Rapport de stage de master II recherche

Segmentation non supervisée d'un flux de parole en syllabes

Stage effectué à l'IRCAM
Institut de Recherche et Coordination Acoustique/Musique
Equipe Analyse/Synthèse

Auteur :
François Lamare

Encadrant :
Nicolas Obin

Responsable pédagogique :
Jean-Dominique Polack

du 01 Mars au 31 Juillet 2012

«Presque tous les hommes sont esclaves faute de savoir prononcer la syllabe : non.»

Sébastien-Roch Nicolas de Chamfort, dit Chamfort,

Maximes et pensées, caractères et anecdotes.

Table des matières

Résumé

L'objectif du stage a été de réaliser une nouvelle méthode de segmentation d'un flux de parole en syllabes. La finalité d'un tel travail serait de pouvoir identifier distinctement les syllabes qui composent ce flux de parole. En effet, la syllabe est l'unité de base de la prosodie. Les identifier permettrait donc de modifier certaines caractéristiques prosodiques de la voix (la fréquence fondamentale par exemple) à une échelle appropriée, ou bien d'intégrer ces caractéristiques dans des systèmes de synthèse de la parole.

Le travail accompli a abouti à deux méthodes de segmentation syllabique non-supervisées. La pierre angulaire de ces méthodes est l'application de critères de voisement, fondés sur l'entropie de Rényi ou une mesure de VUV, à une représentation temps-fréquence multi-bandes d'un signal de parole. L'entropie de Rényi, généralisation de l'entropie de Shannon, permet de quantifier le degré d'organisation du signal. Nous partons alors de l'hypothèse que ce degré d'organisation diffère selon que l'on considère un segment de parole ou un segment sans parole. Le VUV est une autre mesure du degré de voisement dans un signal. L'intérêt d'une telle approche réside dans le fait que l'on peut écarter les trames ou les bandes fréquentielles non pertinentes pour la segmentation en syllabes.

Les performances de segmentation des deux méthodes proposées ont été évaluées et comparées à celles de méthodes déjà existantes de l'état de l'art. L'évaluation s'est faite sur des critères de bonne segmentation, de taux d'insertions/omissions de syllabes et de F-measure, en comparant la segmentation obtenue par l'une des méthodes à une segmentation manuelle de référence.

Les premiers résultats semblent montrer que l'approche retenue, fondée essentiellement sur une analyse multi-bandes à laquelle on applique un critère de voisement, est vraiment pertinente pour la segmentation syllabique, et qu'elle devrait être approfondie.

Remerciements

Je tiens à remercier Axel Roebel pour son accueil au sein de l'équipe Analyse/Synthèse de l'IRCAM.

Je voudrais également remercier mon tuteur Nicolas Obin pour son encadrement, ses remarques, ses conseils, sa gentillesse et la confiance dont il a fait preuve vis-à-vis de mon travail. Il m'a donné l'occasion de travailler avec autonomie, ce dont je lui suis reconnaissant. Le stage qu'il m'a permis d'effectuer a été très riche sur les plans intellectuel et scientifique.

Je remercie de même Marco Liuni pour le temps qu'il m'a consacré et l'aide qu'il m'a apporté sur certains points théoriques. Je remercie aussi Arnaud Dessein qui a eu la gentillesse de me présenter une partie de ses travaux.

Je remercie également les autres membres de l'équipe Analyse/Synthèse, puis le personnel de l'IRCAM pour leur sympathie. Ils ont tous permis à ce stage de se dérouler dans d'agréables conditions.

Enfin, j'adresse des remerciements particuliers à mes collèges stagiaires, qui m'ont transmis leur bonne humeur, et dont les échanges ont été particulièrement enrichissants.

Chapitre 1

Introduction

Les travaux autour de la voix constituent un domaine majeur à l'IRCAM, tant au sein de l'équipe Analyse-Synthèse qu'en production musicale. Dans ce cadre, l'IRCAM s'investit dans de nombreux projets de recherche sur le développement de nouvelles technologies vocales (transformation de la voix, synthèse de la parole à partir du texte, conversion d'identité de la voix, transformation de la voix parlée en voix chantée). Dans ce contexte, la segmentation de la parole constitue la pierre angulaire de l'ensemble de ces technologies pour déterminer des transformations/modélisations en fonction de segments spécifiques, comme par exemple des segments de nature linguistique (phonème, syllabe, mot, etc...).

Deux problématiques importantes en traitement de la parole peuvent être identifiées : la modification de la prosodie de la voix et l'intégration de caractéristiques prosodiques (contours prosodiques, fréquence fondamentale, débit de parole) dans les systèmes de reconnaissance ou de synthèse de la parole. La syllabe constitue justement le segment de base de la prosodie. L'intérêt de la segmentation syllabique parait dès lors évident. La segmentation syllabique permettrait d'accéder au niveau le plus fin de la prosodie ([Obin, 2011]).

De plus, le segment syllabique est commun à un grand nombre de langues, comme les langues romanes ou anglo-saxonnes par exemple. Ainsi, la segmentation syllabique possède un caractère universelle que ne possède pas d'autres types de segmentation, comme la segmentation en mots ou celle en phonèmes. Ces dernières nécessitent en effet la mise en œuvre extrêmement lourde de modèles linguistiques complexes, et bien sûr très dépendants de la langue considérée. Par conséquent, cette caractéristique rend la segmentation syllabique généralisable à un grand nombre de langues.

Enfin, de par sa structure phonétique, la syllabe peut être comparée à des notes de musique, avec une attaque, une partie soutenue (noyau vocalique) et une phase de relâche ou de transition (coda). La syllabe est donc en quelque sorte l'élément musical de base de la voix, qu'elle soit parlée ou chantée. La segmentation syllabique trouve donc des applications musicales directes en production artistique.

Les méthodes de segmentation en syllabe actuelles reposent pour la plupart sur des méthodes signal et l'utilisation de connaissances expertes en parole (onset detection, détection des noyaux vocaliques). Malheureusement, ces méthodes ne sont pas suffisamment robustes pour une intégration à des technologies existantes.

L'objectif principal du stage est de développer une méthode de segmentation d'un flux de parole en syllabes. Cette méthode est non-supervisée pour s'émanciper des contraintes linguistiques habituellement rencontrées en reconnaissance de la parole. Elle devra éga-

lement répondre à des contraintes de robustesse (segmentation de parole spontanée en milieu bruité). Enfin, elle devra être idéalement déclinable en une version temps-réel pour les applications artistiques de l'IRCAM.

Ce rapport s'organise de la façon suivante. Tout d'abord, quelques notions théoriques nécessaires à la bonne compréhension du sujet seront introduites. Un état de l'art non exhaustif des méthodes de segmentation (segmentation de la parole en générale puis segmentation syllabique) sera présenté. Quelques généralités sur la segmentation syllabique seront dès lors mentionnées. Ensuite, les différents travaux réalisés seront décrits. Ces travaux reposent essentiellement sur l'application de critères de voisement, fondés sur l'entropie de Rényi ou le VUV comme nous le préciserons, à une représentation temps-fréquence multi-bandes d'un signal de parole, et cela dans l'objectif d'écarter les trames ou bandes fréquentielles non utiles pour la segmentation syllabique. Les performances de segmentation des méthodes proposées seront exposées, puis comparées à des méthodes de référence. Ce sera alors l'occasion de soulever les difficultés liées à telle méthode ou tel procédé, les améliorations possibles qui pourraient êtres apportées et les pistes qui pourraient être suivies à l'avenir.

Chapitre 2

Notions théoriques et état de l'art

2.1 Définition de la syllabe

La syllabe constitue l'une des unités phonologiques du langage oral, au même titre que le phonème et le mot. Fonctionnellement, la syllabe est largement considérée comme le segment élémentaire de la prosodie - "la musique de la parole". Dans ce cadre, il existe plusieurs façons de définir une syllabe ([Clements and Keyser, 1983] et [Hall, 2006]), suivant le point de vue adopté :

Linguistiquement : une syllabe est constituée d'une et une seule voyelle (le noyau vocalique) autour de laquelle des consonnes sont agrégées (onset et coda).

Physiquement : une syllabe se caractérise par la réalisation d'une cible articulatoire stable (la voyelle), délimitée à gauche et à droite par un relâchement musculaire (articulation et détachement de la syllabe).

Acoustiquement : le noyau vocalique présente un maximum de sonorité, généralement défini en terme d'intensité et de stabilité, tandis que les limites présentent des minima de sonorité.

Une syllabe est formée par un regroupement de plusieurs phonèmes. La phonotactique est la branche de la phonologie linguistique qui étudie la manière dont s'agrègent les phonèmes pour constituer des unités linguistiques de dimensions supérieures (la syllabe, le mot, etc...). La constitution de syllabes se réalise suivant un principe universel de "sonorité", qui se quantifie à partir de quantité mesurables, telles quel l'intensité acoustique ou la quantité d'air extrait des poumons. Elle est croissante jusqu'à un premier pic de sonorité qui représente le noyau vocalique (une voyelle), puis décroit jusqu'à la fin de la syllabe (voir figure 2.1). Le principe de sonorité est universel. Il s'applique à toutes les langues syllabiques du monde, bien que des exceptions puissent exister, comme la règle du "e" muet (le schwa) en français.

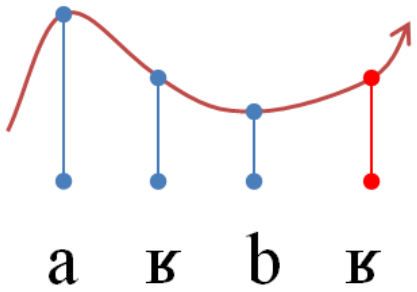


FIGURE 2.1 – Courbe de sonorité d'une syllabe

Dans le contexte de la segmentation d'un flux de parole en syllabe, nous retiendrons principalement la définition acoustique de la syllabe. Avant d'apporter des éléments de réponse au problème spécifique de la segmentation en syllabes, il est pertinent d'introduire le problème plus général de la segmentation de la parole.

2.2 Etat de l'art en segmentation de la parole

2.2.1 Segmentation de la parole

Généralités :

La segmentation de la parole fait référence à des unités variées selon la nature du segment considéré. On peut définir plusieurs types de segmentation (organisés du segment le plus court au segment le plus long) :

- en voisé/non-voisé ;
- en phonèmes ;
- en syllabes ;
- en mots ;
- en groupes inter-pausaux (segments délimités par deux pauses silencieuses) ;
- en locuteurs et tours de parole.

Les méthodes de reconnaissance automatique de la parole ("speech detection", "voice activity detection" ou "endpoint detection") se divisent en deux catégories : les algorithmes supervisées et les algorithmes non-supervisées. Pour les algorithmes supervisés, les paramètres d'un modèle statistique de segmentation sont déterminés (phase d'apprentissage) à partir de bases de données segmentées au préalable. Les méthodes supervisées de segmentation de la parole reposent majoritairement sur des chaines de Markov cachées (HMMs) (HTK -[Young, 1993] et SPHINX - [Lee, 1989], sont les deux outils de référence en segmentation de la parole). En outre, ces méthodes nécessitent un ensemble de connaissances linguistiques (NLPs, soit Natural Language Processing). C'est le cas par exemple de la segmentation phonétique. Malheureusement, les NLPs sont extrêmement dépendants de la langue, ce qui rend le développement d'un système de segmentation "universel" c'est-à-dire indépendant de la langue, très fastidieux.

Dans ce contexte, la segmentation en syllabes pourrait être vue comme une extension des méthodes de segmentation en phonèmes (HMMs) aux segments syllabiques. Cependant,

trois problèmes majeurs se posent :

1. le problème linguistique déjà mentionné.
2. un problème combinatoire : le nombre de modèles statistiques ou de paramètres est déjà très grand pour la segmentation en phonèmes. En effet, pour prendre en compte les phénomènes de coarticulation de la parole, chaque phonème est modélisé par une chaine de Markov cachée "en contexte", c'est à dire en fonction de la nature des phonèmes précédents et suivants. Par exemple, la langue française est composée de 36 phonèmes. Ainsi, une modélisation en contexte d'ordre 3 nécessite l'estimation des paramètres de 36^3 modèles (bien que des techniques d'agrégation ("clustering") permettent de réduire le nombre de modèles à estimer) . Dans le cas de la segmentation syllabique, ce nombre déjà très élevé exploserait.
3. le problème du temps-réel, généralement incompatible avec le paradigme Markovien.

Ces considérations générales nous ont amené alors à considérer seulement des méthodes non-supervisées.

Descripteurs :

Les méthodes de segmentation non-supervisées consistent à segmenter un flux de parole (ou audio) à partir de techniques de segmentation paramétrique ou ad-hoc (mesures de stationnarité ou détection de maxima/minima/discontinuités) à partir d'une représentation acoustique du signal. Dans ce cadre, la représentation acoustique utilisée doit être la plus pertinente possible pour faciliter la segmentation. Par exemple, on peut suivre l'évolution temporelle d'un descripteur du signal. Si la valeur de ce descripteur dépasse un certain seuil, alors de la parole est détectée. Les méthodes non-supervisées de segmentation de la parole présentent en outre l'avantage d'être universelles et sans contrainte a priori de temps-réel.

Un grand nombre de représentations acoustiques ont été utilisées jusqu'à présent. Une famille importante de méthodes de segmentation contient les méthodes basées sur le Zero-Crossing Rate (ZCR), l'énergie à court-terme, où sur l'énergie spectrale ([Savoji, 1989] [Ney, 1981][Gerven and Xie, 1997][Huang and Yang, 2000]). Ces méthodes, en plus d'être assez peu robustes en présence de bruit, ne prennent pas suffisamment en compte les spécificités d'un signal de parole. On peut également avancer la même critique de non robustesse face au bruit pour les descripteurs fondés sur les Coefficients de Prédiction Linéaire (LPCs). Des méthodes d'estimation de la fréquence fondamentale ont également été envisagées, dans la mesure où cette information peut aider directement à mettre en évidence la différence entre segments voisés et non-voisés. Encore une fois, cette estimation de la fréquence fondamentale est délicate en milieu bruité, surtout si les bruits sont de nature harmonique. Les coefficients cepstraux en échelle fréquentielle Mel (MFCCs) sont actuellement la base en reconnaissance de la parole. Plus récemment,([Wang et al., 2011]) propose une méthode de détection d'activité vocale basée sur le calcul d'une distance entre un vecteur MFCC d'une trame de parole et un vecteur MFCC d'une trame de bruit de fond.

2.2.2 Segmentation syllabique

La segmentation en syllabes à proprement parler est un sujet qui a été assez peu abordé pour le moment. Dans [Mermelstein, 1975], Melmerstein a proposé un algorithme

de segmentation de syllabes basée sur l'enveloppe convexe du signal, mesurée sur l'intensité acoustique dans les régions formantiques de la parole (fréquences inférieures à 4 Khz), ou plus précisément de son équivalent perceptif, la sonie [Robinson and Dadson, 1956]. De part la définition dite "acoustique" des syllabes, un noyau vocalique aura une valeur maximale de sonie tandis que les frontières de la syllabe auront des valeurs minimales. Ce paradigme de segmentation demeure encore aujourd'hui une référence, et la plupart des méthodes proposées depuis ne constituent que des variations autour de ce paradigme. Des critères empiriques sont ensuite utilisés pour réaliser la segmentation à partir de la sonie. [Howitt, 1999] et [Villing et al., 2004] présentent des méthodes plus récentes de segmentation syllabique également basées sur la sonie. Une mesure de la périodicité du signal a été introduite pour supprimer les variations d'intensité non-pertinentes pour la segmentation (typiquement, les consonnes fricatives et les plosives). Dans [Xie and Niyogi, 2006] et [De Jong and Wempe, 2009] par exemple, la périodicité est mesurée grâce à la fonction d'autocorrélation. Dans [Villing et al., 2006], Villing montre les limitations des méthodes de segmentation syllabique basées sur la loudness ou l'énergie seules. D'autres critères doivent alors être pris en compte. Les auteurs de [Wang and Narayanan, 2007] exposent une méthode de détection de syllabes pour la mesure de débit de parole (exprimée en syllabe par seconde). Cette méthode repose principalement sur la corrélation croisée de bandes de fréquences d'une analyse temps-fréquence du signal de parole (voir partie 3.5.2), et cela dans le but d'exploiter la structure formantique des noyaux vocaliques.

2.3 Considérations sur la segmentation en syllabes

2.3.1 Phonétique

Les phonèmes se regroupent en différentes classes :

- voyelles ;
- plosives (voisées/non-voisées) ;
- fricatives (voisées/non-voisées) ;
- liquides ;
- glides.

Cette catégorisation se fait selon des critères d'excitation du conduit vocal, d'articulation (mode et lieu d'articulation), de nasalité, de forme des lèvres lors de la prononciation (pour plus de précisions, consulter [1]).

L'annexe partie 5 présente la liste des phonèmes pour le français moderne.

Le coeur de la segmentation en syllabe nécessite en particulier l'identification des noyaux vocaliques (les voyelles). Plusieurs problèmes de nature phonétique font obstacle à cette identification :

1. Les consonnes non-voisées et en particulier les plosives non-voisées (/p/, /t/, /k/) peuvent introduire des maxima d'intensité indésirables. Dans une moindre mesure, les fricatives non-voisées (/f/, /s/, /S/) introduisent également du bruit.
2. Les consonnes voisées, plus précisément les plosives voisées (/b/, /d/, /g/), les liquides (/l/), et dans une moindre mesure les fricatives voisées (/v/, /Z/, /z/), peuvent être confondues avec des noyaux vocaliques.

1. http ://fr.wikipedia.org/wiki/Phonétique_articulatoire

3. Les voyelles muettes, c'est à dire faiblement intenses et peu voisées (exemple du schwa en français) peuvent ne pas être identifiées.
4. la succession de deux voyelles appartenant à deux syllabes distinctes peut poser problème. Ici, c'est le modèle théorique de sonorité qui est remis en cause dans la mesure où aucun minimum de sonorité ne peut être observé entre les 2 voyelles, surtout lorsqu'elles sont co-articulées.

En revanche, la succession de voyelles (glide+voyelle) au sein d'une syllabe ne pose aucune problème.

2.3.2 Phonation

La segmentation syllabique peut s'avérer très difficile pour des voix extrêmement bruitées, comme c'est le cas dans la plupart des registres de qualité vocale (soufflée, craquée, ...), dans les registres extrêmes d'effort vocal (chuchotée, criée), ou bien encore pour des débits de parole très rapides.

2.3.3 Condition d'enregistrement et parole spontanée

La segmentation en syllabe se confronte à la robustesse au bruit environnant, typiquement dans des enregistrements de parole spontanée qui peuvent présenter une très large variété de bruits nuisible pour la segmentation. La présence de voyelles très courtes et/ou très partiellement réalisées est également une source de problème en segmentation syllabique.

2.4 Mesures d'entropie

Un certain nombre de descripteurs basés sur l'entropie de Shannon ont été proposées en segmentation de la parole, mais pas dans le cadre de la segmentation en syllabes. De part ses propriétés, que nous allons expliciter par la suite, nous l'avons jugée très intéressante pour la segmentation syllabique. Par exemple, les mesures d'entropie permettent de donner une représentation du degré d'organisation de la parole - maximum dans le noyau vocalique et minimum aux frontières de la syllabe. Dans le même sens, l'entropie de Rényi, qui est une généralisation de l'entropie de Shannon, va être étudiée dans le cadre de la segmentation syllabique.

2.4.1 Entropie de Shannon

L'entropie, introduite par Claude Shannon dans sa théorie de l'information ([Shannon, 1948]), permet de mesurer la quantité d'information contenue dans un signal aléatoire. Autrement dit, elle mesure le degré d'organisation (l'incertitude) dans ce signal. Elle est définit de la manière suivante :

$$H(x) = \sum_k P(x_k) \cdot \log_2[P(x_k)] \tag{2.1}$$

où $x = \{x_k\}_{0 \leq k \leq N-1}$ est une série temporelle, fréquentielle ou autre, et où $P(x_k)$ est la probabilité d'un certain état x_k.

[Shen et al., 1998] ont été parmi les premiers à utiliser l'entropie dans le cadre de la détection de la parole. Leur étude a montré que l'entropie d'un segment de parole différait significativement d'un segment sans parole. En effet, la structure harmonique de la voix reflète un degré d'organisation qui ne se retrouve pas dans la structure spectrale d'un segment de silence. Plus exactement, l'entropie de Shannon permet de quantifier la "pointicité" du spectre. Dans le cas d'un segment de parole, le spectre est fortement piqué (présence d'un pic principal dû aux résonnances formantiques). Si l'on considère le spectre comme une densité de probabilité, l'allure de la densité de probabilité correspondante traduit un état de certitude élevé (autrement dit une zone de fréquences privilégiée). L'entropie sera dès lors très faible. Inversement, dans le cas d'un bruit blanc par exemple, le spectre est uniforme, ce qui traduit cette fois ci une incertitude totale. L'entropie sera donc très élevée.

2.4.2 Entropie spectrale

La structure harmonique d'un segment de parole n'apparait que dans le spectrogramme, ce qui nécessite par conséquent d'introduire l'entropie spectrale. Toutes les méthodes de segmentation de parole à partir de l'entropie sont en fait plutôt basées sur celle-ci. L'entropie spectrale est défini à partir de la transformée de Fourier à court-terme du signal :

$$S(k,l) = \sum_{n=1}^{N} h(n)s(n-l) \cdot \exp\left(-j2\pi kn/N\right) \tag{2.2}$$

avec $0 \leq k \leq K-1$ et $K = N$ généralement.

$s(n)$ représente l'amplitude du signal au temps n. $S(k,l)$ représente l'amplitude de la k^{ieme} composante fréquentielle, pour la l^{ieme} trame d'analyse. N est le nombre de points fréquentiels considérés pour le calcul de la transformée de Fourier. $h(n)$ est la fenêtre d'analyse.

On peut ensuite définir l'énergie spectrale de la manière suivante :

$$S_{energy}(k,l) = |S(k,l)|^2$$

avec $1 \leq k \leq K/2$.

La probabilité associée à chaque composante spectrale est obtenue en normalisant de la manière suivante :

$$P(k,l) = \frac{S_{energy}(k,l)}{\sum_{i=1}^{N/2} S_{energy}(i,l)}$$

avec $1 \leq i \leq N/2$ et $\sum_i P(i,l) = 1$ pour tout l.

Ce qui mène à l'entropie spectrale pour une certaine trame l :

$$H(l) = \sum_{i=1}^{N/2} P(i,l) \cdot \log_2[P(i,l)] \tag{2.3}$$

Il est souvent commode de considérer l'opposé de l'entropie pour obtenir des profils analogues à ceux de l'intensité.

[Shen et al., 1998] ont montré que l'entropie spectrale outrepassait les performances de détection des algorithmes basées sur l'énergie. En réalité, l'entropie est plus robuste dans le cas de bruits non-stationnaires tandis que l'énergie, de part ses propriétés additives, est robuste face à des bruits stationnaires. L'énergie d'un signal de parole bruité est forcement plus grande que l'énergie du bruit seul (du moins jusqu'à un certain point). A partir de ce constat, [Huang and Yang, 2000] ont proposés un descripteur basé à la fois sur l'énergie et l'entropie, qui permet de ne conserver seulement que les avantages des deux méthodes. Ouzounov dans [Ouzounov, 2005] réalise une étude comparative de plusieurs descripteurs basés sur l'entropie. Une des limitations de l'entropie est que l'entropie d'une trame sans parole bruitée par un bruit blanc coloré peut être équivalente à celle d'une trame avec parole et également bruitée. Une solution est alors proposée dans [Ouzounov, 2005] pour rendre l'entropie plus robuste aux bruits colorés.

D'autres méthodes qui reposent sur une analyse multi-bandes, consistent à calculer une entropie spectrale non plus sur l'ensemble des points fréquentiels d'une trame d'analyse, mais sur des bandes de fréquences intégrées, en ne considérant que les sous-bandes non-bruitées. Cette stratégie a été employée dans l'algorithme ABSE (Adaptive Band-Partitioning Spectral Entropy), présenté dans [Wu and Wang, 2005]. Dans cette méthode, une estimation du niveau de bruit dans chaque bande est réalisée, ce qui permet de choisir de façon adaptative les bandes utiles pour le calcul de l'entropie, c'est-à-dire les bandes non contaminées par le bruit. Les méthodes décrites dans [Liu et al., 2008] et [Jin and Cheng, 2010], basées sur l'ABSE, proposent en plus une phase de suppression du bruit par soustraction spectrale ([Boll, 1979]). Les performances sont ainsi améliorées dans le cas des faibles rapports signal-à-bruit.

2.4.3 Entropie de Rényi

Soit p une densité de probabilité finie, et $\alpha \geq 0 \in \mathbb{R}$ avec $\alpha \neq 1$. L'entropie de Rényi est définie de la manière suivante ([Rényi, 1961]) :

$$H_\alpha(p) = \frac{1}{1-\alpha} \cdot log_2 \sum_k p^\alpha(k) \tag{2.4}$$

Quand $\alpha \to 1$, on se trouve dans le cas particulier de l'entropie de Shannon. Les principales propriétés de l'entropie de Rényi sont les suivantes :

- $H_\alpha(p)$ est une fonction non croissante de α, d'où $\alpha_1 < \alpha_2 \to H_{\alpha_1}(p) \geq H_{\alpha_2}(p)$
- $H_0(p) > H_\alpha(p)$ pour tout $\alpha \neq 0$
- $H_\alpha(p)$ est maximale quand p est uniformément distribué et minimale quand p a une seule valeur non nulle

Dans [Liuni et al., 2011], Liuni met en évidence l'effet qu'a la valeur du paramètre α sur le calcul de l'entropie spectrale. Ils considèrent pour cela un modèle simplifié de spectrogramme. Ils montrent que prendre des valeurs élevées du coefficient α augmente la différence entre l'entropie d'une distribution de probabilité pointue et l'entropie d'une distribution uniforme (bruit blanc dans le cas extrême).

Théoriquement, il peut donc être judicieux de considérer à la place de l'entropie de Shannon l'entropie de Rényi avec un coefficient α assez grand, de telle manière à amplifier les différences entre cas voisé et non-voisé.

2.4.4 Intérêts des mesures d'entropie en segmentation syllabique

A partir de cette section, on prendra toujours l'opposé de l'entropie (entropie maximale pour une distribution avec un seul pic et minimale pour une distribution uniforme).

D'après les propriétés de l'entropie énoncées ci-dessus, l'entropie d'un segment voisé (voyelle par exemple) différera de celle d'un segment non voisé (plosives, fricatives, silence). Nous avons vu que le noyau de la syllabe était caractérisé par un maximum d'énergie. ***Une des idées principales du stage est donc d'utiliser l'entropie pour mesurer le degré de voisement d'une trame. Ainsi, on pourra écarter les pics d'énergie qui ne correspondent pas à un noyau vocalique (une voyelle).*** Nous détaillerons la méthode mise en œuvre dans la partie 3 .

Considérons la phrase suivante : *"Ca me sert à être riche."* . Le signal et le spectrogramme associé sont visibles figure 2.2.

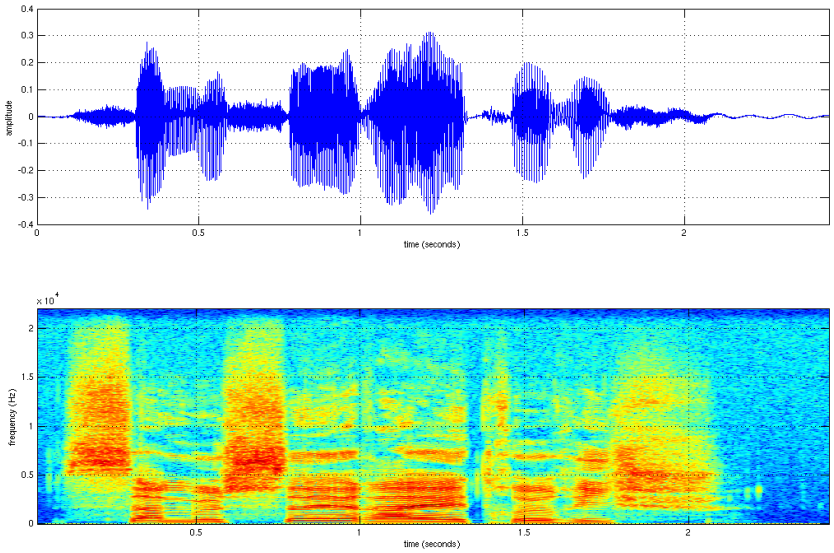


FIGURE 2.2 – Signal et spectrogramme correspondant

Représentons à présent son énergie spectrale et son entropie spectrale de Rényi pour $\alpha = 4$ (figure 2.3) et $\alpha = 0.01$ (figure 2.4). On observe que l'entropie des zones non voisées est significativement plus faible que celle des zones voisées. On remarque également que pour les fricatives (*"/s/"* par exemple), l'entropie de Rényi est plus faible que l'énergie spectrale. On voit aussi que contrairement à ce qui a été dis dans la partie 2.4.3, il est préférable de prendre une valeur de α petite car les profils entropiques sont plus stables dans ce cas particulier et que le contraste entre segments voisés et non voisés est meilleur. En effet, nous avons émis l'hypothèse qu'une valeur élevée de α augmentait la différence d'entropie entre voisé et non-voisé. Cependant, cette différenciation se fait au détriment de la stabilité au bruit. Le fait de prendre une valeur de α grande amplifie certes le pic fréquentiel correspondant à la première région formantique, mais amplifie aussi les pics fréquentiels liés au bruit (bruit de fond, etc...), ce qui peut être très problématique si le bruit en question est coloré. Cette propriété a par ailleurs été également indiquée dans [Obin and Liuni, 2012] pour mesurer le degré de bruit/voisement en reconnaissance automatique de la parole.

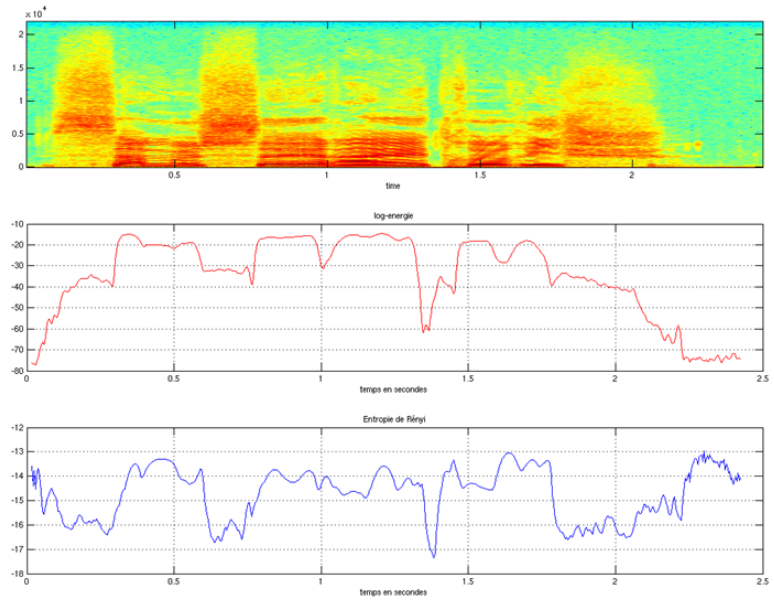


FIGURE 2.3 – Energie spectrale et entropie de Rényi pour alpha = 4.0

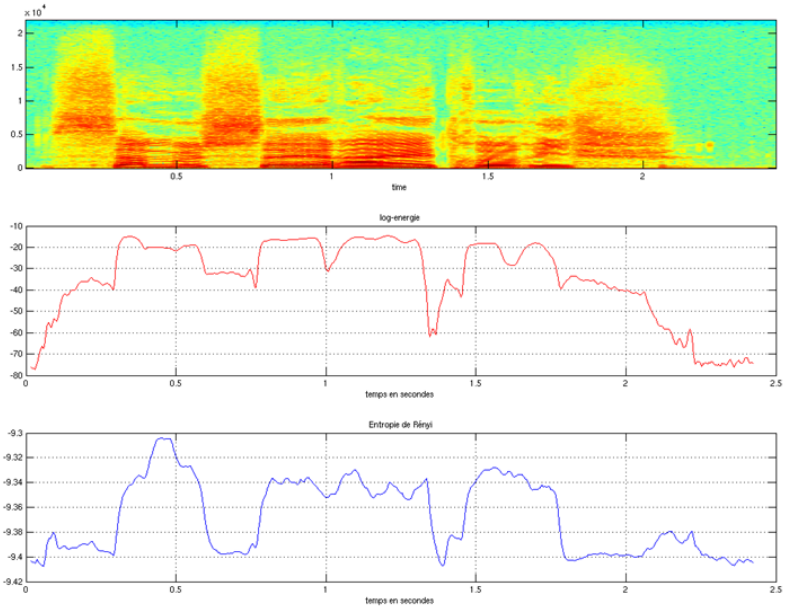


FIGURE 2.4 – Energie spectrale et entropie de Rényi pour alpha = 0.01

2.5 VUV (Voiced/Unvoiced decision)

Le VUV (soft Voiced/Unvoiced measure) est une mesure du degré de voisement contenu dans un signal (ou alternativement de son caractère bruité)([Griffin and Lim, 1985]). Le VUV est définit comme le rapport de l'énergie expliquée par les composantes sinusoïdales sur l'énergie totale d'une trame de signal. Autrement dit, il compare le poids des harmoniques du spectre parmi tous les points fréquentiels de ce spectre. Considérons une analyse multi-bandes d'un signal de parole. Le VUV pour chaque bande de fréquence se mesure à

partir de la formule suivante ([Obin, 2012]) :

$$VUV^{(i)} = \frac{\sum_{k=1}^{K^{(i)}} |A_H(k)|^2}{\sum_{n=1}^{N^{(i)}} |A(n)|^2} \tag{2.5}$$

où i fait référence à l'indice d'une bande de fréquence, où $A_H(k)$ est l'amplitude du k^{ieme} harmonique et où $A(n)$ est l'amplitude du n^{ieme} bin fréquentiel dans la bande considérée. Ainsi, $VUV^{(i)} = 0$ quand la bande i n'a pas de contenu harmonique et $VUV^{(i)} = 1$ quand la bande i est totalement harmonique.

Considérons le signal de parole dont le spectrogramme est représenté figure 2.5. Calculons la VUV par bande de Mel pour ce signal de parole. On prend $N_b = 40$ bandes Mel entre 0 Hz et $f_e/2$ Hz, où f_e est la fréquence d'échantillonnage. Le résultat obtenu est visible figure 2.6(a). Il est utile d'appliquer un filtrage médian 2D sur l'image ainsi obtenu (filtrage 2D avec un voisinage de 5 trames temporelles $\times$ 5 bandes fréquentielles).

Dans le même état d'esprit que l'entropie, on pourra utiliser le VUV pour établir un critère de voisement pour chaque trame, et ainsi écarter les trames qui correspondent à des pics d'énergie non voisés.

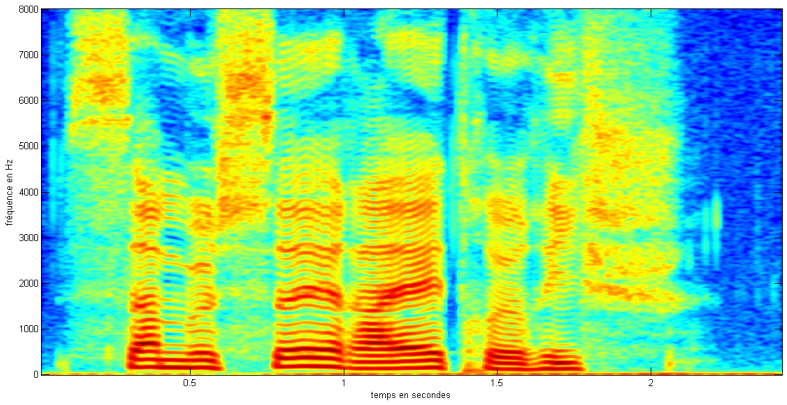
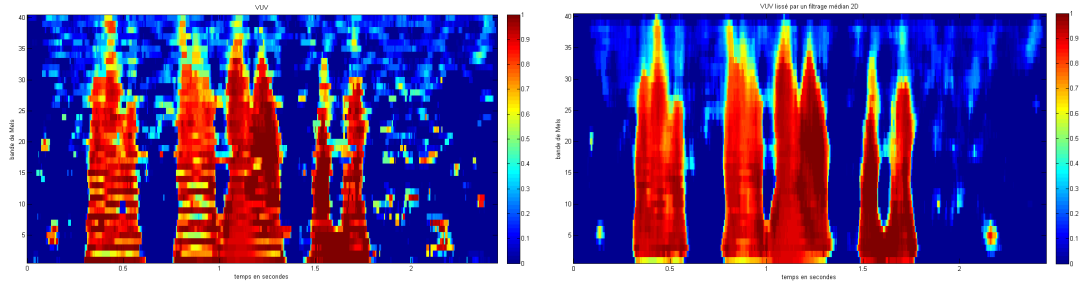


FIGURE 2.5 – Spectrogramme de la phrase : *"Ca me sert à être riche."*



(a) VUV avant filtrage

(b) VUV après filtrage

FIGURE 2.6 – Mesure de voisement par bande de fréquence pour la phrase : *"Ca me sert à être riche."*

Chapitre 3

Présentation des travaux réalisés

L'idée principale de cette étude est de déterminer des profils acoustiques dans un certain nombre de régions fréquentielles, puis de sélectionner/combiner automatiquement les profils utiles pour la segmentation syllabique. La principale contribution de cette étude repose sur l'introduction de 2 nouveaux paradigmes en segmentation syllabique :

1. une analyse fréquentielle multi-bandes (décomposition du spectre de puissance en bandes fréquentielles) ;
2. la combinaison de l'intensité et du degré d'organisation du signal, soit pour déterminer les trames de signal utiles pour la segmentation, soit pour sélectionner les régions fréquentielles utiles pour la segmentation.

L'algorithme de segmentation repose en outre sur un certain nombre de pré et de post-traitements pour optimiser la segmentation en syllabe :

pré-traitement : un algorithme de détection d'activité vocale est utilisé pour segmenter le flux de parole en "groupes de souffle", au sein desquels les syllabes seront identifiées.

post-traitement : deux méthodes de corrélation temporelle et spectrale sont utilisées pour tirer partie de l'organisation temporelle et spectrale des syllabes.

Deux méthodes de segmentation syllabique ont été développées pendant le stage. La première repose principalement sur l'entropie de Rényi et la seconde sur la mesure de VUV. Les deux algorithmes nécessitent une analyse à court-terme du signal. Ils fonctionnent tous deux trame à trame ce qui permettra par la suite une implémentation temps-réel.

On considère encore la même phrase : *"Ca me sert à être riche"*, dont le spectrogramme est visible en figure 3.1.

L'algorithme de référence ([Mermelstein, 1975]) ainsi qu'un algorithme plus récent ([Villing et al., 2004]) ont également été implémentés pour comparaison.

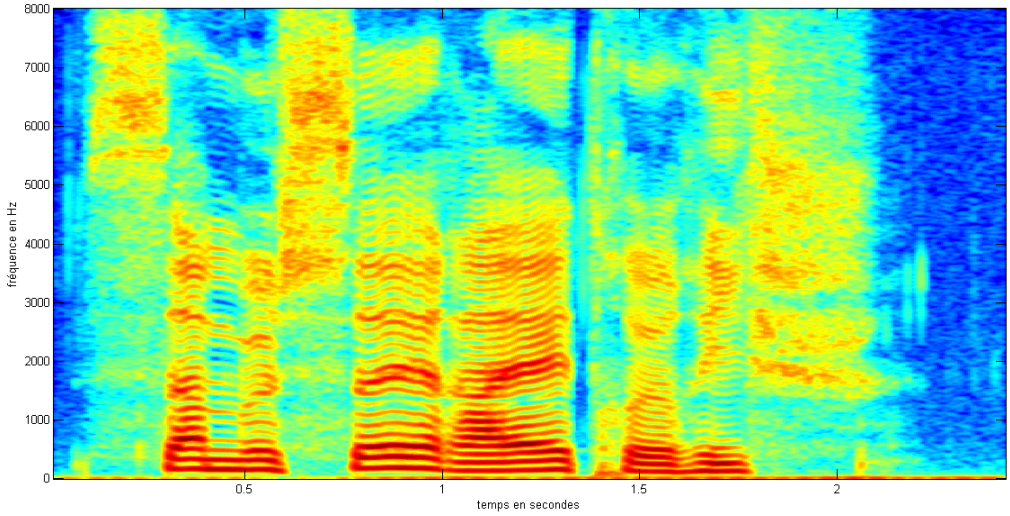


FIGURE 3.1 – Spectrogramme du signal de parole : *"Ca me sert à être riche."*

3.1 Schéma général de l'algorithme

Le schéma général de l'algorithme est présenté en figure 3.2.

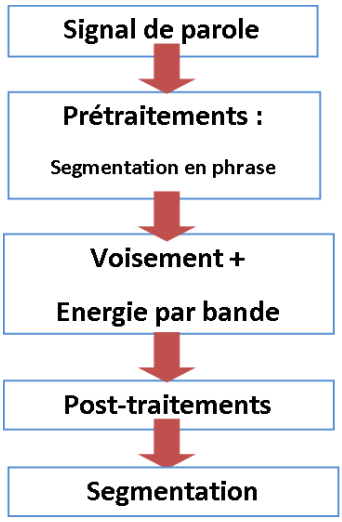


FIGURE 3.2 – Schéma général l'algorithme développé

3.2 Pré-traitement : segmentation en "phrases" (groupes de souffle)

La première étape de l'algorithme consiste à détecter les zones d'activité vocale. Nous avons donc implémenté un algorithme pour segmenter le flux de parole en phrases, dans lesquelles les syllabes seront identifiées par la suite.

L'algorithme retenu ([Wang et al., 2011]) consiste à calculer pour chaque trame une distance entre le vecteur $MFCC$ de cette trame et un vecteur $MFCC_{bruit}$ calculé sur une

portion bruitée du signal (ou bien sur un échantillon représentatif du bruit de fond des fichiers de parole d'un même corpus). Le vecteur $MFCC_{bruit}$ est mis à jour à chaque trame pour prendre mieux en compte les nouvelles spécificités du signal de parole. Si la distance est plus grande qu'un certain seuil, fixe ou adaptatif (voir figure 3.3), alors de la parole est détectée. Les segments de silence dont la durée est inférieure à la durée moyenne d'un silence de plosive (environ $150ms$) sont supprimés. La courbe rouge de la figure 3.4 représente la zone d'activité vocale détecté. L'algorithme est présenté en détail dans [Wang et al., 2011]. La figure 3.5 présente un autre exemple de segmentation de parole pour la phrase : *"La troisième fois, la voici."*.

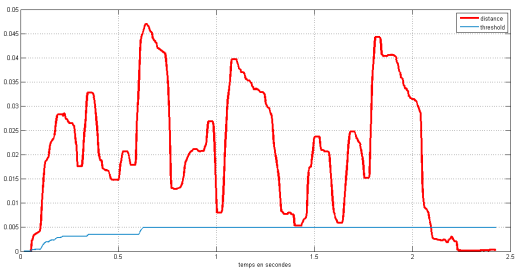


FIGURE 3.3 – Distance de vecteurs MFCC

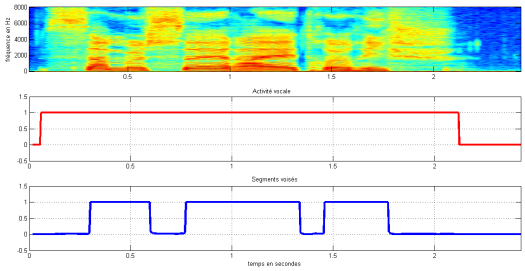


FIGURE 3.4 – En rouge : zone d'activité vocale détectée - en bleu : segments voisés - phrase : "Ca me sert à être riche".

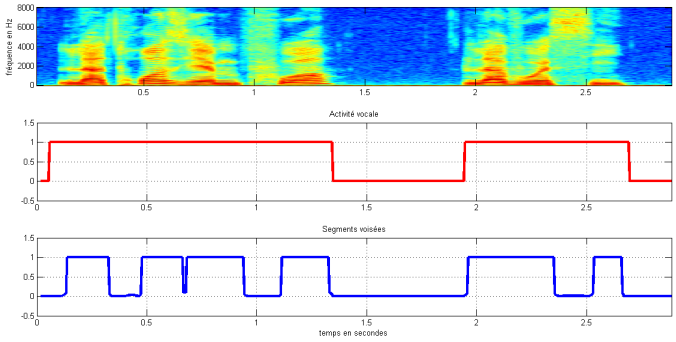


FIGURE 3.5 – En rouge : zone d'activité vocale détectée - en bleu : segments voisés - phrase : "La troisième fois, la voici".

3.3 Analyse fréquentielle multi-bandes

L'étape suivante consiste à effectuer une analyse multi-bandes du signal. On décompose dès lors le signal en N_b bandes de fréquences en échelle de Mel. On calcule ensuite l'énergie dans chaque bande (figure 3.6). On choisit empiriquement $N_b = 40$.

L'une des idées fondatrices des travaux réalisés était de pouvoir sélectionner l'information utile à la segmentation syllabique dans des zones de fréquences déterminées, d'où cette analyse fréquentielle multi-bandes.

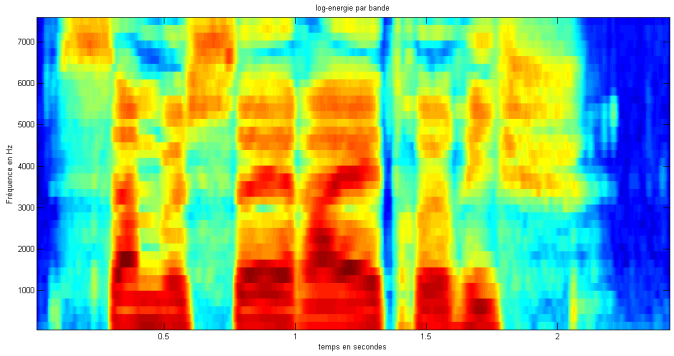


FIGURE 3.6 – Log-énergie par bande

3.4 Mesure de voisement

La mesure de voisement est l'étape clé de l'algorithme. Elle consiste à déterminer pour chaque trame un degré de voisement, et cela afin d'écarter les trames non-voisées lors de la détection des noyaux vocaliques (dans le même principe, les auteurs de[Xie and Niyogi, 2006] proposent un critère de voisement basé sur l'autocorrélation).

Deux critères de voisement sont proposés : le premier est basé sur l'entropie de Rényi, le second sur le VUV.

3.4.1 Critère de voisement fondé sur l'entropie de Rényi

A partir des remarques qui ont été faites sur l'entropie, partie 2.4.2 et dans [Huang and Yang, 2000], il semble pertinent de définir un descripteur C à partir de l'énergie du signal et de l'entropie spectrale de Rényi. On définit C de la manière suivante :

$$C(l) = H_{renyi}(l) \times RMSE(l) \tag{3.1}$$

avec

$$RMSE(l) = \sqrt{\frac{1}{L_w} \sum s(l)^2} \tag{3.2}$$

où l est l'indice de la trame courante, L_w la largeur de la fenêtre d'analyse et $s(l)$ la trame courante. Une valeur de $\alpha = 0.01$ a été choisie pour le calcul de l'entropie de Rényi d'après les études menées dans [Liuni et al., 2011] et [Obin and Liuni, 2012]. $RMSE(l)$(Root Mean Square Energy) est la racine carré de l'énergie à court-terme du signal. L'énergie et l'entropie sont préalablement centrées en 0 pour le calcul de C.

L'intérêt de la RMSE ici est que la moyenne temporelle effectuée pour son calcul rend plus stable et plus robuste face au bruit le coefficient C.

Une fois normalisé entre 0 et 1, le descripteur C s'interprète comme un coefficient de voisement indiquant l'état de voisement d'une trame. On peut ensuite appliquer un seuillage fixe au coefficient. Quand le coefficient dépasse un certain seuil, celui-ci est mis à 1. Sinon, il vaut 0. On peut également appliquer un seuillage plus doux de type "sigmoïde" (voir équation 3.3 et figure 3.7), ou encore, effectuer un seuillage qui conserve les valeurs proche

de zéro afin de ne pas perdre l'information des minima d'énergie qui seront utiles pour la segmentation. Observons à présent les différents profils en question figure 3.8.

$$Sigmoide(x) = \frac{1}{1 + \exp -\lambda \cdot x} \tag{3.3}$$

où λ est le paramètre de la sigmoïde.

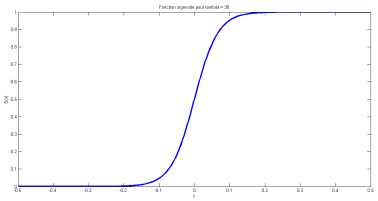


FIGURE 3.7 – Tracé de la fonction sigmoïde pour lambda = 30

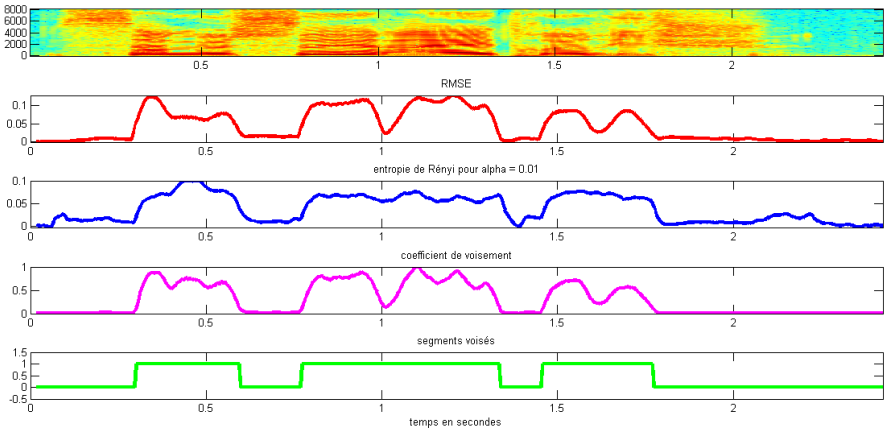


FIGURE 3.8 – Coefficient de voisement du signal de parole : *"Ca me sert à être riche."*

L'une des premières idées formulée durant le stage consistait non pas à calculer une seule entropie pour chaque trame (sur tous les points fréquentiels de la trame) mais à calculer une entropie dans différentes bandes de fréquences (entropie multi-bandes, dans la même philosophie que le VUV par bandes de Mels présenté partie 2.5). On disposait ainsi d'un critère de voisement local à chaque sous-bande. Malheureusement, cette méthode s'est avérée peu robuste dans un premier temps, mais cela reste encore à approfondir. Nous avons donc utilisé à la place de ce critère de voisement par bande, un critère de voisement trame à trame que l'on applique au profil temporel de chaque bande fréquentielle.

3.4.2 Critère de voisement fondé sur le VUV

Nous proposons cette fois-ci un coefficient de voisement basé sur le VUV, présentée partie 2.5. Le calcul du coefficient de voisement C s'effectue de la manière suivante :

1. calcul du VUV pour N_b bandes de fréquence Mels entre 0 Hz et $f_e/2$ Hz. On a choisit empiriquement $N_b = 40$.

2. seuillage du VUV par une fonction sigmoïde : 1 si la bande est considérée comme harmonique, 0 sinon.
3. pour chaque trame l, compter le nombre de bandes harmoniques $N_{bandesUtiles}(l)$.
4. si il n'y a aucune bande harmonique, c'est-à-dire $N_{bandesUtiles(l)} = 0$, alors $C(l) = 0$.
5. à ce stade, deux choix possibles :

   A : si $N_{bandesUtiles(l)} \neq 0$, alors $C(l) = 1$

   B : normaliser entre 0 et 1 $N_{bandesUtiles}(l)$ pour obtenir $C(l)$, et appliquer un seuillage (voir figure 3.9) .

Bien sûr dans le cas de voix expressives peu voisées, comme les voix chuchotées ou craquées (voir [Obin, 2012]), l'utilisation du VUV comme critère de voisement pour écarter les pics d'énergie non significatifs n'est plus pertinente. Cette remarque est aussi valable en partie pour le critère de voisement basé sur l'entropie.

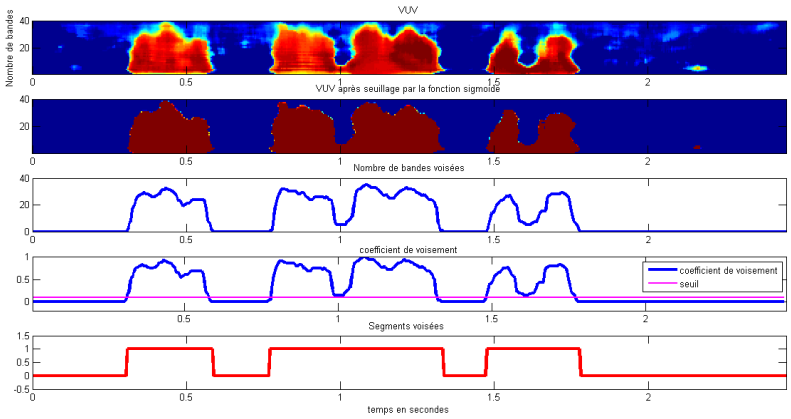


FIGURE 3.9 – de haut en bas : VUV, VUV avec sigmoïde, coefficient de voisement et segments voisés

3.4.3 Application du critère de voisement à la représentation multi-bandes

Nous disposons d'un coefficient de voisement indiquant si une trame est voisée ou non. Nous allons dès lors appliquer ce critère de voisement à la représentation multi-bandes de la partie 3.3 afin d'annuler les bandes des trames non voisées (figure 3.11).

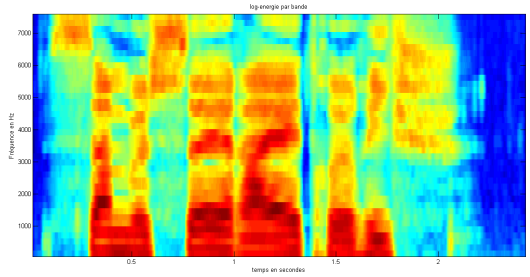


FIGURE 3.10 – Log-énergie par bande

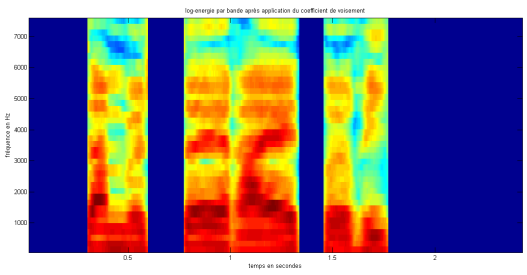


FIGURE 3.11 – Log-énergie après application du coefficient de voisement

3.5 Post-traitements

3.5.1 Corrélation temporelle

L'idée de la corrélation croisée temporelle est d'exploiter la structure temporelle de la parole pour éliminer les variations bruitées du signal.

Un problème classique réside dans le choix de la taille de la fenêtre d'analyse :

- une fenêtre large permet d'obtenir des profils temporels lisses. Cela engendre en contrepartie une perte de détails (maxima significatifs) ;
- inversement, une fenêtre étroite augmente la résolution temporelle (plus de détails) mais la détection des pics significatifs est plus difficile.

L'idée proposée dans [Wang and Narayanan, 2007] consiste à choisir une fenêtre d'analyse étroite puis d'appliquer un filtrage particulier qui prenne en compte le fait que les trames d'une même syllabe sont fortement corrélées. On introduit dès lors la corrélation croisée temporelle :

$$Xcorr_{temp}(l) = \sqrt{\frac{2}{K(K-1)} \sum_{j=0}^{K-2} \sum_{p=j+1}^{K-1} x(t-j) \bullet x(t-p)} \tag{3.4}$$

Le paramètre K est le nombre de trames du voisinage considéré pour le calcul de la corrélation temporelle. On peut par exemple choisir K de telle manière à ce que la fenêtre de corrélation dure un peu moins que la moitié de la durée d'une syllabe (en moyenne 200 ms). $x(l)$ est le vecteur de taille $N_b \times 1$ composé des énergies par bande à l'instant l. $Xcorr_{temp}(l)$ est le vecteur de taille $N_b \times 1$ résultant du filtrage.

En pratique, cela revient à considérer toutes les paires uniques parmi les K trames qui précèdent la trame courante ($2K(K-1)$ paires), à faire le produit des deux éléments de chaque paire, puis à calculer la moyenne des produits.

Ainsi, cette opération est équivalente à un lissage temporel (figure 3.12), mais exploitant la similarité des trames d'une même syllabe. Les auteurs de [Wang and Narayanan, 2007] proposent d'amplifier les discontinuités entre syllabes adjacentes en appliquant une fenêtre de pondération gaussienne centrée au milieu de la fenêtre d'analyse avant l'étape de corrélation temporelle. En effet, dans le cas où la fenêtre gaussienne de taille K est centrée sur le pic d'un noyau vocalique, cette trame aura plus de poids dans le calcul de la corrélation temporelle.

3.5.2 Corrélation spectrale

Le principe de la corrélation croisée spectrale est d'exploiter la structure formantique de la parole pour amplifier les différences entre les trames avec une structure formantique homogène (voyelles) et celles avec une structure formantique hétérogène (typiquement, les consonnes voisées : plosives /b/, /d/, /g/, liquides /l/, et fricatives /v/, /z/).

Nous disposons à ce stade de N_b profils temporels (1 pour chaque bande fréquentielle). L'idée suivante est de regrouper ces profils par zone formantique. Par exemple, pour un signal de parole échantillonné à $44.1kHz$, on considère $N_b = 40$ bandes Mels entre 100 Hz et 22050 Hz que l'on regroupe en 4 zones formantiques (on effectue la moyenne des bandes contenues dans chaque zone formantique) :

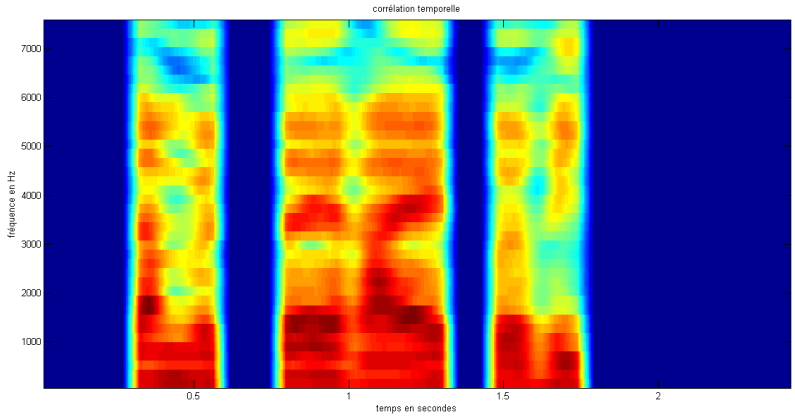


FIGURE 3.12 – Résultat du filtrage par corrélation croisée temporelle

1. zone F_1 : 100 Hz - 800 Hz (12 bandes)
2. zone F_2 : 800 Hz - 1500 Hz (7 bandes)
3. zone F_3 : 1500 Hz - 2500 Hz (8 bandes)
4. zone F_4 : 2500 Hz - 4000 Hz (7 bandes)

On dispose donc de 4 profils temporels notés $b_i(t)$ avec $i = 1 : 4$ (voir figure 3.13). Nous allons une nouvelle fois appliquer une corrélation croisée, mais cette fois-ci dans le domaine spectrale, afin d'exploiter la structure formantique des noyaux vocaliques. Un autre intérêt de la corrélation croisée spectrale est que celle-ci permet de filtrer l'énergie des plosives voisées, /b/ /d/ /g/ (dont l'énergie est concentrée essentiellement dans F_1), qui pourront introduire par la suite des pics non significatifs gênants. On calcule donc la corrélation croisée spectrale ([Wang and Narayanan, 2007]) des 4 profils :

$$Xcorr_{freq}(l) = \sqrt{\frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} b_i(n)b_j(n)} \tag{3.5}$$

avec $N = 4$ le nombre de trajectoires et $M = N(N-1)/2$ le nombre de paires uniques parmi les N trajectoires. Le nouveau profil 1D qui résulte de cette dernière opération est visible figure 3.14.

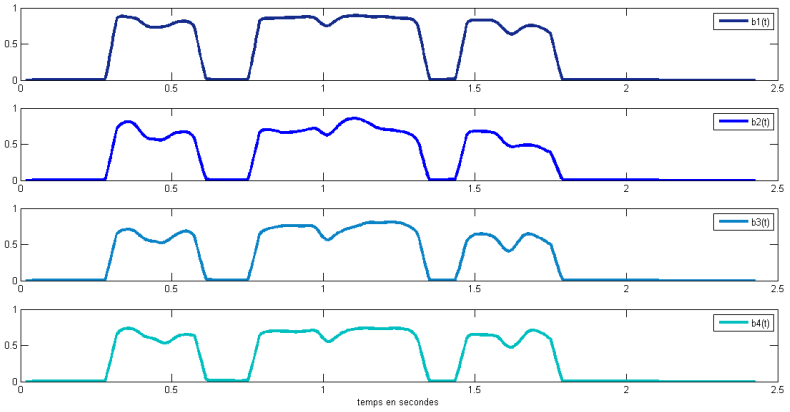


FIGURE 3.13 – Profils temporels des 4 bandes de fréquences

3.6 Exemple de segmentation en syllabe

Nous disposons maintenant d'un profil temporel 1D d'un descripteur du signal. Nous allons appliquer à ce profil un algorithme de détection des minima et maxima en respectant certaines contraintes : les maxima trop proches l'un de l'autre (voisins de moins de 80 ms) et les maxima d'intensité non significative sont écartés. Le résultat de la détection des pics est visible figure 3.14. Les pics bleus d'amplitude 0.5 correspondent aux minima (frontières des syllabes). Les pics bleus d'amplitude 1 correspondent quant à eux aux maxima (noyaux vocaliques).

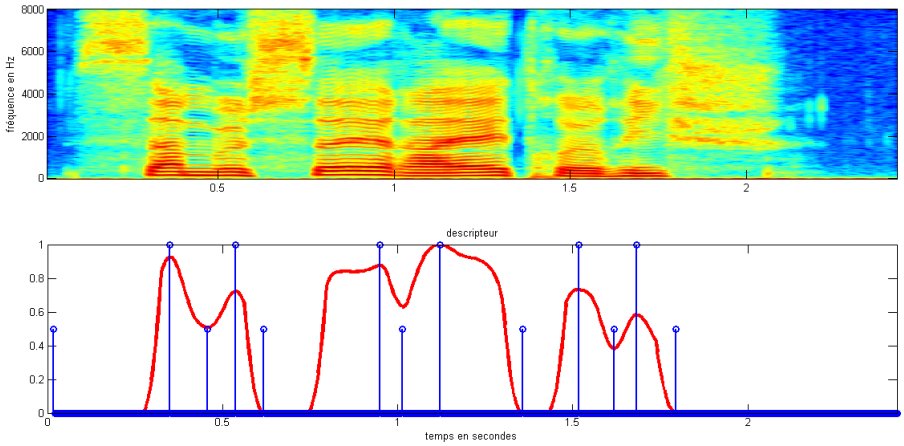


FIGURE 3.14 – Minima et maxima du profil du descripteur

Le résultat de la segmentation est visible figure 3.15. Cet exemple permet d'illustrer une erreur typique de l'algorithme de segmentation : la segmentation a échouée à *"à ê-"* car ces deux syllabes sont fortement co-articulées.

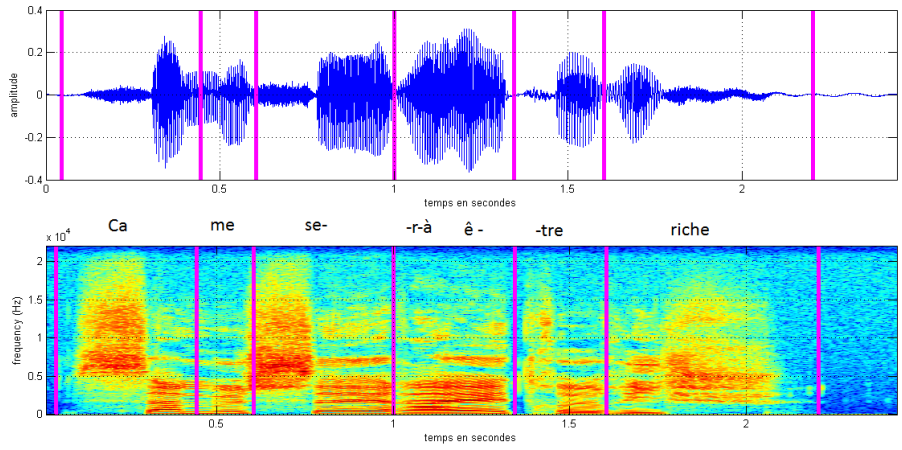


FIGURE 3.15 – Résultat de la segmentation pour : *"Ca me sert à être riche"*

Un autre exemple de segmentation est présenté en figure 3.16. La phrase en question est *"Il arrive demain d'Italie par la route"*.

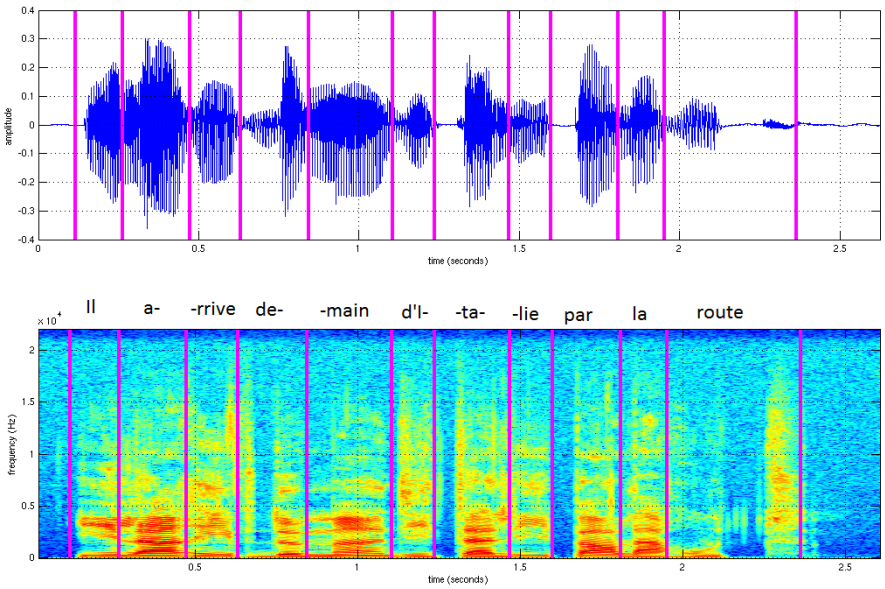


FIGURE 3.16 – Résultat de la segmentation pour : *"Il arrive demain d'Italie par la route"*.

3.7 Schéma récapitulatif de l'algorithme

Un synopsis de l'algorithme est présenté en figure 3.17.

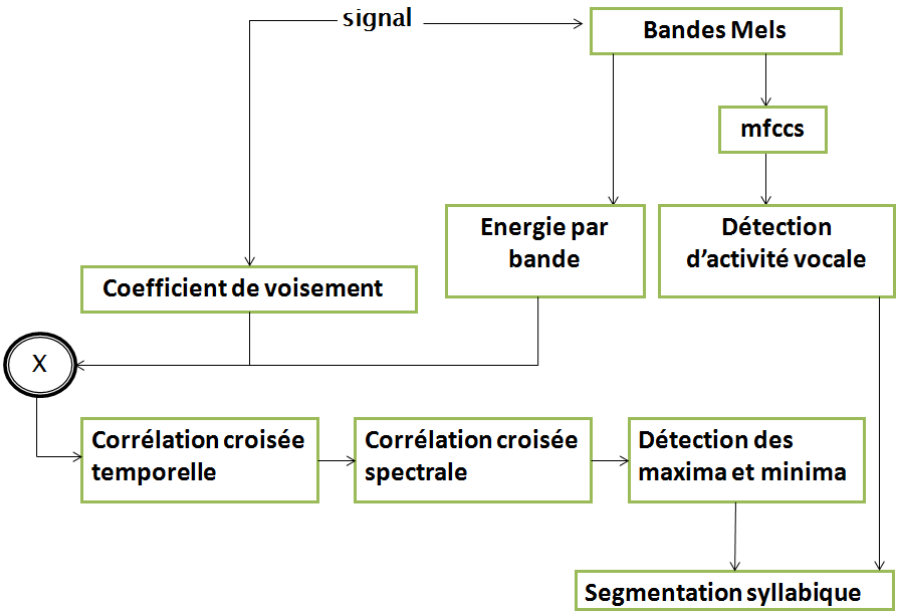


FIGURE 3.17 – Synopsis de l'algorithme développé

Chapitre 4

Evaluation

L'évaluation de la segmentation de la parole en syllabes se décompose généralement en deux parties :

1. la détection des noyaux vocaliques ("vowel landmarks") ;
2. la détection des frontières des syllabes .

Dans la mesure où l'objectif final de la segmentation en syllabes est de déterminer les frontières des syllabes, nous nous concentrerons exclusivement sur cette dernière.

Nous présentons maintenant les performances des méthodes de segmentation en syllabe proposées sur un corpus mono-locuteur en français, puis introduisons les bases d'une évaluation à plus large échelle sur un corpus de référence multi-locuteur en anglais-américain [1](voir 4.1.1).

4.1 Protocole expérimental

Les performances des deux méthodes proposées ont été évaluées et comparées à celles de deux méthodes de segmentation déjà existantes. Nous considérons donc en tout 4 méthodes :

- méthode de Melmerstein ([Mermelstein, 1975]) ;
- méthode de Villing ([Villing et al., 2004]) ;
- méthode basée sur l'entropie (voir 3 et 3.4.1) ;
- méthode basée sur le VUV (voir 3 et 3.4.2).

Nous n'avons pas retenu d'autres méthodes existantes : la méthode proposé dans ([Howitt, 1999]) qui présente des performances proches de celles de la méthode de Melmerstein ([Villing et al., 2004]), les méthodes à base de HMMs([Xie and Niyogi, 2006]), et la méthode proposée dans [Xie and Niyogi, 2006] qui ne permet dans l'état actuel qu'une détection des noyaux vocaliques.

Ces évaluations nécessitent de comparer les résultats de segmentation obtenus à partir de l'une des 4 méthodes à une segmentation de référence, effectuée au préalable (voir partie 4.1.1 et figure 4.1).

1. Cette évaluation sera présentée lors de la soutenance

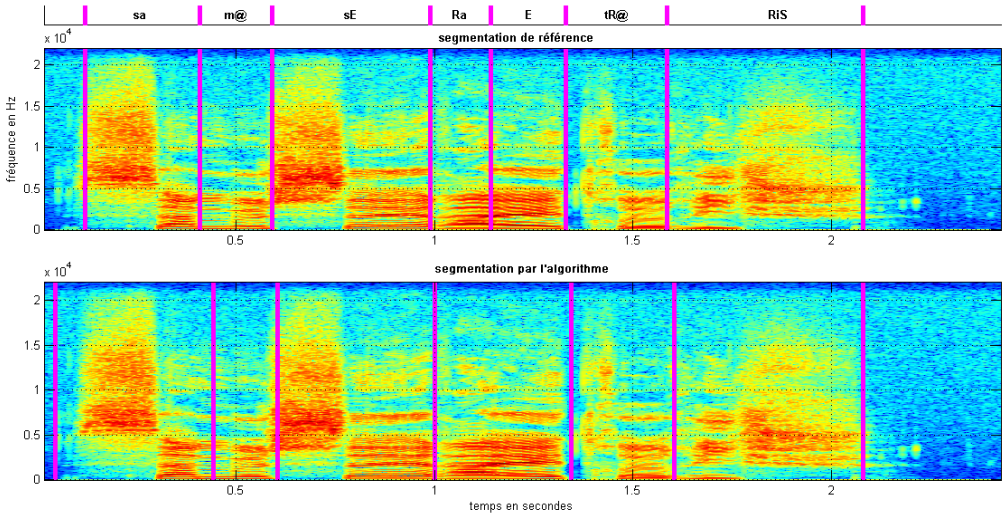


FIGURE 4.1 – Comparaison de la référence et de la segmentation obtenue par la méthode basée sur l'entropie pour la phrase : *"Ca me sert à être riche."*

4.1.1 Bases de données

Deux corpus de parole ont été utilisés pour l'évaluation : une base de données en Français de l'IRCAM et la base de données TIMIT, en anglais-américain.

Base de données de l'IRCAM en Français. FR-IRCAM :

La base de données contient environ 500 phrases (472) extraites d'un ensemble de phrases phonétiquement équilibrée, lues par un seul locuteur masculin dans une chambre anéchoïque. Les enregistrements audio sont encodés en fichiers audio sans compression (wav) et au format 44.1kHz/16bits. La base de donnée est segmentée manuellement en phonèmes, syllabes et mots.

Pour ce corpus de sons, la durée moyenne d'un phonème est de 107 ms tandis que la durée moyenne d'une syllabe est de 240 ms. Une segmentation syllabique de référence est également disponible pour ce corpus.

Base de données TIMIT en anglais-américain. US-TIMIT :

La base de données TIMIT ([Garofolo, 1993]) est un corpus de parole lue qui a été conçue à l'origine pour l'évaluation de systèmes de reconnaissance de la parole. Elle a été développée conjointement par le DARPA - ISTO (Defense Advanced Research Projects Agency - Information Science and Technology Office), le MIT (Massachusetts Institute of Technology), le SRI (Stanford Research Institute) et Texas Instrument.

Elle contient un total de 6300 phrases en anglais-américain, plus exactement 10 phrases prononcées par 630 locuteurs (438 hommes et 192 femmes). Les locuteurs proviennent de 8 régions différentes des Etats-Unis. Les enregistrements audio sont encodés en fichiers audio sans compression (wav) et au format 16kHz/16bits.

La base de données contient initialement des segmentations de référence en mots et en phonèmes, mais malheureusement pas de segmentation en syllabes. La segmentation en syllabe de référence a été obtenue par syllabation des séquences de mots à partir d'un

outil de syllabation de l'anglais-américain [2], puis par ré-alignement temporel des syllabes à partir de l'alignement des phonèmes.

4.1.2 Méthodologie d'évaluation de la segmentation

Nous disposons de deux flux de marques temporelles de segmentation, un flux *résultat* et un flux *référence*, que l'on souhaite comparer. La marque temporelle de segmentation d'une fin de syllabe correspond également à la marque temporelle de segmentation du début de la syllabe suivante.

Pour chacun des marqueurs temporels de début de syllabes (indicés en temps) de l'un des deux flux, on regarde s'il existe un marqueur équivalent dans le second flux. On autorise pour cela un intervalle de tolérance $\delta_{tolerance}$ autour de la marque de segmentation courante. Si la marque courante a bien un équivalent dans le second flux, alors la segmentation sera jugé correcte. Sinon, nous nous trouvons dans le cas d'une omission ou d'une insertion de syllabe, selon que le flux considéré est le flux *référence* ou le flux *résultat* respectivement. Dans le cas où une marque de segmentation d'un flux a plusieurs marques équivalentes dans l'autre flux, dans l'intervalle $\delta_{tolerance}$, on choisira celle qui est la plus proche.

Pour comptabiliser les résultats, on générera deux vecteurs binaires $V_{reference}$ et $V_{resultat}$ de même dimensions. Soit m un indice temporel que l'on incrémente à chaque comparaison de syllabe (voir figure 4.2).

- dans le cas d'une segmentation correcte : $V_{reference}(m) = 1$ et $V_{resultat}(m) = 1$;
- dans le cas d'une omission de syllabe : $V_{reference}(m) = 1$ et $V_{resultat}(m) = 0$
- dans le cas d'une insertion de : $V_{reference}(m) = 0$ et $V_{resultat}(m) = 1$;

Les vecteurs $V_{reference}$ et $V_{resultat}$ permettent ainsi de mesurer le taux d'insertion, le taux d'omission et le taux de bonne segmentation.

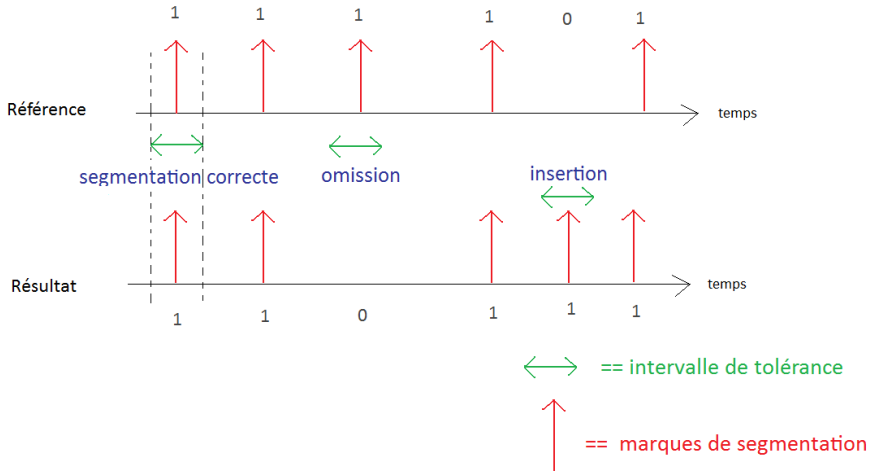


FIGURE 4.2 – Comparaison des deux flux *référence* et *résultat* de marques de segmentation de la phrase : *"Ca me sert à être riche."*

2. http ://aclweb.org/anthology-new/N/N09/N09-1035.pdf

4.1.3 Mesures de performance

On définit la *precision* et le *rappel* à partir des formules suivantes (4.1) :

$$precision = \frac{v_p}{v_p + f_p}, \quad rappel = \frac{v_p}{v_p + f_n} \tag{4.1}$$

Dans notre cas, v_p pour *vrai positif* correspond au nombre de marques correctement segmentées, f_p pour *faux positif* correspond au nombre d'insertions et f_n pour *faux négatif* correspond au nombre d'omissions.

Dans le cadre de la classification binaire, la précision et le rappel sont liés directement aux taux d'insertions et d'omissions par :

$$\%_{insertion} = 1 - precision, \quad \%_{omission} = 1 - rappel \tag{4.2}$$

Dès lors, on introduit la $F_{measure}$(équation 4.3). Elle correspond à la moyenne harmonique de la *precision* (*precision*) et du *rappel* (*recall*) :

$$F_{measure} = 2.\frac{precision.rappel}{precision + rappel} \tag{4.3}$$

Elle présente en outre l'avantage par rapport à la mesure d'"accuracy" (taux de bonne classification) d'être insensible aux différences du nombre d'éléments dans chacune des classes.

4.2 Résultats

Nous présentons dans cette section la comparaison de méthodes de segmentation en syllabes. Nous allons pour cela calculer les taux d'insertions, les taux d'omissions et la $F_{measure}$ pour chaque algorithme. Nous évaluerons également la robustesse des méthodes de segmentation en fonction du rapport signal à bruit (20dB->0dB)

Les résultats pour le corpus US-TIMIT et pour la méthode de segmentation de Villing ([Villing et al., 2004]) ne sont pas encore disponibles. Ils seront présentés lors de la soutenance.

Les paramètres généraux considérés pour l'analyse sont les suivants :

- fenêtre d'analyse de 35 ms ;
- analyse avec une fenêtre de Hanning, avec un pas d'avancement de 4 ms ;
- caractéristiques spectrales intégrées sur un banc de 40 bandes en échelle Mel (entre 100 et $f_e/2$ Hz) ;
- calcul de la log-énergie dans chaque bande Mel.
- durée de la fenêtre pour la corrélation temporelle : 60 ms.

On lance les évaluations sur le premier corpus en français, FR-IRCAM, présenté partie 4.1.1. On choisit $\delta_{tolerance} = +/-50$ ms (50 ms à gauche et à droite de la marque de segmentation courante), ce qui correspond à une précision de l'ordre de la moitié de la durée d'un phonème. Les résultats pour ce corpus sont présentés dans le tableau 4.1.

La méthode basée sur le VUV possède les meilleures performances en termes de taux d'insertions, de taux d'omissions et de $F_{measure}$ sur ce corpus. C'est également cette méthode

qui est la plus robuste et la plus stable face au bruit blanc gaussien. En général, elle dépasse de loin les performances de l'algorithme de Melmerstein. Le reste des évaluations (sur le corpus US-TIMIT et avec l'algorithme de Villing) sera présenté à la soutenance.

Méthodes		**Taux d'insertions**	**Taux d'omissions**	**Fmeasure**
Melmerstein				
	sans bruit	27.6	29.4	71.5
	20 dB	28.6	30.0	70.7
	10 dB	29.7	30.6	69.9
	5 dB	30.0	31.0	69.5
	0 dB	27.7	36.1	67.8
Méthode-entropie				
	sans bruit	24.9	20.2	77.4
	20 dB	26.8	19.4	76.7
	10 dB	29.8	21.6	74.1
	5 dB	31.6	22.6	72.7
	0 dB	34.1	23.5	70.8
Méthode-VUV				
	sans bruit	18.0	13.4	84.2
	20 dB	21.0	13.0	82.8
	10 dB	21.8	13.5	82.1
	5 dB	22.6	14.2	81.3
	0 dB	24.0	15.2	80.1

TABLE 4.1 – Tableau des performances des différentes méthodes pour le corpus FR-IRCAM

Chapitre 5

Conclusion

La syllabe, pour un grand nombre de langues, est la "note de musique" de la voix. Pouvoir accéder à une telle information parait donc essentiel, surtout dans le contexte de création musicale dans lequel baigne l'IRCAM depuis sa fondation. La syllabe est également l'"atome" de la prosodie. Une perspective de modification des caractéristiques prosodiques de la voix ou d'intégration d'éléments prosodiques dans les systèmes de synthèse de la parole nous donnent alors une raison supplémentaire de vouloir les identifier.

Le travail réalisé durant le stage a aboutit à la réalisation de méthodes non-supervisés de segmentation syllabique et à l'implémentation de méthodes déjà existantes. L'avantage des méthodes non-supervisées par rapport aux méthodes supervisées est qu'elles ne nécessitent pas de phase d'apprentissage, souvent lourde à mettre en œuvre et très dépendante de la langue. Elles sont donc plus facilement généralisables.

Les méthodes qui ont été proposées reposent principalement sur l'établissement de deux nouveaux critères de voisement pour la segmentation en syllabe, basés sur l'entropie de Rényi et le VUV. Ces critères de voisement sont alors appliqués sur une représentation temps-fréquence multi-bandes pour écarter les trames non-voisées, ou bien les bandes de fréquences non utiles pour la segmentation syllabique. Les profils temporels acoustiques pour les différentes bandes fréquentielles sont alors exploités pour réaliser la segmentation syllabique.

Une méthode de segmentation syllabique basée sur une analyse multi-bandes semble tout à fait viable dans le sens ou elle permet de prendre en compte la structure formantique des noyaux vocaliques. L'idée d'une corrélation croisée entre ces bandes est judicieuse car elle exploite justement cette structure formantique. Le défaut de cette méthode est que la corrélation croisée spectrale d'une voyelle faiblement énergétique ou dont l'énergie est concentrée dans la première zone formantique (/u/ par exemple) sera faible. Une idée potentiellement intéressante pour exploiter la structure formantique des noyaux vocaliques consisterait à détecter les minima et maxima non pas sur le profil temporel 1D issu de la corrélation spectrale mais sur les N_b profils temporels (voir partie 3.5.2) des N_b bandes fréquentielles, puis d'appliquer un critère de fusion des pics.

L'application d'un critère de voisement pour écarter les trames (ou plus généralement les bandes fréquentielles) non voisées en tant que noyaux vocaliques potentiels est une étape indispensable pour un algorithme de segmentation syllabique. Le coefficient de voisement basé sur la mesure de VUV a prouvé son efficacité dans le cas où la voix est effectivement bien voisée (harmonique). Dans des cas plus extrêmes de voix très expressives, les hypothèses de fonctionnement de cette méthode ne sont plus respectées.

Pour le coefficient de voisement basé sur l'entropie de Rényi, l'hypothèse fondatrice était que le spectre d'une trame voisée était fortement piqué. Cette hypothèse se justifie bien car une part importante de l'énergie d'une trame voisée est contenue dans les basses fréquences (dans la première zone formantique). Cependant, cela n'écarte pas le cas des trames bruitées (non voisées) où l'énergie du bruit est aussi concentrée dans certaines bandes de fréquences (bruits colorés de natures variées...). Il serait alors judicieux de ne pas considérer les points fréquentiels de ces bandes parasites dans le calcul de l'entropie de Rényi. Il faudrait par conséquent trouver un critère pour faire la distinction entre les bandes inutiles de bruit coloré "type" et les bande utiles voisées. Une idée consisterait à utiliser la mesure de VUV pour écarter les points fréquentiels des bandes très énergétiques mais peu voisées. Le calcul de l'entropie de Rényi avec une valeur de *alpha* petite se ferait sur un spectre nettoyé de tous les pics parasites d'intensité élevé. Par contre, cette proposition ne fonctionnera pas si le bruit coloré est de nature harmonique, et dans les limites imposées par le VUV...

Une piste d'amélioration du critère de voisement basé sur l'entropie consisterait à mesurer l'entropie non plus sur le spectre mais sur l'enveloppe spectrale (enveloppe LPC, true envelope...), et cela fin de rendre la mesure plus robuste au bruit. Toujours dans la même otique, mais dans une voie parallèle, nous pourrions, avant le calcul le l'entropie, calculer une estimation du bruit de fond puis soustraire ce bruit du signal de parole (voir état de l'art partie 2.4.2).

La corrélation croisée temporelle permet d'appliquer un filtrage qui prend en compte l'étalement temporel d'une syllabe sur plusieurs trames (voir partie 3.5.1). Le choix de la valeur du paramètre K n'est cependant pas anodin. Sa valeur est en fait assez dépendante de la durée moyenne d'une syllabe et plus généralement du débit de parole. Il faudrait également tester si la pondération par une fenêtre gaussienne permet bien d'amplifier les discontinuités entre syllabes adjacentes, comme cela est suggéré dans [Wang and Narayanan, 2007].

l'étape de détection d'activité vocale (3.2) n'est jamais abordée dans les articles portant sur la segmentation syllabique. Elle est pourtant nécessaire pour délimiter correctement les syllabes voisines d'une période de silence. La détection d'activité vocale est cependant un problème complexe, différent de celui de la segmentation syllabique. L'algorithme retenu est assez simple et donne d'assez bons résultats. Quelques erreurs incontournables de détection, provoquées notamment par des bruits de bouche ou des respirations, peuvent malgré tout se produire. On pourrait aussi envisager d'autres algorithmes plus sophistiqués à la place.

Les performances de segmentation et de robustesse au bruit des deux méthodes proposées, obtenues après les premières évaluations, sont encourageantes. D'autres tests pourront être effectués sur des fichiers de parole bruités avec différents types de bruits naturels, tels que ceux rencontrés dans une salle de spectacle, durant une représentation (bruit de climatisation, etc...). Il existe également des variantes des méthodes proposées qui n'ont pas été présentées, mais dont l'étude devra être poursuivie.

Enfin, une implémentation temps réel en MAX/MSP pourra être envisagée par la suite.

Bibliographie

[Boll, 1979] Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 27(2).

[Clements and Keyser, 1983] Clements, G. and Keyser, S. (1983). Cv phonology. a generative theory of the syllabe.

[De Jong and Wempe, 2009] De Jong, N. and Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2).

[Garofolo, 1993] Garofolo, J. (1993). *Darpa Timit : Acoustic-phonetic Continuous Speech Corps CD-ROM*. US Department of Commerce, National Institute of Standards and Technology.

[Gerven and Xie, 1997] Gerven, S. and Xie, F. (1997). A comparative study of speech detection methods. In *Fifth European Conference on Speech Communication and Technology*.

[Griffin and Lim, 1985] Griffin, D. and Lim, J. (1985). A new model-based speech analysis/synthesis system. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.*, volume 10. IEEE.

[Hall, 2006] Hall, T. (2006). *"Syllable, phonology"*. Encyclopedia of Langage and Linguistics.

[Howitt, 1999] Howitt, A. (1999). Vowel landmark detection. In *Sixth European Conference on Speech Communication and Technology*.

[Huang and Yang, 2000] Huang, L. and Yang, C. (2000). A novel approach to robust speech endpoint detection in car environments. In *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 3. IEEE.

[Jin and Cheng, 2010] Jin, L. and Cheng, J. (2010). An improved speech endpoint detection based on spectral subtraction and adaptive sub-band spectral entropy. In *Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference on*, volume 1. IEEE.

[Lee, 1989] Lee, K. (1989). *Automatic speech recognition : the development of the SPHINX system*. Number 62. Springer.

[Liu et al., 2008] Liu, H., Li, X., Zheng, Y., Xu, B., and Jiang, N. (2008). Speech endpoint detection based on improved adaptive band-partitioning spectral entropy. *Journal of System Simulation*, 20(5).

[Liuni et al., 2011] Liuni, M., Robel, A., Romito, M., and Rodet, X. (2011). Rényi information measures for spectral change detection. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE.

[Mermelstein, 1975] Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *The Journal of the Acoustical Society of America*, 58.

[Ney, 1981] Ney, H. (1981). An optimization algorithm for determining the endpoints of isolated utterances. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'81.*, volume 6. IEEE.

[Obin, 2011] Obin, N. (2011). *MeLos : Analysis and Modelling of Speech Prosody and Speaking Style*. PhD. Thesis, Ircam - UPMC.

[Obin, 2012] Obin, N. (2012). Cries and whispers-classification of vocal effort in expressive speech. In *Interspeech*.

[Obin and Liuni, 2012] Obin, N. and Liuni, M. (2012). On the generalization of Shannon entropy for speech recognition. In *IEEE workshop on spoken language technology*.

[Ouzounov, 2005] Ouzounov, A. (2005). Robust features for speech detection-a comparative study. In *Int. Conf. on Computer Systems and Technologies*.

[Rényi, 1961] Rényi, A. (1961). On measures of entropy and information. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability*.

[Robinson and Dadson, 1956] Robinson, D. and Dadson, R. (1956). A re-determination of the equal-loudness relations for pure tones. *British Journal of Applied Physics*, 7.

[Savoji, 1989] Savoji, M. (1989). A robust algorithm for accurate endpointing of speech signals. *Speech Communication*, 8(1).

[Shannon, 1948] Shannon, C. (1948). A mathematical theory of communications. *Bell System Technical Journal*, 27(3).

[Shen et al., 1998] Shen, J., Hung, J., and Lee, L. (1998). Robust entropy-based endpoint detection for speech recognition in noisy environments. In *Fifth International Conference on Spoken Language Processing*.

[Villing et al., 2004] Villing, R., Timoney, J., and Ward, T. (2004). Automatic blind syllable segmentation for continuous speech.

[Villing et al., 2006] Villing, R., Ward, T., and Timoney, J. (2006). Performance limits for envelope based automatic syllable segmentation. In *Irish Signals and Systems Conference, 2006. IET*. IET.

[Wang and Narayanan, 2007] Wang, D. and Narayanan, S. (2007). Robust speech rate estimation for spontaneous speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(8).

[Wang et al., 2011] Wang, H., Xu, Y., and Li, M. (2011). Study on the mfcc similarity-based voice activity detection algorithm. In *Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on*. IEEE.

[Wu and Wang, 2005] Wu, B. and Wang, K. (2005). Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments. *Speech and Audio Processing, IEEE Transactions on*, 13(5).

[Xie and Niyogi, 2006] Xie, Z. and Niyogi, P. (2006). Robust acoustic-based syllable detection. In *Ninth International Conference on Spoken Language Processing*.

[Young, 1993] Young, S. (1993). The htk hidden markov model toolkit : Design and philosophy. *Department of Engineering, Cambridge University, UK, Tech. Rep. TR*, 153.

Annexes

Liste XSAMPA des phonèmes du français moderne

Le code XSAMPA est une extension de la norme SAMPA de transcription en code ASCII de l'alphabet standard API (Alphabet Phonétique Intenrational) [1].

Symbole ascii	Exemple	Transcription
p	pont	po~
b	bon	bo~
t	temps	ta~
d	dans	da~
k	quand	ka~
g	gant	ga~
f	femme	fam
v	vent	va~
s	sans	sa~
z	zone	zon
S	champ	Sa~
Z	gens	Za~
m	mont	mo~
n	nom	no~
J	oignon	oJo~
N	camping	ka~piN
l	long	lo~
R	rond	Ro~
w	coin	kwe~
H	juin	ZHe~
j	pierre	pjER
i	si	si
e	ses	se
E	seize	sEz
a	patte	pat
A	pâte	pAt
O	comme	kOm
o	gros	gRo
u	doux	du
y	du	dy
2	deux	d2
9	neuf	n9f
@	justement	Zyst@ma~
e~	vin	ve~
a~	vent	va~
o~	bon	bo~
9~	brun	bR9~

1. http ://fr.wikipedia.org/wiki/X-SAMPA