façon dont la parole est créée. Nous présentons ici les principales techniques de synthèse de parole, qui sont associées à la façon¸ dont la parole est représentée pour être synthétisée (une présentation complète des techniques d'analyse et de synthèse de la parole peut être trouvée dans le livre [Boeffard et d'Alessandro, 2002], ainsi que de nombreux exemples sonores dans le livre de [d'Alessandro et Tzoukermann, 2001]). Ces techniques sont actuellement étudiées à différents niveaux de recherche, de l'acquisition de connaissances sur la phonation au développement de systèmes performants.

4.1.1 Synthèse articulatoire

La synthèse articulatoire simule le fonctionnement physique de l'appareil phonatoire. Un modèle articulatoire simule le conduit vocal et l'écoulement de l'air pour calculer le signal résultant. Les paramètres de commandes sont la pression subglottale, la tension des cordes vocales et la position relative des articulateurs. Cette technique apporte beaucoup d'informations sur le mécanisme de phonation, mais la grande quantité de paramètres qu'elle requiert la rend encore difficilement exploitable [Mermelstein, 1973, Maeda, 1982].

4.1.2 Synthèse à formants

Cette synthèse est fondée sur un modèle prenant en compte l'information perceptive principale associée aux sons voisés : les formants de la voix. Les trois premiers formants peuvent suffir au codage spectral des phonèmes. Il faut ensuite ajouter des bandes de bruits pour produire les sons non voisés, et reconstituer l'effet du canal nasal. Une douzaine de

paramètres permettent ainsi une bonne restitution [Flanagan, 1972, Holmes, 1973, Klatt, 1980, Hertz, 1991, Hanson, 1997].

4.1.3 Prédiction linéaire

Le codage par prédiction linéaire (en anglais LPC, Linear Prediction Coding) est basée sur un modèle de production de la parole. La parole est vue comme un signal "autorégressif" : à un instant t, le signal est une combinaison linéaire des p échantillons précédents pour un modèle d'ordre p. L'algorithme calcule alors les coefficients de la combinaison linéaire de façon à minimiser l'erreur quadratique moyenne entre le signal original et le signal prédit sur une fenêtre donnée [Atal et David, 1979, Makhoul, 1975, Rabiner et Schafer, 1978].

4.1.4 Synthèse harmoniques + bruits

Les modèles hybrides harmoniques+bruits (ou harmoniques/stochastiques) modélisent le signal de parole en la somme d'une série d'harmoniques ayant pour référence la fréquence fondamentale F0, associée principalement aux sons voisés, et une composante apériodique de bruits pour les sons non voisés ou certaines phases transitoires. On trouve ainsi le modèle à excitation multibande (MBE, [Griffin, 1987]), le modèle HNM ([Stilianou et al., 1995]) mais aussi d'autres modèles présentés dans [Abrantes et al., 1991, d'Alessandro et al., 1998].

4.1.5 Synthèse directe et modifications prosodiques

Nous avons vu une technique basée sur un modèle de production de la parole : la synthèse articulatoire. Nous avons vu des techniques fondées sur des modèles de perception de la parole, d'abord la synthèse par formants, qui utilise le modèle le plus simple, puis la synthèse LPC, qui offre un modèle un peu plus complexe de la parole et la synthèse harmoniques et bruit qui enrichie encore le modèle. Nous arrivons maintenant à la synthèse "directe", sans modèle, par concaténation d'unités pré-enregistrées : le signal temporel entier est stocké. Il est nécessaire, la plupart du temps, de modifier la prosodie des unités concaténées. La technique TD-PSOLA, "addition-recouvrement de fenêtres temporelles synchrones au pitch" (en anglais Time-Domain Pith-Synchronuous Overlap and Add) a été mise au point à France Télécom R&D (alors appelé CNET) [Charpentier et Moulines, 1989, Moulines et Charpentier, 1990, Moulines et Laroche, 1995, Faucheur et al., 1991]. Elle permet de modifier efficacement le pitch et la durée des segments pour les concaténer et appliquer la prosodie cible. Elle consiste à décomposer temporellement le signal à l'aide de fenêtres synchrones au pitch, et à le recomposer avec les nouvelles périodes associées aux marqueurs de pitch cibles. La technique PSOLA s'applique donc surtout aux parties voisées du signal.

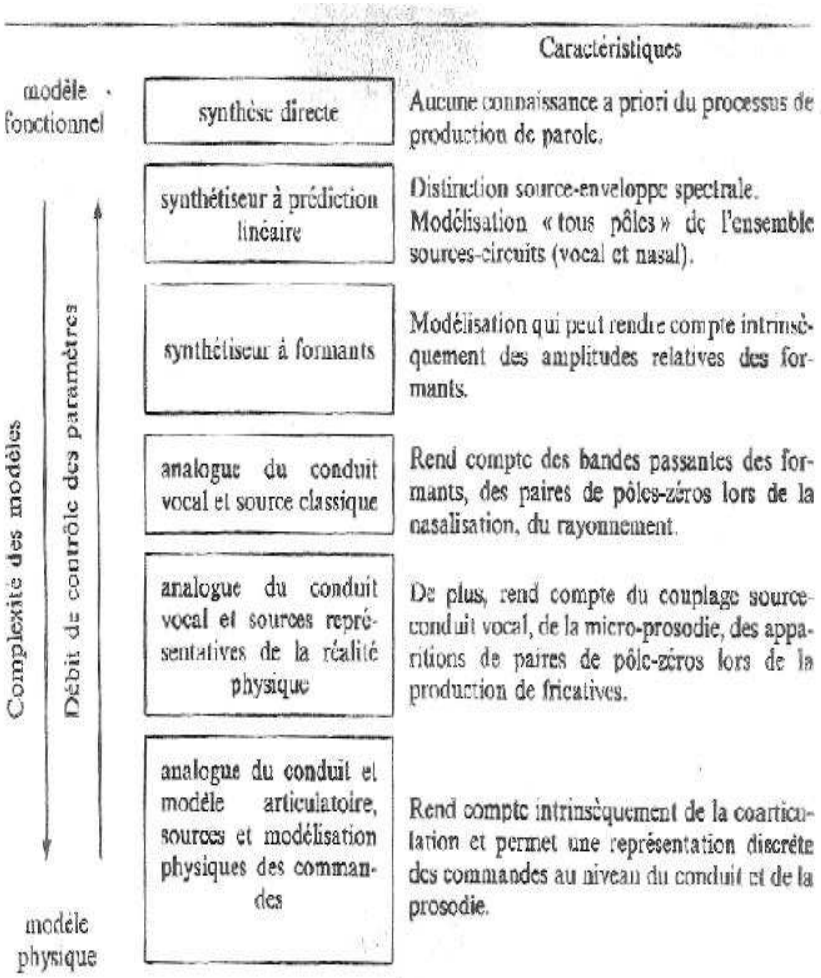

FIG. V.7 – *classification des techniques de synthèse (d'après [7])*

Chapitre 5

Synthèse de la parole à partir du texte

5.1 Contexte du stage

La synthèse musicale par sélection d'unités consiste à choisir dans une large base de données les unités sonores les plus appropriées pour construire, par concaténation et modification, la phrase musicale à produire. La thèse de D. Schwarz sur ce sujet s'est terminée en 2004.
Elle présente :

- Constitution d'une large base de données par alignement de partitions.
- Création d'un système de gestion et de sélection : CATERPILLAR
- Applications musicales.

Conséquemment à ses travaux, une application en voix parlée a été aussi envisagée dans le cadre d'un projet de reconstitution de la voix d'un locuteur disparu. Ce stage s'inscrit dans le projet de synthèse de la parole de haute qualité : TALKAPILLAR. Ce projet vise à synthétiser la voix de locuteurs spécifiques (Jean Cocteau et Gille Deleuze) pour rendre audibles des textes jamais prononcés par ces locuteurs. La génération de la prosodie par sélection d'unités est à mi-chemin entre la synthèse de la parole et la synthèse musicale. D'ailleurs, on parle également de prosodie instrumentale lorsqu'on veut décrire des nuances de pitch expressives (vibrato, pitch bend...) ou des variations de durée propres à l'interprétation d'un instrumentiste.

5.2 Analyse syntaxique

L'analyse syntaxique ainsi que la phonétisation automatique sont réalisées, dans le cadre de notre travail, par le logiciel EULER développé par l'université polytechnique de Mons (Belgique) qui s'inscrit dans un project de TTS plus vaste. Le module d'analyse du texte est composé de :

- Un module de pré-traitement, qui transforme les phrases données en groupes de mots strucurés. Il identifie les nombres, les abréviations, les acronymes ... et les transcrit en toutes lettres si besoin est. Un problème important se pose lorsqu'il s'agit de la ponctuation créant des ambiguités. Nous avons décidé de résoudre ce problème en remplaçant tous les signes de ponctuation par des points (, ? ! : ; ...) formant ainsi des phrases plus courtes mais dont le traitement est rendu plus aisé pour la suite.
- Un module d'analyse morphologique, dont la tâche est de proposer toutes les catégories linguistiques possibles pour chaque mot pris individuellement, sur la base de leur prononciation. Les mots infléchis, dérivés ou composés sont décomposés en leurs unités graphémiques élémentaires (leurs morphèmes) par de simples règles grammaticales exploitant des lexiques radicaux et affixes.(voir le programme de conversion TTS du CNET [Larreur et al. 89]).
- Le module d'analyse contextuelle prend en compte les mots suivant leur contexte, ce qui permet de réduire la liste des catégories linguistiques possibles à un nombre restreint d'hypothèses hautement probables, étant donné les catégories linguistiques possibles pour les mots voisins. Ceci peut être réalisé grâce aux n-grammes [voir Kupiec 92, Willemse & Gulikers 92], qui decrivent les dépendances syntaxiques locales sous la forme d'un automate probabiliste à états finis (comme un modèle de Markov par exemple), ou à un moindre degré des perceptrons multi-couches (commme les réseaux de neurones), tous les deux nécessitant un apprentissage.
- Enfin, un analyseur syntaxico-prosodique, qui examine les espaces de recherche restants et trouve la structure du texte (organisation en proposition principale et subordonnée par exemple) qui repose plus précisément sur sa réalisation prosodique attendue (voir plus loin)

5.3 Phonétisation automatique

Le module de synthèse de lettres (Letter-To-Sound en anglais) s'occupe de la détermination de la transcription phonétique d'un texte donné. Cela signifie, en premier lieu, que son rôle est aussi simple que de réaliser l'équivalent d'une organisation de dictionnaire. Si on examine de plus près cependant, on réalise rapidement qu'une bonne partie des mots du langage parlé apparaissent selon plusieurs transcriptions phonétiques, dont la plupart ne sont même pas mentionnées dans des dictionnaires de prononciation.
A savoir :

- Les dictionnaires de prononciation se réfèrent uniquement à la racine du mot. Elles ne tiennent pas explicitement compte des variations morphologiques (comme le féminin pluriel, les conjugaisons par exemple et spécialement pour une langue inflexive comme le français), qui de plus doivent être traitées grâce à une discipline de la phonologie, appelée morphophonologie.
- Certains mots correspondent en fait à plusieurs entrées dans le dictionnaire, ou plus généralement à plusieurs analyses morphologiques, avec elle-mêmes différentes prononciations. C'est typiquement le cas pour les homographes hétérophones (comme

les mots qui se prononcent différemment bien qu'ils s'épellent de la même manière comme "abdomen" et "examen"), ceci consistue de loin les plus sérieuses ambiguités de prononciation. Leur prononciation correcte dépend généralement de leur catégorie linguistique.

- Les dictionnaires de prononciation fournissent plutôt quelque chose de proche d'une transcription phonémique que phonétique (ils se réfèrent aux phonèmes plutôt qu'aux phones). Comme décrit par Withgott and Chen [1993] : "Alors qu'il est relativement facile de construire des modèles de programmation pour les phénomènes morpho-phonologique, comme la production du dictionnaire de prononciation de "électricité" étant donné la forme de base "électrique", c'est un autre problème de modèliser comment cela se prononce exactement". Les consonnes, par exemple, peuvent être réduites ou ignorées dans les groupes de consonnes, phénomène appelé simplification de groupe de consonnes, comme dans le mot "schéma" dans lequel le s et le c fusionnent en une seule prononciation.
- Les mots considérés dans des phrases ne sont pas prononcés de la même façon que lorsqu'ils sont isolés. De manière assez surprenante, les différences ne proviennent pas uniquement des variations des frontières des mots (comme pour les liaisons phonétiques), mais aussi des alternances basées sur l'organisation de la phrase en unités non-lexicales, soit en groupes de mots (comme pour la longueur phonétique) soit en parties non-lexicales (de nombreux processus phonologiques, par exemple, sont sensible à la structure syllabiques).
- Finalement, on ne peut pas trouver tous les mots dans un dictionnaire phonétique : la prononciation des mots nouveaux et de beaucoup de noms propres doit être déduite de celle de mots déjà connus.

Clairement, les points 1 et 2 dépendent énormément de l'analyse préliminaire morphosyntaxique (et parfois sémantique) de toutes les phrases à lire. Dans une moindre mesure, il arrive également que ce soit le cas pour le point 3 aussi, puisque les processus de réduction ne sont pas seulement une affaire de phonétisation contextuelle, mais reposent aussi sur la structure morphologique et sur les groupements de mots, autrement dit sur la morpho-syntaxe. Le point 4 nécessite une analyse poussée de la phrase, qu'elle soit syntaxique ou métrique, et enfin le point 5 peut être partiellement résolu en classant la structure morphologique en trouvant les analogies graphémiques entre les mots.

Il est possible ensuite d'organiser la tâche du module LTS de plusieurs manières, souvent classé sommairement par des stratégies basées sur des dictionnaires ou des règles, bien que beaucoup de solutions intermédiaires existent.

Les solutions basées sur des dictionnaires consistent à stocker le maximum de caractéristiques phonologiques dans un lexique. Afin de garder une taille raisonnablement petite, les entrées sont généralement restreintes aux morphèmes, et la prononciation des différentes formes est réalisée par des règles d'inflexion, de dérivation, et de composés morphophonémiques qui décrivent comment les transcriptions phonétiques de leur constituants morphémiques sont modifiés lorsqu'ils sont combinés dans des mots. Les morphèmes qui ne peuvent pas être trouvés dans le lexique sont transcrit par des règles. Après qu'une première traduction phonémique de chaque mot a été obtenue, des post-traitements phonétiques

sont appliqués, afin de produire le phénomène de lissage co-articulatoire. Cette approche a été suivie par le système MITTALK [Allen et al. 87] depuis le premier jour. Un dictionnaire de 12 000 morphèmes couvre environ 95% des mots d'entrée. Le système des laboratoires AT&T Bell suit la même ligne de conduite [Levinson et al. 93], avec un lexique de morphèmes augmenté à 43 000 morphèmes [Coker 85].
Une stratégie quelque peu différente est adoptée par les systèmes de traduction basés sur une approche par règles, qui transfère la majeure partie de la compétence phonologique des dictionnaires en un jeu de règles de lettre vers son (ou graphème vers phonèmes). Cette fois, seuls les mots prononcés de telle manière qu'ils constitue une règle par eux-mêmes sont stockés dans un dictionnaire d'exceptions. Notons que, comme de nombreuse exceptions sont trouvées dans les mots les plus fréquents, un dictionnaire d'exceptions de taille raisonnable suffit pour une large portion de mots d'un texte donné. En anglais, par exemple, 2000 mots suffisent typiquement à couvrir 70% des mots d'un texte [Hunnicut 80].
Il a été question ces dernières années de méthodes basées sur des dictionnaires qui soient capables d'obtenir une meilleure précision que des règles lettres vers son [Coker et al. 90], étant donnée la disponibilité de très grands dictionnaires phonétiques sur ordinateur. D'un autre coté, des efforts considérables ont été réalisés récemment à l'élaboration de jeu de règles dont la couverture est plus grande (en partant de dictionnaires numériques et en ajoutant des règles et des exceptions jusqu'à ce que tous les mots soient représentés, comme pour le travail de Daelemans & Van den Bosch [1993] ou celui de Belrhali et al.[1992]). Clairement, des inter-dépendances sont inévitables. De plus, le compromis dépend de la langue considérée, étant données les différences évidentes de fiabilité des correpondances lettre vers son pour différents langages.

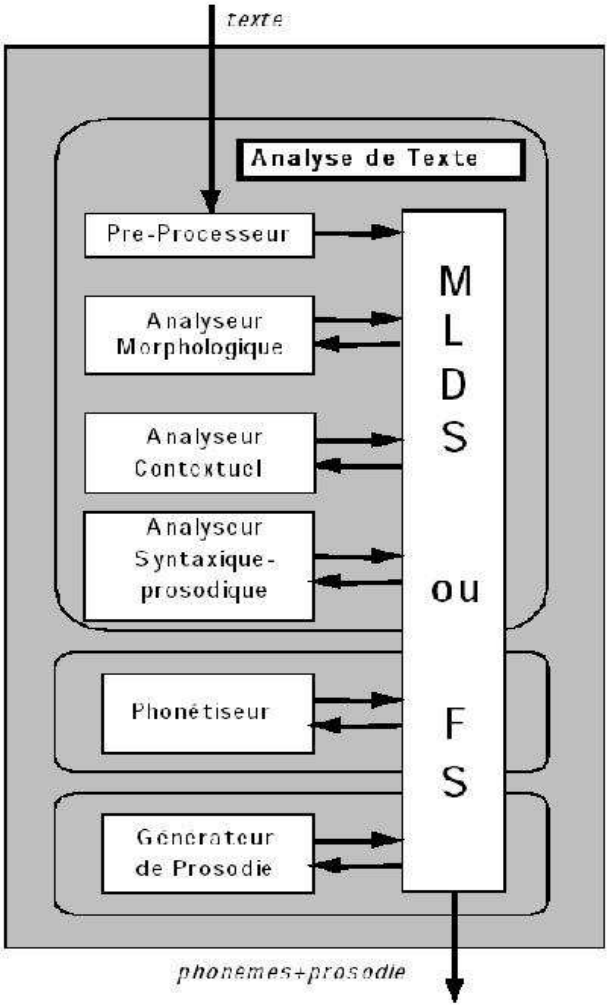

FIG. V.2 – *Module de traitement de texte (d'après [6])*

5.4 Synthèse par concaténation

Dans un système de synthèse par sélection d'unités, des segments audio de tailles variables sont sélectionnés dans un grand corpus de parole puis concaténés pour synthétiser un signal de parole extrêmement naturel. La première étape indispensable est l'indexation et la segmentation de la source. La deuxième étape consiste en l'évaluation du meilleur candidat correspondant le mieux possible avec la cible.
On distingue actuellement deux tendances pour le système de sélection :
La première, issue des travaux de Black, Hunt, et Campbell (1996) et utilisée principalement par ATT (US) et ATR (Japan), procède par minimisation dynamique d'une fonction

de coût, estimée à partir de la phrase à produire (et de ses caractéristiques linguistiques) et des phrases enregistrées dans une base de données (ces phrases étant elles-mêmes analysées en fonction des mêmes critères linguistiques que la phrase à produire). La base de données n'est pas organisée de façon particulière. Les unités disponibles ne sont pas regroupées en fonction de leurs similitudes spectrales. Cette approche, utilisée par Diemo Schwarz est la base de CATERPILLAR.
La seconde, qui résulte d'une thèse de doctorat déposée par Robert Ed. Donovan à Cambridge en 1996, organise au contraire la base de données de façon à pouvoir choisir rapidement l'unité requise, à partir de ses critères linguistiques. Le plus souvent, il s'agit d'une classification en arbre, effectuée une fois pour toutes, lors de la conception du synthétiseur. La taille de l'arbre est représentative de la finesse de la modélisation et peut donc être adaptée à l'inventaire des segments disponible. La sélection d'unités ne se fait qu'entre classes dont les contextes sont adéquats par opposition à une sélection globale.

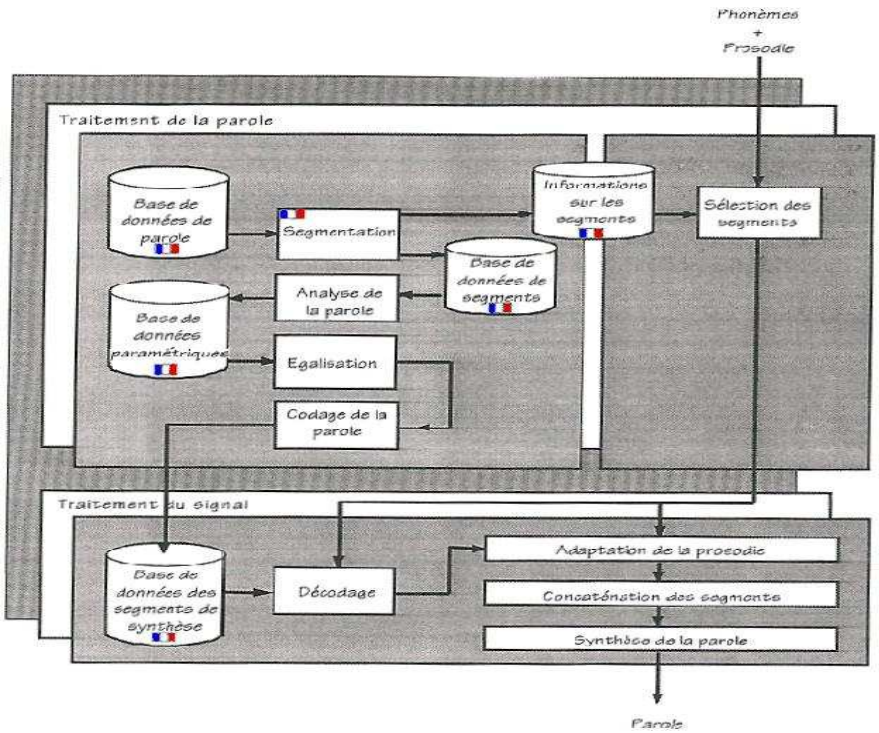

FIG. V.10 – *Schéma général d'un synthétiseur par concaténation. Les opérations qui dépendent de la langue sont indiquées par un drapeau.(d'après [6])*

5.5 Analyse prosodique

5.5.1 Définition

La structure prosodique résulte d'interactions complexes entre différents niveaux d'organisation sémantico-pragmatiques, syntaxique et rythmique. Elle se manifeste par le jeu simultané de plusieurs paramètres acoustiques : la fréquence fondamentale F_0, le timbre, l'intensité, la durée des phonèmes. Perceptivement, la hauteur et son évolution, le rythme et le tempo (débit), le registre et le timbre mais aussi les pauses et les silences nous permettent la compréhension d'informations au-delà des mots prononcés. C'est cette deuxième partie du double codage de la parole qui lui confère un caractère "naturel" et évite la monotonie. Elle permet entre autre de véhiculer des informations ectolinguistiques ou phonostylistiques (expressivité, sentiments), de lever des ambiguïtés de sens entre deux phrases phonétiquement similaires et de structurer l'énoncé. La variation de hauteur est certainement l'indice acoustique le plus important dans la prosodie. Le registre couvert par la plupart des locuteurs est souvent divisible en 4 niveaux perceptivement distinguables.
Nous les nommerons :

- H+H+ : niveau le plus haut
- HH
- LL
- L-L- : niveau le plus bas

La fréquence fondamentale F_0 évolue dans ce registre. Son évolution au cours du temps décrit des contours. Une phrase est généralement composée d'une suite de contours qui ne suivent pas nécessairement la même orientation de pente. On observe cependant une déclinaison générale qui correspond à un abaissement de F_0 du début à la fin de l'énoncé. La hauteur la plus basse correspond donc à la fin de cet énoncé et constitue ainsi un bon indice de segmentation. Ce phénomène à priori universel est de nature physiologique, mais il est géré par le locuteur à des fins linguistiques ; il permet de délimiter la fin d'une phrase syntaxique. Il faut remarquer que l'on ne peut évaluer cette fréquence fondamentale que sur les segments voisés (voyelles et quelques consonnes...). Aussi, nous extrapolons celle-ci durant les segments non voisées afin d'avoir des contours continus.

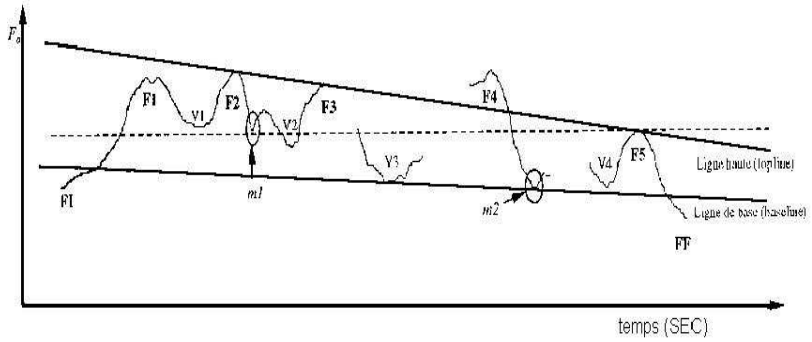

FIG. V.4 – *Présentation des principaux paramètres permettant de caractériser les événements mélodiques présents lors d'une analyse acoustique ; FI = fréquence initiale, FF = fréquence finale, Vx = vallées, Fx =pics mélodiques, mx = creux micromélodiques (d'après [3])*

5.5.2 La prosodie du Français

Le français est une langue à accent fixe ou accent de groupe (de mot). Elle se distingue ainsi des langues à accent libre comme l'anglais. L'anglais est une langue très musicale, caractérisée par de fortes variations de hauteurs et couvrant une large tessiture. Il utilise principalement les variations de hauteur et d'intensité. Les tons mélodiques sont très difficiles à acquérir pour les Français dont la tessiture est restreinte. D'autre part, l'organisation rythmique de l'anglais est complètement différente de celle du français. L'anglais est une langue "stress timed" (Pike, 1947) où l'accent n'est pas prédictible, mais l'espace entre deux pics accentuels est à peu près stable. A l'inverse, la place de l'accent tonique en français est totalement prédictible puisqu'elle affecte toujours la dernière syllabe du groupe

rythmique.
On distingue deux types d'accents qui mettent en relief la phrase :

- L'accent primaire (ou tonique) se traduit par un allongement de durée et une variation significative de Fo. Il a une fonction structurante et peut se déduire de la syntaxe.
- L'accent secondaire se manifeste par des variations plus subtiles de Fo et de l'intensité. Il a une fonction focalisante, rhétorique ou expressive.

Cette distinction est fondamentale car elle met en valeur la différence fonctionnelle de ces deux accents. Ils sont les marqueurs temporels et acoustiques de deux types de groupes :

- Les groupes intonatifs (qui se terminent par un accent primaire). Ils expriment la modalité de la phrase. Ils ne sont pas congruents à la syntaxe mais la syntaxe est congruente à l'intonation.
- Les groupes accentuels (qui contiennent un accent secondaire). Ils mettent en relief des mots.

Les groupes intonatifs comprennent généralement un ou plusieurs groupes accentuels. Mais cet imbriquement et cette différence de durée n'impliquent en rien une hiérarchisation entre ces deux éléments car ils ne possèdent pas la même fonction. L'intonation permet de manifester la modalité de la phrase en français :

- phrase assertive : contour descendant du niveau haut au niveau moyen
- phrase impérative : contour descendant linéairement du niveau haut au niveau bas.
- question partielle ou interrogation : contour courbe descendant du niveau haut au niveau bas
- question totale : contour courbe montant du niveau bas au niveau haut

Le laboratoire de morphosyntaxe de Paris III (1991-1997) propose d'affiner cette description par contours en positionnant le locuteur sur ce qu'il dit : Chaque niveau de hauteur étant le reflet de ce positionnement :

- H+H+ : Mise en place de la co-énonciation
- HH : Consensualité acquise
- LL : Niveau neutre
- L-L- : Rupture de la co-énonciation, égocentrage

L'attitude monologale (contour descendant du LL au L-L-) et l'attitude dialogale (contour montant du HH au H+H+) deviennent des descripteurs de contour efficients pour exprimer les modalités suivantes :

- l'incise (parenthèse) : accélération du débit
- la négation : discordance de point de vue désengagement du locuteur
- la question :
  - valeur neutre et consensuelle
  - changement de thème, demande de confirmation
  - suscite une réaction, énonciative
- l'exclamation :
  - appel à une convergence de point de vue
  - ironie, égocentrage suivant un consensus
  - surprise, discontinuité dans le fonctionnement de la pensée

5.5.3 les modèles accentuels de la phrase française

Les recherches sur la prosodie ont aboutit à de nombreux modèles accentuels de la phrase française. Tous ces modèles sont issus d'observations et aboutissent pour la plupart à des jeux de règles. Qu'ils partent d'analyses syntaxique, phonologique ou rythmique (psycho-acoustique), ils permettent de mieux comprendre d'ou proviennent les paramètres acoustiques de la prosodie. Cependant, il convient de se demander s'ils sont adaptés à la prédiction prosodique pour une génération automatique qui se veut naturelle et surtout personnalisée. Peux t on envisager une construction de la prosodie par règles dans le cadre de notre mission artistique ? L'élaboration de tous ces modèles visent à obtenir une vision globale et généraliste de la structuration prosodique. Dans tous les cas, ces modèles ont été élaborés dans l'optique de prédire l'évolution des paramètres acoustiques de n'importe quel locuteur. Cela revient à dire que, par conception, ces règles ne peuvent aboutir qu'au caractère normalisé de notre expression. En effet, de nombreux modèles ne cherchent à prédire que l'apparition des accents primaires, qui sont les indices de la modalité (frontières des groupes intonatifs). Elles ne mènent que rarement aux marqueurs accentuels (accents secondaires) propres à l'expressivité et dont les apparitions révèlent la "personnalité prosodique" de chacun. Une approche par règles nous est donc prohibée si nous voulons restituer dans des phrases synthétisées, la personnalité d'un locuteur spécifique. En l'occurence, il se trouve que Gilles Deleuze est particulièrement expressif de part son intonation.

5.5.4 La prosodie dans la synthèse de la parole

La synthèse de la prosodie apparait clairement indispensable pour tout système TTS (Text To Speech) qui désire véhiculer des informations que ne peut contenir les mots seulement. On distingue dans la littérature, trois méthodes pour la génération de la prosodie :

- L'approche par règles
- L'approche basée sur l'apprentissage à partir de corpus :
  - par réseaux de neuronnes
  - par HMM (Hidden Markov Models)
  - par d'autres méthodes statistiques...
- L'approche par sélection d'unités.

La connaissance de patrons intonatifs ou contours types permet aux domaines de la reconnaissance et de la synthèse de la parole d'élaborer des modèles de l'intonation française :

- Au CNET (1977-1989) : On étudie un corpus pour en extraire un jeu de règles qui attribue un patron intonatif en fonction de la syntaxe.
- Chez IBM (1971-1980) : On construit un jeu de règles statuant 9 contours types selon le nombre de syllabes, le nombre de mots... On distingue quatre niveaux dans une phrase : phrase, proposition, groupe et mot. L'auteur précise que les niveaux phrase et groupe suffisent pour la majorité des énoncés. Cela revient un peu à négliger les accents secondaires.
- G. Bailly (Grenoble) (1983) : Il segmente aussi la phrase en groupes de respiration, de phonation, de sens. Leurs tailles est généralement comprises entre 8 et 12 syllabes.

Pour genérer les contours, il utilise le modèle de H. Fujisaki. La continuité des contours est de nature physiologique. Ils répondent à des commandes discrètes :

- commande de groupe : réponse d'un 2nd ordre à un Dirac.
- commande d'accent : réponse d'un 2nd ordre à un Echelon.

Ce second ordre modélise le muscle crico-thyroïdien (en translation et en rotation). L'avantage de cette modélisation et qu'elle présente des coefficients constants adaptables pour chaque locuteur). Seuls l'amplitude et le temps de déclenchement varient. Les trois commandes de groupe sont : initialisation, réinitialisation, finalisation. Ils correspondent à l'expression de la modalité.

- V.Aubergé (Grenoble) (1991-1997) : Création d'un lexique de contours. Il part de l'autonomie entre syntaxe et prosodie. Grâce à un réseau de neurones entrainé sur un corpus, il crée un lexique faisant le lien entre syntaxe et contours prototypiques. C'est aujourd'hui le modèle le plus abouti.
- F.Beaugendre (LIMSI) (1994) : Reconnaissance de contours perceptivement pertinents ; 30 règles pour la génération de mouvements standards.

Dans le contexte d'une synthèse par concaténation d'unités (allant du semi-phone au mot ou plus), il semble "logique" de sélectionner aussi des unités prosodiques. Mais ce choix vient en fait de motivations plus profondes. En effet, cette approche permet tout d'abord une plus grande variété prosodique que les approches par règles. De plus elle permet de refléter le "caractère prosodique" de l'individu (chacun ayant ses modes d'intonation, registre...), ce qui est essentiel compte tenu de notre but artistique. Enfin, l'introduction de contours réels de Fo sur des blocs de parole permet de conserver la structure micro-mélodique.

5.5.5 Génération automatique de prosodie utilisant la sélection d'unités supra-segmentales (Malfrère, Dutoit et Mertens) 1998

Le système de l'université polytechnique de Mons que nous allons décrire repose sur la sélection d'unités prosodiques. Il utlise le générateur LIPSS du projet EULER qui génére une description symbolique de la prosodie à partir d'un texte (fichier .txt.mlc) :

- une étude syntaxique donne les accents finaux qui délimitent les unités :
  - NA : syllabe non accentuée
  - AF : syllabe accentuée (accent final=accent primaire)
  - UNDEFINED : pause (silencieuse)
- une étude de la modalité donne la hauteur du ton final :
  - déclaration : L-L-
  - interrogation : HH
  - exclamation : H+H+
  - temps de pause : P1 ou P2

Ce générateur est appliqué aux phrases de la source comme à celles de la cible. Il permet de créer des descripteurs d'unités prosodiques de longueurs variables et dont les frontières sont les accents finaux. Ainsi, chaque unité descripteur prosodique possède une clé propre

représentant :

- L'index de l'unité dans la phrase
- les tons des accents finaux de début (qui appartient à l'unité précédente) et de fin d'unité
- le nombre de syllabes neutres, inaccentuées dans l'unité

Cette clé peut ressembler par exemple à : "FA1NA1NA2NA3FA2" ou FA1 et FA2 prennent leur valeurs dans HH, HH, L-L-, H+H+,N dans lequel N représente le début d'une phrase. On ajoute aux clés des unités de la source des marqueurs en liens avec le fichier audio aligné qui nous permettent de retrouver les paramètres acoustiques comme l'évolution réelle de Fo durant l'unité. Le choix de l'unité optimale s'effectue en minimisant une fonction de coût. Comme pour le choix d'unités segmentales, cette fonction de coût résulte de l'addition de deux coûts :

- coût de distance à la cible :
  - les tones des premier et dernier accents doivent correspondre
  - une pondération est ajustée en tenant compte du nombre de syllabes
  - une autre est fonction de la position de l'unité dans la phrase

  On obtient ainsi une présélection de plusieurs unités candidates.
- coût de concaténation : Il est seulement basé sur la proximité des valeurs moyennes de Fo de deux unités consécutives. On aboutit grâce à l'algoritme de Viterbi à la sélection finale des unités en choisissant celles dont l'enchaînement présente le coût le plus faible. Puis on va extraire des unités suprasegmentales de la source choisies, les paramètres acoustiques (l'évolution de Fo). Ensuite, on les fournit à l'organe de synthèse pour que celui-ci applique des transformations élémentaires à l'enchaînement des unités segmentales choisies en parallèle. Ainsi la phrase synthétisée présente une courbe intonative semblable à celle qu'aurait pu produire le locuteur lui-même.

Chapitre 6

Résultats et discussion

Le but de mon stage était principalement d'automatiser le processus de segmentation et d'alignement. Nous avons eu également le besoin d'unifier les différents outils à notre disposition afin de faciliter à la fois le pré-traitement et l'importation des diphones dans la base de données.
Mon travail s'est donc échelonné en plusieurs étapes :

- Tout d'abord, la nécessité de pouvoir, en parallèle avec EULER (qui nous fournit la description phonétique et prosodique du texte), segmenter un fichier audio en autant de fichiers qu'il y a de phrases dans le texte. Pour ce faire, j'ai utilisé un programme en C qui permet dans un fichier audio de repérer les silences dans celui-ci. Ce programme renvoit à l'utilisateur les positions dans l'audio des silences et de leur durées. Plusieurs paramètres rentrent alors en ligne de compte : la durée du silence et le volume du silence. La durée du silence en premier lieu est primordiale et afin d'obtenir un découpage le plus précis possible nous avons procédé à un enregistrement en chambre anéchoïque du texte à segmenter en marquant des pauses entre chaque phrase. Ainsi, en fixant une durée de silence au programme (typiquement 0,5 s) celui-ci nous repère précisément tous les silences supérieurs à cette valeur. En ce qui concerne le volume du silence, ce dernier est à fixer empiriquement étant donné qu'il est relatif au niveau de bruit de l'enregistrement considéré. Grâce à la donnée de ses deux facteurs, il devient alors possible de réaliser la segmentation en phrase de manière adéquate. A ce stade du processus, nous disposons donc d'autant de fichiers audios que l'on a de phrases dans le texte.

- Ensuite, il a été nécessaire de constituer le dictionnaire de base (bootstrap) requis par les outils d'alignement afin de synthétiser des phrases cibles permettant d'être comparées avec les phrases sources en vue de la création des marqueurs de diphones. Nous pensions au départ pouvoir éventuellement utiliser le bootstrap déjà implémenté mais il s'est avéré que la comparaison du fichier source avec des diphones provenant d'un autre locuteur (avce des conditions d'enregistrement tres mauvaises : bruit, réverbération ...) ne donnait pas de résultats probant quant à la qualité de l'alignement. Nous avons donc du segmenter manuellement ces diphones dans un fichier

audio que nous avions enregistré dans ce but. Ce fichier contenant au moins une occurence de chaque diphone (papap,popop,...) avec des frontières bien définies nous donnait de bons exemplaires des diphones dont nous avions besoin. Il est à noter que plus le dictionnaire de base est bien réalisé (nombre de diphones présents, plusieurs contextes différents pour chaque diphones ...), plus l'alignement est bon. En effet, lorqu'il manque des diphones, l'alignement pour les diphones précédent et suivant est faussé.

- Nous allons pouvoir désormais décrire de façon plus détaillée les différentes étapes de l'alignement :
  - Création d'un fichier audio par concaténation de diphones : Après avoir extrait du fichier généré par EULER les informations relatives à la phonétisation de la phrase que l'on souhaite aligner, on synthétise cette phrase en mettant bout à bout les diphones présents dans le texte dont on connait les positions temporelles dans le dictionnaire "bootstrap". Ce fichier audio synthétisé "à la volée " nous donne déjà une première idée de la façon dont une synthèse concaténative est réalisée. Cependant, aucune modification prosodique et rythmique n'est appliquée et par conséquent les diphones se suivent sans grande cohérence (variation brutale de rythme ou de fréquence fondamentale).
  - Analyse MFCC des fichiers audios source (la phrase enregistrée) et cible (la phrase concaténée). Le programme nous renvoie pour la phrase considérée un vecteur contenant les valeurs des coefficients cepstraux source et cible.
  - Mise en place de l'algorithme de DTW (Dynamic Time Warping) de la manière suivante : à partir des coefficients cepstraux obtenus, on calcule la distance quadratique entre chaque position de la source et de la cible. On obtient alors une matrice de distances locales dont les valeurs (s_i, c_j) sont celles de la distance entre chaque position s_i de la source avec chaque position c_j de la cible. On crée alors un autre matrice qui va nous calculer le chemin le moins coûteux par addition incrémentale de chacune des distances locales. Il ne reste plus alors qu'à décrire le chemin dont le coût est minimal à chaque étape. Cela ne nécessite donc finalement le calcul que du chemin le moins coûteux.(one-pass DTW)
    Le programme nous renvoit alors un vecteur nous donnant la position des horizontales et des verticales (traduisant une accélération ou un ralentissement de la source par rapport a la cible).
  - Il ne reste alors plus qu'à écrire les marqueurs temporels de diphones sur le fichier audio source afin d'obtenir l'alignement.
- A ce stade du processus, les phrases sont segmentées et prêtes à être importées dans la base de données. Toutefois, il est nécessaire d'importer ces diphones dans la base de données accompagnés de leurs descripteurs prosodiques (donnés par EULER) et acoustiques (donnés par CATERPILLAR). Ainsi, les diphones présents dans la nouvelle base de données comporte à la fois des renseignements acoustiques (fréquence fondamentales, moyenne ...) et prosodiques (position dans la phrase, dans le mot ...). Il ne reste alors plus qu'à fixer les coûts de distances acoustiques, nécessaires pour

la sélection du bon diphone (on doit selectionner un “pa” lorsque l’on cherche un “pa”), et les coûts de distances prosodiques (sélectionner parmis tous les “pa” celui qui correspond le mieux de par sa position avec les diphones suivants et précédents permettant par exemple lorsque l’on dispose dans le fichier audio du mot “accapparé” et que l’on souhaite synthétiser le mot “accapparé” que le système sélectionne tous les diphones de ce mot réduisant ainsi les concaténation non adéquates - les fameux clics présents lors d’une telle synthèse).
Lors de l’élaboration du système plusieurs points de faiblesse se sont révélés :

– Premièrement, la présence dans le texte de noms propres est à proscrire car en effet ces mots ne seront pas transcrit correctement par EULER, un nom propre pouvant se prononcer d’une manière complètement différentes des régles usuelles et surtout imprédictible. En outre, les liaisons entre les mots, principalement pour des “s” ou des “t” finaux pose un problème sérieux quant à l’alignement car la trancription phonétique ne sera pas fidèle à l’audio et créera par conséquent des erreurs de marqueurs dans les diphones à synthétiser. Afin de résoudre ce problème, il est nécessaire de modifier la transcription phonétique afin d’obtenir la phonétisation souhaitée.
– Deuxièmement, l’établissement des coûts de distance doit être réglé de manière judicieuse et précise. Une étude plus appronfondie dans ce domaine doit être menée afin d’obtenir des règles d’apprentissage plutôt que de fixer ces valeurs empiriquement (gain de temps).
– De par la structure même du programme d’alignement, il n’est pas encore possible de pouvoir augmenter au fur et à mesure le dictionnaire de bootstrap. Il serait bienvenu que l’on puisse le faire car cela créerait un cercle vertueux : plus on dipose de diphones au départ, meilleur est l’alignement, meilleurs sont les diphones importés ce qui améliore encore l’alignement en terme de justesse.

Chapitre 7

Conclusion

Ce stage m'a apporté beaucoup tant au niveau scientifique que personnel. En effet, j'ai été amené à m'intéresser de près aux différentes méthodes d'analyse de la parole et de ce fait ce stage m'a permis de mettre un pied dans la recherche effectuée dans ce domaine très vaste qui couvre à la fois l'acoustique (pour la compréhension de sa production), la perception (pour la partie de modèles prosodiques) et le traitement du signal (pour l'analyse et la synthèse du signal de la parole). En outre, bien que les recherches dans ce domaine soient anciennes, la thématique n'en reste pas moins actuelle tant la caractérisation de la parole est difficile du fait de sa structure individuelle (chaque locuteur possède une voix différente d'un autre).

En ce qui concerne les objectifs de mon stage, à savoir automatiser la segmentation et l'alignement de la parole, celui-ci a été atteint avec toutefois des améliorations a apporter pour l'alignement, principalement du à la nécessité d'avoir un dictionnaire de base fourni et de bonne qualité (ou les diphones sont découpés de manière précise et judicieuse). En outre, une étude plus appronfondie du programme EULER serait souhaitable dans la mesure ou il est difficile de prévoir a priori la gestion des exceptions linguistiques par ce dernier. Il est ainsi possible, moyennant un enregistrement de bonne qualité (avec suppression de noms propres et des liaisons adéquates principalement), de pouvoir segmenter un fichier audio de voix parlée et de pouvoir l'aligner avec les diphones du dictionnaire de base dont la constitution actuelle est de l'ordre de 1100 diphones sur 1444 diphones possibles théoriquement sachant que pour la langue française certains diphones n'apparaîtront jamais.

En outre, les résultats de la sélection d'unités n'ont pas encore donné de résultats probants en terme de synthèse par semi-phones. En effet, du fait que les diphones sont découpés dans la partie stable du signal pour les voyelles et dans la partie bruitée pour les consonnes il est préférable d'utiliser une sélection sur les diphones plutôt que sur des semi-phones car cela provoque une sorte de doublement des voyelles. Il serait intéressant également de pouvoir disposer de triphones afin de réduire les concaténations mais leur nombre élévé (de l'ordre de 50000 pour la langue française) rend leur utilisation fastidieuse et nécessiterait pour cela d'effectuer au préalable une discrimination des tri-phones les plus fréquents afin de réduire le travail de segmentation du dictionnaire de base.

Au delà de cela, ce stage m'a permis d'apprendre le language PERL (très utile pour

la manipulation de chaînes de caractères), de travailler sous l'environnement Linux, particulièrement adapté pour les applications audio et de consolider mes connaissances en MATLAB et en C.

Chapitre 8

Remerciements

Je tiens tout d'abord à remercier Mr Xavier Rodet pour ces conseils utiles et son soutien tout au long de mon stage à l'IRCAM. Je remercie également Mme Nicole Gache pour son suivi au cours de mes six mois de stage de fin d'études. Je remercie Gregory Beller avait qui je travaillais en collaboration sur ce projet et avec qui j'avais de nombreuses discussions sur la manière de mener à bien ce projet en essayant de construire un système le plus cohérent possible. En outre, je remercie particulièrement l'équipe d'Analyse/Synthèse pour son aide et sa disponibilité à chaque fois que j'en avais besoin et notamment Axel Röbel, Geoffroy Peeters, Damien Tardieu et Joseph Escribe pour leurs conseils en C et en Matlab notamment. Mes remerciements vont également à l'ensemble des personnes présentes à l'IRCAM avec qui j'ai pu avoir des discussions instructives et fructueuses sur la musique de manière générale.

Bibliographie

[1] Christophe Blouin. *Sélection des unités pour la synthèse vocale par concaténation*. PhD thesis, Université Paris 11 Orsay, 2003.

[2] Brigitte Zellner Erick Keller. *Les défis actuels en synthèse de la parole*. Etudes de lettres, 1998.

[3] James L. Flanagan. *Speech Analysis Synthesis and Perception*. Springer-Verlag, 1972.

[4] Hélène François. *Synthèse de la parole par concaténation d'unités acoustiques : construction et exploitation d'une base de parole continue*. PhD thesis, Université de Rennes 1, 2002.

[5] Orsten Kärki. Rapport de stage, système talkapillar, 2003.

[6] Biing-Hwang Juang Lawrence Rabiner. *Fundamentals of Speech Recognition*. Prentice Hall PTR, 1993.

[7] Pierre-Yves Le Meur. *Synthèse de la parole par unités de taille variable*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, 1996.

[8] Hervé Bourlard René Boite, Thierry Dutoit. *Traitement de la parole*. Presses Polytchniques et Universitaires Romandes, 2000.

[9] Gaël Richard. Traitement de la parole, 2003.

[10] Diemo Schwarz. *Data-driven concatenative sound synthesis*. PhD thesis, Université Paris 6, 2004.

[11] Calliope Tubach J.P. *La parole et son traitement automatique*. Masson, 1989.

Table des matières