Contributions à l'étude et à la reconnaissance automatique de la parole en Fongbe

Frejus Adissa Akintola Laleye

Université d'Abomey-Calavi **Université du Littoral Côte d'Opale**

Contributions à l'étude et à la reconnaissance automatique de la parole en Fongbe

THÈSE

présentée et soutenue publiquement le 10 décembre 2016

pour l'obtention du

Doctorat délivré conjointement par l'Université d'Abomey-Calavi et l'Université du Littoral Côte d'Opale

(spécialité : Génie Informatique, Automatique et Traitement du Signal)

par

Fréjus A. A. Laleye

Composition du jury

Président :	Prof. Vianou Antoine, Université d'Abomey-Calavi, Bénin
Rapporteurs :	Prof. Gouton Pierre, Université de Bourgogne, France Prof. Pinti Antonio, Université d'Orléans, France Prof. Gbaguidi Julien, Université d'Abomey-Calavi, Bénin
Examinateurs :	Prof. Wira Patrice, Université de la Haute Alsace, France Prof. Dagba K. Théophile, Université d'Abomey-Calavi, Bénin
Directeurs de Thèse :	Prof. Eugène C. Ezin, Université d'Abomey-Calavi, Bénin Prof. Cina Motamed, Université du Littoral et Côte d'Opale, France

Unité de Recherche d'Informatique et Sciences Appliquées
&
Laboratoire d'Informatique, Signal et Image de la Côte d'Opale

Remerciements

En premier lieu, je souhaite remercier chaleureusement mon encadrant Eugène C. EZIN. C'est en toute sincérité que je lui adresse toute ma reconnaissance, pour m'avoir donné l'opportunité de travailler avec lui. Je lui suis reconnaissant pour tout le temps conséquent qu'il m'a accordé pendant les trois années de thèse, sa rigueur, sa franchise, sa sympathie et surtout son amour. Il a su me soutenir et m'encourager tant sur le plan professionnel que sur le plan personnel. Mes remerciements vont particulièrement aussi à l'endroit de Monsieur Cina MOTAMED pour m'avoir offert le cadre de travail et donné l'opportunité de travailler dans un environnement idéal pour de meilleurs rendements.

J'adresse mes remerciements à Monsieur Laurent BESACIER, pour m'avoir accueilli, pendant quelques jours, dans le Laboratoire d'Informatique de Grenoble (LIG), pour sa disponibilité, ses orientations et sa collaboration.

Un grand remerciement à tous les membres du Laboratoire Informatique Signal et Image de la Côte d'Opale (LISIC-ULCO), pour leur accueil et leur sympathie et à tout le personnel de l'Institut de Mathématiques de Sciences Physiques (IMSP-UAC).

J'adresse tout particulièrement mes remerciements à Mikael A. MOUSSE, mon compagnon de tous les jours. Merci pour tous ces moments de joies et de peines partagés. Plus qu'un ami, tu es un frère.

Je tiens également à adresser mes sincères remerciements à Camus LANMADOU-CELO, Christian AKPONA et à Roméo AZANDEGBE, mes camarades de lutte qui n'ont jamais cessé de me soutenir pendant ces trois années de thèse. Voici le résultat de vos soutiens, prières et encouragements.

Mes sincères remerciements à l'Agence universitaire de la Francophonie (Bureau de l'Afrique de l'Ouest) qui m'a soutenu pendant la réalsation de cette thèse.

Enfin, je voudrais exprimer mes plus profonds remerciements à tous ceux qui, de près ou de loin, ont contribué à la réussite de ce travail ; puisse l'éternel vous bénir.

Je dédie cette thèse
à mes parents,
à mon épouse, Cyrielle G. M. ANAGO ...

Sommaire

Table des figures

Introduction générale

Introduction générale

"*agban e glo ségbanhentɔ́ ɔ́,*
jegbanhentɔ́ sixú hen a"

En quarante ans, le domaine de la reconnaissance automatique de la parole a considérablement progressé et a révélé d'innovants algorithmes et techniques pour le traitement statistique de la parole. L'usage du langage naturel dans le dialogue homme/machine positionne ce domaine au centre d'intérêts des chercheurs des sciences comme la phonétique, le traitement du signal, la linguistique, l'informatique et l'intelligence artificielle. Malheureusement, malgré l'incroyable progrès des connaissances, techniques et algorithmes, la reconnaissance automatique de la parole reste un sujet de recherche actif car les résultats obtenus jusque là sont encore loin de l'idéal et les applications demeurent fortement dépendantes de la langue cible.

Parmi les 6000 langues parlées dans le monde, très peu sont dotées d'un système de reconnaissance automatique de la parole ou de synthèse vocale. Elles ne présentent pas soit d'intérêts majeurs (langues minoritaires), soit ne disposent pas de ressources linguistiques utilisables pour la réalisation de technologies vocales (langues peu dotées). En dehors des langues très répandues (l'anglais, le français, le chinois etc.), un grand nombre de langues ne sont pas informatisées et manquent de ressources numériques.

Dans le cadre de cette thèse, nous avons choisi d'étudier une langue nationale majoritairement parlée au Bénin (le Fongbe). Pour cette langue, les ressources numériques comme un corpus de textes et un corpus audio pour une tâche de reconnaissance de la parole, ne sont pas disponibles.

L'ensemble des activités de recherche décrites dans ce manuscrit de thèse visent à étudier le Fongbe d'un point de vue acoustique et le doter de méthodes et algorithmes pour l'élaboration d'un système de reconnaissance automatique de la parole. Les travaux sont effectués en phase avec l'avancée des outils et techniques employés dans la littérature. Nos recherches ont exploré tous les domaines liés à la réalisation d'un système de reconnaissance automatique de la parole dont l'analyse de signaux (extraction de caractéristiques

acoustiques), la segmentation du signal de la parole, la classification et la reconnaissance automatique de la parole.

L'originalité de cette thèse vient de la volonté et des efforts à mettre à la disposition de la communauté scientifique :

- des corpora audio et textuels en Fongbe [2] ;
- des algorithmes de segmentation automatique [3, 4] et de classification [5] qui tiennent compte des spécificités acoustiques des sons du Fongbe et de la variabilité des signaux de paroles prononcées par des locuteurs du Fongbe ;
- et le développement des premiers modèles acoustiques et du langage pour le Fongbe en vue de l'élaboration d'un premier système de reconnaissance en Fongbe [2].

Le présent manuscrit s'articule autour de trois parties :

- la première partie présente sur 2 chapitres le contexte d'étude, l'état de l'art général et les ressources linguistiques collectées. Le premier chapitre aborde la problématique des langues africaines en général, et celle du Fongbe en particulier. Il présente aussi la mesure d'informatisation du Fongbe calculée à partir de la méthode développée par V. Berment dans [6]. Il se termine par un état de l'art général sur les procédés et méthodes de la segmentation automatique de la parole et sur les systèmes de reconnaissance automatique de la parole. Le deuxième chapitre présente le protocole adopté pour collecter les ressources linguistiques telles que les deux corpora (audio et texte) qui ont servi aux travaux d'analyse, de segmentation, de classification et de reconnaissance.
- la deuxième partie expose nos travaux sur la reconnaissance automatique des phonèmes isolés du Fongbe. Elle comprend les chapitres 3 et 4 du manuscrit. Le chapitre 3 expose notre étude sur l'analyse acoustique, dans un contexte isolé, des différents sons du Fongbe. Dans le chapitre 4, nous avons proposé une approche basée sur une fusion intelligente et adaptative de décisions dans un contexte de multiclassification pour la reconnaissance automatique des phonèmes du Fongbe.
- la dernière partie du manuscrit présente nos travaux sur la segmentation et la reconnaissance automatique de la parole continue en Fongbe. Elle comprend les chapitres 5 et 6. Nous proposons, dans le chapitre 5, deux méthodes innovantes permettant de segmenter la parole continue en de petits segments contenant des phonèmes ou syllabes. La première méthode exploite les propriétés géométriques du signal de

parole et l'entropie de Rényi que nous avons choisi calculer à partir l'énergie à court-terme du signal. La deuxième méthode utilise des connaissances expertes induises par l'approche des règles et ensembles flous. Ces deux méthodes présentent de bonnes performances sur le corpus de parole continue en Fongbe comparées à quelques autres méthodes choisies dans l'état de l'art. Dans le cinquième chapitre, nous avons proposé un premier système de reconnaissance de la parole continue en Fongbe à partir du développement de modèles acoustiques et du langage à l'aide de la boîte à outils Kaldi ASR.

Première partie

Cadre théorique

Chapitre 1

Contexte d'étude et état de l'art

Hun ɖe ɖyɔ ɔ wefɔ tɔn nɔ ɖyɔ

Sommaire

Introduction

Depuis environ quarante ans, la reconnaissance automatique de la parole est un domaine d'étude qui a captivé l'attention des chercheurs informaticiens, linguistes, mathématiciens et même télécoms. Elle trouve ses applications dans plusieurs domaines et se situe au croisement du traitement du signal numérique et du traitement du langage. C'est donc une thématique de recherche qui conduit l'informaticien ou le mathématicien à développer des algorithmes de traitement du signal de parole et du langage qui satisfont les principes lexicologiques, syntaxiques et sémantiques d'une langue développés par un linguiste. Ces algorithmes utilisent pour la plupart la parole comme vecteur d'information et sont utilisés pour produire des applications capables de traiter efficacement la parole naturelle. Malgré l'avancée fulgurante des technologies et l'évolution des ordinateurs ces dernières années, la communauté scientifique ne dispose pas encore d'un système de reconnaissance idéal pour les langues très connues du monde comme le français et l'anglais. Des applications dans le domaine émergent et sont fortement liées à la langue d'étude. De façon générale, les performances d'un système de reconnaissance de la parole dépendent fortement de la langue étudiée, de sa capacité à modéliser les connaissances acoustiques et linguistiques, ainsi que de la quantité et la qualité des données utilisées pour l'apprentissage des modèles du système. A son niveau plus élémentaire, il comprend des algorithmes pour traiter le signal depuis sa production jusqu'à sa reconnaissance. Il s'agit des algorithmes pour la détection et l'extraction des caractéristiques acoustiques pour l'analyse de la parole, la détection de régions stables pour la segmentation du signal de parole en unités phonétiques, syllabiques ou en mots et la reconnaissance des différentes unités pour la transcription. Tous ces éléments sont abordés dans cette thèse dans le but de fournir un premier système de reconnaissance de la parole en Fongbe (une langue locale du Bénin).

Le contexte multilingue du Bénin (comme la plupart des pays africains) est un facteur favorisant l'analphabétisation de la population béninoise et la marginalisation lors des rendez-vous devant impliquer des personnes de diverses langues. Ceci limite l'employabilité des langues nationales dans certains secteurs comme les administrations publiques et l'éducation. Dans l'idée de ne pas faire d'un handicap la multiplicité des langues nationales, le développements d'applications de communication entre l'homme et la machine répondant aux principes du système de reconnaissance de la parole serait un atout. La langue visée dans cette thèse (le Fongbe) se présente comme étant la langue majoritaire-

ment parlée du Bénin et pratiquée par près de 50% de la population. Son choix se justifie aussi par le fait que, c'est la langue qui jusque là a déjà fait objet de plusieurs études linguistiques (phonologie, lexique et syntaxe).

L'idée de cette thèse est de contribuer à l'étude du Fongbe en proposant des algorithmes de segmentation, de classification et de reconnaissance de la parole en Fongbe. Elle s'est déroulée en cotutelle entre l'Université d'Abomey-Calavi (UAC) du Bénin et l'Université du Littoral de Côte d'Opale (ULCO) de la France dans le cadre de la collaboration entre l'équipe de recherche URISA (Unité de Recherche en Informatique et en Sciences Appliquées) et le LISIC (Laboratoire d'Informatique Signal et Image de la Côte d'Opale).

Dans ce chapitre, il est présenté le cadre théorique dans lequel les travaux ont été effectué : le contexte d'étude (l'introduction), la problématique des langues africaines, la description du Fongbe et l'état de l'art sur la segmentation et la reconnaissance de la parole.

1.1 Problématique des langues africaines

Les pays africains disposent d'un système linguistique organisé autour d'une langue officielle et des langues nationales. La langue officielle (souvent le français ou l'anglais pour les pays de l'Afrique de l'ouest) favorise l'accès à une communication entre plusieurs ethnies. Elle constitue la langue de travail et est enseignée dans les écoles. Les langues nationales sont souvent propres à une ethnie et servent souvent à l'alphabétisation des personnes âgées. Elles ne disposent pas, pour la plupart, de ressources linguistiques pour le développement des TIC. Ce qui n'est pas le cas des langues officielles qui ont connu de grands travaux linguistiques et informatiques.

Selon "Ethnologue : Languages of the world" [1] ($19^{\text{ième}}$ Edition, 2016), l'Afrique dispose de 2139 langues vivantes parlées (et parfois écrites) pour 847.791.487 locuteurs. Elle est classée au deuxième rang, après l'Asie (2.296 langues pour 3.929.931.706), des continents qui disposent plus de langues vivantes. A ce titre, les langues africaines représentent le tiers des langues parlées dans le monde et sont donc importantes pour le patrimoine linguistique de l'humanité. La figure 1.1 présente une répartition [2] des langues vivantes par pays sur le continent africain. Malgré leur nombre important, peu d'entre elles (14) possèdent un statut de langue officielle : le français dans 23 états, l'anglais dans 19, l'arabe dans 10, le portugais dans 5, l'afrikaans, le swahili et l'espagnol dans 2. Parmi les

1. http ://www.ethnologue.com
2. http ://www.axl.cefan.ulaval.ca/Langues/1div_cont_Afrique.htm

langues officielles, on retrouve des langues nationales comme le swahili au Kenya et en Tanzanie, l'amharique en Éthiopie etc. Ce sont ces langues nationales ajoutées au Wolof, Haussa et au yoruba qui ont fait objet de travaux dans la reconnaissance automatique de la parole parce qu'elles disposent de grandes quantités de ressources électroniques. La figure 1.1 révèle que les pays les plus multilingues sont d'abord le Nigeria (527 langues), le Cameroun (280 langues), le Congo-Kinshasa (212 langues), la Tanzanie (126 langues). Il existe aussi des pays qui présentent une homogénéité linguistique comme le Rwanda (5 langues), le Lesotho (5 langues), l'île du Cap-Vert (4 langues), le Burundi (4 langues).

FIGURE 1.1 – Répartition des langues africaines par pays

Plus de 70% des langues nationales africaines sont peu dotées car, d'une part, elles ne sont pas totalement décrites, pour la plupart, et ne sont pas complètement codifiées (grammaire, syntaxe et lexique) et d'autre part, elles ne disposent pas de ressources électroniques pour des traitements comme :

- la saisie de texte ;
- la correction orthographique ou grammaticale ;
- la synthèse vocale et reconnaissance de la parole ;
- la traduction automatique ;
- la reconnaissance optique de caractères.

Face aux progrès des technologies dans le monde, les langues africaines ont accumulé un retard considérable dans l'implantation des Technologies de l'Information et de la

Communication (TIC) pour leur usage. Ce retard constitue pour l'Afrique un handicap pour le développement des systèmes de reconnaissance de la parole (et autres technologies comme la traduction automatique) malgré l'émergence de quelques unes des langues.

Pour combler ce retard, plusieurs projets sont récemment nés pour revaloriser les langues africaines en les dotant de ressources linguistiques. Au nombre des ces projets, nous pouvons citer entre autres le projet DiLaf [3] qui vise à informatiser des dictionnaires langues africaines-français (bambara, haoussa, kanouri, tamajaq, songhai-zarma) afin de pouvoir les diffuser plus largement et étendre leur couverture. Il y a aussi le projet ALFFA [4] (*African Languages in the Field speech Fundamentals and Automation*) dont l'objectif est le développement de technologies vocales pour les téléphones mobiles en Afrique. Il aborde deux aspects fondamentaux de la langue à savoir : les aspects fondamentaux de l'analyse du langage parlé (description des langues, phonologie, dialectologie) et les technologies de la parole (reconnaissance et synthèse) pour les langues africaines. Il constitue un partenariat entre le Laboratoire d'Informatique d'Avignon (LIA-Université d'Avignon), le Laboratoire d'Informatique de Grenoble (LIG-Université Joseph Fourier), le Laboratoire Dynamique Du Langage (DDL-CNRS) et VOXYGEN.

Les langues africaines, pour certaines comme l'amharique en Éthiopie, ont une ancienne tradition écrite et pour d'autres une tradition plus récente basée sur l'alphabet arabe ou latin utilisant les caractères de l'Alphabet Phonétique International (API). Elles ont pour la plupart un caractère tonal qui impose l'usage des signes diacritiques avec certains caractères latins. C'est donc le cas du Fongbe qui a deux tons lexicaux (haut et bas) qui, à leur tour, peuvent être modifiés en trois autres tons phonétiques. Certaines langues peuvent se transcrire dans plusieurs systèmes d'écriture [7]. Par exemple, en wolof (langue nationale du Sénégal), l'écriture arabe permet la transcription de textes religieux, mais des règles d'écriture en lettres latines augmentées de caractères API permettent une autre transcription. Pour l'orthographe des mots, une langue peut avoir plusieurs orthographes possibles. Il faut noter que la plupart des langues nationales ont une tradition orale et ne possèdent toujours pas une convention d'écriture fixe. Depuis les années 60, d'énormes efforts ont été faits pour certaines langues, qui désormais possèdent une orthographe fixe et officielle permettant, à partir de l'alphabet, une transcription de tous les mots. Ces efforts favorisent aujourd'hui l'intégration de ces langues nationales sur internet par une forme d'écriture standard et contribuent à leur informatisation.

Nous pouvons aussi remarquer des alternances tonales pour beaucoup de familles de langues africaines. La forme d'un mot peut varier de façon très importante le contexte

3. http ://www.dilaf.org/dilaf/Home.po
4. http ://alffa.imag.fr/

phonologique. Leur traitement informatique pour des système de reconnaissance ou de traduction automatique devient complexe et nécessite de travaux supplémentaires intégrant des algorithmes de diacritisation pour permettre les réalisations tonales. Vu la particularité des langues nationales africaines et leur différence avec les langues européennes comme le Français et l'anglais, les méthodes et algorithmes existants doivent être révisés pour une adaptation au contexte phonologique et syntaxique des langues africaines.

1.2 Le Fongbe

Le Fongbe est une langue nationale africaine parlée majoritairement au Bénin. Le Bénin compte 55 langues nationales (dont 50 autochtones et 5 non-autochtones), selon "*Ethnologue : Languages of the world*" [5] (19ième Edition, 2016), qui sont pour la plupart liées aux différentes ethnies. La figure 1.2 montre une répartition démographique des langues nationales parlées au Bénin selon le recensement général de la population et de l'habitation (2002).

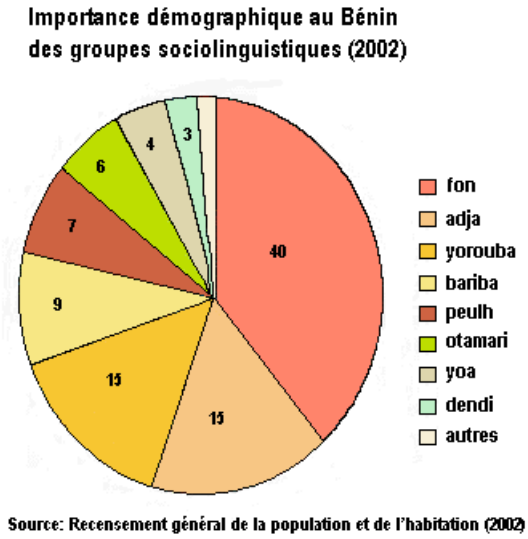


FIGURE 1.2 – Importance démographique des groupes sociolinguistiques.

Il se révèle que depuis 2002 à nos jours, plus de la moitié de la population béninoise parle le Fongbe. Il est aussi parlé au Togo et au Nigeria. Il fait partie du groupe des dialectes *Gbe* et est parlé par le plus grand groupe ethnique du Bénin [8]. Il est assez répandu dans les médias et est aussi utilisé pour l'éducation et l'alphabétisation des adultes. Il est classé dans la catégorie des langues *Kwa* de la famille nigero-congolaise [9].

5. http ://www.ethnologue.com

Depuis 1975, son alphabet est connu officiellement et son écriture est basée sur le latin avec les caractères de l'API.

Le Fongbe est une langue tonale qui dispose d'un système complexe avec deux tons lexicaux, haut et bas, qui peuvent être modifiés pour générer trois autres tons phonétiques : bas-haut montant, haut-bas descendant et moyen [8]. Les signes diacritiques sont utilisés pour transcrire ces différents tons.

Le système vocalique du Fongbe est bien adapté au timbre vocalique dessiné par les premiers phonéticiens. Il comprend douze (12) timbres : sept (7) voyelles orales avec 4 degrés d'aperture et cinq (5) voyelles nasales avec 3 degrés d'aperture. Le système consonantique comprend 22 phonèmes. Les tableaux 1.1 et 1.2 présentent la classification des différents sons voyelles et consonnes du Fongbe selon leur mode d'articulation et leur lieu d'articulation.

TABLE 1.1 – Inventaire des voyelles

Oral			Nasal		
Front	Central	Back	Front	Central	Back
i		u	ḭ		ṵ
e		o	ɛ̰		ɔ̰
ɛ		ɔ		a̰	
	a				

TABLE 1.2 – Inventaire des consonnes

Orale			Nasale
Voisantes sourdes	Voisantes sonores	Non-voisantes	
f	v	b	m
t	d	ɖ	n
c	j	y	ŋ (ny)
s	z	l	
kp	gb	w	
k	g		
x	ɣ (h)		

On peut aisément remarquer que le Fongbe partage les mêmes sons avec le français à la différence des voyelles /ḭ/ et /ṵ/ et des consonnes / kp/, / c/ et / x/ qui lui sont propres. Le Fongbe est manipulé avec l'ensemble de ces sons voyelles, consonnes et tons. Son écriture est basée sur un ensemble de conventions qui se traduisent par les règles pratiques suivantes :

- toutes les voyelles qui viennent après une consonne nasale sont systématiquement nasalisées, ainsi la marque de nasalisation n'est plus écrite. Exemple : [nṵ̃] ("*chose*" en français) s'écrit |nŭ| ;

- une voyelle nasale est écrite en remplaçant le tilde (˜) par la consonne /n/. Exemple : [tá̰] ("*marigot*" en français) s'écrit |tán|.

- la seule nasale syllabique du système phonologique du Fongbe /n̄/ s'écrit en combinant la voyelle /u/ avec la consonne /n/. Exemple : [n̄ wá] ("*je suis venu*" en français) s'écrit |un wá|.

- toute voyelle qui ne porte pas de ton est intonnée moyen ; Exemple : [mḭ̃] ("*vous*" en français) s'écrit |mi |.

- la voyelle /a/ en position initiale dans un mot est toujours intonnée bas /à/ ; Exemple : [àḍŭ] ("*dent*" en français) s'écrit | aḍŭ |.

Il faut noter que, lorsque les tons sont utilisés dans des phrases, ils changent l'orthographe des mots. Ainsi pour maîtriser l'écriture du Fongbe, il convient d'observer les mots du point de vue de la structure interne des mots. L'ensemble des mots peut être regroupé en trois différentes structures syllabiques : monosyllabique, dissyllabique et trisyllabique.

Les études scientifiques, dans le domaine linguistique, ont commencé en 1963 avec la publication du premier dictionnaire Fon-Français [10]. Depuis 1976, plusieurs chercheurs linguistes ont travaillé sur le Fongbe et de nombreux articles ont été publiés sur les aspects linguistiques de la langue. Contrairement à la plupart des langues occidentales (français, Anglais, Espagnol etc.), asiatiques (Chinois, Japonnais, etc.) et africaines (Wolof, Swahili, Haussa), le Fongbe souffre d'un manque très important de ressources linguistiques sous forme numérique (corpus audio et de textes) et ceci malgré les nombreux travaux linguistiques (phonologie, lexique et syntaxe).

1.3 Mesure d'informatisation du Fongbe

Pour mesurer le niveau d'informatisation de la langue d'étude (le Fongbe), nous avons adopté le protocole développé par V. Berment dans sa thèse [6] qui nous a semblé être une bonne méthode pour évaluer le degré d'informatisation d'une langue. Dans son protocole, V. Berment a défini 3 niveaux d'informatisation à partir d'un indice$-\sigma$ qui permet de distinguer trois groupes de langues selon la disponibilité des ressources linguistiques. Cet indice est calculé à partir des valeurs de criticité et de notes attribuées à chaque ressource,

par un nombre représentatif de locuteurs, d'une langue. Les ressources sont répertoriées dans le tableau 1.3. La criticité C_k définit l'importance relative d'une ressource pour un groupe d'évaluation donné [6].

TABLE 1.3 – Tableau d'évaluation du niveau d'informatisation d'une langue.

	Services/ressources	**Criticité C_k (0 à 10)**	**Note N_k (/20)**	**Note pondérée (C_kN_k)**
Traitement du texte				
	Saisie simple			
	Visualisation/impression			
	Recherche et remplacement			
	Sélection du texte			
	Tri lexicographique			
	Correction orthographique			
	Correction grammaticale			
	Correction stylistique			
Traitement de l'oral				
	Synthèse vocale			
	Reconnaissance de la parole			
Traduction				
	Traduction automatisée			
ROC				
	Reconnaissance optique de caractères			
Ressources				
	Dictionnaire bilingue			
	Dictionnaire d'usage			
Total		$\sum C_k$		$\sum C_kN_k$
Moyenne (/20)				$\sum C_kN_k / \sum C_k$

Ainsi on distingue les trois groupes de langues suivantes selon la valeur (moyenne pondérée) de l'indice$-\sigma$:

- langues-π ont une moyenne entre 0 et 9,99 (langues peu dotées) ;
- langues-μ ont une moyenne entre 10 et 13,99 (langues moyennement dotées) ;
- langues-τ ont une moyenne entre 14 et 20 (langues très bien dotées).

Pour évaluer le Fongbe, nous avons évalué chaque ressource selon nos connaissances sur les logiciels existants et sur les travaux réalisés sur le Fongbe. Nous avons ensuite complété notre évaluation par une enquête auprès des institutions d'alphabétisation et du département de langue de l'Université. Le tableau 1.4 renseigne sur les niveaux de criticité de chaque ressource, leur note et de l'indice$-\sigma$ calculé.

TABLE 1.4 – Tableau d'évaluation du niveau d'informatisation du Fongbe.

	Services/ressources	**Criticité (/10)**	**Note (/20)**	**Note pondérée (Criticité*Note)**
Traitement du texte				
	Saisie simple	10	14	140
	Visualisation/impression	10	14	140
	Recherche et remplacement	8	15	120
	Sélection du texte	6	15	90
	Tri lexicographique	5	0	0
	Correction orthographique	5	0	0
	Correction grammaticale	4	0	0
	Correction stylistique	0	0	0
Traitement de l'oral				
	Synthèse vocale	5	0	0
	Reconnaissance de la parole	5	0	0
Traduction				
	Traduction automatisée	6	0	0
ROC				
	Reconnaissance optique de caractères	9	0	0
Ressources				
	Dictionnaire bilingue	10	5	50
	Dictionnaire d'usage	10	2	20
Total		93		560
Moyenne (/20)				**6,0215053763**

L'indice-σ ($\sigma = 6,02$), obtenu après évaluation, permet d'affirmer que le Fongbe est une langue peu dotée ($\sigma < 9,99$). Ainsi le Fongbe fait partie des langues-π. On remarque aisément dans le tableau que les ressources liées au traitement automatique de la parole sont inexistantes et peuvent justifier notre intérêt à vouloir doter la langue d'un premier système de reconnaissance de la parole. Cela pourrait aussi augmenté son $indice - \sigma$ et fait d'elle une $langue - \mu$.

Pour construire le système de reconnaissance de la parole en Fongbe, l'une des premières contributions est la réalisation des ressources linguistiques numériques. Ces ressources sont d'une importance indéniable pour le développement des applications de traitement automatique des langues. Dans notre cas d'étude, elles constituent un pré-requis pour le développement d'applications à hautes valeurs ajoutées pour le Fongbe. Les ressources que nous avons collectées dans le cadre de nos travaux sur le Fongbe sont :

- un corpus audio d'unités phonémiques ;
- un corpus de parole continue ;

- un corpus de textes extraits du web ;
- un vocabulaire ;
- et un dictionnaire de prononciation.

1.4 Segmentation de la parole

La segmentation de la parole est un ensemble de procédés d'identification d'unités variées dans un signal de parole selon la nature du segment considéré. Selon cette nature, on distingue les formes de segmentation suivantes :

- la segmentation en sons voisés ou non ;
- la segmentation en phonèmes ;
- la segmentation en syllabes ;
- la segmentation en mots ;
- la segmentation en locuteurs.

La tâche de segmentation de la parole est indispensable pour l'apprentissage des modèles acoustiques d'un système de reconnaissance de la parole et de synthèse vocale. Il existe deux formes de segmentation : la segmentation manuelle (dépendant du corpus) et la segmentation automatique (indépendant du texte). Les grands corpora des langues moyennement et très bien dotées (Français, anglais ou l'italien) disposent déjà d'un étiquetage phonétique des signaux audio : ce qui n'est pas le cas des langues peu dotées. Il est donc bien indiqué une segmentation automatique de bonne précision pour des corpora non annotés car elle accélère les procédures de vérification manuelle qui prend suffisamment de temps et coûtent chères. Malgré le coût élevé en temps et en argent, la segmentation manuelle est beaucoup plus précise que la segmentation automatique.

1.4.1 Segmentation manuelle de la parole

La segmentation peut être effectuée manuellement par l'examen de la forme d'onde du signal de parole à l'aide du spectrogramme. La segmentation manuelle, en plus du processus de segmentation du corpus, intègre un processus d'étiquetage des unités par une annotation linguistique. Ceci désigne la notation descriptive des unités audio ou textuelles permettant de les interpréter pour des fins de reconnaissance (pour notre cas d'étude)

ou de synthèse vocale. L'analyse du spectrogramme du signal de parole ressort une série d'éléments sonores distincts qui peuvent être des zones spectralement homogènes [11] (voir figure 1.3). On peut remarquer sur la figure 1.3, que les transitions entre certains sons ne sont pas franches. Ceci démontre, selon S. Nefti [11], le paradoxe entre la perception des segments de parole et la variabilité acoustique de cette dernière et montre aussi que la segmentation est un problème fondamentalement complexe.

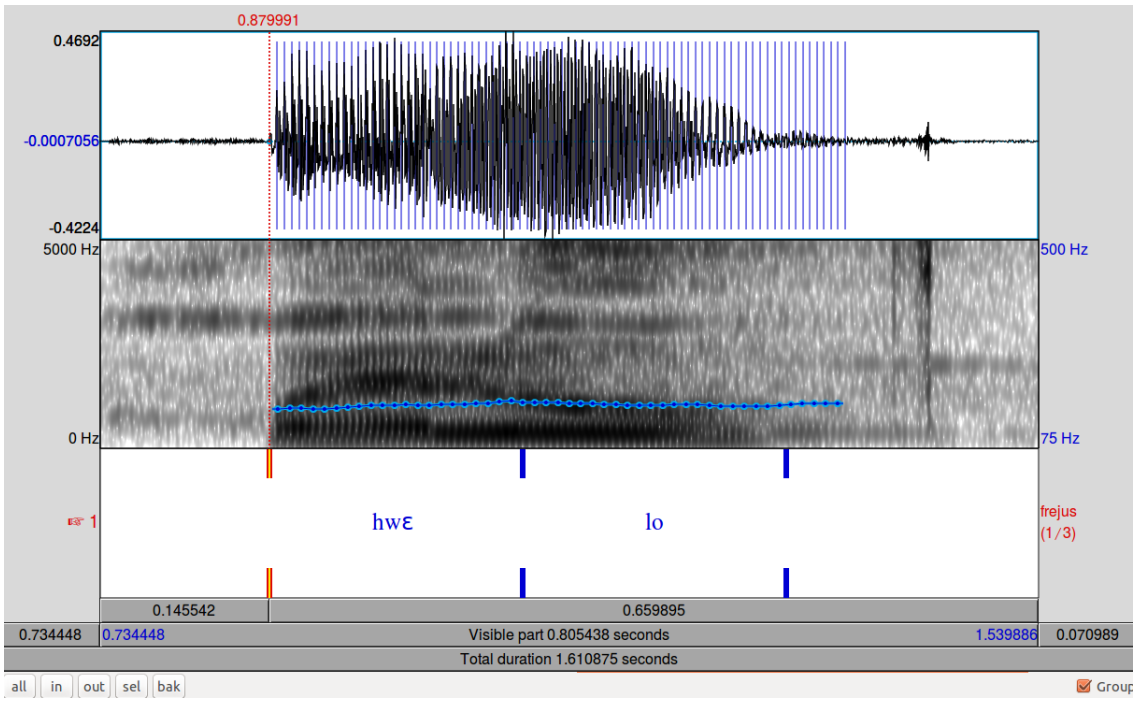


FIGURE 1.3 – Forme d'onde et spectrogramme d'un énoncé en Fongbe : "hw ɛ lo"

La segmentation manuelle, couplée à l'étiquetage, est une tâche souvent réalisée par des experts phonéticiens de la langue qui, aux plus petits détails, font correspondre une description linguistique aux unités acoustiques contenues dans un énoncé. Pour sa réalisation, elle prend en compte les informations sur les conditions d'enregistrement du corpus de parole à segmenter, sur les caractéristiques du locuteur enregistré et sur le style d'élocution. Lorsque plusieurs experts interviennent sur un même corpus, l'idéal est qu'ils se partagent la segmentation afin d'intégrer au maximum les connaissances sur la langue et les différences entre segmentation manuelle. Une segmentation prise isolément doit inclure en grande partie la subjectivité de l'expert phonéticien qui l'a effectué [12].

La segmentation manuelle peut se réaliser, soit par un seul phonéticien, soit par plusieurs phonéticiens qui se la partagent. Ainsi pour rendre plus précise la segmentation, une méthode complémentaire consiste à réaliser une segmentation automatique que tous les experts vérifient et corrigent manuellement. Cette stratégie est adoptée de nos jours dans

des projets à grands financements capables d'employer plusieurs experts afin de gagner en temps de réalisation. Cette forme de segmentation est dénommée semi-automatique.

Plusieurs outils existent aujourd'hui pour permettre aux experts de réaliser la segmentation manuelle. Parmi ces outils, il en existe qui sont utilisés par les experts informaticiens (Julius [6], Sphinx [7], HTK [8], etc.), par les experts linguistes (Praat + plugins praat [9], wafesurfer [13], Snorri [10], etc.) et aussi par ces deux types d'experts (SPPAS [11]). D'autres outils tels que SLAM [12] et *The Aligner* [14] sont utilisés pour effectuer la segmentation semi-automatique.

La forme d'onde et le spectrogramme du signal d'un énoncé révèlent des informations très pertinentes qui permettent d'étiqueter et de délimiter les unités (qu'elles soient des phonèmes, syllabes ou mots) contenues dans l'énoncé. La délimitation conduit à ressortir les transitions entre unités et à créer les différentes frontières existantes même si celles-ci ne sont toujours pas systématiquement visibles pour l'expert. Cette opération reste encore fastidieuse pour l'expert car il n'existe pas encore de méthodes conventionnelles efficaces à 100% lui permettant d'affronter la variabilité des différents sons d'une langue et de marquer avec précision les transitions entre segments dans un signal. Comparée à la segmentation automatique, la segmentation manuelle reste la méthode de segmentation la plus précise pour les applications de reconnaissance de la parole quand elle est effectuée par les experts phonéticiens de la langue. La segmentation automatique trouve son sens, d'une part, dans l'intérêt de réduire le coût d'une segmentation manuelle et d'autre part, dans le soucis d'introduire des algorithmes pour automatiser les tâches d'étiquetage de segmentation.

1.4.2 Segmentation automatique de la parole

La segmentation automatique se présente comme la deuxième façon (opposée à la segmentation manuelle) de segmenter un énoncé en de petites unités acoustiques à partir de méthodes automatiques sans trop d'interventions manuelles. Elle se base sur les connaissances acoustico-phonétiques des unités (phonétiques ou syllabiques pour la plupart) pour identifier des segments plus petits dans une parole continue. Elle intervient, soit à l'étiquetage, soit à la vérification et à la correction.

6. http ://julius.osdn.jp/en_index.php
7. http ://cmusphinx.sourceforge.net/
8. http ://htk.eng.cam.ac.uk/
9. www.praat.org
10. http ://www.loria.fr/ laprie/WinSnoori/
11. http ://www.sppas.org/
12. http ://www.pd.istc.cnr.it/pages/slam.htm

Deux grandes catégories de méthodes de segmentation automatique existent [15] :

- les méthodes de segmentation sans contraintes linguistiques : ce sont des méthodes de segmentation du signal de parole qui ne nécessitent pas une connaissance a priori du contenu linguistique du signal à segmenter ;
- les méthodes de segmentation avec contraintes linguistiques : ce sont des méthodes de segmentation du signal de la parole qui intègre, pour leur fonctionnement, une description linguistique du signal à segmenter.

Les méthodes de segmentation classées dans la première catégorie sont aussi appelées des méthodes non-supervisées tandis que celles de la deuxième catégorie sont appelées des méthodes supervisées compte tenu de leur mode de fonctionnement.

1.4.2.1 Segmentation sans contraintes linguistiques : non-supervisée

La segmentation Sans Contraintes Linguistiques (SCL) est effectuée indépendamment de la langue donc sans lien a priori avec les connaissances linguistiques. Elle consiste à segmenter un signal de parole à l'aide de méthodes paramétriques ou ad-hoc (mesure de stationnarité ou détection de maxima/minima/discontinuités) à partir d'informations pertinentes extraites de la représentation acoustique du signal. Elle opère sur la dynamique physique du signal pour calculer une mesures de distance entre vecteurs acoustiques ou modèles statistiques afin de détecter ou de délimiter des zones d'homogénéité spectrale (segments acoustiques du signal).

De nos jours, la plupart des méthodes de segmentation SCL sont basées sur le calcul du *Zero Crossing Rate* (ZCR), de l'énergie à court-terme [16, 17, 18]. Ce sont des descripteurs temporels du signal qui sont mis en exergue par la représentation acoustique et permettent de séparer des modèles acoustiques selon les distances calculées. On y retrouve aussi les descripteurs fondés sur les coefficients de prédiction linéaire (LPC), sur l'estimation de la fréquence fondamentale et sur les coefficients cepstraux à l'échelle fréquentielle de Mel (MFCC) [19]. Sans connaissances a priori de l'étiquetage du signal de parole, les méthodes de segmentation SCL ne fournissent pas un étiquetage linguistique des segments acoustiques qu'elles délimitent. Ces segments donnent néanmoins la réalité physique du signal en entrée. Dans le domaine de la reconnaissance de la parole, nous pouvons citer les méthodes de segmentation SCL [12] suivantes :

- **détection de rupture de stationnarité dans le temps -** cette méthode vise à détecter des ruptures correspondant à des discontinuités de stationnarité. Les travaux

[20] et [21] ont exploité le fonctionnement de cette méthode pour la segmentation automatique d'un flux de parole.

- **détection d'activité vocale -** la détection d'activité vocale vise à localiser avec précision les zones contenant de la parole à partir des échantillons du signal de parole. Cela revient à séparer le silence de la parole dans un signal de parole. Elle se base, soit sur la comparaison des amplitudes du signal de parole avec le niveau de bruit [22], soit sur la fonction d'énergie à court-terme calculée par la somme du signal multiplié par une fonction de fenêtrage sur N trames [23].

- **détection de voisement -** les méthodes de détection de voisement considèrent un segment voisé, soit en calculant les valeurs de la mesure HNR (*Harmonic-to-Noise Ratio*) locale, de l'énergie, d'un coefficient de corrélation, soit en utilisant l'information sur le nombre de passages par zéro. Dans le premier cas, on distingue les méthodes du domaine temporel [24, 25] et les méthodes du domaine fréquentiel [26, 27]. Dans le second cas, le principe est de construire une courbe qui passe par les milieux des segments, puis de détecter les passages par zéro de cette courbe [28, 22].

- **segmentation fricatif/non-fricatif -** les méthodes de segmentation fricatif/non-fricatif exploitent une statistique du nombre de passages par zéro de la dérivée du signal pour déterminer un bruit de friction. La segmentation est basée sur l'identification de ce bruit [29, 30].

- **segmentation par ondelettes -** ce sont des méthodes qui segmentent le signal de parole à partir de son analyse temps/fréquence. On retrouve, entre autres, la segmentation par paquets d'ondelettes [31], et par ondelettes de Malvar [26].

- **détection des variables spectrales -** ces méthodes sont basées sur le calcul de la fonction de variation spectrale SVF (*Spectral Variation Function*) définie comme une mesure de corrélation ayant pour but de localiser des changements spectraux rapides [32, 33].

Certaines de ces méthodes seront abordées et bien détaillées dans le chapitre traitant de la segmentation syllabique de la parole en Fongbe.

1.4.2.2 Segmentation avec contraintes linguistiques : supervisée

La segmentation Avec Contraintes Linguistiques (ACL) utilise une séquence de symboles linguistiques du signal pour délimiter des trames acoustiques d'un énoncé. Ces symboles sont pour la plupart des phonèmes, syllabes ou mots et constituent la description

linguistique de l'énoncé. Les méthodes de la segmentation ACL ont pour but de déterminer des frontières entre unités acoustiques conformément aux différentes étiquettes pré-établies à l'étiquetage du signal. Pour une segmentation syllabique, les marques de frontières entre les unités acoustiques représentent les marques de transitions entre les syllabes constituant l'énoncé contenu dans le signal.

Comme la segmentation SCL, il existe plusieurs catégories de méthodes qui permettent une segmentation du signal de parole en se basant sur une séquence de symboles linguistiques contenu dans le signal. Nous pouvons citer les méthodes :

- **basées sur le *Dynamic Time Warping* (DTW) -** la segmentation par DTW produit des marques de segmentation à partir de la phonétisation connue du signal à segmenter. Elle exploite l'algorithme basé sur la programmation pour comparer le signal de parole à segmenter à un signal de référence produit par son système de synthèse de parole. L'algorithme vise à minimiser la distorsion spectrale entre les séquences de trames acoustiques des deux signaux alignés [34].

- **basées sur les Modèles de Markov Cachés (MMCs) -** la segmentation de la parole par les MMCs s'effectuent en deux étapes : l'apprentissage des modèles (MMC) des unités acoustiques et le décodage (SCL ou ACL) ou l'alignement [35, 36].

- **basées les réseaux de neurones -** la segmentation par les réseaux de neurones, comme la précédente, est basée sur l'emploi de modèles pour déterminer les frontières entre unités acoustiques des signaux de la parole. La segmentation est effectuée en procédant d'abord à une estimation des paramètres des modèles d'apprentissage du corpus et ensuite à un alignement entre la séquence des trames du signal à segmenter et la séquence des modèles associés au contenu linguistique de l'énoncé. Pour ce cas, nous pouvons citer le travail de Vorstermans et al. [37] qui est basé sur l'utilisation des réseaux de neurones pour estimer les probabilités a posteriori des marques de frontières phonétiques et des classes phonétiques larges de la langue.

1.5 Reconnaissance automatique de la parole

Dans une chaîne de traitement automatique de la parole, l'étape qui suit la segmentation de la parole est la reconnaissance de la parole. La reconnaissance permet d'extraire les informations lexicales contenues dans des segments de parole obtenus à la segmentation. Dans cette section, nous décrivons les principes et modules de traitements d'un Système de Reconnaissance de la Parole (SRP) en précisant les algorithmes utilisés pour chaque méthode de reconnaissance.

1.5.1 Description générale

Généralement on distingue 3 types de systèmes de reconnaissance selon le mode d'élocution :

- système de reconnaissance de mots isolés - lorsque le locuteur marque une pause entre chaque mot prononcé dans une phrase ;
- système de reconnaissance de mots connectés - reconnaissance de mots déjà prédéfinis ;
- système de reconnaissance de la parole continue - reconnaissance d'un flux de parole (lorsque le locuteur s'exprime naturellement).

Le dernier type de SRP est celui qui nous a intéressé dans le cadre des contributions au Fongbe. En plus du mode d'élocution, les SRPs peuvent être aussi classés selon :

- la taille du vocabulaire utilisé et la difficulté de la grammaire (complexité du langage) ;
- la dépendance plus ou moins grande vis-à-vis du locuteur ;
- l'environnement protégé ou non (la robustesse aux conditions d'enregistrement).

Ils sont généralement conçus pour des applications spécifiques (commandes vocales, dictée automatique, identification du locuteur, compréhension d'un énoncé, etc.). Les plus décrits dans la littérature se basent sur une modélisation statistique pour transcrire sous forme textuelle un signal de parole. La figure 1.4 présente le schéma descriptif d'un SRP qui se résume généralement en deux unités principales : le module de décodage acoustico-phonétique (module acoustique sur le schéma) et le module de modélisation du langage (module lexical et syntaxique sur le schéma).

Les principales entités du système de reconnaissance par modélisation statistique sont :

- l'extraction de paramètres acoustiques - le premier traitement effectué sur le signal de parole consiste à extraire des paramètres caractéristiques qui sont mis en entrée au module acoustique ;
- le décodage acoustico-phonétique - le décodage est effectué par le module acoustique. Il permet, à partir de l'analyse du signal à reconnaître, de définir quel est l'élément acoustique (phonème, syllabe, mot, etc.) qui est le plus probablement produit. Pour chaque segment de parole, il produit une ou plusieurs hypothèses phonétiques (pour le cas présenté dans le schéma) associées à une valeur de probabilité.

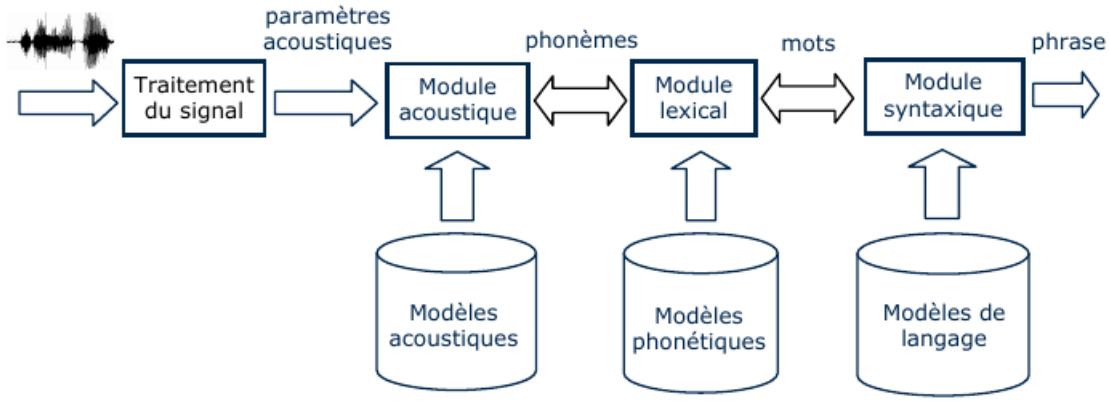


FIGURE 1.4 – Reconnaissance automatique de la parole par modélisation statistique [1]

- le module lexical - il permet de reconnaître les mots (information sur leur positionnement) contenus dans le signal de parole par des techniques soit à base de grammaire (représentée par les modèles phonétiques - dictionnaire de prononciation), soit purement statistiques, soit à base de grammaires probabilistes.

- le module syntaxique - il intègre des modèles du langage qui permettent de reconnaitre les dépendances grammaticales entre les mots. Elles sont définies par des contraintes syntaxiques qui sont mises à la sortie du module lexical pour extraire les relations grammaticales que les mots entretiennent entre eux.

Par la suite, nous allons présenter l'approche probabiliste adoptée par la plupart des Systèmes de Reconnaissance Automatique de la Parole (SRAP), les modules intégrés dans un système RAP et les mesures d'évaluation d'un système RAP.

1.5.2 Reconnaissance automatique de la parole par l'approche bayésienne

Le principe de la reconnaissance automatique de la parole par une approche probabiliste est de rechercher, étant donné un signal de parole X émis par un locuteur, la séquence de mots W^* la plus vraisemblable parmi un ensemble de séquences possibles W. Cela se traduit par la détermination de la séquence W^* qui maximise la probabilité d'émission de W sachant X. L'équation bayésienne appliquée à ce problème se traduire par l'équation suivante :

$$W^* = \arg\max_W P(W|X) \tag{1.1}$$

avec $P(W|X)$ la probabilité d'émission de W sachant X. L'application du théorème de Bayes permet d'obtenir l'équation suivante :

$$W^* = \arg\max_W \frac{P(X|W)P(W)}{P(X)} \tag{1.2}$$

avec $P(X)$ la probabilité d'observation de la séquence acoustique X. N'étant pas a priori calculable, cette probabilité peut être retirée de l'équation. L'expression revient donc à l'équation 1.3.

$$W^* = \arg\max_W P(X|W)P(W) \tag{1.3}$$

La reconnaissance de la parole consiste donc à maximiser le produit des probabilités :

- $P(X|W)$, probabiité d'une séquence d'observations acoustiques X sachant une séquence de mots W ;
- $p(W)$, probabilité a priori d'une séquence de mots.

La probabilité $P(X|W)$ est fournie par les modèles acoustiques tandis que la probabilité $p(W)$ est fournie par les modèles du language. Le module acoustique recherche les suites de phonèmes les plus probables à partir des observations acoustiques. Le module du langage via les modèles de langage va sélectionner la séquence qui a la plus grande probabilité d'apparition parmi toutes les séquences de mots. La figure 1.5 présente une vue d'un système de RAP probabiliste. On y retrouve les deux modules et le dictionnaire de prononciation de l'ensemble des mots disponibles d'une langue donnée.

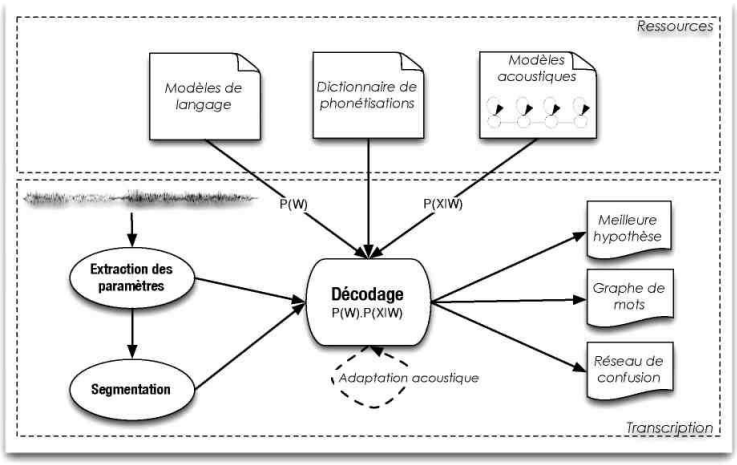


FIGURE 1.5 – Vue d'ensemble d'un système de RAP probabiliste

1.5.3 Module acoustique

Le module acoustique se charge des premiers traitements sur le signal recueilli dans un système RAP. Son rôle principal est de produire, à partir de paramètres acoustiques, des hypothèses phonétiques associées à une probabilité pour chaque unité de segment de parole. Il est aussi connu sous le nom de module de décodage acoustico-phonétique. Pour reconnaître les unités obtenues à la phase de segmentation de la parole (phonème pour la plupart), le module doit avoir appris à quoi ressemblent les réalisations acoustiques de ces unités en terme de vecteurs acoustiques. A partir de ces vecteurs, un modèle statistique est construit pour chaque unité considérée selon leur distribution. L'ensemble des modèles statistiques pour chaque unité contenue dans la parole constitue un modèle acoustique de la parole qui sera stocké dans le système RAP. Le module de décodage acosutico-phonétique est généralement composé de deux sous-modules. Le premier consiste à extraire les paramètres caractéristiques choisis pour représenter le signal et le deuxième consiste à apprendre les modèles d'unités acoustiques à partir des jeux de paramètres.

1.5.3.1 Extraction de paramètres

Dans un système RAP, le premier traitement effectué sur le signal de parole est l'extraction de paramètres acoustiques. Il consiste à calculer une séquence de vecteurs de caractéristiques qui fournissent une représentation compacte d'un signal de parole donné. Il est généralement réalisé en trois étapes. La première étape, appelée analyse de la parole, effectue l'analyse temporelle du spectre du signal de parole et génère les premières caractéristiques qui décrivent l'enveloppe du spectre de puissance d'intervalle de parole courte. La deuxième étape réalise une compilation de vecteurs de caractéristiques étendues composés de caractéristiques statiques et dynamiques. Enfin la dernière étape transforme ces vecteurs de caractéristiques étendues dans des vecteurs plus compacts et robustes constituant des paramètres représentatifs du signal de parole. Les principaux paramètres les plus utilisés dans la littérature sont :

- **Energie du signal -** elle est évaluée sur plusieurs trames de signal successives mettant en évidence les différentes variatons du signal. Elle correspond à la puissance du signal et se calcule comme suit :

$$E(\text{fenêtre}) = \sum_{n \in \text{fenêtre}} |n|^2 \tag{1.4}$$

- **Taux de passage par zéro (*zero crossing rate* en anglais) -** il correspond au nombre de fois que le signal, dans sa représentation amplitude/temps, passe par la

valeur zéro [38]. Pour un segment de parole donné, il est lié à la fréquence moyenne et est nul pour un segment silencieux. Il est calculé à partir de l'équation 1.5.

$$Z_m = \sum_n |sign[x(n)] - sign[x(n-1)]|w(m-n) \tag{1.5}$$

La fonction *sign* est définie comme suit :

$$sign[x(n)] = \begin{cases} 1 & x(n) \geq 0 \\ -1 & x(n) < 0 \end{cases} \tag{1.6}$$

$$w(n) = \begin{cases} 1/2N & 0 \leq n \geq N-1 \\ 0 & autrement \end{cases} \tag{1.7}$$

$x(n)$ est le signal de la fenêtre n.

- **Mel-Frequency Cepstral Coefficients (MFCCs)-** ce sont des coefficients cepstraux basés sur une échelle non linéaire de perception appelée Mel. C'est l'une des techniques d'extraction de caractéristiques les plus populaires utilisées dans la reconnaissance de la parole basée sur l'échelle de l'oreille humaine (échelle de Mel). C'est une technique qui extrait des paramètres de parole similaires à ceux utilisés par l'homme pour la parole auditive. Les MFCCs sont considérés comme étant du domaine fréquentiel et sont plus précis que les coefficients du domaine temporel. Les coefficients MFCCs sont une représentation du cepstre d'un signal fenêtré de courte durée dérivée de la transformée de Fourier rapide (FFT) du signal. Ils sont robustes et fiables contre la variation de locuteurs et des conditions d'enregistrement. Le signal de parole est d'abord divisé en trames temporelles composées d'un nombre arbitraire d'échantillons. Chaque trame temporelle est ensuite fenêtrée avec la fenêtre de Hamming pour éliminer les discontinuités au niveau des bords [39]. Les coefficients $w(n)$ d'une fenêtre de Hamming de longueur n sont calculés selon la formule :

$$w(n) = \begin{cases} 0.54 - 0.46cos(\frac{2\pi n}{N-1}) & 0 \leq n \leq N-1 \\ 0 & \text{autrement} \end{cases}$$

N est le nombre total d'échantillons et n l'échantillon courant. Après le fenêtrage, la transformée de Fourier rapide (FFT) est calculée sur chaque trame pour extraire des composantes fréquentielles du signal dans le domaine temporel. Cette transformée

est utilisée pour accélérer les traitements et conduit à l'obtention du spectre. Un banc de filtres logarithmiques est appliqué aux fenêtres transformées pour couvrir chacun une fréquence. A l'aide des filtres, chacune des fenêtres obtenues est convertie à l'échelle de Mel. Cette échelle est approximativement linéaire jusqu'à 1 kHz, et logarithmique à de plus grandes fréquences. La relation entre la fréquence de la parole et l'échelle de Mel s'établit comme suit :

$$M_{mels} = x \times log(1 + \frac{f_{Hz}}{y})$$

Dans la littérature, plusieurs valeurs ont été données aux variables x et y. Les plus couramment utilisées sont $x = 2595$ et $y = 700$. La dernière étape de calcul des coefficients MFCCs est le calcul de la transformation discrète de cosinus (DCT) sur les sorties du banc de filtre. Ceci conduit à l'obtention des coefficients MFCCs. Ainsi, pour chaque trame de parole, un ensemble de coefficients MFCCs est calculé. Cet ensemble est appelé vecteur acoustique et représente les caractéristiques phonétiquement importantes de la parole. Il est très utile pour une analyse plus approfondie et le traitement dans la reconnaissance de la parole.

- **Linear Predictive Coding (LPC) -** il consiste à synthétiser des échantillons de signal de parole à partir d'un modèle de système de production vocale et d'excitation. Il permet de prédire une valeur future du signal à partir d'une combinaison des valeurs précédentes. Le codage LPC est l'une des techniques les plus puissantes d'analyse de la parole qui a gagné en popularité en tant que technique d'estimation de formants [40]. Les coefficients LPC sont calculés en découpant le signal de la parole en de petites fenêtres de courte durée. La fenêtre de Hamming est ensuite appliquée sur les différentes portions de signal obtenues. L'application de la fenêtre de Hamming permet de diminuer la distorsion spectrale. Avec l'équation 1.8, le signal à l'instant n est prédit à partir des p échantillons précédents.

$$s(n) = \sum_{k=1}^{p} a_k \times s(n-k) + e(n) \tag{1.8}$$

La moyenne que constitue la somme pondérée du signal sur p pas de temps introduit une erreur car la parole ne constitue pas un processus parfaitement linéaire. Cette erreur est corrigée par l'introduction du terme $e(n)$. Le codage par prédiction linéaire consiste donc à déterminer les coefficients a_k qui minimisent l'erreur $e(n)$, ceci en fonction d'un ensemble de signaux constituant un corpus d'apprentissage.

- **Linear Frequency Cepstral Coefficient (LFCC) -** il s'agit d'une variante des MFCCs. La différence vient de l'utilisation d'un banc de filtres linéaire, contrairement à l'échelle de Mel des MFCCs.

- **Perceptual Linear prediction (PLP) -** la prédition linéaire perceptuelle PLP modélise la parole humaine en se basant sur le concept psychophysique de l'audition. Elle est développée par Hermansky [41]. La paramétrisation du signal en coefficients PLP est identique à celle du codage LPC, sauf que les caractéristiques spectrales du PLP sont transformées pour correspondre aux caractéristiques du système auditif de l'homme.

- **Autres paramètres -** au-delà des principales approches pour la paramétrisation citées ci-dessus, d'autres peuvent être trouvées dans la littérature telles que NPC (*Neural Predictive Coding* - extension non linéaire du codage LPC [42]), LSF (*Line Spectral Frequencies* - les fréquences de raies spectrales, issues du LPC,[43])...

1.5.3.2 Modélisation acoustique

La modélisation acoustique, dans un système RAP, conduit à la réalisation de modèles dont le rôle est d'estimer la probabilité qu'une séquence d'unités linguistiques (phonèmes, diphones, syllabes, mots, etc...) particulière ait généré le signal émis. Deux approches fondamentales sont couramment utilisées pour la modélisation acoustique : les approches statistiques et les modèles probabilistes. Des années 70 à nos jours, ces approches ont connu une nette amélioration, des performances remarquables, une robustesse au bruit et à la variabilité des locuteurs [36]. Il ressort de la littérature que les plus performantes et très utilisées sont celles basées sur les modèles de Markov cachés.

Les premiers systèmes sont basés sur la reconnaissance de mots à partir de patrons de vecteurs acoustiques [44]. Ce sont des systèmes qui apprennent les vecteurs acoustiques, à partir d'un corpus bien fourni, de plusieurs exemples de mots à reconnaître. La phase de reconnaissance est alors effectuée à l'aide de l'algorithme DTW [45]. Face à la grande variabilité que démontre un signal de parole, il est en effet peu probable de mesurer deux signaux de parole totalement semblables même si des mots identiques sont prononcés par le même locuteur. Cette variabilité due aux locuteurs et aussi au contexte d'enregistrement des données audio rend complexe cette approche et limite sa performance sur de très grands dictionnaires.

Par la suite, des unités acoustiques plus petites (en terme de durée) que les mots sont apparues et prises en compte pour déduire des modèles acoustiques plus performants que des exemples de mots. En plus de les considérer pour couvrir beaucoup plus de variations,

les modèles probabilistes ont aussi fait leur apparition dans la modélisation acoustique des systèmes RAP. Le principe consiste donc à déduire des modèles de phonèmes (ou combinaisons de phonèmes) plutôt que d'utiliser la reconnaissance par exemple.

Les modèles de Markov cachés sont aujourd'hui les plus utilisés pour la modélisation statistique acoustique. Le principe est de modéliser chaque unité de parole par un MMC. Ils sont beaucoup plus utilisés sur des modèles de phonèmes (ou polyphones) afin de limiter le nombre de paramètres à estimer. Nous ne présenterons pas l'approche markovienne mais une explication détaillée pourra être retrouvée dans [35]. Le formalisme des MMCs, a permis de développer, de nos jours, des algorithmes performants de modélisation acoustique. Ces algorithmes ont prouvé leur efficacité dans de nombreux domaines de la RAP et sont utilisés dans les outils de construction de systèmes de RAP les plus répandus tels que Kaldi RAP [13], HTK Toolkit [14], Julius [15] etc...

1.5.4 Modèle de language

Le modèle de language accompagne le modèle acoustique pour créer une cohérence linguistique entre les différents éléments acoustiques prononcés. Il permet d'introduire dans le système de RAP des connaissances linguistiques pour un décodage cohérent des phrases en entrée du point de vue syntaxique ou grammatical. On distingue, dans la littérature, les types de modèles de language suivants :

1. **les modèles à base de grammaires formelles -** ils sont réalisés par les experts linguistes et formulent des réponses en oui/non.

2. **les modèles probabilistes -** ils opèrent sur un corpus et décrivent automatiquement un langage à partir de l'observation du corpus. Ils sont beaucoup utilisés dans les systèmes de RAP pour leur réponse probabiliste qu'ils génèrent.

3. **les modèles hybrides -** ils combinent les approches à grammaires formelles et probabilistes. Ils sont aussi appelés les modèles à grammaire probabiliste.

Le modèle probabiliste est couramment utilisé dans les systèmes de reconnaissance automatique de la parole au détriment du modèle à base de grammaires car la génération manuelle d'un ensemble de règles décrivant une langue est un processus long, difficile et coûteux [46]. De plus, il s'intègre bien dans le processus de décodage et possède le formalisme statistique du problème de la reconnaissance automatique de la parole. Le

13. http ://kaldi-asr.org/
14. http ://htk.eng.cam.ac.uk/
15. http ://julius.osdn.jp/en index.php

modèle le plus largement utilisé dans la littérature est les modèles n-grammes. Ce sont des modèles à base de chaîne de Markov dont le rôle est de déterminer la probabilité *a priori* d'une séquence d'unités acoustiques (phonèmes ou mots). L'estimation de la probabilité s'effectue avec l'équation 1.9.

$$P(W_1^k) = P(w_1)\Pi_{i=2}^{k} P(w_i/h_i) \tag{1.9}$$

où $P(W_1^k)$ représente la probabilité de la suite d'unités h_i correspond à l'historique de l'unité considérée. L'idée des modèles n-grammes est d'estimer la probabilité *a priori* de la suite de mots w_k. On obtient ainsi l'équation 1.10 où l'historique du mot est représenté par les (n-1) mots qui le précèdent.

$$P(W_1^k) = P(w_1)\Pi_{i=2}^{n-1} P(w_i/w_1, ..., w_{i-1}) \tag{1.10}$$

où $P(w_i/w_1, ..., w_{i-1})$ est la probabilité d'avoir le mot w_i sachant les observations $w_1, ..., w_{i-1}$. A cause de la longueur de l'historique, le calcul de la probabilité $P(w_i/w_1, ..., w_{i-1})$ devient impossible si l'historique du mot ne se réduit pas à un sous-historique de taille réduite et fixe n. Généralement dans les système de RAP, n prend les valeurs 2 ou 3 et on parle de modèle bi-grammes ou tri-grammes. Dans le cas d'un modèle tri-grammes, l'équation précédente peut être réécrite de la manière suivante :

$$P(W_1^k) = P(w_1).P(w_2/w_1)\Pi_{i=3}^{k} P(w_i/w_{i-2}w_{i-1}) \tag{1.11}$$

La probabilité d'apparition d'un mot est alors obtenue par le critère de maximum de vraisemblance qui est défini comme suit :

$$p(w_i/w_{i-n+1}, ..., w_{i-1}) = \frac{\#(w_{i-n+1}, ..., w_{i-1}, w_i)}{\#(w_{i-n+1}, ..., w_{i-1})} \tag{1.12}$$

#(...) désigne le nombre d'occurrences de la suite de mots $w_1...w_k$ dans le corpus d'apprentissage. Le calcul de la probabilité de l'équation 1.11 peut rencontrer des problèmes si les séquences de mots à reconnaître n'apparaissent pas dans le corpus d'apprentissage pouvant conduire ainsi à des probabilités nulles. Ce problème trouve sa solution dans le lissage des probabilités, par interpolation ou par repli [47].

1.5.5 Evaluation des systèmes de reconnaissance automatique de la parole

Pour une comparaison de performances entre deux ou plusieurs systèmes RAP, il convient de les évaluer sur les mêmes données de test et dans les mêmes conditions de réalisation. Généralement, ils sont évalués en terme d'erreur mot (*Word Error Rate - WER*). Le *WER* est l'une des mesures les plus utilisées dans l'évaluation de performances d'un SRAP. Il prend en compte la sortie obtenue avec le système et la transcription de référence obtenue manuellement. Pour son calcul, le WER intègre les erreurs :

- de substitution (s) : mot reconnu à la place d'un mot de la transcription manuelle ;
- d'insertion (i) : mot reconnu inséré par rapport à la transcription de référence ;
- de suppression (d) : mot de la référence oublié dans l'hypothèse fournie par le système de RAP.

Le calcul s'effectue selon la formule suivante :

$$WER = \frac{s + i + d}{\text{nombre de mots de la référence}} \tag{1.13}$$

Le WER est calculé pour évaluer la capacité du système à décoder la parole absente de la base d'apprentissage. Ainsi les données utilisées pour l'évaluation ne devraient être pas utilisées pendant la phase d'apprentissage. C'est une fonction d'erreur dont la valeur idéale est 0. $WER = 0$ si les hypothèses obtenues en sortie sont identiques aux références. $WER = 100$ signifie que chacun des mots entre les hypothèses et les références est différent à condition que le nombre de mots à la sortie du système soit égal au nombre de mots dans la référence.

Une mesure alternative telle que le taux d'erreur phrase (SER) est aussi utilisée pour l'évaluation de la qualité d'un système RAP. Le SER mesure le nombre de phrases décodées qui correspondent aux références dans un jeu de données test. Le WER et le SER admettent une corrélation qui est faible quand la performance globale du système est faible. Dans ce cas, le SER est généralement proche de 100% et les changements dans le WER ont peu d'influence sur le SER.

Conclusion

Dans ce chapitre, il a été introduit le cadre d'étude des travaux réalisés au cours de cette thèse et l'état de l'art sur la segmentation et la reconnaissance automatique de la parole. Il

a été aussi présenté une description détaillée du Fongbe et sa mesure d'informatisation qui révèle un manque important de ressources linguistiques liées au traitement automatique de la parole. Son indice ($\sigma = 6,02$) justifie notre interêt à le doter d'un système de reconnaissance de la parole.

Chapitre 2

Collecte des ressources linguistiques du Fongbe

"e kpɔ́n nu gbannya gbannya nyönaxo tɔ́n ɔ́, e na xwle zö xi a"

Sommaire

Introduction

Le développement du système de reconnaissance automatique de la parole continue est réalisé à partir d'une grande quantité de données qui doivent contenir à la fois des signaux de parole (pour la modélisation acoustique du système) et les données textuelles (pour le modèle de langage du système). Il devient un défi et très difficile quand la langue cible est une langue qui ne dispose pas des ressources linguistiques (langue peu dotée). Dans ce chapitre, nous décrivons la méthodologie utilisée pour recueillir les textes et les signaux audio de la langue Fongbe et les différents corpora réalisés pour la construction du système de reconnaissance automatique de la parole en Fongbe.

2.1 La linguistique pour la reconnaissance automatique de la parole

2.1.1 Le vocabulaire

Le vocabulaire est utilisé pour la transcription automatique dans un système RAP. Il comprend une liste close de N unités lexicales (phonèmes, graphèmes ou mots) pouvant être reconnus par un système RAP. Il influence le système de reconnaissance de part sa couverture lexicale. Cette dernière représente le taux d'unités lexicales du vocabulaire présente dans un corpus de textes. Elle est meilleure quand le contenu du corpus est bien fourni en nombre important d'occurences de chaque unité lexicale.

La plupart des méthodes de recueil du vocabulaire sont manuelles et consistent à sélectionner les unités lexicales en appliquant des seuils sur leurs fréquences observées sur des corpora d'apprentissage [48]. Le recueuil consiste donc à sélectionner les N mots les plus fréquents. Pour maximiser la couverture lexicale, plusieurs corpus sont utilisés sans que les contenus ne se chevauchent. Le recueil est soumis à deux contraintes fondamentales que sont la taille N et le choix des unités lexicales. Il est recommandé de rechercher un bon compromis entre la couverture lexicale et l'augmentation de la taille du vocabulaire pour limiter le nombre de ressources requises pour un système de reconnaissance. Un autre problème des systèmes RAP est la reconnaissance des mots dits inconnus qui ne sont pas présents dans les corpora d'apprentissage. Ils sont appelés Mots-Hors-Vocabulaire (MHV). Un mot absent du vocabulaire n'est pas reconnu par le système et est remplacé par un ou plusieurs mots connus qui s'en rapprochent acoustiquement. Le problème est que l'insertion de mots approchants peut engendrer des problèmes si les contextes de ce qui a été produit ne sont pas compatibles avec le mot à l'origine. Pour pallier à cela, il est

possible d'inclure des mots courants pour minimiser le nombre de mots hors vocabulaire que sont susceptibles de prononcer les utilisateurs.

Pour évaluer un vocabulaire V, on se concentre sur la couverture lexicale en calculant le MHV sur un corpus de textes C. L'équation pour l'évaluation est donnée par l'expression 2.1

$$T_{MVH}(V, C) = \frac{\text{nombre d'occurences dans C} \notin V}{\text{nombre d'occurences total dans C}} \tag{2.1}$$

La constructon manuelle du vocabulaire pour les sytèmes de reconnaissance est devenue une tâche fastidueuse qui de nos jours laisse place aux méthodes automatiques. Elles assurent une construction rapide du vocabulaire en se basant sur des contextes locaux appliqués aux corpora d'apprentissage [48].

2.1.2 Dictionnaire de prononciation

Le dictionnaire de prononciation contient les prononciations de toutes les formes de mots des données de transcription du corpus de textes et de l'audio. Il regroupe tous les mots les plus fréquents d'une langue et leurs prononciations. L'objectif est de transcrire en petites unités lexicales (phonèmes ou graphèmes) tout ce qui peut être dit. Le dictionnaire permet donc de pallier à toutes les éventualités en termes de prononciations. Chaque entrée consiste en un mot et sa prononciation séparés par un espace. La prononciation consiste en une concaténation des symboles de phones séparés par des espaces. Les prononciations multiples par mot peuvent être incluses dans le dictionnaire pour représenter explicitement la variation de prononciation.

Dans la littérature, plusieurs approches sont utilisées pour construire un dictionnaire de prononciations pour une langue donnée. On distingue l'approche guidée par les données qui utilise généralement des données d'apprentissage et l'approche à base de règles qui necessite une expertise linguistique [49]. La première utilise un corpus d'apprentissage pour apprendre automatiquement des liens entre les graphèmes pour écrire un mot et les phonèmes pour représenter la prononciation du mot. Elle s'appuye sur un dictionnaire de mots déjà phonétisés manuellement. La deuxième ne necessite pas de techniques d'apprentissage mais plutôt une expertise sur la langue et ses règles de phonétisation. Elle a comme avantage de pouvoir mieux contrôler la qualité de la construction du dictionnaire de prononciations. En cas d'erreur, il est possible d'ajouter une nouvelle règle de phonétisation.

2.2 Corpus de texte

Un corpus de textes est une collection de données textuelles en grande quantité utiles pour le développement des systèmes RAP. Il est indispensable pour la modélisation du language. Sa construction consiste à collecter les données textuelles, sur un grand nombre de sources d'informations (généralement des sites web), dans une langue cible. Une fois extraites des sources, ces données sont ensuite filtrées selon des critères fixés pour être incluses dans un corpus. Le problèmes avec les langues peu dotées comme le Fongbe, est qu'il n'existe presque pas de sources d'informations à grandes échelles, comme les langues bien dotées, pouvant permettre une collecte considérable de données textuelles.

2.2.1 Recueil d'un corpus de textes pour des langues peu dotées

Un problème majeur, constituant un obstacle lors de l'élaboration d'un système RAP pour une langue peu dotée, est le manque de données textuelles. Selon l'IWS [16] (Institut World Stats), seules une dizaine de langues (avec l'anglais en tête) disposent d'une large diffusion sur Internet avec des sites web bien fournis.

L'Internet constitue la première et principale source de collecte des données textuelles pour construire un corpus de textes. Cependant avec les langues peu dotées, on rencontre des problèmes comme : un nombre peu important de sites Internet, une très faible vitesse de transmission (problème lié aux faibles bandes passantes dans ces pays) qui ne facilitent pas la collecte et rend le corpus moins riche. Plusieurs travaux ont été réalisés dans ce cadre pour permettre la construction d'un corpus de textes dans une langue quelconque. Parmi ces travaux, nous pouvons citer ceux qui proposent l'utilisation d'un robot web [50] et des moteurs de recherche [51]. L'auteur dans [50] présente un robot qui, à partir d'un jeu d'heuristiques et de filtres, parcours des pages web dans le but de collecter des informations pertinentes exploitables en français. *CorpusBuilder*[17] est une application de collecte de textes proposée par [52]. Elle implémente une méthode de recueil automatique de textes par des requêtes lancées sur un moteur de recherche.

2.2.2 Collecte de données textuelles en Fongbe

Il a été montré dans les sous-sections précédentes que la manière standard la plus courante pour construire un corpus de textes est la collection des textes depuis des sites web. Comme il a été justifié dans le chapitre 1, le Fongbe est une langue peu dotée qui

16. Source : www.Internetworldstats.com/stats.htm
17. http ://www.cs.cmu.edu/ TextLearning/corpusbuilder/

dispose d'un nombre très limité de sites web comparé aux langues telles que le Wolof, le haoussa et aussi l'arabe, qui ont une très grande couverture sur Internet et ne souffrent pas d'un manque de données textuelles. Pour collecter les donnés textuelles en Fongbe pour en faire un corpus, deux approches ont été utilisées pour extraire sur Internet des textes. Dans un premier temps, nous avons recherché sur Internet, avec des requêtes formatées sur le Fongbe, des sites qui ont des contenus (expressions ou phrases) en Fongbe. La couverture sur intenet du Fongbe se limite à quelques sites web qui renferment des textes provenant de la Bible traduite en Fongbe et aussi de la vie courante. Le site de la Bible est le seul retrouvé qui dispose d'un nombre considérable de textes (environ 30.000 phrases) et d'une bonne vitesse de transmission. On n'y retrouve aussi des sites web abordant des sujets comme l'éducation, les chants, les contes et la déclaration universelle des droits de l'homme. Une fois les sites web recensés, l'approche RLAT (Rapid Language Adaptation Toolkit) [53] a été utilisée pour explorer et extraire les textes en Fongbe depuis une page web donnée. RLAT fournit des méthodes novatrices et des outils web interactifs pour permettre aux utilisateurs de développer des modèles de traitement de la parole, pour recueillir des données de parole et de texte appropriées pour construire ces modèles, ainsi que d'évaluer les résultats permettant des améliorations itératives.

Pour améliorer la quantité de textes obtenus à partir des liens HTML des sites web, nous avons ajouté des textes obtenus à partir des fichiers PDF qui couvrent un grand nombre de citations et chansons en Fongbe. Après avoir extrait tout le contenu textuel des pages web et fichiers pdf, il a été procédé à un nettoyage et à une normalisation des textes suivant les tâches suivantes :

1. la suppression des balises et codes HTML ;
2. la supression des lignes vides et des ponctuations ;
3. la conversion des textes en Unicode [18] ;
4. la suppression des pages et lignes contenant des textes en une langue autre que le Fongbe ;
5. la transcription des caractères spéciaux et des chiffres ;
6. la suppression des lignes en double.

Nous avons construits des outils spécifiques pour réaliser chacune de ces tâches de récupération et de traitement du corpus de textes pour le Fongbe. L'Unicode est utilisé

18. http ://www.unicode.org

à cause de son expansion à travers les systèmes d'exploitations et les navigateurs. Les caractères Fongbe extraits des sites web ont été encodé avec le format UTF-8 [19] de la norme Unicode. L'UTF-8 est le format de transformation des caractères Unicode en ASCII le plus commun pour les applications liées à l'Internet. Il assure aussi une compatibilité avec les manipulations simples de chaînes en ASCII dans les langages de programmation.

Pour réaliser les travaux dans le cadre de cette thèse, nous avons construit un corpus de textes en Fongbe d'une taille de 8, 7 Mo contenant environ 34.653 phrases recueillies à partir des quelques sites web et documents disponibles en Fongbe (voir tableau 2.1). De façon distincte, 10.130 mots ont été collectés pour servir à la construction du vocabulaire et du dctionnaire de prononciations.

TABLE 2.1 – Contenu du corpus de textes.

Source	Sites web	Textes	Nombre de phrases
1	`http://www.fongbe.fr`	Textes de la vie courante	1,500
2	`http://unicode.org/udhr/d/udhr_fon.txt`	Déclaration universelle des droits de l'homme	92
3	`http://ipedef-fongbe.org/`	Textes sur l'éducation, chansons et contes	2,200
4	`http://www.vodoo-beninbrazil.org/fon.html`	Textes sur l'éducation	1,055
5	`https://www.bible.com/fr/bible/813/dan`	La Bible	29,806

2.3 Corpus audio

Deux différents corpora audio ont été construits pour permettre la réalisation des différentes tâches de reconnaissance automatique de la parole. L'un contient les phonèmes du Fongbe et l'autre, les paroles continues. Dans cette section nous décrivons dans un premier temps la procédure de collecte des données de parole et dans un second temps la description des différents corpus réalisés.

2.3.1 Protocole de collecte

Pour construire les corpora audio, nous avons enregistré un échantillon de personnes choisies suivant des critères comme l'âge, l'origine et le sexe. Pour approcher les réalités de l'élaboration d'un système RAP, nous avons choisi d'enregistrer des locuteurs natifs ou

19. Unicode Transformation Format : http ://www.ietf.org/rfc/rfc2279.txt

étrangers du Fongbe dans les environnements de la vie courante (salle de classe, chambre à coucher, bureaux, etc.) loin du contexte de laboratoire. Les non natifs (étrangers) ont été considérés pour intégrer dans les corpora toutes les formes de prononciations des phonèmes et mots du Fongbe et assurer une fluidité dans la couverture lexicale. La réalisation des deux corpora a suivi le même critère de choix mais avec des locuteurs différents. Le premier corpus contient un ensemble de 32 phonèmes prononcés par les élèves de différents niveaux scolaires, les étudiants du département des langues de l'UAC et d'autres locuteurs analphabètes tels que des commerçants et artisans. Chacun des locuteurs a prononcé les 32 phonèmes en marquant une courte pause entre deux prononciations. Ensuite, avec le module d'acquisition et de traitement, les signaux recueillis ont été segmentés manuellement en de petits signaux contenant un phonème. Des traitements ont été effectués sur chaque signal afin d'éliminer les bruits de fond additionnels et les signaux indésirables.

le deuxième corpus renferme des signaux de parole continue en Fongbe. Il a été extrait d'un site web [20] un ensemble de signaux audios (environ 1000 phrases lues). Ces signaux ont été ré-échantillonnés et introduits dans le module d'acquisition pour permettre aux locuteurs d'effectuer un "re-speaking" (écouter et répéter). Ceci à conduire à l'obtention d'un corpus de parole continue d'une grande taille d'environ 10 heures. Les locuteurs enregistrés sont aussi des élèves de divers âges, des étudiants et un grand nombre d'analphabètes qui, pour la plupart, sont des natifs de la langue. la figure 2.1 montre une répartition selon la proximité du locuteur avec la langue, l'âge et le niveau scolaire des locuteurs enregistrés.

2.3.2 Construction du corpus d'unités phonémiques : *FongbePhones-FLDataset*

Le corpus d'unités phonémiques (*FongbePhones-FLDataset*) comprend des enregistrements de 32 phonèmes prononcés par les locuteurs étrangers et natifs de la langue Fongbe. Les enregistrements, effectués avec le logiciel *Audacity* [21] dans des environnements réels de la vie courante, ont été digitalisés à 44.100Hz en deux canaux de voix et sont rangés en deux classes (les consonnes et les voyelles) selon les identifiants des locuteurs. Nous avons donc obtenu un corpus de 4929 signaux (classés par données d'apprentissage et de test - voir tableau 2.3) de parole prononcée (environ 4 heures de lecture) pour 174 locuteurs dont les âges sont compris entre 9 et 45 ans incluant 53 femmes (enfants et adultes) et 119 hommes (enfants et adultes). Le tableau 2.2 présente le nombre d'échantillons enregistrés

20. www.fongbe.fr

21. Audacity est un logiciel open source permettant d'enregistrer et de mixer différents sons - www.audacity.fr

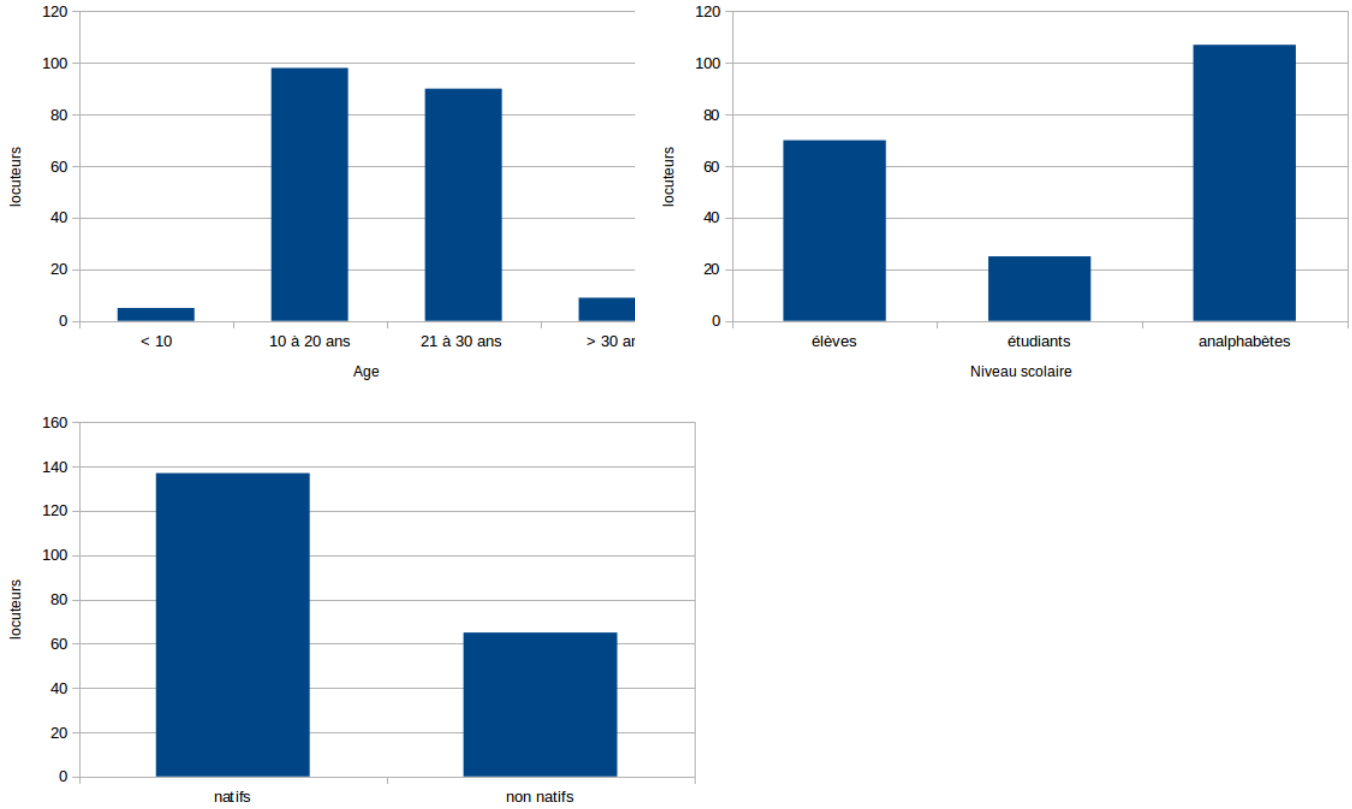


FIGURE 2.1 – Répartition des locuteurs selon leur proximité avec la langue, l'âge et le niveau scolaire.

par phonème de chaque classe.

TABLE 2.2 – Nombre d'échantillons par phonème.

Voyelle		Consonne			
Phoneme	Nombre	Phoneme	Nombre	Phoneme	Nombre
/ i/	172	/ f/	156	/ ŋ/ (ny)	155
/ e/	143	/ v/	161	/ c/	174
/ ɛ/	160	/ b/	167	/ j/	157
/ u/	142	/ t/	170	/ x/	125
/ḭ/	114	/ s/	169	/ l/	173
/ɛ̰/	140	/ z/	169	/ ɣ/ (h)	133
/ ɔ/	157	/ m/	152	/ d/	150
/ṵ/	118	/ n/	154	/ gb/	154
/ a/	164	/ kp/	158	/ g/	163
/a̰/	138	/ gb/	154	/ k/	167
/ o/	157	/ w/	160	/ p/	157
/ɔ̰/	158				

TABLE 2.3 – Nombre d'échantillons par classes de phonèmes.

	Nombre d'échantillons	
	C_1	C_2
Données d'apprentissage	2831	1112
Données de test	493	493

Les signaux du corpus sont pour la plupart dégradés par la présence d'un bruit additif de fond pour avoir été enregistrés dans les environnements réels de la vie courante. Après traitement avec *Audacity*, les signaux ont subi un débruitage supplémentaire dans le but d'améliorer le rapport signal sur bruit et de réduire l'effet du bruit de fond, nous avons utilisé la transformée en ondelettes de Daubechies basée sur la méthode de seuillage multi-niveaux détaillée dans [54]. Le débruitage par la transformée en ondelettes est basé sur deux fonctions de seuillage dont les fonctions *Hard* et *Soft*. Le seuillage *Hard* est le processus habituel de mise à zéro des coefficients dont les valeurs absolues sont inférieures au seuil. Le seuillage *Soft* est une extension du seuillage *Hard*. En plus de la mise à zéro des coefficients dont les valeurs absolues sont inférieures au seuil, les coefficients non nuls sont diminués vers zéro. Pour débruiter les signaux des phonèmes enregistrés, nous avons utilisé la méthode *Soft* avec un seuil basé sur les coefficients détaillés de la transformée en ondelettes de Daubechies. La fonction est définie comme :

$$X_{soft} = \begin{cases} sign(X) & (|X| - |\tau|) \quad \text{si} |x| > \tau \\ 0 & autrement \end{cases} \tag{2.2}$$

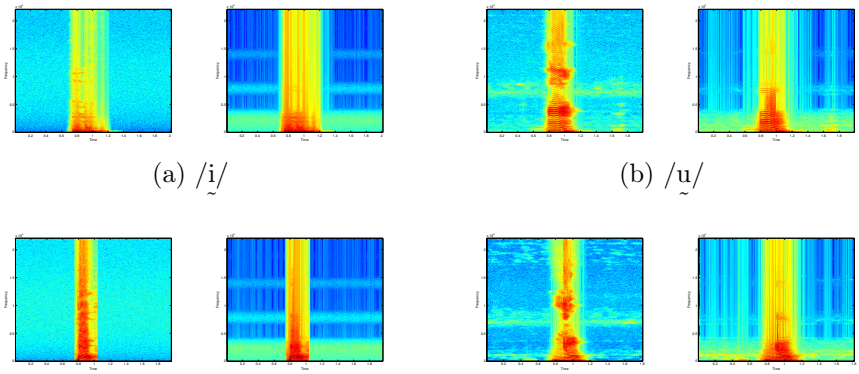
L'algorithme de débruitage est résumé comme suit :

- appliquer la transformée en ondelettes jusqu'à 8 niveaux au signal de bruit pour produire les coefficients d'ondelettes bruités ;
- appliquer le seuillage *Soft* aux coefficients d'ondelettes détaillés avant de choisir un seuil approprié ; le seuil utilisé est celui de Donoho et Johnstone [55] qui est défini comme suit :

$$\tau = \sigma * \sqrt{2log(N)} \tag{2.3}$$

  avec N la longueur du signal et σ le niveau de bruit ;
- appliquer la transformée en ondelettes discrète inverse aux coefficients d'ondelettes qui sont obtenus à l'étape précédente ; Ceci produit le signal débruité ;

La figure 2.2 montre la représentation fréquentielle des signaux /ḭ/, /ṵ/ de la classe des voyelles et / kp/, / ŋ/ de la classe des consonnes. Les images à gauche montrent les signaux enregistrés avec le bruit de fond et les images à droite montrent les signaux débruités avec la méthode des ondelettes de Daubechies.



(a) /ḭ/ (b) /ṵ/

(c) /kp/ bruité (d) / ŋ/ bruité

FIGURE 2.2 – Représentation fréquentielle des signaux bruités (à gauche) et débruités (à droite) des phonèmes /ḭ/, /ṵ/, /kp/, / ŋ/.

2.3.3 Construction du corpus de parole continue : *FongbeSpeech-FLDataset*

Comme un corpus audio de parole continue n'est pas disponible pour le Fongbe, nous avons procédé à la collecte des signaux de parole pour des fins de construction du premier système de reconnaissance automatique de la parole en Fongbe. Nous avons ainsi effectué la tâche fastidieuse de l'enregistrement des textes (extraits du corpus de textes) prononcés par des locuteurs natifs (dont 8 femmes et 20 hommes) du Fongbe dans des environnements non bruyants. Nous avons enregistré, à 16 Khz, 28 locuteurs natifs qui ont prononcé environ 1500 phrases (de la vie quotidienne) regroupés en 3 catégories. Une catégorie de textes est lue par plusieurs locuteurs et contient des textes différents des contenus des autres catégories. Les enregistrements ont été réalisés avec l'application Android LigAikuma [56] développée par l'équipe GETALP du Laboratoire d'Informatique de Grenoble (France). Lig-Aikuma est une version étendue de l'application Aikuma [22] qui permet d'enregistrer de la parole et de la re-prononcer. Lig-Aikuma, destiné aux informaticiens et linguistes, est conçu pour rendre plus facile et aisée la collecte des données.

22. http ://www.aikuma.org/

Il a été installé sur les mobiles des locuteurs possédant un smartphone afin de leur permettre d'enregistrer les phrases une fois chez eux et de nous envoyer les fichiers audio et les transcriptions. Dans l'ensemble, 10 heures de données de parole recueillies. Les données obtenues ont été divisées par catégories de textes menant ainsi à une première configuration du corpus (FSC1). FSC1 contient deux catégories de textes pour les données d'apprentissage (environ 8 heures) et une catégorie pour les données de test (environ 2 heures). La deuxième configuration du corpus (FSC2) est le résultat de la division des données par locuteurs dont 20 locuteurs (environ 8 heures) pour les données d'apprentissage et 8 locuteurs (environ 2 heures) pour les données de test (voir tableau 2.4). Les données ont été divisées de cette façon afin d'une part, de nous assurer assurer que la catégorie de textes qui apparait dans les données de test ne sera pas présente dans les données d'apprentissage et d'autre part, de réduire le risque d'avoir des locuteurs qui se chevauchent entre les données d'apprentissage et de test.

TABLE 2.4 – Contenu du corpus de parole Fongbe.

	Segments de parole	Phrases	Durée	Categories	Locuteurs
		FSC1 - config			
Données d'apprentissage	8.234	879	7 heures 35 mn	C2 & C3	25
Données de test	2.168	542	1 heure 45mn	C1	4
		FSC2 - config			
Données d'apprentissage	8.651	1.421	8 heures	C1, C2 & C3	21
Données de test	1.751	1.410	2 heures	C1, C2 & C3	7

2.4 Conclusion

Il est présenté dans ce chapitre les deux corpora audio (*FongbePhones-FLDataset* et *FongbeSpeech-FLDataset*)réalisés pour rendre possible la construction du premier système de reconnaissance automatique de la parole continue en Fongbe. Le premier corpus est utilisé pour les tâches d'analyses acoustiques des différents sons du Fongbe et de la reconnaissance automatique des phonèmes isolés du Fongbe [57, 5]. Le deuxième corpus est utilisé pour des tâches de segmentation automatique de la parole continue en de petits segments syllabiques [3, 4] et de la reconnaissance automatique de la parole continue avec l'outil Kaldi ASR [2]. Il a été aussi présenté la méthodologie adoptée pour la collecte des données et la description du corpus de textes construit pour la modélisation du language.

Deuxième partie

Contributions à la reconnaissance automatique des phonèmes isolés du Fongbe

Chapitre 3

Analyse acoustique des sons du Fongbe

"*nyönú đé klɔ́ dokwín bö asú*
tɔ́n be yi sa : nuklɔ́sá hán
me đé a"

Sommaire

Introduction

Le Fongbe à l'instar des autres langues dispose d'un inventaire de sons (phonèmes) caractérisés par des configurations articulatoires ou acoustiques communes à toutes les langues du monde. Ses sons, définis comme des éléments irréductibles de la chaine parlée, sont regroupés en de classes (voyelles et consonnes) dans lesquelles ils sont décrits selon les critères articulatoires du latin et de l'Alphabet Phonétique International (IPA). Ainsi, le fongbe est caractérisé par une série de voyelles (orales et nasales), de consonnes (orales et nasales) et de tons utilisés pour la production des mots et des énoncés. Dans ce chapitre, nous nous sommes interressés à l'étude des sons du Fongbe dans un contexte isolé

partant de l'analyse formantique des voyelles à la description acoustique des consonnes. Notre objectif est de produire un panorama de référence de caractéristiques acoustiques permettant de distinguer l'ensemble des sons du Fongbe.

3.1 Description acoustique des voyelles

Une voyelle est reconnue comme étant un son musical presque pur et se caractérise principalement par la présence de zones de fréquences où les harmoniques sont particulièrement intenses (formants). Les valeurs des formants diffèrent d'une voyelle à une autre selon la configuration des cavités bucco-pharyngales du locuteur. Il est donc reconnu que la production des voyelles du Fongbe dépend des facteurs comme : le locuteur, le débit, l'attitude du locuteur, son état émotif, le ton, etc. Le caractère tonal du Fongbe fait que chacune de ses voyelles dans un mot ou dans un énoncé est affectée d'un ton pertinent. Ce qui va occasionner des valeurs spécifiques de formants. Les caractéristiques formantiques calculées sur l'ensemble des occurences de chaque voyelle présente dans le corpus *FongbePhones-FLDataset* sont représentatives vu qu'elles prennent en compte 174 locuteurs de différents âges, dans différentes attitudes et émotions différentes.

3.1.1 Le timbre vocalique

Le timbre vocalique est caractérisé par la position fréquentielle et la variation au sein des trois premiers formants du profil formantique d'une voyelle. Nous avons caractérisé les voyelles du Fongbe par les formants F1, F2 et F3 qui représentent respectivement le degré d'aperture de la mandibule, la position de la langue et la position des lèvres. Le système vocalique du Fongbe comprend douze timbres : sept (7) voyelles orales avec quatre (4) degrés d'aperture et cinq (5) voyelles nasales pour trois (3) apertures (voir tableau 1.1). Il ressort de ce système vocalique, dix (10) voyelles qui possèdent les mêmes configurations articulatoires que le français :

- trois voyelles orales antérieures qui se prononcent en déplaçant le bout de la langue vers l'avant de la bouche ;
  - / i/ prononcée comme /i/ français
  - / e/ prononcée comme /é/ fermé français
  - / ɛ/ prononcée comme /è/ ouvert français
- trois voyelles orales postérieures qui se prononcent en déplaçant la langue à l'arrière de la bouche ;

  - / u/ prononcée comme /ou/ français
  - / o/ prononcée comme /o/ fermé français
  - / ɔ/ prononcée comme /o/ ouvert français

- une voyelle centrale ouverte (/ a/ prononcée comme /a/ français) qui se prononce en ouvrant largement la bouche et en positionnant horizontalement la langue ;
- trois voyelles nasales qui se prononcent en laissant passer de l'air dans la bouche et dans le nez.
  - /ɛ̰/ (ant.) prononcée comme / ɛn/ français
  - /ɔ̰/ (post.) prononcée comme / ɔn/ français
  - /a̰/ (cent.) prononcée comme /an/ français

Les deux dernières voyelles nasales (ant. /ḭ/ prononcée comme /in/ et post. /ṵ/ prononcée comme /un/) du système vocalique restent typiques au Fongbe. Les voyelles du fongbe peuvent être prononcées de quatre façons différentes (dans un mot ou une phrase) selon les quatres formes tonales (bas, moyen, haut ou modulé bas-haut) et peuvent se retrouver au début ou à la fin d'une syllabe.

3.1.2 Les fréquences formantiques

La configuration articulatoire de chaque voyelle occasionne de valeurs spécifiques de formants correspondant à la forme que prend le tractus vocal pour chaque voyelle. La forme du tractus vocal va induire une fréquence de résonnance dans la production des voyelles : plus sa taille est importante, plus sa fréquence est basse. Le tableau 3.1 présente les valeurs moyennes des fréquences formantiques calculées à partir des signaux de phonèmes du corpus *FongbePhones-FLDataset*. Nous avons caractérisé chaque voyelle du Fongbe par les valeurs formantiques moyennées sur toutes les occurences des différents locuteurs présents dans le corpus afin de prendre en compte les différentes configurations articulatoires.

Ces résultats (voir tableau 3.1) montrent une large variation des trois premiers formants de la configuration acoustique de chaque voyelle. L'aperture de la mandibule (**F1**) varie de 311 Hz (ant. / i/) à 681 Hz (cent. / a/) pour les voyelles orales et de 315 Hz (ant. /ḭ/) à 610 Hz (cent. /a̰/) pour les voyelles nasales. Le deuxième formant (**F2**) indiquant le mouvement de la langue en position avant ou arrière prend des valeurs allant de 697 Hz (post. / u/) à 2238 Hz (ant. / i/) pour les voyelles orales et de 406Hz (post. /ṵ/) à 1630Hz (post. /ḭ/). Le troisième formant (**F3**) influencé par l'arrondissement des lèvres

TABLE 3.1 – Valeurs moyennes des formants F1, F2 et F3 pour chaque voyelle du système vocalique du Fongbe

			Moyenne sur F1	Moyenne sur F2	Moyenne sur F3
Voyelles orales	ant.	i	311	2238	3174
		e	353	2042	3096
		ɛ	530	1255	3081
	post.	u	326	697	3083
		o	380	822	2727
		ɔ	555	1051	2696
	cent.	a	681	1404	2789
Voyelles nasales	ant.	ḭ	315	1630	2738
		ɛ̰	405	1362	2870
	post.	ṵ	325	406	2720
		ɔ̰	501	705	2711
	cent.	a̰	610	1015	2769

présente des valeurs qui varient de 2696Hz (post. / ɔ/) à 3174 Hz (ant. / ɔ/) pour les voyelles orales et de 2711 Hz (post. /ɔ̰/) à 2870 Hz (ant. /ɛ̰/).

Au niveau de **F1**, les voyelles orales (ant. et post.) / i/, / u/, / e/ et / o/ ont des valeurs plus basses pendant que les voyelles orales (ant. et post.) / ɛ ɔ/ et la voyelle orale centrale / a/ ont les valeurs les plus hautes. A chaque niveau d'aperture, une voyelle antérieure présente un F1 plus grand que celui d'une voyelle postérieure. Les voyelles / i/ et / a/ sont respectivement les extrémités basses et hautes de l'intervalle de fréquences de l'aperture du mandibule (**F1**). Ces mêmes remarques sont observées avec les voyelles nasalisées. Ceci indique que les voyelles avant sont / i/, / e/, / ɛ/, /ḭ/ et /ɛ̰/ et les voyelles arrières répertoriées sont / u/, / o/, / ɔ/, /ṵ/ et /ɔ̰/. Au niveau du formant **F2**, les voyelles antérieures présentent des fréquences plus élevées par rapport aux voyelles postérieures qui se réalisent avec des fréquences basses. La seule différence entre les voyelles de la même classe (ant. ou post.) est la valeur du troisième formant **F3**. Le **F3** des voyelles / u/ ($F3 = 3083Hz$) et /ṵ/ ($F3 = 2720Hz$) est respectivement plus bas que celui des voyelles / i/ ($F3 = 3174Hz$) et /ḭ/ ($F3 = 2738Hz$). Ce qui justifie l'arrondissement des lèvres pendant la production des voyelles / u/ et /ṵ/. On en déduit donc les voyelles arrondies qui sont / u/, / o/, / ɔ/, /ṵ/ et /ɔ̰/ et les voyelles étirées / i/, / e/, / ɛ/, /ḭ/ et /ɛ̰/.

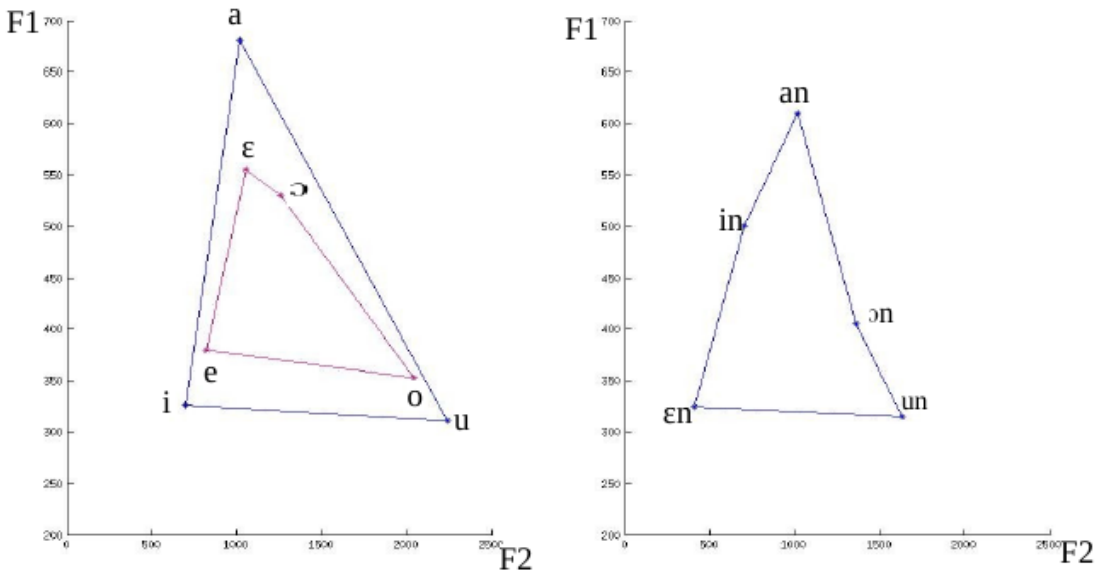


FIGURE 3.1 – Triangle vocalique sur le plan F1-F2, **Gauche-** Les voyelles orales, **Droite-** Les voyelles nasales

La figure 3.1 présente le triangle vocalique des voyelles du Fongbe. On remarque que dans le plan F1-F2, les voyelles occupent un large espace acoustique. Les voyelles orales / a/, / i/ et u/ et les voyelles nasales /a̰/, /ḭ/ et /ṵ/ sont bien disposées aux extrémités du triangle vocalique pointé vers le haut. Ce triangle pointé vers le bas montre la configuration articulatoire du système vocalique du Fongbe présenté dans le tableau 1.1.

A ces trois fréquences formantiques, nous avons aussi ajouté une quatrième dimension portée par la durée moyenne de prononciation de chaque voyelle pour la description acoustique du système vocalique. Ces durées moyennes répertoriées dans le tableau 3.2 sont calculées sur l'ensemble des occurences de chaque voyelle présente dans le corpus *FongbePhones-FLDataset*. Il faut noter que la durée relative d'une voyelle Fongbe, comme les voyelles d'autres langues, dépend fortement de l'environnement de prononciation et de la vitesse d'élocution. Il en ressort des valeurs obtenues, que les voyelles nasalisées sont caractérisées par une durée moyenne plus grande que celle des voyelles orales. Les voyelles centrales présentent des durées moins importantes par rapport aux voyelles orales et nasales.

3.2 Structure acoustique des consonnes

Le système consonnantique du Fongbe est composé de vingt-deux (22) phonèmes [58] regroupées en deux (2) classes (consonnes orales et nasales) selon le resserrement du

TABLE 3.2 – Durée moyenne en secondes de chaque voyelle du système vocalique du Fongbe. Les valeurs entre paranthèses constituent les marges d'erreurs autorisées pour chaque durée calculée.

			Durée moyenne
Voyelles orales	ant.	i	0.3678 (0.06)
		e	0.3239 (0.02)
		ɛ	0.3511 (0.06)
	post.	u	0.3701 (0.03)
		o	0.3768 (0..07)
		ɔ	0.3485 (0.04)
	cent.	a	0.1980 (0.09)
Voyelles nasales	ant.	ḭ	0.3853 (0.05)
		ɛ̰	0.4085 (0.05)
	post.	ṵ	0.3724 (0.04)
		ɔ̰	0.3751 (0.05)
	cent.	a̰	0.1993 (0.09)

conduit vocal (voir le tableau 3.3). Les consonnes orales regroupent des sous-ensembles comme les voisantes sourdes, les voisantes sonores et les non voisantes qui sont caractérisés par l'influence de la vibration du conduit vocal. Cette classe est constituée de phonèmes très différents aussi bien du point de vue articulatoire qu'acoustique.

TABLE 3.3 – Le système consonnantique du Fongbe [10].

			Labiales	Apicales	Dorsales	Alvéolaires	Labiovélaires	Vélaires	Uvulaires
Orales	Voisantes	sourdes	f	t	c	s	kp	k	x
		sonores	v	d	j	z	gb	g	ɣ/ (h)
	Non-voisantes		b	ɖ	y	l	w		
Nasales			m	n	ŋ (ny)				

Contrairement à la configuration acoustique des voyelles qui est caractérisée par la présence de fréquences formantiques, l'analyse acoustique des consonnes a consisté à calculer trois autres marqueurs acoustiques : la fréquence fondamentale (pitch) qui matérialise la vibration périodique des cordes vocales, la durée des segments de chaque consonne et l'intensité qui matérialise la pression sous-glottique. Généralement, on distingue trois grandes classes pour l'étude acoustique des consonnes : les occlusives, les fricatives et les

sonnantes. Les consonnes en français sont étudiées suivant les occlusives, les fricatives et les consonnes vocaliques [59]. En Fongbe, on retrouve en plus des occlusives, fricatives et vocaliques, deux autres classes qui sont les consonnes nasales et les semi-consonnes.

TABLE 3.4 – Marqueurs acoustiques par consonne.

Phonèmes (Fongbe)	Moyenne sur f0 (Pitch Hz)	Moyenne sur Intensité (db)	Moyenne sur durée (s)
b	272	60,02	0,7308
c	330	52,68	0,7458
d	266	56,87	0,7929
Ð	271	59,01	0,7291
f	372	51,72	0,7629
g	276	57,33	0,7839
gb	280	56,45	0,7219
h	294	57,94	0,7582
j	263	55,3	0,7329
k	351	52,6	0,7195
kp	319	55,13	0,7474
l	245	57,22	0,9925
m	233	63,24	0,8092
n	220	56,98	0,8005
ny	224	61,33	0,794
p	350	58,56	0,6414
s	326	52,3	0,8023
t	330	50,27	0,7121
v	287	53,44	0,7293
w	272	57,47	0,8073
x	339	54,47	0,7525
z	250	58,32	0,8089

3.2.1 Les occlusives Fongbe

Cette catégorie de consonnes présente la même configuration articulatoire que la plupart des autres langues dont le français. On distingue des occlusives sonores et sourdes qui sont essentiellement composées de consonnes orales voisantes et non voisantes. Leur configuration articulatoire indique la fermeture du conduit vocal lors de l'émission de la consonne. Les occlusives sonores du Fongbe sont :

- les consonnes orales non-voisantes / b/ et /d̠/ ;
- les consonnes orales voisantes sonores / d/, / g/, / gb/ et / j/ ;

Elles sont caractérisées par des fréquences fondamentales (**f0**) qui varient moyennement entre les voisantes sonores (263 Hz pour / j/ et 279 Hz pour / gb/). Les non-voisantes se réalisent avec des valeurs de pitch moins élevées que les voisantes à la différence de / d/ et de / j/. La prononciation des occlusives sonores du Fongbe dure en moyenne 0,73 s avec un pic observé sur le phonème / j/ à l'exception des phonèmes / g/ et / d/ dont les durées moyennes respectives sont 0,7839 s et 0,7929 s. Les fréquences fondamentales calculées sur l'ensemble des locuteurs présents dans le corpus *FongbePhones-FLDataset* sont présentées sur la figure 3.2 et les durées moyennes de chaque consonne dans le Tableau 3.4. L'intensité moyenne la plus petite de la voix pendant la prononciation isolée des occlusives sonores s'observe avec la consonne voisante sonore / d/ (56 dB) pendant que la plus élevée s'obtient avec la non-voisante / b/ (60 dB) (voir les valeurs moyennes sur la figure 3.3).

Les occlusives sourdes du Fongbe sont composées de certaines consonnes orales sourdes dont / c/, / k/, / kp/, / t/ et aussi / p/ qui n'est pas considéré comme étant une consonne phonémique mais uniquement retrouvé dans certains mots d'emprunt. A la différence des occlusives sonores, les occlusives sourdes du point de vue articulatoire sont produites sans aucune vibration des cordes vocales et sont réalisées dans les zones dorsales, vélaires, labio-vélaires, bilabiales et apicales de la bouche. Elles sont caractérisées par un silence durant toute la durée de l'occlusion et sont dites "les occlusives apériodiques". Une analyse du tableau 3.4 montre que les occlusives apériodiques (sourdes) ont une durée moyenne de prononciation courte comparées aux occlusives périodiques (sonores). L'intensité de la voix pendant la prononciation d'une occlusive sourde varie entre 53 dB (pour la consonne / c/) et 59 dB (pour la consonne / p/).

3.2.2 Les fricatives Fongbe

La configuration articulatoire des fricatives indique la présence d'un bruit durant toute leur prononciation. Ce bruit intervient pendant le rétrécissement du passage d'air par le resserrement du conduit vocal. Comme les occlusives, on distingue les fricatives sonores (périodiques) et les fricatives sourdes (apériodiques) en Fongbe. Les fricatives sonores sont composées des consonnes orales voisantes sonores / h/, / v/, / z/ et les fricatives sourdes regroupent les consonnes orales voisantes sourdes / f/, / s/, / x/. Il est à remarquer que la fréquence fondamentale de la fricative sonore / h/ (250 Hz) qui se produit dans la zone uvulaire de la bouche est moins élevée que celle des fricatives périodiques alvéolaires / z/ (293 Hz) et labiales / v/ (286 Hz). Les fricatives sourdes se réalisent avec des valeurs d'intensité faibles comparées aux fricatives sonores dont les pics sonores s'observent respectivement sur les consonnes / x/ (54 dB) et / z/ (58 dB). Les fricatives sourdes sont

plus brèves dans la prononciation que les fricatives sonores à l'exception des consonnes / v/ et / f/ (voir tableau 3.4).

3.2.3 Les consonnes nasales et les semi-consonnes Fongbe

On distingue trois consonnes nasales en Fongbe. Il s'agit des consonnes / m/, / n/ et / ŋ/ qui sont respectivement produites dans les zones bilabiales, alvéolaires et palatales de la bouche. Leur configuration articulatoire indique le passage de l'air par les fosses nasales pendant que le voile du palais est abaissé. Ce qui les rends comparables aux occlusives sonores du point de vue articulatoire [59]. Ainsi elles se réalisent avec les fréquences fondamentales les plus faibles comparées aux occlusives sonores. Comparées aux autres consonnes du système consonantique du Fongbe, les consonnes nasales sont produites avec l'intensité la plus faible de la voix et constituent avec les semi-consonnes les moins brèves dans la prononciation. La réalisation des semi-consonnes s'effectuent avec presque les mêmes rythmes de vibration périodique des cordes vocales (valeurs de pitch) et la même pression sous-glottique (intensité) que les occlusives sonores.

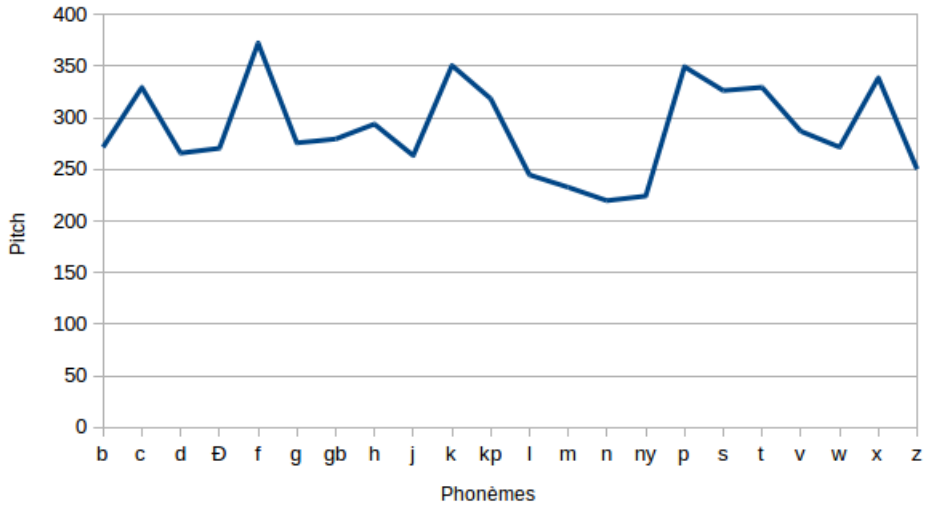


FIGURE 3.2 – Les valeurs pitch par consonnes.

Conclusion

Cette étude s'est intéressée à la production des sons du Fongbe en contexte isolé. La configuration acoustique des voyelles révèle une large variation des trois premiers formants dont les fréquences dépendent fortement de la configuration articulatoire de chaque

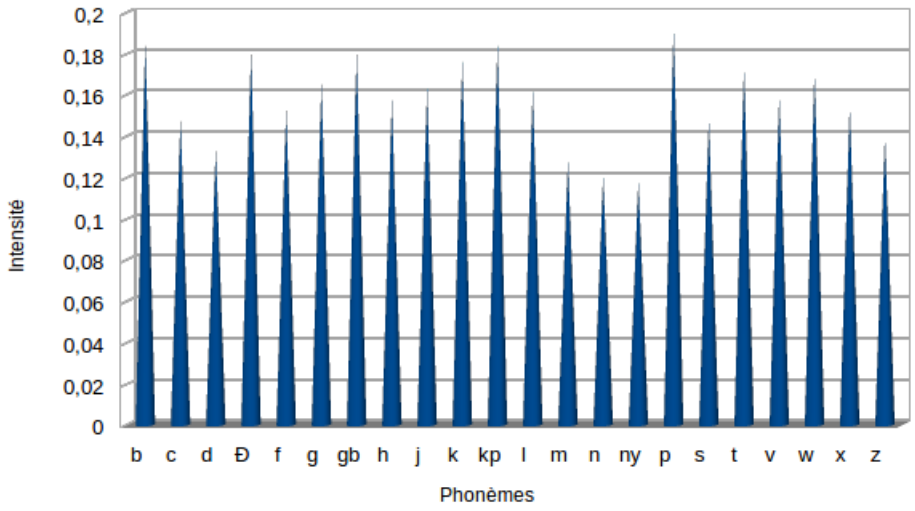


FIGURE 3.3 – L'intensité de chaque phonème consonne.

voyelle. Le triangle vocalique reste conforme à la configuration articulatoire du système vocalique du Fongbe. Il montre que les voyelles du Fongbe occupent un large espace acoustique. Pour décrire acoustiquement les consonnes du Fongbe, nous avons basé notre étude sur la vibration périodique des cordes vocales, la durée de prononciation et la pression sous-glottique. L'étude réalisée dans ce chapitre permet l'élaboration d'une référence sur les caractéristiques acoustiques des sons du Fongbe. Cette référence a été, par la suite, exploitée pour proposer une recette complète d'algorithmes depuis la segmentation de la parole à la reconnaissance de phonèmes dans un signal de parole continue [57].

Chapitre 4

Reconnaissance automatique des phonèmes du Fongbe dans un contexte isolé

"*adö tɔ́ bö nya đé yí jonɔ́*
azön atön bö asitön gbe ado
e ma kple bo wà à ɔ́, e nö kplé
bo đu a"

Sommaire

Introduction

La Reconnaissance de phonèmes est un processus intégré au système de reconnaissance automatique de la parole spontanée ou continue. Depuis les années 60, des progrès de recherche significatifs, liées au développement de méthodes statistiques et de techniques d'intelligence artificielle, ont essayé de surmonter les problèmes d'analyse et de caractérisation du signal de parole. L'un des problèmes majeurs est la spécificité acoustique et linguistique des différentes langues.

L'objectif d'un système de reconnaissance de la parole est de convertir le signal acoustique en un ensemble de mots à partir d'une segmentation du signal en petites unités phonétiques ou syllabiques. La reconnaissance de phonèmes, une composante du système de reconnaissance de la parole, est un processus par lequel on caractérise un segment de signal parlé par une empreinte phonétique. Ainsi pour obtenir de bonnes performances de reconnaissance, la reconnaissance de phonèmes doit être bien réalisée dans le but de produire des connaissances acoustiques des phonèmes d'une langue donnée. Il faut noter que, la reconnaissance de phonèmes est utilisée dans de nombreuses applications telles que la reconnaissance de la parole et du locuteur, l'indexation du locuteur, la synthèse vocale etc.

Dans ce chapitre, nous proposons une approche de reconnaissance de phonèmes isolés du Fongbe en utilisant plusieurs classifieurs. Elle traite particulièrement de la fusion de décisions provenant de deux classifieurs différents à savoir : le bayésien naîf et la quantification vectorielle à apprentissage (LVQ). Depuis les années 60, la combinaison de classifieurs a été l'un des axes de recherche les plus soutenus dans le domaine de la reconnaissance de formes. Depuis lors, des methodes de fusion de décisions ont été appliquées avec succès dans divers domaines tels que la reconnaissance et la vérification de signatures, la reconnaissance et l'identification de visages ou encore l'analyse d'images médicales. Elle a été introduite dans la reconnaissance automatique de la parole pour reconnaitre les phonèmes, la parole, l'âge et le genre d'un locuteur et pour identifier avec de bonnes performances une langue donnée. L'idée derrière la proposition dans ce chapitre est de construire un système discriminatoire robuste de phonèmes consonnes et voyelles à partir d'une combinaison intelligente de classifieurs basée sur la fusion de décisions [60, 5]. Pour ce faire, nous avons étudié la performance, des approches paramétriques (utilisant les DNNs) et

non-paramétriques (utilisant une combinaison pondérée) et une approche adaptative utilisant la logique floue que nous avons proposée. Cette fusion intelligente de décisions utilise les coefficients acoustiques MFCC, PLP et Rasta-PLP qui ont été d'abord fusionné avant d'etre appliqués aux classifieurs afin de fournir une identité phonétique des sons du Fongbe. Pour finir, les expériences ont été réalisées sur le corpus *FongbePhones-FLDataset* dont les résultats présentés dans ce chapitre ont révélé de meilleures performances avec la méthode de fusion floue proposée.

4.1 Etat de l'art

Le travail présenté dans ce chapitre traite de deux questions différentes à savoir : la fusion intelligente de décisions dans un contexte de multi-classification et la classification de phonèmes. Dans cette section, nous présentons un bref aperçu des théories et recherches récentes liées à la fusion de décisions et à la classification de phonèmes dans lesquelles nous positionnons l'approche que nous proposons.

4.1.1 La classification de phonèmes

Dans cette première partie de l'état de l'art, les travaux de recherches récentes, menées sur la classification des phonèmes appliquée à différentes langues parlées du monde, sont présentés comme suit.

Wong et. al. ont proposé, dans [61], une approche de classification de phonèmes anglais qui a montré de meilleures performances avec le corpus TIMIT. Dans ce travail, ils ont introduit une méthode de normalisation de la longueur du tractus vocal (VTLN) pour compenser la variation inter-locuteur du signal du locuteur en appliquant une distorsion, spécifique au locuteur, de l'échelle de fréquence du banc de filtres. L'interet visé dans ce travail est de comparer chacun des résultats de reconnaissance de phonèmes obtenus à partir de différentes valeurs prises pour facteur de distorsion de fréquences. Les résultats ont montré de meilleures performances de reconnaissance de phonèmes avec un facteur de distorsion de 1,4 sur un intervalle de fréquences de 300-5000Hz et qui ensuite a été appliqué à une reconnaissance de mots.

Dans les domaines de caractéristiques linéaires, l'adaptation de bruit additif est exacte et peut conduire à une classification plus précise que les représentations impliquant des traitements non-linéaires et une réduction de dimensionnalité. Se basant sur cette approche, Ager et. al., dans [62], ont développé une librairie pour la classification de phonèmes isolés utilisant des caractéristiques linéaires dans le but d'améliorer la robustesse

face au bruit additif. Dans ce travail, les auteurs ont réalisé leurs expériences sur tous les phonèmes présents dans le corpus TIMIT afin d'étudier les avantages potentiels de la classification de phonèmes basée sur des caractéritiques linéaires directement liées aux ondes acoustiques. Le but final est de mettre en œuvre l'adaptation exacte du bruit additif. Ainsi, il en ressort qu'en présence d'un bruit additif, la classification avec la librairie proposée devient plus performante qu'un classifieur PLP adapté au bruit. Outre les méthodes basées sur l'exploitation directe de paramètres acoustico-articulatoires comme dans [61, 62], la classification de phonèmes est aussi réalisée avec des techniques d'apprentissage automatique. Ces techniques utilisent des paramètres acoustiques calculées sur les signaux de parole pour fournir une identitié acoustique aux phonèmes contenus dans les signaux. Genussov et. al., dans [63], ont exploité et intégré une technique d'apprentissage de collecteur non linéaire, nommée *Diffusion maps*, à un système de classification de phonèmes. Le collecteur non linéaire construit un graphe à partir des vecteurs de paramètres utilisés afin de faire correspondre des distances euclidiennes aux différentes connexions dans le graphe. Les auteurs ont ainsi utilisé les distances euclidiennes pour la classification une fois que le graphe ait été contruit et que les différentes correspondances non linéaires éffectuées deviennent optimales. Les expériences réalisées sur plus de 1100 phonèmes isolés, extraits du corpus TIMIT, ont montré que les *Diffusion maps* permettent la réduction de la dimensionnalité et améliorent les résultats de classification.

Comme techniques d'apprentissage automatique, nous avons aussi les réseaux de neurones artificiels qui sont très souvent utilisés pour la classification de phonèmes isolés. L'approche des réseaux de neurones à convolution est étudiée avec succès par Palaz et. al., dans [64], pour classer des phonèmes contenus dans des signaux de paroles brutes avec des expériences réalisées sur le TIMIT et le corpus WSJ. Dans sa thèse [65], Ezin C. Eugène a étudié des architectures de réseaux de neurones artificiels intégrant à la fois les potentialités d'un réseau de neurones et d'un système d'inférence floue pour traiter des signaux de parole dans différents environnements bruités. Il ressort de ses travaux, l'utilisation des architectures de réseaux de neurones de temporisation [66, 67], pour la construction des systèmes de classification de phonèmes dans un environnement bruité. Les expériences sont effectuées sur les consonnes [b,d,g,p,t,k] de l'anglais et ont montré la performance de ces architectures qui permettent de détecter et de classer correctement les paramètres acoustiques des signaux de parole.

Des techniques comme les machines à vecteurs de support (SVM) permettent de classer des phonèmes avec un bruit additif en exploitant les paramètres acoustiques pour rendre robuste le système de classification [68].

Nous pouvons donc remarquer que de nombreux travaux ont été réalisés sur le corpus

TIMIT (anglais) qui constitue une base de données de référence dans la communauté scientifique. D'autres langues aussi ont fait objet de travaux sur la question de classification de phonèmes. Nous pouvons citer les travaux [69, 70, 71, 72] dans lesquels les auteurs ont respectivement effectués des expériences sur le Vietnamien, l'Afrikaans, l'Anglais, le Xhosa, le Hausa et l'Anglais américain.

4.1.2 Méthodes de fusion de décisions

La deuxième question traitée dans ce chapitre est la fusion de décisions pour une classification optimale de phonèmes isolés du Fongbe. La combinaison de décisions de classifieurs pour atteindre une décision optimale et une plus grande précision de classification est devenue, depuis les années 90, un sujet de recherche important. Dans la littérature, il y a des travaux qui ont combiné plusieurs classifieurs [73, 74, 75] et d'autres sont portés sur la combinaison de plusieurs *experts* dont les décisions s'expriment sous forme de distribution de probabilités [76, 77].

Parmi les méthodes de fusion de décision, il y a ceux qu'on appelle des méthodes non paramétriques (les décisions des classifieurs sont combinées dans un système dont les paramètres restent invariants) et les méthodes avec apprentissage dont l'interet est d'apprendre et d'adapter, sur les données disponibles, les paramètres nécessaires à la fusion. En reconnaissance automatique de la parole, plusieurs chercheurs ont adopté avec succès la fusion de décisions pour reconnaître par une machine soit les phonèmes, soit la parole continue, soit l'âge et le sexe d'un locuteur. Par exemple, nous pouvons citer le travail dans [78], où les auteurs ont effectué une combinaison de décisions de plusieurs classifieurs pour la reconnaissance et l'analyse de l'expression émotionnelle d'une personne. Certains auteurs ont adopté des méthodes non paramétriques comme la moyenne pondérée [79, 80, 81] et le vote majoritaire [82, 83]. D'autres ont adopté les méthodes paramétriques comme l'inférence bayésienne [84, 85, 86] et la méthode de Dempster-Shafer [87].

Dans le travail que nous proposons dans ce chapitre, nous avons adopté les deux approches de fusion de décision afin de comparer leurs performances pour une classification optimale des phonèmes du Fongbe. D'abord, nous avons effectué une moyenne pondérée (méthode non paramétriques) des décisions obtenues après classification. Cette méthode a besoin d'une valeur seuille que nous avons choisie judicieusement par expérimentation à l'étape d'apprentissage. Ensuite, la deuxième méthode de fusion que nous avons utilisée, est une méthode paramétrique avec apprentissage automatique basée sur les réseaux de croyance profonds (DBN). Les réseaux de croyance profonds ont récemment montré des performances impressionnantes dans la fusion de décision et les problèmes de classi-

fication [88]. En plus de ces deux méthodes de fusion, nous avons également utilisé une approche adaptative basée sur la logique floue. La logique floue est souvent utilisée pour les problèmes de classification et a récemment montré de bonnes performances dans la reconnaissance de la parole [89]. En effet, l'une des limite de l'utilisation d'une valeur seuille qu'exige la méthode pondérée, est que cette valeur est fixe et ne fournit pas une flexibilité face à la variation des données en entrée du système. Ainsi, pour surmonter cette limite et le temps que prend un processus d'apprentissage avec les réseaux de croyance profonds, nous avons proposé une troisième approche basée sur la logique floue qui peut imiter la décision de l'homme en codant ses connaissances sous la forme de règles linguistiques. La logique floue est capable de simuler les capacités de la pensée humaine face à l'incertitude dans la prise de décisions donc nécessite en utilisant des connaissances expertes.

4.2 Algorithmes et méthodes de classification

Cette partie détaille les algorithmes implémentés pour la classification des phonèmes voyelles et consonnes.

4.2.1 Le classifieur bayésien naïf

Le classifieur bayésien naïf est une méthode d'apprentissage surpervisé probabilistique basée sur le théorème de Thomas Bayes qui repose sur une hypothèse d'indépendance conditionnelle entre des descripteurs. Il apparaît dans la reconnaissance de la parole pour résoudre les problèmes de classification multi-classes. Il calcule explicitement les probabilités pour des hypothèses précises et se montre robuste face aux données bruitées en entrée au système de classification. Malgré sa simplicité, le classifieur bayésien naïf surclasse parfois certaines méthodes de classification plus complexes.

Dans un processus de classification, le classifieur naïf attribue une classe $c(x)$ à une donnée d'entrée x selon la règle de Bayes suivante :

$$p(c|x) = \frac{p(c,x)}{p(x)} \tag{4.1}$$

$$= \frac{p(c)p(x|c)}{\sum_{c'} p(c')p(x|c')} \tag{4.2}$$

où $p(c)$ est la probabilité a priori de la classe c, et $p(x|c)$ est la probabilité conditionnelle de x sachant la classe c. Considérant un ensemble de descripteurs $X = \{x_1, x_2, ..., x_n\}$, X

est classé dans $C = +$ si et seulement si,

$$F(X) = \frac{p(C = +|X)}{p(C = -|X)} \geq 1 \tag{4.3}$$

$F(X)$ représente le classifieur bayésien.

Le bayésien naïf est la forme la plus simple d'un réseau bayésien où l'on suppose que tous les attributs sont indépendants compte tenu de la classe [90].

$$p(X|c) = p(x_1, x_2, ..., x_n|c) = \prod_{i=1}^{n} p(x_i|c) \tag{4.4}$$

Le classifieur bayésien naïf est obtenu par :

$$F_{nb}(X) = \frac{p(C = +|X)}{p(C = -|X)} \prod_{i=1}^{n} \frac{p(x_i|C = +)}{p(x_i|C = -)} \tag{4.5}$$

Cette implémentation est très utilisée dans la communauté des chercheurs. Parce que d'une part elle est très facile à programmer, sa mise en œuvre est aisée ; d'autre part, parce que l'estimation de ses paramètres, dans la construction du modèle, est très rapide sur de très grandes bases de données, que ce soit en nombre de variables ou en nombre d'observations. Elle a été utilisée dans [91] pour reconnaitre les fricatives isolées, voyelles et consonnes nasales du corpus TIMIT. Le modèle bayesien naïf est un classifieur linéaire qui propose un biais de représentation similaire à celui de l'analyse discriminante, de la régression logistique ou des machines à vecteurs de support. Ceci explique sa robustesse en prédiction face aux techniques complexes car il a été prouvé que son biais n'est vraiment pas préjudiciable à la performance de la classification des phonèmes [92].

4.2.2 La quantification vectorielle à apprentissage ou LVQ

La quantification vectorielle à apprentissage (LVQ) est une version surpervisée de la quantification vectorielle dont l'architecture a été proposée en 1988 par Kohonen [93] pour les problèmes de classification. Elle est utilisée dans les tâches de reconnaissance de formes, de classification multi-classes, de la compression de données, de la reconnaissance de la parole et du traitement d'images. Les réseaux LVQ sont des réseaux hybrides qui utilisent un apprentissage semi-surpervisé [94]. Comme illustré dans la figure 4.1, un réseau LVQ comporte deux couches : la première est une couche compétitive et la deuxième, une couche linéaire. Chaque neurone de la première couche apprend un vecteur prototype ce qui permet de classer une région donnée de tout l'espace d'entrée. Ainsi, pour une entrée

p donnée, on obtient :

$$y_i^{(1)} = \|w_i^{(1)} - p\| \tag{4.6}$$

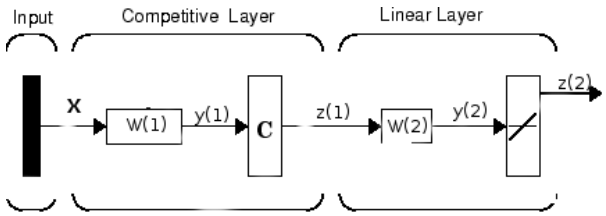


FIGURE 4.1 – Représentation d'un réseau LVQ

La sortie $z_m^{(1)}$ du vecteur, pour lequel la valeur de $y_1^{(1)}$ est la plus petite, prendra la valeur 1 et les autres la valeur 0. A noter que la première couche peut posséder plus de neurones que la deuxième ; ce qui fait que les prototypes de la même classe, mais de sous-classes différentes, peuvent activer différents neurones. La seconde couche combine les sous-classes en classes uniques. Pour ce faire, les colonnes de la matrice $w_i^{(2)}$ représentent les sous-classes et les lignes les classes. Cette matrice a un seul 1 par colonne, les autres éléments étant nuls.

$$y^{(2)} = W^{(2)} z^{(1)} \tag{4.7}$$

$$y^{(2)} = \sum_i w_{ij}^{(2)} z_j^{(1)} \tag{4.8}$$

$w_{ij}^{(2)} = 1$: indique que la sous-classe j fait partie de la classe i, sinon $w_{ij}^{(2)} = 0$. Ce processus de combinaison de sous-classes permet au réseau LVQ de créer des classes avec des délimitations complexes. La règle d'apprentissage est une règle de Kohonen.

L'algorithme d'apprentissage se résume comme suit.

1. Initialiser les poids $w_{ij}^{(1)}$ à des valeurs aléatoires comprises entre 0 et 1.
2. Ajuster le coefficient d'apprentissage $\eta(t)$.
3. Pour chaque prototype p, trouver le neurone d'indice i^* dont le vecteur poids $w_{i^*}^{(1)}$ est le plus proche de p.
4. Si la classe indiquée en sortie du réseau, pour le neurone d'indice i^*, correspond à celle du prototype d'indice i, alors faire :

$$w_{i^*}^{(1)}(t+1) = w_{i^*}^{(1)}(t) + \eta(t)(p(t) - w_{i^*}^{(1)}(t)) \tag{4.9}$$

sinon :

$$w_{i^*}^{(1)}(t+1) \quad = \quad w_{i^*}^{(1)}(t) - \eta(k)(p(t) - w_{i^*}^{(1)}(t)) \tag{4.10}$$

5. Si l'algorithme a convergé avec la précision souhaitée, alors arrêter, sinon aller en 2 en changeant de prototype.

Plusieurs études en reconnaissance de la parole ont fait usage de la méthode LVQ pour reconnaître les phonèmes, la parole ou le locuteur. J. Mantysalo et al ont appliqué la méthode LVQ pour effectuer la reconnaissance de la parole dépendant du locuteur en finlandais [95]. Un algorithme LVQ-MMC hybride a été évalué pour examiner une reconnaissance de mots japonais utilisant un large vocabulaire et une reconnaissance de la phrase japonaise dans [96].

4.3 Architecture proposée pour la classification de phonèmes

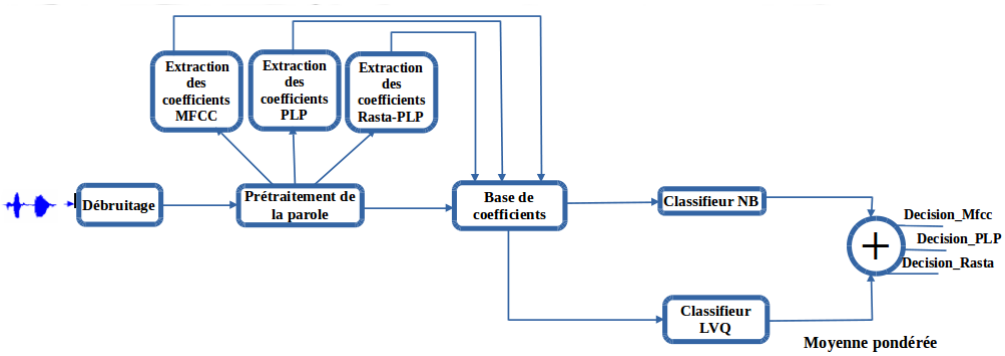
4.3.1 Vue globale du système proposé

Notre système de fusion intelligente se résume en deux modules qui sont chacun subdivisés en sous-modules :

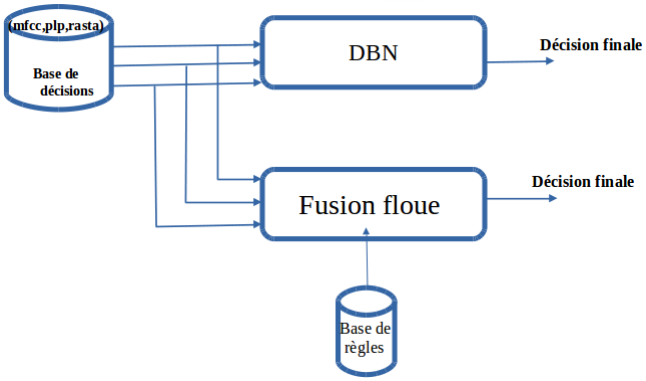
1. la fusion de paramètres acoustiques et classification : le premier module effectue d'abord une fusion des coefficients acoustiques extraits des signaux qui sont par la suite utilisés comme données en entrée aux classifieur bayésien naïf et LVQ. Cette première classification produit des sorties (premières décisions). Ce module renferme les sous-modules que sont :
   - le débruitage des signaux (i) ;
   - l'extraction des paramètres acoustiques (MFCC, PLP, Rasta-PLP) (ii) ;
   - la fusion de paramètres et la classification avec le bayésien naïf et LVQ (iii).
2. la fusion de décisions et la prise de décision optimale : le deuxième module effectue, en parallèle, la fusion de décisions avec l'approche floue que nous proposons et la méthode avec apprentissage basée sur le DBN (v).

Les deux modules sont séparés par un sous-module intermédiaire qui est la normalisation des décisions des classifieurs par le calcul de la moyenne pondérée des sorties des

classifieurs (iv). Les sorties du premier module sont combinées pour produire des décisions envoyées au module de fusion pour une prise de décision optimale. Les différentes étapes sont représentées sur la figure 6.1.



(a) Classification et normalisation.



(b) Fusion de décisions utilisant la logique floue et le DBN.

FIGURE 4.2 – Vue d'ensemble du système de classification

Trois techniques différentes ont été utilisées pour extraire les caractéristiques acoustiques de nos signaux de phonèmes afin de réaliser l'apprentissage de notre corpus. Il s'agit des coefficients MFCC, PLP et Rasta-PLP qui sont, après classification, fusionnés pour servir de données d'entrée au module de fusion de décisions de notre système. Les MFCCs exploitent le domaine fréquentiel des signaux utilisant l'échelle de Mel [97].La prédiction linéaire perceptuelle est basée sur le concept de psychophysique de l'audition : la résolution spectrale de bande critique et la loi de puissance de l'intensité sonore [98]. Rasta-PLP est une extension du PLP qui le rend plus robuste face aux distorsions spectrales linéaires.

L'avantage pour nous d'utiliser ces trois types de coefficients est d'élargir l'échelle de variation des données d'entrée du système de classification. Ceci permet à notre système d'apprendre plus d'informations acoustiques des phonèmes du Fongbe. Ces trois types de

coefficients ont permis aux deux classifieurs de suffisamment apprendre les phonèmes de la base d'apprentissage et de produire des décisions qui seront fusionnées par le module adaptatif proposé. Les treize (13) premières valeurs cepstrales des coefficients ont été extraites sur des fragments de 32 ms des signaux des phonèmes.

4.3.2 Fusion de décisions par simple moyenne pondérée

L'intérêt de cette étape intermédaire est de normaliser les sorties des classifieurs avant de les introduire comme entrées au module de fusion. D'abord, nous avons calculé la valeur moyenne pondérée des deux sorties des classifieurs pour chaque type de coefficient en affectant à chaque sortie un poids compte tenu du taux de reconnaissance de chaque classifieur. Pour ce faire, nous avons utilisé l'expression (4.11). Ensuite, pour appliquer la logique floue et la technique des réseaux de croyance profonds dans la fusion de décisions de chaque classifieur, nous avons utilisé l'expression 4.12, développée dans [99], pour déterminer la valeur seuille pouvant permettre aux classifieurs de décider de la classe du phonème en entrée.

$$input_1 = \frac{S^{naivebayes} \times \tau^{naivebayes} + S^{lvq} \times \tau^{lvq}}{\tau^{naivebayes} + \tau^{lvq}} \tag{4.11}$$

où S^A représente la décision du classifieur A et τ^A le taux de reconnaissance du classifieur A.

$$\tau = -1,2 \sum_i C_i + 2,75(\sum_k w_k^1 \lambda_1 + \sum_k w_k^2 \lambda_2) \tag{4.12}$$

C_i est le numéro de la classe i, w_k^i le poids du classifieur k lié à la classe i λ_1 et λ_2 sont des valeurs qui peuvent prendrent 0 ou 1 dépendamment de la classe. Par exemple, pour la classe des consonnes : $\lambda_1 = 1$ et $\lambda_2 = 0$. Les résultats de ces premières décisions des deux classifieurs seront par la suite comparés avec ceux des méthodes avec la logique floue et les réseaux de croyance profonds.

4.3.3 Fusion de décisions basée sur la logique floue

La logique floue est une approche mathématique-linguistique introduite par L.A. Zadeh pour généraliser la logique booléenne. Elle fournit un moyen simple pour arriver à une conclusion définitive sur la base des informations vagues, ambigués, imprécises, bruitées

ou manquantes. Les modèles de logique floue se composent d'un certain nombre de règles conditionnelles "if-then". Les systèmes flous convertissent ces règles en leurs équivalents mathématiques. La figure 4.3 montre l'architecture d'un système de logique floue où les entrées sont constituées des variables x_1, x_2, ..., x_n et la sortie est constituée de la variable y. Cette architecture présente les trois parties importantes d'un système de logique floue : la fuzzification (Réel →Flou), les règles d'inférence et la défuzzification (Flou →Réel).

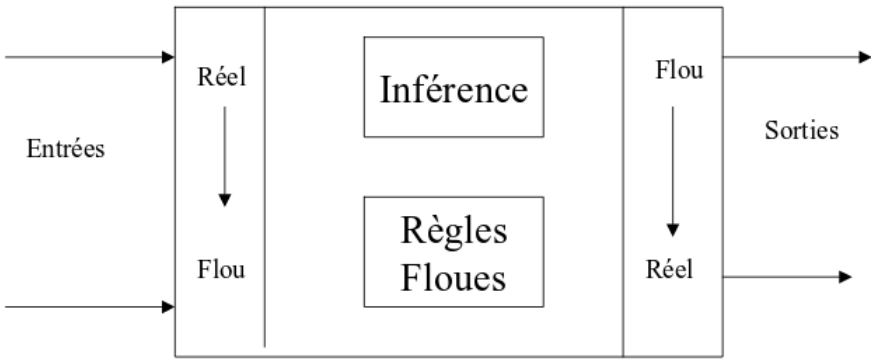


FIGURE 4.3 – Architecture d'un système de logique floue

1. La fuzzification : Les entrées sont converties en un ensemble flou avec un degré d'appartenance correspondant. La fonction d'appartenance est le composant le plus important dans la fuzzification et est utilisée pour décrire la relation entre les entrées et les ensembles flous.

2. L'inférence : A ce stade, les règles sont appliquées.Chaque règle, qui est composée de plusieurs instructions "*IF*" et une ou plusieurs conséquences "*THEN*", établit une relation entre les valeurs linguistiques par le biais d'une instructions"*IF-THEN*"

3. La défuzzification : C'est par ce processus que les degrés d'appartenance des variables linguistiques de sortie sont converties en valeurs numériques.

La nature des résultats obtenus à la première étape nous permet d'appliquer la logique floue sur quatres fonctions d'appartenance. Les entrées à notre système de logique floue sont les décisions provenant de la classification à partir de chacun des coefficients MFCC, PLP et Rasta-PLP et la sortie obtenue est le degré d'appartenance en valeur numérique d'un phonème à la classe consonne ou voyelle. Les variables d'entrée (les décisions) sont fuzzifiées en quatre ensembles complémentaires à savoir : *faible*, *moyen*, *fort* et *très fort*. La sortie est fuzzifiée en deux esnembles à savoir : consonne et voyelle. Aussi, nous obtenons pour les différents coefficients en tenant compte des valeurs obtenues après la moyenne pondérée :

- mfcc : faible - moyen - fort - très fort
- plp : faible - moyen - fort - très fort
- rasta : faible - moyen - fort - très fort

Après fuzzification, les règles floues ont été générées dans le cadre de la classification des phonèmes du Fongbe et sont présentées dans le tableau 4.1. Les régles ont été générées avec les résultats de classification sur les données de test et ont été par la suite appliquées à tout le corpus audio utilisé.

TABLE 4.1 – Les règles floues générées.

Règles No	Entrée			Sortie
	mfcc	rasta	plp	
1	faible	faible	faible	consonne
2	faible	faible	moyen	voyelle
3	faible	faible	fort	consonne
4	faible	moyen	faible	voyelle
5	faible	fort	faible	consonne
6	faible	fort	fort	consonne
7	faible	très fort	faible	voyelle
8	faible	très fort	très fort	voyelle
9	moyen	faible	faible	voyelle
10	moyen	faible	très fort	voyelle
11	moyen	très fort	faible	voyelle
12	moyen	très fort	très fort	voyelle
13	fort	faible	faible	consonne
14	fort	faible	fort	consonne
15	fort	fort	faible	consonne
16	fort	fort	fort	consonne
17	très fort	faible	faible	voyelle
18	très fort	faible	moyen	voyelle
19	très fort	faible	fort	consonne
20	très fort	faible	très fort	voyelle
21	très fort	moyen	faible	voyelle
22	très fort	moyen	très fort	voyelle
23	très fort	fort	fort	consonne
24	très fort	très fort	faible	voyelle
25	très fort	très fort	moyen	voyelle
26	très fort	très fort	très fort	voyelle

Tout d'abord, les données d'entrée (les décisions) ont été arrangées dans un intervalle $[Xmin..Xmax]$. Les différentes fonctions d'appartenance ont été obtenues en subdivisant

les intervalles formés par chaque entrée du système et en examinant la distribution locale des échantillons de phonèmes pris dans les deux classes (Voir Figure 4.4). L'examen de la distribution locale a induit les quatres sous-ensembles en fonction de la variation des données d'entrée et la sortie est obtenue en fontion de la nature des données et de manière surpervisée. Par exemple, si l'on donne en entrée au système les valeurs des décicions obtenues avec MFCC, PLP et Rasta, la sortie consonne ou voyelle est obtenue en fonction des sous-ensembles des données en entrée. En raison de la linéarité des valeurs contenues dans les sous-ensembles, une courbe en triangle simple (trimf) est utilisée pour les fonctions d'appartenance *faible* et *moyen* et une courbe de forme trapèze est utilisée pour les fonctions d'appartenance *fort* et *très fort*.

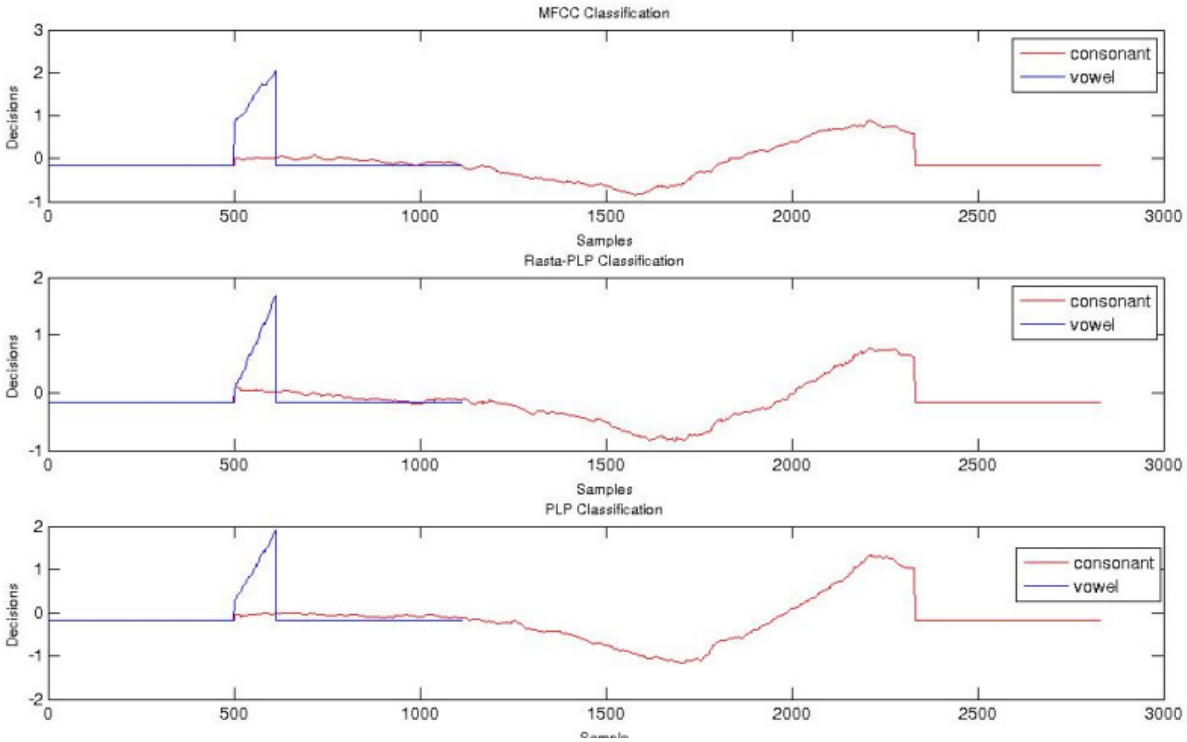


FIGURE 4.4 – **Haut-** distribution locale des décisions provenant de la classification des coefficients MFCC, **Milieu-** distribution locale des décisions provenant de la classification des coefficients Rasta-PLP, **BAS-** distribution locale des décisions provenant de la classification des coefficients PLP,

4.3.4 Fusion de décisions basée sur les DBNs

Dans cette sous section, nous décrivons la deuxième méthode employée pour la fusion de décisions dans le but d'adapter la classification finale à une prise de décision optimale. Cette méthode basée sur l'utilisation des réseaux de croyance profonds (DBNs) necessite

une étape d'apprentissage pour une bonne adaptation des décisions à la sortie du système. Les DBNs sont des modèles probabilistes multi-couches qui sont construits comme des hiérachies de modèles graohiques probabilistes plus simples et connectés de manière réccurente [100, 101]. Ils sont appelés : machines de Boltzmann restreinte (RBM). Chaque RBM est constitué de deux couches de neurones, une couche cachée et une couche visible. En utilisant l'apprentissage non surpervisé, chaque RBM est entrainé pour coder dans une matrice de poids une distribution de probabilités qui prédit l'activité de la couche visible depuis l'activité de la couche cachée [88]. Le travail, dans [101], fournit plus d'informations sur le fonctionnement du DBN et ses méthodes d'apprentissage. Pour la fusion de décisions de notre système, nous avons utilisé les paramètres DBN du tableau 4.2 pour l'apprentissage des différentes machines de Boltzmann.

TABLE 4.2 – Les paramètres DBN.

RBM Couche 1	200 units
RBM Couche 2	200 units
Pas d'apprentissage	0.01
Nombre d'itérations	100
Taille du batch	8

Les algorithmes 1 et 2 résument les différentes parties du classifieur proposé et implémenté sous matlab. Les noms des fonctions donnent l'idée de l'opération effectuée et les phrases commençant par // représentent des lignes de commentaires dans l'algorithme. Par exemple, final_decision_2 $\leftarrow fusiondbn(toutes\_les\_entrees)$ signifie que la décision optimale donnée par les fusion avec les DBNs est sauvegardée dans la variable final_decision_2.

4.4 Evaluation de performances : résultats expérimentaux

Nous présentons dans cette section les différents résultats obtenues après apprentissage, les tests effectués avec les deux classifieurs, les résultats de la fusion de décision avec l'approche de la logique floue et les réseaux de croyance profonds. Les expériences ont été réalisées avec le corpus audio *FongbePhones-FLDataset* sous l'environnement Matlab avec un processeur Intel Core i7 CPU L 640 @ 2.13GHz $\times$ 4 et 4GB de mémoire RAM.

Algorithm 1: Classification avec le Bayésien naïf et LVQ.

Data: Signaux de phonèmes
Result: Décision de chaque classifieur par rapport à chaque technique d'extraction.

débruitage du signal;
pour *signal* ∈ *base_phonemes* **faire**
  signal←débruitage(signal);
  base←insérer(signal)
fin
Extraction de caractéristiques;
pour *signal* ∈ *base* **faire**
  m←mfcc_calcul(signal);
  p←plp_calcul(signal);
  r←rasta_calcul(signal);
  base_mfcc←insérer(m);
  base_plp←insérer(p);
  base_rasta←insérer(r);
fin
apprentissage←insérer(m,p,r);
//Classification avec le Bayésien naïf et LVQ;
pour $i \leftarrow 1$ **a** $taille(apprentissage)$ **faire**
  if $i <= taille(base\_mfcc)$ **then**
    $bayes\_mfcc\_decision$←$bayes(apprentissage(i))$;
    lvq_mfcc_decision←lvq(apprentissage(i));
  end
  if $i > taille(base\_mfcc)$ **ET** $i <= taille(base\_mfcc) + taille(base\_plp)$ **then**
    $bayes\_plp\_decision$←$bayes(apprentissage(i))$;
    lvq_plp_decision←lvq(apprentissage(i));
  end
  if $i > taille(base\_mfcc) + taille(base\_plp)$ **ET**
  $i <= taille(base\_mfcc) + taille(base\_plp) + taille(base\_rasta)$ **then**
    $bayes\_rasta\_decision$←$bayes(apprentissage(i))$;
    lvq_rasta_decision←lvq(apprentissage(i));
  end
fin

4.4.1 Première étape - résultats des classifications

Les phonèmes sont classés en deux catégories : les consonnes et les voyelles. Nous avons utilisé 80% des données de chaque classe du corpus audio *FongbePhones-FLDataset* pour construire la base d'apprentissage et 20% pour les données de test. Cela conduit à 2831 échantillons de consonnes et 1112 d'échantillons de voyelles pour l'apprentissage et 493 échantillons de chaque classe pour les tests. Les paramètres LVQ utilisés pour obtenir de

Algorithm 2: Fusion de décisions avec la logique floue et les DBNs.

Data: Décision de chaque classifieur en fonction de la technique d'extraction.
Result: Décision finale

//calcul du taux de reconnaissance;
for $j \leftarrow 1$ **to** $taille(classes)$ **ET** $k \leftarrow 1$ **to** $taille(classifieurs)$ **do**
  $\tau \leftarrow -1,2 \sum_i C_i + 2,75(\sum_k w_k^1 \lambda_1 + \sum_k w_k^2 \lambda_2)$;
end
//calcul des valeurs de la moyenne pondérée comme entrées au système flou;
for $l \leftarrow 1$ **to** 3 **do**
  $entree_i \leftarrow \frac{S^{naivebayes} * \tau^{naivebayes} + S^{lvq} * \tau^{lvq}}{\tau^{naivebayes} + \tau^{lvq}}$;
  $toutes\_les\_entrees \leftarrow inserer(entree_i)$;
end
final_decision_1 $\leftarrow systemelogiquefloue(toutes\_les\_entrees)$;
final_decision_2 $\leftarrow fusiondbn(toutes\_les\_entrees)$;

bonnes performances de classification sont :

- nombre de neurones cachés : 60
- pourcentage de la première et la seconde classe : 0.6 et 0.4
- coefficient d'apprentissage : 0.005
- nombre d'itérations : 750

Une distribution normale a été utilisée pour la classification avec le bayésien naïf. Le tableau 4.3 présente les différents taux de reconnaissance obtenus à l'apprentissage et à la phase test des deux classifieurs.

TABLE 4.3 – Résultats d'apprentissage. Les valeurs sont estimées en pourcentage.

Classifieur	MFCC		RASTA-PLP		PLP	
	C_1	C_2	C_1	C_2	C_1	C_2
Résultats de l'apprentissage						
Bayésien naïf	88,66	51,53	90,43	59,17	88,2	68,25
LVQ	98,09	47,44	97,32	40,65	97,35	51,53
Résultats de test						
Bayésien naïf	92,29	38,34	91,48	46,04	93,10	60,24
LVQ	98,78	24,95	98,58	21,70	97,97	20,89

D'après le tableau 4.3, le classifieur bayésien naïf reconnait 90,43% des consonnes utilisant les coefficients Rasta-PLP et 68,25% des voyelles utilisant les coefficients PLP. Le classifieur LVQ montre de forts taux de reconnaissance des consonnes et de faibles taux

des voyelles. Nous avons obtenu 98,09% avec les coefficients MFCC pour les consonnes et 51,53% avec les coefficients PLP pour les voyelles. L'évolution des taux de reconnaissance se confirme avec les résultats obtenus aux tests et ceci dans les deux classes.

4.4.2 Seconde étape - fusion de décisions des classifieurs

L'objectif de cette fusion est d'obtenir plus d'informations à partir des décisions de chaque classifieur. Le problème rencontré est que la classification des consonnes donne de meilleurs taux pendant qu'avec les voyelles, les classifieurs ne sont pas très efficaces en raison du manque important de signaux de voyelles. Ainsi, l'objectif es d'équilibrer la classification pour permettre une bonne reconnaissance des voyelles. Le tableau 4.4 montre les résultats des différentes méthodes de fusion utilisées.

TABLE 4.4 – Résultats des fusions de décisions.

Méthodes de fusion	Consonne	Voyelle
Moyenne pondérée	99,73%	54,02%
Logique floue	95,54%	83,97%
Réseaux de croyance profonds	88,84%	84,79

4.4.3 Analyse de performance

Plusieurs mesures ont été développées pour traiter le problème de classification [102]. Les valeurs du Vrai Positif (TP), du Vrai Négatif (TN), du Faux Positif (FP) et du Faux Négatif (FN) ont été calculées après la fusion de décisions. Ces valeurs ont été utilisées pour calculer les paramètres de performances tels que la sensibilité (SE- *sensivity* en anglais), la spécificité (SP- *specifity* en anglais), le rapport de vraisemblance positive (LRP- *Likehood Ratio Positive* en anglais), le rapport de vraisemblance negative (LRN- *Likehood Ratio Negative* en anglais), l'exactitude (Ac - *Accuracy* en anglais) et la précision (Pr- *Precision* en anglais). Trois autres mesures importantes ont été utilisées comme paramètres d'évaluation : F-mesure, G-mesure et le temps d'exécution. F-mesure considère à la fois la précision (Pr) et la sensibilité (SE) pour calculer un score. Ce score représente la moyenne harmonique de la précision et de la sensibilité. Il mesure la capacité du système à donner toutes les solutions pertinentes et à refuser les autres. G-mesure est définie par la sensibilité et la spécificité et donne la mesure d'une performance équilibrée de l'apprentissage entre les classes positives et négatives. Le temps d'exécution mesure le temps de calcul de chacune des méthodes de fusion à l'étape de *testing*. Ces différents

paramètres de performance sont calculés à partir des équations suivantes :

$$SE = \frac{TP}{TP + FN} \tag{4.13}$$

$$SP = \frac{TN}{FP + TN} \tag{4.14}$$

$$LRP = \frac{SE}{1 - SP} \tag{4.15}$$

$$LRN = \frac{1 - SE}{SP} \tag{4.16}$$

$$Ac = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.17}$$

$$Pr = \frac{TP}{TP + FP} \tag{4.18}$$

$$F\text{-mesure} = \frac{2 \times Pr \times SE}{Pr + SE} \tag{4.19}$$

$$G\text{-mesure} = \sqrt{SE \times SP} \tag{4.20}$$

Nous avons utilisé le même corpus audio pour évaluer la performance des classifieurs bayésien naïf et LVQ et les méthodes de fusion de décisions employées pour la classification des consonnes et voyelles du Fongbe. Dans le tableau 4.4, nous remarquons qu'en considérant l'équilibre des classes de phonèmes dans le corpus audio, l'approche floue pour la fusion de décisions réalise de meilleures performances, même si la moyenne pondérée et les réseaux de croyance profonds classent respectivement les consonnes et voyelles mieux que l'arhictecture de logique floue proposée.

Dans le tableau 4.5, nous remarquons que le système de logique floue combine efficacement les décisions et fournit la décision optimale, mais avec un temps d'exécution qui s'augmente de 6% comparé aux réseaux de croyance profonds. Les résultats présentés dans le tableau 4.5 montrent les meilleures performances du système de logique floue avec les

paramètres d'exactitude, de F-mesure et de G-mesure qui sont les indicateurs choisis pour évaluer la performance des méthodes de fusion de décisions. Ces meilleures performances confirment que l'ajout de connaissances expertes améliore la prise de décision.

TABLE 4.5 – Analyse de performance. Valeurs en gras sont soulignées pour la comparaison des performances.

Parametres	Bayésien naïf	LVQ	Myenne pondérée	Système de logique floue proposé	DBNs
SE	0.93	0.99	0.99	0.95	0.88
SP	0.60	0.25	0.38	0.84	0.86
LRP	2.36	1.32	1.60	5.94	6.28
LRN	0.12	0.04	0.03	0.06	0.14
Ac	0.77	0.62	**0.69**	**0.90**	**0.87**
Pr	0.70	0.57	0.62	0.86	0.88
F-measure	0.80	0.72	**0.76**	**0.90**	**0.88**
G-measure	0.75	0.50	**0.61**	**0.89**	**0.87**
Execution time (seconds)	-	-	**0.10**	**0.7**	**0.04**

Conclusion

Dans ce chapitre, nous avons évalué la performance de trois méthodes de fusion de décisions par la combinaison intelligente de classifieurs pour résoudre le problème de reconnaissance des phonèmes du Fongbe. L'évaluation de la performance a été réalisée sur les méthodes telles que la moyenne pondérée, les réseaux de croyance profonds et l'approche de la logique floue. Ces méthodes ont opéré sur les décisions provenant des classifieurs bayésien naïf et LVQ qui ont entrainé trois différents coefficients acoustiques tels que MFCC, PLP et Rasta-PLP. L'idée principale est de fournir une décision optimale à partir des décisions obtenues avec chaque classifieur.

Les résultats des paramètres de performance tels que l'exactitude, F-mesure et G-mesure présentés dans le tableau 4.5, montrent la meilleure performance que réalise notre système de logique floue proposé qui utilise le raisonnement humain pour fonctionner.

Ce chapitre met en évidence deux résultats principaux :

1. la comparaison de performances de trois méthodes de fusion de décisions dans un problèmes de classification de phonèmes avec plusieurs classifieurs (i) ;

2. la proposition d'un système robuste de classification de phonèmes du Fongbe qui intègre une fusion, basée sur l'approche floue, des classifieurs bayésien naïf et LVQ (ii).

Cette proposition résulte de la perfomance réalisée par notre système de logique floue comparé à l'approche des réseaux de croyance profonds et surtout en raison des limites de la valeur seuille fixe de la fusion pondérée.

Troisième partie

Contributions à la reconnaissance automatique de la parole continue en Fongbe

Chapitre 5

Segmentation indépendante du texte de la parole continue en Fongbe

"*awasagbe mö ajinakú fö lobo*
đekɔ́ : e sí me đé nukú me ɔ́, e nö
le sí gudó tön"

Sommaire

Introduction

Pour construire un système efficace de reconnaissance automatique de la parole, l'étape de segmentation de la parole en unités plus petites doit être correctement réalisée afin de faciliter la reconnaissance des phonèmes, des syllabes ou des mots dans une phrase prononcée. La segmentation syllabique permet l'identification de segments syllabiques dans une parole continue. Elle est utilisée pour détecter les frontières appropriées de début et de fin d'une syllabe. Les techniques récentes utilisées pour la segmentation sont pour la plupart reliées au corpus de données et necessitent la disponibilité de transcriptions bien alignée des unités de parole. Il se révèle que les techniques manuelles de segmentation sont plus précises que les techniques automatiques même si elles necessitent beaucoup de temps et d'argent pour segmenter un très grand large corpus.

Plusieurs techniques existent pour la segmentation de la parole en unités syllabiques et sont basées sur les caractéristiques du domaine temporel et fréquentiel telles que la transformée en ondelettes, l'auto-corrélation, l'énergie à court-terme ou l'énergie spectrale, le taux de passage par zéro, la fréquence fondamentale, la fréquence de Mel, le centroid spectral et le flux spectral.

Dans ce chapitre nous présentons deux nouvelles approches proposées pour la segmentation syllabique automatique. Elles se basent sur des caractéristiques temporelles pour réaliser une segmentation de la parole continue en unités syllabiques.

5.1 Etat de l'art

Cette section présente les travaux recensés dans l'état de l'art sur la segmentation non supervisée de la parole en syllabe. La segmentation non supervisée consiste à découper

un signal de parole, avec des méthodes paramétriques, en de petits segments syllabiques. Plusieurs méthodes ont été proposées pour cette forme de segmentation syllabique de la parole naturelle. Ce sont pour la plupart des mesures de stationnarité, des mesures basées sur le taux de passage par zéro, sur l'énergie à court-terme ou sur l'énergie spectrale pour estimer les limites des différentes unités syllabiques. Il y a aussi des méthodes d'estimation de la fréquence fondamentale qui permettent de différencier les segments voisés des non voisés et les caractéristiques spectrales [103]. Il faut noter que, la segmentation syllabique peut être effectuée en utilisant trois types de caractéristiques : les caractéristiques du domaine temporel, les caractéristiques du domaine des fréquences et une combinaison des deux.

La segmentation syllabique a été traitée pour la première fois par Mermelstein à l'aide d'un algorithme basé sur l'enveloppe convexe mesurée sur l'intensité acoustique dans les régions de formants [104]. Zhao et al., dans [105] ont combiné la variation spectrale et l'analyse de l'enveloppe convexe sur l'énergie du signal pour construire une méthode hybride de segmentation syllabique automatique. Ils ont utilisé l'énergie de courte durée pour diviser un énoncé de parole en plusieurs parties sur lesquelles ils ont appliqué l'enveloppe convexe et la méthode du taux de passage par zéro. Ces deux méthodes ont été utilisées pour corriger les erreurs de rejet dans la détection des syllabes.

D'autres algorithmes du domaine temporel pour la segmentation syllabique ont été développés par Pfitzinger et al. [106] et par Jittiwarangkul et al. [107]. Dans [106], les auteurs ont utilisé les maxima locaux du contour de l'énergie pour trouver les noyaux syllabiques et les positions des noyaux par rapport aux transcriptions manuelles. Différents types de fonctions d'énergie temporelle et les méthodes de lissage pour la détection des frontières syllabiques sont présentées dans [107]. Dans [108], Wu et al. ont opté pour le traitement syllabique afin améliorer la précision dans la reconnaissance de la parole. En utilisant des techniques de filtrage en deux dimensions, ils ont calculé les spectres lissés de la parole afin d'améliorer les variations d'énergie de l'ordre de 150 ms. Ainsi, ils obtiennent des caractéristiques qu'ils combinent avec le log-rasta pour en faire des données en entrée d'un classifieur neuronal pour estimer les frontières syllabiques. Dans le même sens, Massimo et al. ont proposé l'utilisation de deux calculs différents de l'énergie tels que l'énergie d'origine et l'énergie du signal filtré à l'aide d'un filtre passe-bas pour détecter l'énergie maxima la plus pertinente [109]. Leur algorithme de segmentation syllabique est basé sur l'analyse de la forme temporelle de l'énergie du signal de parole. Il a été testé sur l'anglais et l'italien. Dans [110], Sheikh et al. ont utilisé le lissage flou pour proposer un algorithme de segmentation de la parole connectée en unités syllabiques. Ils ont appliqué le lissage flou sur la fonction d'énergie à court-terme pour lisser toute fluctuation

locale afin de préserver l'information contenue dans le segment syllabique. La méthode a été testée sur 200 énoncés pris dans la base de données de parole en Farsi. Les limites syllabiques ont été détectées sur les minima de l'énergie. Tous les algorithmes de segmentation syllabique mentionnés ci-dessus sont des méthodes qui utilisent des caractéristiques du domaine temporel. Les méthodes, du domaine temporel, pour la segmentation syllabique utilisent principalement l'énergie à court-terme (STE) pendant que les méthodes du domaine fréquentiel utilisent des caractéristiques cepstrales.

Nous retrouvons dans l'état de l'art, les méthodes du domaine des fréquences telles que la théorie fractale et la représentation temps-fréquence pour la segmentation syllabique. La théorie fractale a été utilisée par Pan et al. dans [111], pour déterminer les points de début et de fin de chaque syllabe dans la parole continue. Dans leur travail, ils ont proposé un algorithme basé sur la trajectoire de la dimension fractale pour réaliser la segmentation. Dans le but d'améliorer les performances de la segmentation, ils ont combiné la dimension fractale et la transformée en ondelettes en utilisant l'algorithme de seuillage dynamique pour débruiter les signaux de parole. Les résultats de simulation ont montré que, dans le cas d'un faible rapport bruit signal (SNR), l'algorithme de segmentation syllabique basée sur la dimension fractale obtient une grande précision relative (88%) comparé aux autres. Dans [112], les auteurs ont utilisé la représentation temps-fréquence du signal pour exploiter les caractéristiques de la parole en fonction de la région de fréquences et de la fusion des mesures d'intensité et de voisement à travers différentes régions fréquentielles pour la segmentation automatique de la parole en syllabes. Dans [113], Villing et al. ont proposé un algorithme de segmentation syllabique qui utilise l'enveloppe dans trois bandes fréquentielles pour identifier les limites de la syllabe dans la parole. Chou et al. dans leur travail [114], ont utilisé la technique MFCC comme méthode d'extraction de caractéristiques pour construire un vecteur de caractéristiques des syllabes provenant des chants d'oiseaux.

D'autres auteurs ont adopté les méthodes basées sur les MMCs et les Réseaux de Neurones Artificiels (RNA) qui sont largement utilisées pour la tâche de segmentation dans la reconnaissance automatique de la parole [115, 116, 62, 15]. On retrouve dans [115] l'utilisation d'un réseau de flux temporel pour le calcul d'une fonction ayant des pics locaux à noyaux syllabiques. L'intérêt était d'utiliser un autre type de RNA pour trouver des noyaux syllabiques dans un signal de parole. Ching et al., dans [116], ont proposé l'utilisation d'un réseau de neuro-flou hiérarchique pour segmenter un énoncé en une séquence de consonnes silencieuses. A cet effet, il a été utilisé une règle de concaténation syllabique pour détecter les frontières entre syllabes. En plus des méthodes basées sur les RNA, on retrouve aussi l'utilisation des approches basées sur les MMC dépendants

d'un modèle du langage pour la segmentation en syllabes tonales [117] et la segmentation syllabique dans une tâche de synthèse vocale [117].

Nos deux méthodes, comparées aux méthodes décrites ci-dessus, se basent sur des approches du domaine temporel pour générer des caractéristiques pour la détection d'unités syllabiques, et donc ne dépendent pas des régions de fréquence du signal en entrée et l'enveloppe des bandes de fréquences comment dans les travaux [111], [112] et [113]. Comparée aux approches basées sur les réseaux de neurones artificiels et les modèles de Markov cachés, notre première méthode de segmentation ne nécessite pas une phase d'apprentissage. Ces deux méthodes de segmentation que nous proposons dans ce chapitre sont indépendantes du modèle de langage pour la segmentation de la parole continue (en Fongbe) en de petites unités syllabiques donc se différencient des méthodes comme celles proposées par Makashay et. al dans [62]. Contrairement aux approches du domaine temporel citées dans cette section, nous nous sommes basés sur l'étude des informations pertinentes que fournit l'analyse de la distribution locale des exposants de singularité (comme [118]) pour obtenir des segments de phonèmes ou de syllabes.

5.2 Approche basée sur l'entropie de Rényi pour la segmentation syllabique

Dans cette section, nous présentons la première approche proposée pour la segmentation en syllabes d'un énoncé en Fongbe [3]. L'approche est basée sur le calcul des exposants de singularité et de l'entropie de Rényi. La figure 5.1 montre les différentes étapes de réalisation de la segmentation syllabique à partir de l'approche proposée. Ces étapes se résument comme suit :

1. la suppression des zones silencieuses du signal de parole ;

2. le calcul des exposants de singularité dans le but d'exploiter leur distribution pour l'analyse des dynamiques temporelles des segments de parole obtenus précédemment. Ceci conduit à l'obtention des segments dits "candidats" succeptibles de contenir une syllabe ou non.

3. le calcul de l'énergie à court-terme dans chaque segment candidat généré par l'analyse des exposants de singularité ;

4. l'introduction de la mesure basée sur l'entropie de Rényi calculée sur la matrice des énergies à court-terme qui exploitent au mieux les exposants de singularité pour

améliorer la précision dans la détection des frontières entre syllabes. Ceci conduit au rejet des faux segments candidats.

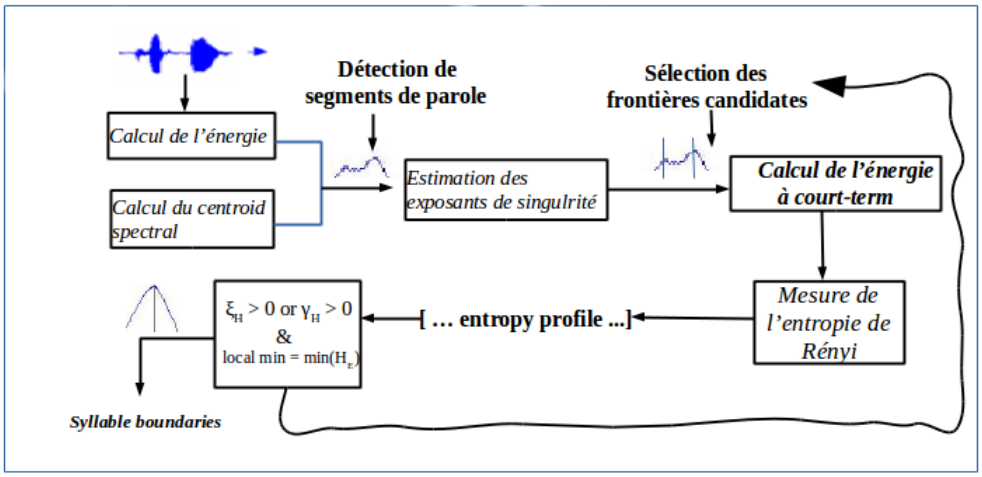


FIGURE 5.1 – Vue d'ensemble des étapes de l'approche proposée

5.2.1 Définition d'une unité syllabique

Une syllabe est une unité phonologique plus petite contenue dans un discours. Elle a longtemps été considérée comme une unité robuste dans le traitement de la parole. Elle est généralement composée d'un noyau (en général, une voyelle) et éventuellement entourée par des groupes de consonnes (à gauche et/ou à droite) [119]. Elle est acoustiquement définie par le principe de la sonorité qui est supposée être maximale dans le noyau et minimale au niveau de ses limites [112]. Les segments syllabiques à la sortie de notre méthode sont soit des unités phonétiques soit une simple concaténation de consonnes et de voyelles, conduisant à une forme monosyllabique.

5.2.2 Les exposants de singularité

L'estimation précise des Exposants de Singularité (EdS) dans les systèmes multi-échelles permet de caractériser les paramètres pertinents et d'identifier le contenu d'information [120]. Nous avons utilisé ces exposants pour extraire des caractéristiques temporelles du signal de parole considérée comme un système dynamique complexe. Les exposants de singularité permettent la prévisibilité et l'analyse précise du signal de parole et sont estimés avec les méthodes dérivées des principes issus de la physique statistique.

Dans cette sous section, nous présentons l'approche de calcul des exposants de singularité basée sur le Formalisme Microcanonique Multiéchelle (FMM) proposée dans [118]. Le FMM est basé sur le calcul des exposants locaux d'un signal donné dont la distribution est la quantité clé définissant sa dynamique intermittente [121].

Etant donné un signal s, la relation donnée par l'équation (5.1) doit être valable pour chaque instance t et pour de petites échelles r.

$$\Gamma_r(s(t)) = \alpha(t) r^{h(t)} + O(r^{h(t)}) \qquad r \longrightarrow 0 \tag{5.1}$$

Dans l'équation 5.1, $h(t)$ représente les EdS du signal s à l'instant t et Γ_r une fonctionnelle dépendante de l'échelle r. Le but est de faire un bon choix de Γ_r de sorte qu'une estimation précise des EdS soit atteinte. Plus $h(t)$ est petit, moins le système est prévisible pour l'instant t. Avec l'équation 5.2, la fonctionnelle Γ_r est définie à partir du gradient s' du signal.

$$\Gamma_r(s(t)) := \frac{1}{r} \int_{t-r}^{t+r} d\tau |s'(\tau)| \tag{5.2}$$

Γ_r peut être projeté en ondelettes pour obtenir des interpolations continues des données échantillonnées discrètes. Avec une ondelette Ψ, la projection à l'instant t est donnée par l'équation 5.3.

$$\tau_\Psi \Gamma_r(t) = \frac{1}{r} \int_{-r}^{r} |s'(\tau)| \Psi(\frac{t-\tau}{r}) d\tau \tag{5.3}$$

Il a été prouvé dans [122] que si s satisfait l'équation 5.1 alors la mesure s' calculée avec l'équation 5.3 satisfera une équation similaire avec les mêmes EdS $h(t)$. Cette condition conduit à une simple estimation des EdS par la régression log-log sur une projection en ondelettes à chaque instant t [123]. Pour plus de détails sur l'estimation des EdS et leur calcul, on peut se référer à [118].

Plutôt que de l'utiliser pour la segmentation phonétique comme les auteurs l'ont fait, nous nous sommes basés sur l'analyse de la distribution des EdS pour sélectionner des unités contenant des informations monosyllabiques. A cet effet, nous avons modifié les instances temporelles t utilisées dans leur travail pour considérer des segments plus larges que les unités phonétiques (> 32 ms). Ainsi, nous avons calculé les exposants de singularité sur un fenêtrage de longueur 100 ms autour du temps t.

5.2.3 Sélection des segments candidats

Pour la sélection des segments candidats du signal de parole, nous avons appliqué le calcul des EdS détaillé précédemment sur chaque segment de parole obtenu après suppression des zones silencieuses ou bruitées. Par une analyse de la distribution locale des EdS

et avec une interprétation des divers changements de niveau, nous avons obtenu de petits segments qui contiennent soit un phonème (consonne ou voyelle, C-V) ou un ensemble de consonnes et voyelles (CV, par exemple). Nous avons également obtenu des frontières brutes entre syllabes avec des segments contenant des syllabes qui se chevauchent. Les résultats obtenus avec l'analyse des EdS sont présentés en image dans la Figure 5.2. La sélection des segments candidats peut produire des erreurs telles que les frontières mal identifiées et la reconnaissance manquée de syllabes courtes. Ces erreurs sont corrigées dans la section suivante avec la mesure entropique que nous proposons dans ce travail.

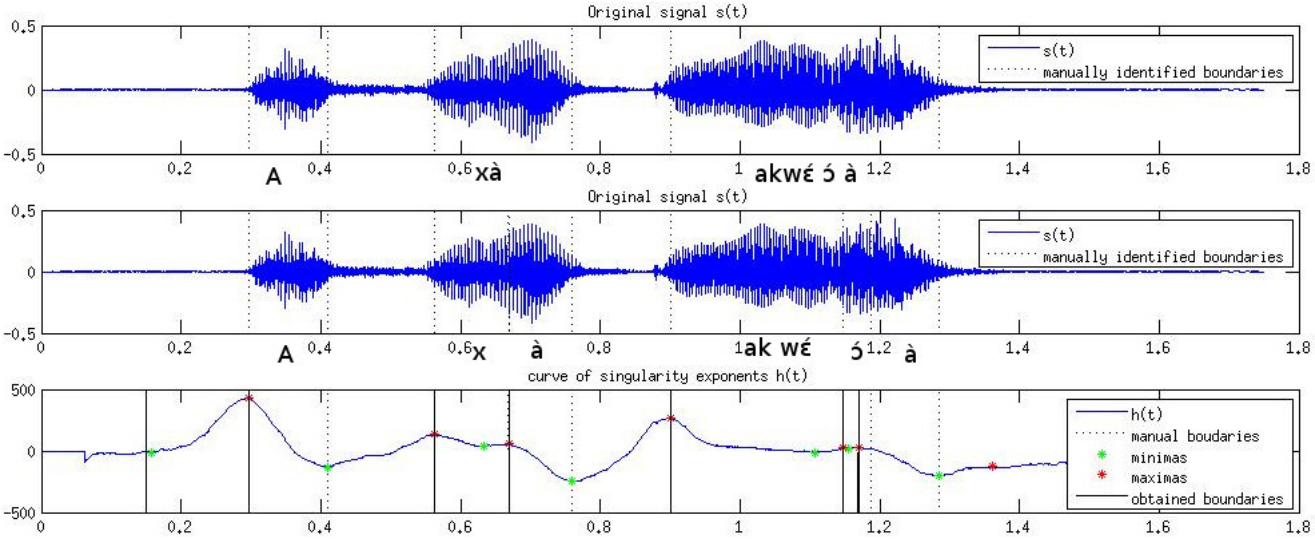


FIGURE 5.2 – **Haut-** Le signal original de parole de l'énoncé "*A xa kwɛ a ?* ", la transcription d'origine tracée en traits verticaux en pointillés, **Milieu-** Le même signal de parole d'origine avec la transcription manuelle syllabique tracée en traits verticaux en pointillés, **BOTTOM-** Courbe des changements de niveau des EdS $h(t)$ avec les minima tracés en vert et les maxima en rouge, les frontières obtenues en lignes verticales solides et les frontières identifiées manuellement en lignes verticales en pointillés.

5.2.4 Détection de frontières syllabiques

Nous décrivons dans cette section la méthode proposée pour la détection efficace des frontières entre syllabes dans chaque segment candidat. Cette méthode permet de mieux exploiter les changements de niveau EdS et comprends deux niveaux de calcul : le calcul de l'énergie à court-terme et de l'entropie.

Plutôt que de calculer l'entropie à partir de la densité spectrale ou d'une échelle fréquentielle, nous l'avons calculé à partir de l'énergie à court-terme du signal dans un

segment de parole afin d'exploiter la variation temporelle des énergies dans les segments qui résultent de la distribution des EdS.

L'énergie à court-terme est calculée à partir de l'équation 5.4, où $s(n)$ est un signal de parole à temps discret, m le taux en nombre d'échantillons, l la longueur d'une fenêtre et w_H la fonction de la fenêtre de Hamming.

$$E(m) = \frac{1}{l} \sum_{n=1}^{l} [s(n) w_H(m-n)]^2 \tag{5.4}$$

Il est à remarquer que les segments voisés ont une énergie supérieure à celle des segments non voisés. Ce qui agit sur le degré de voisement que nous calculons par la mesure entropique. Les valeurs presque nulles de l'énergie entre deux segments permettent un début de délimitation des frontières qui sont par la suite améliorées par l'interprétation linéaire de l'entropie dans les segments.

En raison de leurs propriétés avantageuses, les mesures d'entropie permettent d'obtenir une vue sur le niveau de distribution de la parole et de mesurer le degré d'incertitude dans un signal de parole. Plusieurs descripteurs, basés sur l'entropie de Shannon, ont été proposés pour la segmentation de la parole mais n'ont pas été implémentés pour la segmentation en unités syllabiques. Ce sont des mesures qui prennent des valeurs maximales dans le noyau vocalique et minimales aux frontières des syllabes. Par la suite, nous allons décrire quelques mesures d'entropie et présenter la méthode de détermination des frontières avec l'entropie de Rényi.

5.2.4.1 Entropie de Shannon

L'entropie a été introduite par Claude Shannon pour mesurer la quantité d'information contenue dans un signal aléatoire [124]. Son utilisation est répandue dans la théorie de l'information. Elle est définie comme suit :

$$H(x) = \sum_{k} P(x_k) \times log_2[P(x_k)] \tag{5.5}$$

où $x = \{x_k\}_{0 \leq k \leq N-1}$ est une variable aléatoire discrète (temporelle, fréquentielle ou autre), et $P(x_k)$ la probabilité d'un certain état x_k. L'entropie est exprimée en bit. Si la base du logarithme est $b \neq 2$, on note l'entropie par $H_b(X)$. Elle dépend seulement de la fonction de probabilité $p(x_k)$ de la variable aléatoire x_k. Elle est utilisée au travers des différentes propriétés suivantes :

1. l'entropie de Shannon est toujours positive : $H(x) \geq 0$;

2. l'entropie conditionnelle est majorée : $H(X|Y) \leq H(X)$; Avec égalité si et seulement si X et Y sont indépendants.

3. l'entropie conditionnelle se généralise et en considérant 3 variables aléatoires X, Y, Z, on a : $H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$;

4. $H_b(X) = log_b(a)H_a(X)$; cette propriété permet de changer la valeur de l'entropie d'une base en une autre ;

En traitement de parole, l'entropie de Shannon est utilisée pour la détection de la parole sur la base qu'un segment de parole diffère significativement d'un segment sans parole [125]. En effet, l'entropie de Shannon permet de quantifier la "pointicité" du spectre d'un signal aléatoire. Le spectre est fortement piqué dans un segment de parole et reste uniforme dans un segment sans parole. on obtient donc une entropie faible dans le premier cas et très élevée dans le second cas (incertitude totale).

5.2.4.2 Entropie spectrale

L'entropie spectrale caractérise le désordre présent dans le spectre fréquentiel d'un signal. Elle est définie à partir de la transformée de Fourier à court-terme du signal. Avec un nombre de points fréquentiels N, une fenêtre d'analyse $h(n)$, l'amplitude $S(k, l)$ de la $k^{\text{ième}}$ composante fréquentielle pour la $l^{\text{ième}}$ trame d'analyse est définie comme suit :

$$S(k, l) = \sum_{n=1}^{N} h(n)s(n - l).exp(-j2\pi kn/N) \tag{5.6}$$

avec $0 \leq k \leq K - 1$ et $K = N$. Cette amplitude sert à calculer l'énergie spectrale avec l'expression de l'équation 5.7.

$$S_{energy}(k, l) = |S(k, l)|^2 \tag{5.7}$$

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

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

$$P(k, l) = \frac{S_{energy}(k, l)}{\sum_{i=1}^{N/2} S_{energy}(i, l)} \tag{5.8}$$

avec $1 \leq i \leq N/2$ et $\sum_i P(i, l) = 1$ pour tout l. Les $P(k, l)$ constituent une distribution de probabilités de l'énergie prise pour une densité. Ainsi pour une trame l, l'entropie

spectrale se définit comme suit :

$$H(l) = \sum_{i=1}^{N/2} P(i,l).log_2[P(i,l)] \tag{5.9}$$

$H(l)$ donne l'information sur la répartition spectrale du signal de parole. L'entropie spectrale, dans la littérature, est souvent combinée à l'énergie pour produire des descripteurs robustes pour la caractérisation des signaux de parole bruités ou non [18].

5.2.4.3 Entropie de Tsallis

L'entropie de Tsallis est une généralisation de l'entropie de Shannon définie par Constantino Tsallis [126]. Elle est définie par :

$$H_\alpha = \frac{1}{\alpha - 1}\Big(1 - \sum_{i=1}^{n} P_i\Big) \tag{5.10}$$

où $\alpha \geq 0$. En utilisant l'entropie de Tsallis, pour $\alpha \geq 1$, les composantes avec une forte probabilité contribuent plus à la valeur de l'entropie que les composantes avec une faible probabilité. L'entropie de Tsallis est indiquée comme utile lorsque le système à étudier possède des états distincts avec une forte corrélation.

5.2.4.4 Entropie de Rényi

L'entropie de Rényi, comme l'entropie de Tsallis, est une généralisation de l'entropie de Shannon qui, pour $\alpha = 1$ est réduite à l'entropie de Shannon. L'entropie de Rényi est définie par l'équation 5.11.

$$H_\alpha(p) = \frac{1}{1-\alpha} log_2 \sum_k p^\alpha(k) \tag{5.11}$$

où p est une distribution de probabilités.

5.2.4.5 Détermination des frontières avec l'entropie de Rényi

Dans le cadre de ce travail, il a été judicieux pour nous de considérer l'entropie de Rényi avec un coefficient α assez grand à la place de l'entropie de Shannon et Tsallis afin de pouvoir amplifier les différences entre segments voisés ou non voisés. L'entropie spectrale ne peut être considérée car elle est calculée à partir de composantes fréquentielles et ne peut être appliquée à l'énergie à court-terme du signal. L'entropie de Rényi mesure la complexité du signal quand elle est appliquée à la distribution temps-fréquence [127].

Une distribution temps-fréquence introduite par Wigner et étendue par Ville aux signaux analytiques [128], a été traitée comme une fonction de densité de pseudo-probabilité pour un signal non stationnaire auquel l'entropie de Rényi a été appliquée comme mesure de complexité du signal.

L'idée intuitive de notre travail est d'obtenir une segmentation syllabique dans le domaine temporel en considérant, comme fonction de densité de probabilité, la distribution d'énergies à court-terme dans les segments candidats obtenus (voir équation 5.12).

$$H_\alpha(E) = \frac{1}{1-\alpha} log_2 \sum_k E^\alpha(k) \tag{5.12}$$

Cela renforce les informations temporelles recueillies avec les EdS et distingue clairement les changements entre syllabes. Avec les valeurs d'entropie calculées sur tous les segments candidats, nous avons dégagé un profil entropique pour chaque segment candidat c_i défini par l'expression 5.13.

$$H(E) = [H_{1\alpha}(E) \quad H_{2\alpha}(E) \quad ... \quad H_{k\alpha}(E)] \tag{5.13}$$

Pour exploiter les changements dans l'évolution du profil entropique, la mesure ξ_H a été définie pour toujours prendre le négatif de la valeur de l'entropie. La mesure est donnée par l'équation 5.14.

$$\xi_H = -\int_{]t_1,t_2[} H(E) \tag{5.14}$$

Ainsi une frontière est clairement détectée quand ξ_H est supérieur à 0 et représente la minimale du profil. Une frontière candidate c_i est alors maintenue comme frontière effective quand sa mesure ξ_H répond à la condition, dans le cas contraire, c_i est supprimé des segments candidats. En plus, ξ_H ne change pas de valeur en face des segments candidats qui se chevauchent. Ceci conduit, pendant la segmentation effective, à éliminer les segments qui se chevauchent en établissant de nouvelles frontières par l'analyse linéaire de la mesure ξ_H. L'algorithme 3 résume toute la procédure de la détection des frontières effectives.

Il ressort de cet algorithme que pour chaque itération n, nous recherchons la valeur de la mesure ξ_{iH} qui minimise le profil entropique du segment candidat i. Ce minimum existe s'il est logé entre deux valeurs $[\xi_{\alpha H}\ \xi_{\beta H}]$, si $\xi_{\alpha H}$ et $\xi_{\beta H}$ sont positifs (l'un presque nul) ou de signes opposés et si α et β, étant différents de i, sont deux itérations consécutives. La plus petite valeur ξ_{iH} devient une limite effective du segment si sa valeur est suppérieure à 0. La position effective d'une limite (frontière) dans un segment candidat dépend fortement de la détection de ξ_{iH} qui minimise son profil entropique. Dans le cas contraire, le segment candidat est supprimé après toutes les itérations.

Algorithm 3: Détection automatique de frontières de syllabes.

Data: Sélectionner les segments candidats
Result: c_i : frontières de syllabe

while $i < NbreDeSegments$ **do**
  for $n \leftarrow 1$ **to** l **do**
    $E_k[n] \leftarrow E_k[n-1] + (s[n] \times w_H(m-n))^2$;
    $E_k \leftarrow [E_k \;\; E_k[n]]$;
  end
  $E(n) \leftarrow 1/l \times (E_k)$;
  for $j \leftarrow 1$ **to** k **do**
    $E_\alpha \leftarrow E_\alpha(j-1) + E_\alpha(j)$;
    $H_{j\alpha}(E) \leftarrow 1/(1-\alpha) \times log_2.E_\alpha$;
    $H(E) \leftarrow [H(E) \;\; H_{j\alpha}(E)]$;
  end
  $\xi_{iH} \leftarrow -\int H(E)$;
  $\xi_{iH} \leftarrow \text{argmin } \xi_{iH}$;
  if $\xi_{iH} > 0$ **then**
    $c_i \leftarrow \xi_{iH}$;
  end
end

5.3 Approche basée sur la logique floue pour la segmentation syllabique

Dans cette section, nous présentons la deuxième approche proposée pour la segmentation indépendante du texte de la parole continue en unités syllabiques [4]. En plus d'intégrer les deux premières étapes de la première approche (la suppression des zones silencieuces ou bruitées), elle comprend les étapes suivantes :

- le calcul du taux de passage par zéro et l'énergie à court-terme dans chaque segment candidat généré par l'analyse locale des EdS.

- la génération des ensembles et règles flous pour la phase d'adaptation afin d'améliorer la précision dans la détection des frontières de début et de fin d'un phonèmes ou syllabe. Nous avons employé un système de logique floue pour la phase d'adaptation des segments candidats pour obtenir une précision dans la détermination des frontières.

La figure 5.3 montre la variation temporelle de chaque paramètre calculé sur la phrase prononcée "*A xa kwɛ a ?* ".

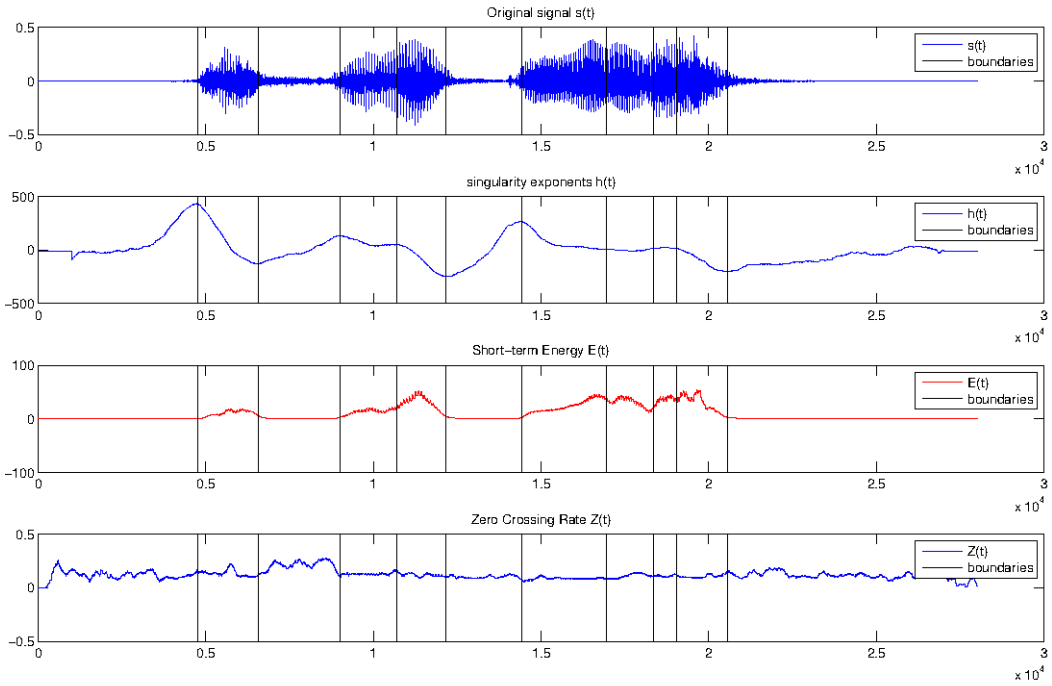


FIGURE 5.3 – **HAUT-** Signal original s(t) et sa transcription manuelle avec les frontières, **MILIEU-1** en bleu, courbe des changements de niveaux des EdS $h(t)$, **MILIEU-2** en rouge, variation temporelle de l'énergie à court-terme E(t), **BAS-** variation temporelle du taux de passage par zéro Z(t).

5.3.1 Les paramètres acoustiques utilisés

Comme la première, la deuxième approche a exploité les informations pertinentes à travers la dynamique du signal de parole donnée par les EdS pour dégager de premières frontières de segments entre minima et maxima. Sur ces segments candidats, sont calculés les deux autres paramètres pour constituer les données à l'entrée du système de logique floue construit pour reconnaitre un segment de phonème ou de monosyllabe. Les équations 5.4 et 1.5 ont servi à calculer respectivement les énergies à court-terme et le taux de passage par zéro. Les coefficients des paramètres calculés constituent ainsi la base de données d'apprentissage et de test pour la phase d'adaptation basée sur le système de logique floue.

5.3.2 La phase d'adaptation basée sur la logique floue

5.3.2.1 Architecture du système de logique floue

La phase d'adaptation de cette approche pour la segmentation automatique de la parole continue en Fongbe a été réalisée en utilisant un modèle de logique floue comprenant

un nombre de règles "if-then". L'architecture utilisée est celle d'un système d'inférence flou décrite dans la section 4.3.3 du chapitre 4. Les entrées du système sont les EdS (désigné SE), l'énergie à court-terme (STE) et le taux de passage par zéro (ZCR) et sa sortie est le degré d'appartenance d'une variable d'entrée à un segment de phonème, de syllabe ou de silence. Les variables d'entrées sont fuzzifiées, selon la variation des coefficients, en trois ensembles complémentaires à savoir : *faible*, *moyen* et *fort*. La variable de sortie est aussi fuzzifiée en trois ensembles complémentaires : *silence*, *phonème*, *syllabe*. En considérant les valeurs des différents coefficients, nous avons obtenu pour les paramètres les ensembles suivants :

- STE : faible - moyen - fort
- ZCR : faible - moyen - fort
- SE : faible - moyen - fort

5.3.2.2 Architecture du DBN utilisée

L'architecture DBN utilisée pour l'adaptation des ensembles flous et la génération automatique des règles floues est celle proposée dans le chapitre 4 (section 4.3.4). Le récapitulatif des paramètres d'apprentissage sont répertoriés dans le tableau 5.1.

TABLE 5.1 – Les paramètres DBN.

RBM Couche 1	200 units
RBM Couche 2	200 units
Pas d'apprentissage	0.01
Nombre d'itérations	100
Taille du batch	8

5.3.2.3 Génération automatique des ensembles flous et des règles floues

Les systèmes de logique floue sont considérés comme des systèmes à base de connaissances. Par le système d'inférence flou et les fonctions floues d'appartenance, la connaissance humaine est incorporée dans leurs connaissances. Le système d'inférence flou et les fonctions floues d'appartenance sont généralement construits par des décisions subjectives ayant une grande influence sur les performances du système. Dans la plupart des applications existantes, les règles floues sont générées par des experts surtout pour des problèmes de contrôle avec seulement quelques entrées. Dans notre approche, nous avons opté pour une génération automatique à partir d'une procédure développée pour ajuster automatiquement les ensembles flous de chaque paramètre. La procédure est la suivante :

1. Chaque variable d'entrée est partitionnée en de sous ensembles de variables linguistiques. Pour un signal $X(t)$, les entrées $X_i, i \in 1, 2, 3$ sont obtenues correspondant aux paramètres (STE, ZCR, SE) calculés. Les sous ensembles sont :

$$\begin{aligned} X_1 \text{ partitionné en } \Big[[M_0...M_k][M_{k+1}...M_{k+1+l}][M_{k+2+l}...M_{k+2+m}]\Big] \\ X_2 \text{ partitionné en } \Big[[N_0...N_k][N_{k+1}...N_{k+1+l}][N_{k+2+l}...N_{k+2+m}]\Big] \\ X_3 \text{ partitionné en } \Big[[P_0...P_k][P_{k+1}...P_{k+1+l}][P_{k+2+l}...P_{k+2+m}]\Big] \end{aligned}$$

avec $k < l < m$. Ces sous ensembles correspondent aux variables linguistiques *faible*, *moyen* et *fort* énumérées ci-dessus. La base d'apprentissage pour le système d'inférence flou est construite comme indiqué dans la matrice suivante :

$$\begin{pmatrix} [M_0...M_k] & [N_0...N_k] & [P_0...P_k] \\ [M_0...M_k] & \cdots & [P_{k+1}...P_{k+1+l}] \\ \vdots & \ddots & \vdots \\ [M_{k+1}...M_{k+1+l}] & [N_{k+1}...N_{k+1+l}] & [P_{k+1}...P_{k+1+l}] \\ \vdots & \ddots & \vdots \\ [M_{k+2+l}...M_{k+2+m}] & \cdots & [P_{k+1}...P_{k+1+l}] \\ [M_{k+2+l}...M_{k+2+m}] & [N_{k+2+l}...N_{k+2+m}] & [P_{k+2+l}...P_{k+2+m}] \end{pmatrix}$$

Ce qui correspondant à la matrice suivante dont les valeurs sont les variables linguistiques :

$$\begin{pmatrix} faible & faible & faible \\ faible & \cdots & moyen \\ \vdots & \ddots & \vdots \\ moyen & moyen & moyen \\ \vdots & \ddots & \vdots \\ fort & \cdots & moyen \\ fort & fort & fort \end{pmatrix}$$

Chaque colonne des matrices représente les valeurs d'une caractéristique donnée.

2. Sélectionner le sous ensemble à réajuster : *faible*, *moyen* ou *fort*.

3. Calculer la valeur moyenne de l'intervalle déduit du sous ensemble ; par exemple pour la variable linguistique *fort* de l'entrée X_i, nous obtenons $\hat{M}_{im} = \frac{M_{k+2+m}+M_{k+2+l}}{2}$.

4. Calculer la valeur moyenne de l'intervalle formé par les valeurs allant de la borne inférieure de l'intervalle précédent à sa valeur moyenne calculée $[M_{k+2+l}...\hat{M_i}m]$;

5. Réduire le sous ensemble courant à $[\hat{M}_{im}...M_{k+2+m}]$. Les valeurs de M_{k+2+l} à $\hat{M}_{im}-1$ sont ajoutées au sous ensemble de la variable linguistique adjacente. Ceci lui assigne un nouvel intervalle $[M_{k+1}...\hat{M}_{im}-1]$. Une nouvelle base d'apprentissage est créée comme suit :

$$\begin{pmatrix} [M_0...M_k] & [N_0...N_k] & [P_0...P_k] \\ [M_0...M_k] & \cdots & [P_{k+1}...P_{k+1+l}] \\ \vdots & \ddots & \vdots \\ [M_{k+1}...\hat{M}_{im}-1] & [N_{k+1}...N_{k+1+l}] & [P_{k+1}...P_{k+1+l}] \\ \vdots & \ddots & \vdots \\ [\hat{M}_{im}..M_{k+2+m}] & \cdots & [P_{k+1}...P_{k+1+l}] \\ [\hat{M}_{im}..M_{k+2+m}] & [N_{k+2+l}...N_{k+2+m}] & [P_{k+2+l}...P_{k+2+m}] \end{pmatrix}$$

6. Une nouvelle configuration de variables d'entrée est obtenue et utilisée comme entrée au système DBN pour l'apprentissage. Le système DBN, pendant la phase d'apprentissage, ajuste les sorties en minimisant les erreurs entre Y_d (sortie désirée) et Y_i (sortie observée). A la fin de l'apprentissage, les règles floues sont déduites pour la configuration courante des variables d'entrée. Les ensembles flous sont attribués à chaque variable de sortie.

7. Appliquer les règles floues obtenues à partir des ensembles flous au système d'inférence pour trouver une sortie nette d'estimation des fonctions d'appartenance. Chaque ensemble flou donné par une variable X_j est divisé en I_j ensembles équidistants et triangulaires. De la même façon, la sortie Y_d est granulée en I_y ensembles flous triangulaires. Ainsi, pour chaque $(X, Y) = (x_1, ..., x_p, y)$

   - le degré d'appartenance à chaque ensemble flou est calculé par : $min\{\mu_{k_1}, 1(x_1), ..., \mu_{k_p}, p(X_$
   - $1 \leq k_j \leq I_j$ et $1 \leq k_y \leq I_y$;
   - μ_{k_j} indique la fonction d'appartenance de la $K_j -$ ème valeur linguistique de X_j ;
   - l'ensemble flou $(k_1, ..., k_p, k_y)$ avec la fonction d'appartenance MAX génère la règle
     $R_{(k_1,...,k_p)}$: si x_1 est $\mu k_1, 1$ et ... et x_p est $\mu k_p, p$ alors y est μ_{k_y} ;

- un degré d'appartenance sera attribué à chaque règle comme la valeur pondérée $\beta_{(k_1,\ldots,k_p)}$.

A la fin de cette étape, nous calculons le taux de frontières correctement détectées lié à l'apprentissage de la configuration courante des variables d'entrée que sont STE, ZCR et SE ;

8. Répéter les étapes allant de 3 à 7 pour une nouvelle configuration des ensembles flous jusqu'à obtenir $\hat{M}_{im} = M_{k+l+2}$;

9. Sélectionner la variable linguistique suivante avec ses sous ensembles correspondants et répéter les étapes de 3 à 7 ;

10. Arrêter la procédure une fois tous les ensembles de toutes les variables linguistiques ont été réajustées. Pour finir, on détermine les ensembles flous et règles qui optimisent le taux de frontières correctement détectées en minimisant l'erreur en sortie.

A la fin, nous obtenons les meilleurs ensembles flous avec des règles floues qui fournissent avec précision les frontières effectives des segments de parole dégagés par les EdS. Commençant par une base d'apprentissage contenant 27 règles, la phase d'apprentissage a généré 8 règles floues formelles énumérées dans le tableau 5.2 [129]. Dans ce tableau, x est une variable qui peut prendre les valeurs *faible*, *myen* or *fort*.

TABLE 5.2 – Les règles floues générées.

No Règles	Entrée			Sortie
	STE	ZCR	SE	
1	faible	x	x	silence
2	x	faible	x	silence
3	x	x	faible	silence
4	moyen	moyen	moyen	phonème
5	moyen	moyen	fort	phonème/syllabe
6	x	fort	x	silence
7	fort	moyen	moyen	phonème
8	fort	moyen	fort	phonème/syllabe

La figure 5.4 montre les résultats de la phase d'adaptation. Ces résultats sont liés aux différentes valeurs des coefficients STE et ZCR en fonction de SE. Pour chaque meilleur ensemble de SE, ZCR est tracé sur l'axe des abscisses et STE sur l'axe des ordonnées. Les conclusions issues des résultats sont :

- quand SE est dans l'intervalle $[-0,0184...0,4]$, STE dans l'intervalle $[-0,0025...0,015]$ et ZCR dans l'intervalle $[0...0,089]$, le segment en entrée contient du silence ou du bruit (voir le tracé en haut sur la figure 5.4) ;

- quand SE est dans l'intervalle $[0,35...0,9]$, STE dans l'intervalle $[0...0,032]$ et ZCR dans l'intervalle $[0,009...0,1]$, le segment en entrée contient un phonème (voir le tracé au milieu sur la figure 5.4) ;

- quand SE est dans l'intervalle $[0,8...1,07]$, STE dans l'intervalle $[0...0,032]$ et ZCR dans l'intervalle $[0,009...0,1]$, le segment en entrée contient soit un phonème soit un monosyllabe (voir le tracé en bas sur la figure 5.4) ;

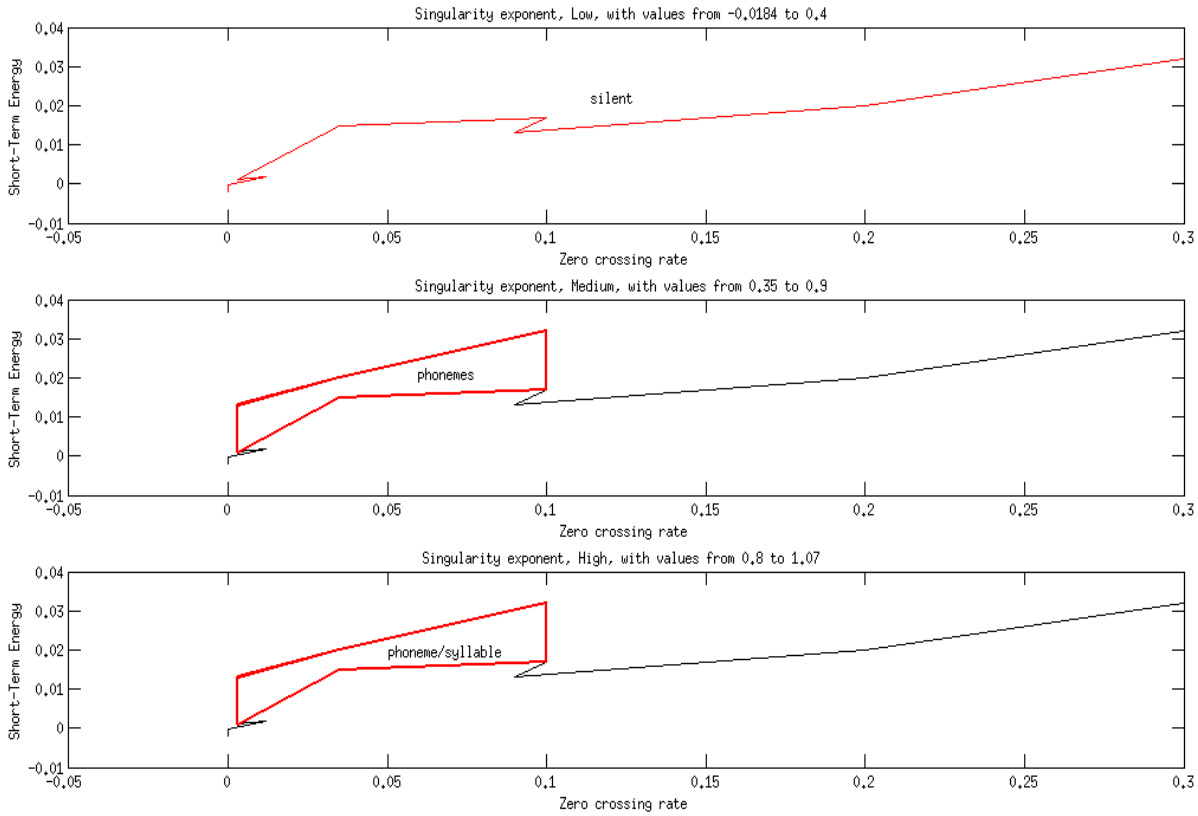


FIGURE 5.4 – Résultats de la phase d'adaptation.

5.4 Evaluation de performances

5.4.1 Description des métriques d'évaluations

Les deux algorithmes que nous avons proposés dans ce travail ont été appliqués au corpus de parole continue en Fongbe dont la description est présentée au chapitre 2 (section 2.3, tableau 4.3). Nous avons comparé les frontières détectées automatiquement par les méthodes proposées contre les frontières de la transcription manuelle des signaux de parole pris en exemple depuis le corpus de parole. L'évalution est réalisée avec les métriques de base couramment utilisés dans l'état de l'art. Il s'agit du taux de succès (TS), du taux de fausses alertes (FA) et du taux de sur-segmentation (TSS). Ils sont définis comme suit :

- TS (Taux de succès) : représente le taux de frontières correctement détectées et défini avec l'expression N_c/N_T. Il utilise le nombre de frontières correctement détectées N_c et le nombre total de frontières dans le signal N_T.

- FA (taux de Fausses Alarmes) : représente le taux de frontières détectées par erreur et défini avec l'expression $(N_{T_d} - N_c)/N_{T_d}$. Il utilise le nombre total de frontières détectées N_{T_d} et le nombre de frontières correctement détectées N_c.

- TSS (Taux de Sur-Segmentation) : montre combien (plus ou moins) est le nombre total des détections de l'algorithme par rapport au nombre total des limites de référence provenant de la transcription $(N_{T_d} - N_T)/N_T$.

Nous décrivons la qualité générale de l'algorithme par le calcul des scores des mesures F-*value* et R-*value*. F-*value* est une valeur scalaire habituellement utilisée pour l'évaluation d'une méthode de segmentation. Elle tient compte simultanément du taux de frontières correctement détectées et du taux de frontières détectées par erreur et se définit par l'expression $F_1 = (2 \times (1 - FA) \times TS)/(1 - FA - TS)$. La mesure R-*value*, similaire à F-*value*, évalue la proximité de l'algorithme à la segmentation idéale. Elle a été proposée par Rasanen et. al [130] et est considérée comme plus précise que F-*value*. Elle tient compte simultanée du taux de succès TS et du taux de sur-segmentation TSS et est définie à partir des valeurs r_1 et r_2 dont les expressions sont :

$$r1 = \sqrt{(1 - TS)^2 + OS^2} \tag{5.15}$$

$$r2 = (TS - OS - 1)/\sqrt{2} \tag{5.16}$$

$$R = 1 - (|r1| + |r2|)/2 \tag{5.17}$$

Pour situer nos deux approches dans l'état de l'art, nous avons procédé à une comparaison entre les méthodes proposées et quelques unes de l'état de l'art couramment utilisées pour la segmentation de la parole utilisant le corpus TIMIT [23]. Pour ce faire, nous avons appliqué les méthodes recensées sur notre base de données afin d'évaluer leurs performances contre celles choisies dans l'état de l'art. Les métriques utilisées pour la comparaison sont le nombre de frontières correctes, les erreurs d'insertion et les erreurs de suppression. Elles sont calculées avec les équations suivantes :

- détections correctes = $\frac{\text{nombre de détections correctes}}{\text{nombre de segments obtenus automatiquement}}$
- erreurs d'insertion = $\frac{\text{nombre d'insertions}}{\text{nombre de segments obtenus automatiquement}}$
- erreurs de suppression = $\frac{\text{nombre de suppression}}{\text{nombre de segments dans la référence}}$

5.4.2 Performance de l'approche basée sur l'entropie de Rényi

Le tableau 5.3 présente les résultats obtenus avec différentes durées prises entre les fenêtres (FSS) appliquées aux signaux de parole du dataset Fongbe [24]. Pour les trois mesures utilisées et pour toutes les durées FSS, on réalise que ξ_H surpasse les performances des EdS. La performance globale de cette approche est obtenue à $0,2s$ considéré comme la durée moyenne de la syllabe. Nous pouvons remarquer une amélioration significative du taux de fausses alarmes (F_A). Cela montre que ξ_H identifie correctement une frontière à chaque fois qu'une frontière est détectée et supprime efficacement les insertions introduites par les EdS (voir figure 5.5). La figure 5.6 montre l'évolution de TS et FA par rapport aux durées FSS tracées en abscisses. Une observation est que plus FSS est petit, plus le taux de frontières mal identifiées est élevé. Pour les meilleures performances de la mesure proposée ξ_H comparée aux EdS, nous obtenons une augmentation du taux de succès (24%) et une réduction du taux de fausses alarmes (26%) pour la même durée FSS (0,2s).

En plus d'utiliser les métriques standards, nous avons évaluer l'approche avec les mesures F-*value* et R-*value* qui prennent en compte les performances globales de la méthode de segmentation. Le tableau 5.4 présente ces performances selon les méthodes comparées. On pourra remarquer une amélioration significative des mesures $Fvalue$ et $Rvalue$. Pour $FSS > 0,15$, $Fvalue$ et $Rvalue$ ont pratiquement les mêmes valeurs quand nous appliquons la mesure ξ_H. Ceci peut être clairement vu sur la figure 5.7. Les performances

23. https ://catalog.ldc.upenn.edu/LDC93S1
24. Corpus audio de parole continue en Fongbe

TABLE 5.3 – Les mesures de performance (TS, F_A et TSS) pour la détection de frontières avec la simple méthode des EdS et avec ξ_H en fonction de la durée FSS.

FSS	Metriques	SE	ξ_H
		Dataset Fongbe	Dataset Fongbe
0,05 s	TS	0,31	0,51
	F_A	0,85	0,65
	TSS	0,69	0,49
0,1 s	TS	0,44	0,74
	F_A	0,77	0,42
	TSS	0,56	0,26
0,15 s	TS	0,51	0,77
	F_A	0,67	0,4
	TSS	0,49	0,23
0,2 s	TS	0,57	0,81
	F_A	0,6	0,34
	TSS	0,43	0,19
0,25 s	TS	0,56	0,8
	F_A	0,62	0,39
	TSS	0,444	0,2
0,3 s	TS	0,54	0,79
	F_A	0,61	0,39
	TSS	0,46	0,21

obtenues à $FSS = 0,2s$ montre que dans cette étude, notre algorithme est bien adapté pour la détection avec haute précision des frontières syllabiques.

Cela conduit à confirmer que l'entropie de Rényi calculée sur une distribution d'énergies à court terme améliore efficacement la détection de frontière entre syllabes pour les énoncés du Fongbe. Une détection de base avec les exposants de singularité permet de déterminer les niveaux de changement dans la dynamique temporelle du signal de parole. La figure 5.5 montre la segmentation du signal de l'énoncé "$A\ \ xa\ \ kw\varepsilon\ \ a\ ?$ " en unités syllabiques. Sur cette figure, on remarque aisément que les exposants de singularité ont une variance plus élevée dans les points contenant du silence et dans des pauses marquées. Cela rend donc difficile l'utilisation seule des exposants de singularité pour la détection automatique des frontières parce qu'ils incluent parfois ces différents points dans les segments délimités. Par contre, les énergies ont une variance faible dans ces points et sont sensibles dans les pauses de parole. Ainsi l'utilisation des énergies à court-terme pour le calcul de l'entropie de Rényi nous a conduit à améliorer la détection automatique des frontières. Les performances globales ont été déterminées avec la mesure proposée ξ_H. La

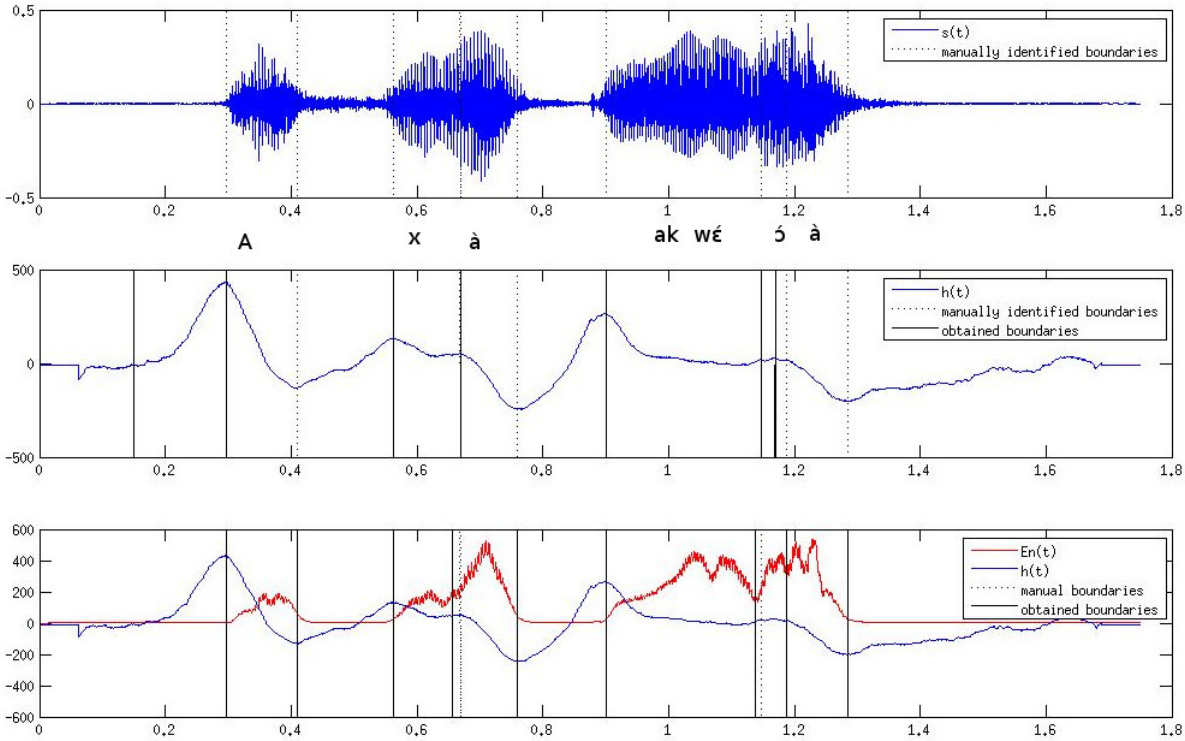


FIGURE 5.5 – **HAUT-** Signal original et sa transcription avec les frontières entre syllabes de la phrase prononcée "*A xa kwɛ a ?* ", **MILIEU-** Courbe des changements de niveaux des EdS $h(t)$, **BAS-** L'énergie court-terme En tracée en rouge avec la distribution des EdS pour montrer la différence dans les variations et l'information additionnelle fournie par le calcul de l'entropie de Rényi. Ceci permet de retrouver facilement les frontières entre syllabes qui sont tracées en noir.

langue Fongbe étant une langue tonale, les syllabes sont affectées par le ton de certains phonèmes qui s'ajoute parfois à la prononciation d'un son pendant un souffle. Il est à remarquer que la prononciation avec le ton a un impact sur la détection des pauses entre syllabes et augmente la durée du son dans chaque FSS.

Notre approche a été aussi comparée avec quatre autres méthodes de l'état de l'art. Une comparaison est faite avec la méthode de base de Mermelstein qui est incluse dans certains travaux comme [131] pour la segmentation syllabique dont les performances sont similaires. Nous avons également implémenté la méthode de Villing [113] et une autre méthode basée sur le calcul des caractéristiques MFCC et étudiée dans [114]. Avec cette dernière méthode, nous découpons le signal original en de courtes trames chevauchantes, et pour chaque trame, nous calculons un vecteur de caractéristiques composé de coefficients MFCC. La dernière méthode implémentée pour la comparaison est une méthode indépendante des modèles acoustiques ou d'alignement forcé pour une transcription ma-

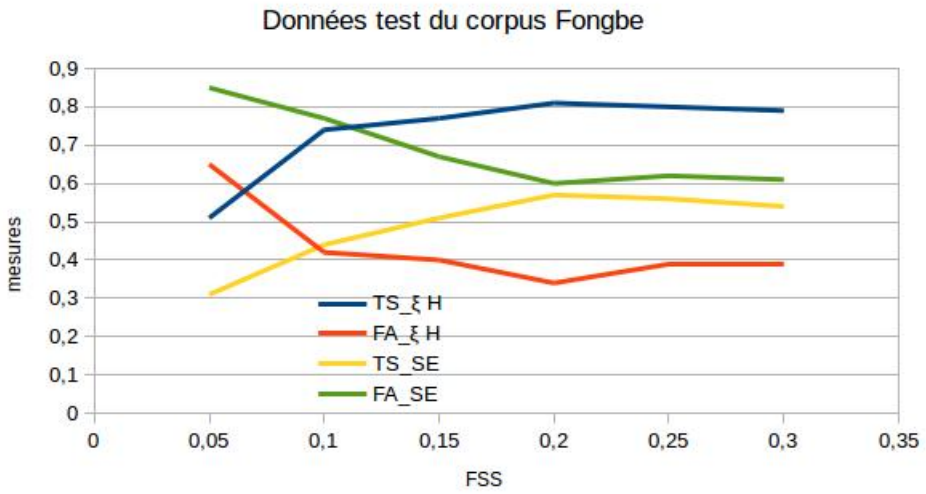


FIGURE 5.6 – Courbes de TS et F_A des EdS et de ξ_H en fonction des durées FSS.

TABLE 5.4 – Les mesures de performance ($Fvalue$ et $Rvalue$) pour la détection de frontières avec la simple méthode des EdS et avec ξ_H en fonction de la durée FSS.

FSS	Metriques	SE	ξ_H
		Dataset Fongbe	Dataset Fongbe
0,05 s	$Fvalue$	0,20	0,42
	$Rvalue$	0,02	0,31
0,1 s	$Fvalue$	0,30	0,65
	$Rvalue$	0,21	0,63
0,15 s	$Fvalue$	0,40	0,67
	$Rvalue$	0,31	0,67
0,2 s	$Fvalue$	0,47	0,73
	$Rvalue$	0,39	0,73
0,25 s	$Fvalue$	0,45	0,70
	$Rvalue$	0,38	0,72
0,3 s	$Fvalue$	0,45	0,69
	$Rvalue$	0,35	0,70

nuelle, mais fonctionne à base d'un réseau de neurones profond (DNN) pour la détection des frontières entre syllabes. Pour ce faire, nous avons adopté la procédure décrite dans [15] mais avec une longueur de fenêtrage adaptée à une syllabe.

Le tableau 5.5 présente les scores obtenus pour chaque méthode implémentée. Il montre clairement que la performance globale, en fonction des trois critères choisis, est obtenue avec l'algorithme proposé, même si la méthode DNN présente des performances similaires à l'exception des erreurs d'insertion.

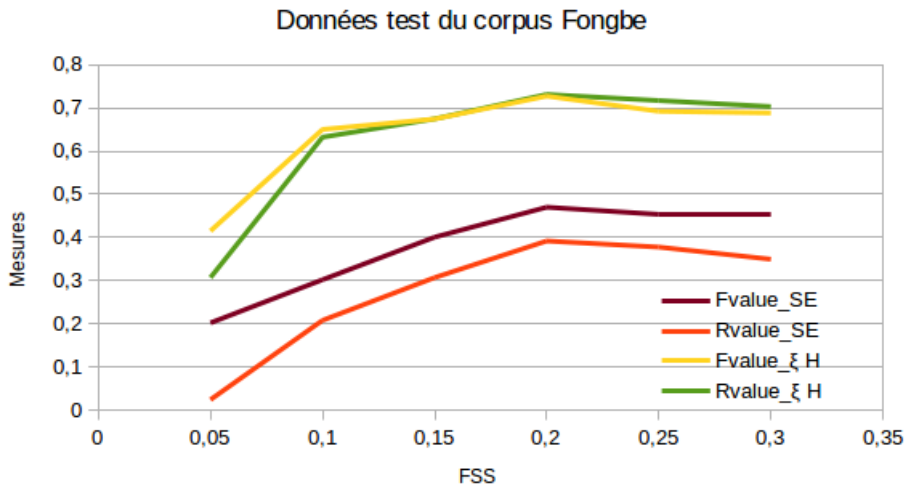


FIGURE 5.7 – Courbes de Fvalue et Rvalue de chaque méthode

TABLE 5.5 – Les scores de chaque algorithme de détection de frontières entre syllabes.

Scores	Mermelstein [104]	Villing [113]	MFCC [114]	DNN [15]	Algorithme proposé
Détections correctes	75,2%	87,7%	82,5%	88,4%	89,9%
Détections d'omissions	24,8%	12,3%	17,5%	11,6%	10,1%
Détections d'insertions	17,6%	0%	1,4%	1,7%	0,17%

5.4.2.1 Performance de l'approche basée sur la logique floue

Les expériences, pour cette approche, ont été réalisées avec une durée de 30 ms entre les fenêtres appliquées aux signaux de parole. Cette durée est considérée comme la durée moyenne d'un phonème prononcé. En plus des méthodes basées sur les EdS, les coefficients MFCC et le DNN, nous avons implémenté deux autres méthodes intéressantes rencontrées dans les travaux [132] et [133]. Dans la première méthode [132], les auteurs ont utilisé l'alignement forcé avec l'algorithme de Viterbi basé sur les Modèles de Markov Cachés (MMCs) pour émettre des hypothèses sur les frontières de transition entre les sons dans une parole. Ensuite, ils exploitent la technique des SVMs pour affiner les hypothèses émises et dégager les frontières effectives. Dans la deuxième méthode [133], les auteurs se sont basés sur une analyse de régression pour la fusion de prévisions de plusieurs moteurs indépendants de segmentation. La fonction principale de la méthode est basée sur la Fusion de Régression de Prévisions des Frontières (FRPF) pour la segmentation automatique de la parole. Toutes ces méthodes ont été implémentées pendant nos expérimentations et appliquées aux données audio du corpus de parole continue en Fongbe.

Le tableau 5.6 présente les résultats de l'évaluation de performance des différentes méthodes implémentées. On peut remarquer que l'approche proposée avec l'adaptation

TABLE 5.6 – Mesures de performance de chaque méthode.

Metriques	FRPF [133]	SVM [132]	DNN [15]	S.E	MFCC	Méthode proposée
TS	0,55	0,61	0,85	0,81	0,73	0,87
F_A	0,35	0,29	0,2	0,2	0,22	0,2
TSS	0,45	0,39	0,15	0,19	0,27	0,13
$Fvalue$	0,59	0,66	0,82	0,80	0,75	**0,83**
$Rvalue$	0,36	0,45	0,79	0,73	0,62	**0,82**

des frontières basée sur la logique floue donne les meilleures performances ($Fvalue = 0,83$, $Rvalue = 0,82$) dans la détection des frontières entre phonèmes ou syllabes. Il est donc clair que l'utilisation des connaissances expertes additionnelles et la procédure automatique de génération des ensembles flous et de règles floues améliorent la détection des frontières comparée à la méthode des EdS.

TABLE 5.7 – Résultats des algorithmes.

Scores	FRPF [133]	SVM [132]	DNN [15]	S.E	MFCC	Algorithme proposé
Détections correctes	65,9%	76,4%	91,6%	89,6%	87,2%	**92,7%**
Détection d'omissions	34,1%	23,6%	8,4%	10,4%	12,8%	**7,3%**
Détections d'insertions	12,2%	9,4%	0,7%	2,6%	5,8%	**1,3%**

Le tableau 5.7 présente les résultats des scores obtenus avec les métriques tels que le nombre de détections correctes, les erreurs d'insertion et les erreurs de suppression. On remarque que la performance globale sur les trois critères d'évaluation est obtenue avec l'algorithme proposé qui, comparé aux autres, nécessite un longue durée pour la génération des ensembles et règles flous. Nous mettons l'accent sur le fait que ces performances obtenues sont valables sur les données audio du Fongbe incluant sa complexité tonale. La méthode utilisant le DNN montre des performances similaires avec un bon score pour la détection des insertions. Elle est donc adaptée pour la segmentation automatique de la parole continue en Fongbe.

Conclusion

Dans ce chapitre, nous avons proposé deux méthodes pour la segmentation automatique de la parole continue en langue Fongbe. Les deux méthodes utilisent une approche non-linéaire pour l'analyse du signal de parole. Elles exploitent les caractéristiques du domaine temporel du signal tels que les propriétés géométriques des exposants de singularité détaillées dans [118] , l'énergie à court terme du signal et le taux de passage par zéro. La première méthode calcule, dans un premier temps, les exposants de singularité dans chaque point du signal de parole. Les exposants sont ensuite combinés à une mesure

entropique calculée sur l'énergie à court-terme du signal pour améliorer la précision dans la détection des frontières entre syllabes. L'algorithme proposé dans ce chapitre a été détaillé et décrit comme étant un algorithme très simple à mettre en œuvre et moins coutêux en calcul. Compte tenu des performances obtenues, nous pouvons conclure que l'entropie Rényi peut être appliquée à une distribution d'énergies et incluse dans un algorithme de segmentation automatique de la parole par l'usage d'une analyse de singularité locale. La deuxième méthode exploite les connaissances expertes que fournissent les mécanismes de la logique floue pour opérer une segmentation de la parole continue en de petites unités contenant soit un phonème soit une monosyllabe. Ces deux méthodes proposées ont été comparées aux méthodes interressantes dans la littérature pour une évaluation de performances globales. Cette comparaison positionne très bien les méthodes proposées dans l'état de l'art en montrant leurs performances sur les signaux de parole en Fongbe.

Chapitre 6

Reconnaissance automatique de la parole continue en Fongbe

"*xan é ma yon sé a é non ji do min si non koho ta haa*"

Sommaire

Introduction

La reconnaissance automatique de la parole (RAP) est une technologie qui permet à un ordinateur d'identifier des mots prononcés par une personne dans un microphone. Un système de reconnaissance automatique de la parole convertit la parole en texte afin d'en extraire la signification sémantique du texte. Il est, de nos jours, de plus en plus utilisé, grâce notamment à l'arrivée de produits commerciaux grand public de bonne qualité et financièrement accessibles. Les applications RAP ont été réalisées avec succès pour la plupart des langues occidentales telles que l'Anglais, le Français, l'Italien etc., pour la plupart des langues asiatiques comme le Chinois, le Japonais, l'Indien etc., parce qu'elles disposent d'une grande quantité de ressources linguistiques numériques [56]. En Afrique, malgré les 2.000 langues parlées, la technologie de la reconnaissance de la parole est moins répandue à cause du manque ou de l'indisponibilité des ressources linguistiques numériques de la plupart de ses langues (vernaculaires pour la plupart). Aussi, pour la plupart du temps, elles ne sont pas écrites (pas de grammaire formelle, nombre limité de dictionnaire, manque de linguistes). Malgré les insuffisances, certaines ont fait objet d'études et disposent désormais de ressources linguistiques pour permettre la réalisation de systèmes de reconnaissance vocale. Par exemple, dans le contexte du projet ALFFA [25], des systèmes de RAP ont été réalisés pour quatre (4) langues de l'Afrique Subsaharienne (le swahili, le hausa, l'amharique et le wolof) [134].

L'objectif principal du travail présenté dans ce chapitre est d'introduire un premier système de reconnaissance automatique de la parole en Fongbe [2]. Comme il a été démontré dans le chapitre 1, le Fongbe est une langue peu dotée ne disposant pas de ressources linguistiques numériques telles qu'un corpus audio et un corpus de texte adaptés à un système de reconnaissance automatique. Le chapitre 2 a présenté une description des données collectées dans le cadre de la construction de ces deux ressources importantes pour un système RAP en Fongbe. Elles sont utilisées ici pour construire des modèles acoustiques et de langage pour le décodage de la parole continue en Fongbe. La modélisation acoustique a été élaborée à un niveau graphémique et la modélisation du langage a donné lieu à deux modèles pour des fins de comparaison. La boîte à outils Kaldi ASR a été utilisée pour entrainer les modèles acoustiques sur les données de paroles collectées. Pour la modélisation du langage, nous avons utilisé la boîte à outils SRLIM [26] pour construire des modèles du langage tri-grammes entraînés sur les données textuelles collectées. En post traitement, pour améliorer la performance de notre système de reconnaissance qui intègre

25. `http://alffa.imag.fr`
26. www.speech.sri.com/projects/srilm/

les différents modèles développés, les voyelles ont été transformées pendant une phase de normalisation des différents tons du Fongbe. Nous avons décidé d'utiliser la boîte à outils Kaldi [135] parce que ses modules de reconnaissance sont suffisamment rapides dans le processus de reconnaissance.

6.1 Kaldi et la reconnaissance automatique de la parole

Dans cette section, nous abordons les techniques employées par Kaldi pour la reconnaissance automatique de la parole.

6.1.1 La reconnaissance automatique statistique de la parole

Le but de la RAP statistique est de décoder les séquences de mots dans une parole donnée. Le terme *décodage* trouve son origine dans la terminologie des MMCs. En reconnaissance de la parole, il est équivalent à la *reconnaissance* d'une séquence de mots $W*$ étant donné les observations acoustiques comme décrit dans l'équation 6.2. La meilleure séquence de mots ne dépend pas des probabilités des caractéristiques acoustiques $P(a)$.

$$w^* = argmax_w\{P(w|a)\} = argmax_w \left\{ \frac{P(a|w) \times P(w)}{P(a)} \right\} \quad (6.1)$$

$$= argmax_w\{P(a|w) \times P(w)\} \quad (6.2)$$

La modélisation acoustique est l'estimation des paramètres θ d'un modèle de sorte que la probabilité $P(a|w;\theta)$ soit la plus précise que possible. La probabilité $P(w)$ intervient dans la modélisation du langage. Le processus de décodage comprend la paramétrisation de la parole, le calcul des caractéristiques acoustiques sur de petites fenêtres extraites du signal de parole et le décodage lui-même à l'aide de la méthode *beam search*.

L'amélioration de la précision du moteur de la reconnaissance de la parole dépend principalement de l'amélioration des modèles acoustiques et du langage.

6.1.2 Paramétrisation de la parole

L'objectif de la paramétrisation de la parole est de réduire les influences environnementales négatives sur la reconnaissance de la parole. La parole varie selon un certain nombre d'aspects :

- les différences entre la prononciation des locuteurs varie selon le sexe, le dialecte, la voix etc. ;
- les bruits environnementaux ;
- le canal d'enregistrement ; par exemple, le signal téléphonique est réduit à une bande de fréquences comprise entre 300 à 3000 Hz ; la qualité du signal du téléphone mobile influe aussi sur la qualité du signal audio ;

Différentes méthodes de paramétrisation existent et peuvent permettre d'améliorer la robustesse du système de reconnaissance de la parole pour différentes conditions d'enregistrement.

La paramétrisation de la parole extrait, de l'onde brute du signal, des caractéristiques acoustiques distinctives. Les deux méthodes efficaces les plus utilisées pour la paramétrisation de la parole sont les coefficients MFCCs [136] et la prédiction linéaire perceptuelle (PLP) [41]. Sur un signal audio échantillonné et quantifié, les transformations MFCCs ou PLP calculent efficacement sur une fenêtre donnée les statistiques avec une dimension réduite. Nous avons choisi dans notre travail d'utiliser les coefficients MFCCs car ils sont efficacement calculés par la boîte à outils que nous avons choisi pour notre système de reconnaissance. Les coefficients MFCC statistiques sont calculés pour chaque fenêtre du signal échantillonné. Ainsi, pour une fenêtre de 25ms décalée de 10 ms avec une fréquence d'échantillonnage de 16 Khz, nous obtenons 13 coefficients cepstraux statiques à partir des $16000 \times 0,025 = 400$ échantillons d'une fenêtre donnée (voir la section 1.5.3.1 pour le processus de calcul des MFCCs). Les coefficients MFCC statiques sont généralement étendus par leurs dérivées temporelles $\triangle + \triangle\triangle$. En conséquence, MFCC $\triangle + \triangle\triangle$ extrait $13 + 13 + 13 = 39$ caractéristiques acoustiques pour une trame. Le vecteur original pour les 400 échantillons audio dans une trame est réduit au vecteur des 39 caractéristiques acoustiques MFCC $\triangle + \triangle\triangle$.

6.1.3 Les transformations de caractéristiques

Les transformations de caractéristiques sont généralement appliquées en plus de la paramétrisation MFCC ou PLP. Elles sont appliquées par trame et prennent généralement compte du contexte de plusieurs trames précédentes (contexte gauche) et consécutives (contexte droite).

Les transformations linéaires ou affines sont respectivement exprimées par les multiplications matricielles Ax et Ax^+. La matrice A représente la matrice. x est le vecteur d'entrée et Ax représente les caractéristiques transformées. Les transformations affines

utilisent un vecteur étendu $(x^+)^T = (x_1, \cdots, x_n, 1)$ et la matrice $A : (n+1) \times (n+1)$. En fonction des données acoustiques, l'objectif est de choisir, parmi la grande variété de transformations disponibles, la plus appropriée. Certaines transformations sont estimées discriminatives, d'autres utilisent des modèles génératifs. Il en existe qui sont dépendants ou non du locuteur. La boîte à outils Kaldi implémente un nombre varié de transformations de caractéristiques dont voici quelques unes :

- HLDA (*Heteroscedastic Linear Discriminant Analysis*, [137]) ;
- LDA (*Linear Discriminant Analysis*, [138]) ;
- MLLT (*Maximum Likelihood Linear Transform*, [138]) ;
- ET (Exponential Transform, [139]) ;
- CMVN (Cepstral Mean and Variance Normalisation, [140]) ;

La combinaison LDA + MLLT applique la transformation de caractéristiques en deux étapes : LDA réduit la dimension des caractéristiques et MLLT applique une transformation linéaire simple [138]. Le HLDA estime la réduction de la dimension et la transformation de l'espace en une seule étape [137]. La combinaison LDA + MLLT est effectuée de la même façon que le HLDA et gagne une amélioration significative sur la transformation $\triangle + \triangle\triangle$.

6.1.4 La modélisation acoustique

La modélisation acoustique est sans doute le cœur de la reconnaissance de la parole. Le modèle acoustique estime la probabilité $P(a|w;\theta)$ de génération des caractéristiques acoustiques a pour un mot w donné.

Les méthodes de modélisation acoustique les plus réussies n'estiment pas directement la probabilité $P(a|w)$, mais plutôt la probabilité $P(a|f_1f_2f_3f_4)$ de la génération des caractéristiques a pour les phones $w = f_1f_2f_3f_4$ qui forment la prononciation du mot w. Le phone est l'unité contrastive la plus petite de la parole. Les caractéristiques acoustiques d'un phone dépendent de son contexte. Le phone précédent et le suivant influencent fortement le son du phone au milieu.

Le triphone est une séquence de trois phones qui capturent le contexte d'un seul phone. Par conséquent, les propriétés acoustiques des triphones varient beaucoup moins en fonction du contexte du phone. Notons que certaines combinaisons de préfixes ont le même effet sur le phone exemple, par exemple t et p ont le même effet sur i. Ainsi

ces triphones seront mis en ensemble afin de réduire le nombre de triphones pour la modélisation acoustique.

La méthode statistique basée sur les modèles de Markov cachés est très puissante dans la caractérisation des échantillons de données observés d'une série temporelle discrète avec un état inconnu [141]. Dans le cas de la reconnaissance de la parole, les états cachés du MMC représentent généralement les monophones ou triphones et nous observons les échantillons de caractéristiques acoustiques. Les MMCs ont deux types de paramètres : les probabilités de transition entre les états et la distribution probabilistique pour générer l'observation dans un état donné. Ces paramètres doivent être estimés dans l'apprentissage du modèle acoustique.

La probabilité de transition est une probabilité de changement de l'état q à l'état u. Chaque transition est représentée comme un arc $e = qu$ entre les états q et u. La probabilité est généralement représentée par le poids w_e de l'arc e. Le modèle de Markov émet une observation lors de la traversée sur ses arcs. Le modèle de Markov caché émet l'observation stochastique basée sur la distribution probabilistique liée à l'état visité. En reconnaissance de la parole, une distribution gaussienne multivariée est généralement utilisée pour modéliser les probabilités des états MMC. Les paramètres de la distribution gaussienne sont individuellement estimés pour chaque état. Cependant, les états sont généralement regroupés pendant l'apprentissage du modèle acoustique et les états au sein d'un même groupe partagent les mêmes paramètres pour la distribution gaussienne.

La boîte à outils Kaldi utilise l'algorithme d'apprentissage Viterbi pour entraîner les modèles acoustiques MMC. Elle modélise les probabilités d'observation à l'aide de la distribution gaussienne multivariée avec la dimension des caractéristiques acoustiques a.

En règle générale, les probabilités de transition sont initialisés avec des valeurs uniformément distribuées. Les probabilités d'observation sont initialisées par la distribution gaussienne multivariée de μ et σ définis sur la moyenne globale et la matrice de covariance globale estimée sur toutes les données acoustiques de l'apprentissage.

L'apprentissage Viterbi des modèles acoustiques Kaldi applique le critère Viterbi dans l'attribution de l'observation acoustique aux états MMC. Les travaux récents ont révélé que l'apprentissage Viterbi est tout aussi efficace pour la reconnaissance de la parole continue que l'algorithme de Baum-Welch [142].

Etant donné un ensemble d'observations d'apprentissage O^r, $1 \leq r \leq R$ et une séquence d'état MMC $1 < j < N$, la séquence d'observations est alignée à la séquence d'états via un alignement. Le meilleur alignement T résulte de la maximisation de l'équation 6.3

pour $1 < i < N$.

$$\phi N(T) = max[\phi_i(T)a_i N] \tag{6.3}$$

La variable $\phi_i(o_t)$ est calculée récursivement à l'aide de l'équation 6.4.

$$\phi_i(o_t) = b_j(o_t)max \begin{cases} \phi_j(t-1)ajj \\ \phi_{j-1}(t-1)aj-1j \end{cases} \tag{6.4}$$

Les conditions initiales sont $\phi_1(1) = 1$ et $\phi_j(1) = a_{1j}b_j(o_1)$, pour $1 < j < N$. Initialement, les paramètres du modèle sont mis à jour en fonction du meilleur alignement de l'observation individuelle des états et des composants gaussiens au sein des états. Ensuite les probabilités de transition sont estimées à partir des fréquences relatives avec l'équation 6.5, où Aij désigne le nombre de transitions de l'état i à l'état j.

$$\hat{a}_{ij} = \frac{A_{ij}}{\sum_{k=2}^{N} A_{ik}} \tag{6.5}$$

La fonction d'indication $\psi_{jm}^r(t)$ est utilisée pour mettre à jour les moyennes et la matrice de covariance à partir des statistiques. Elle renvoie 1 si o_t^r est associé avec le composant m de l'état j et 0 partout ailleurs. Le vecteur des moyennes et la matrice de covariance sont mis à jour selon les équations 6.6 et 6.2.

$$\hat{\mu jm} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\psi_{jm}^r(t)o_t^r}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\psi_{jm}^r(t)} \tag{6.6}$$

$$\sum_{\hat{jm}} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\psi_{jm}^r(t)(o_t^r - \hat{\mu jm})(o_t^r - \hat{\mu jm})'}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\psi_{jm}^r(t)} \tag{6.7}$$

Enfin, les poids sont calculés en fonction du nombre d'observations attribuées à chaque composant.

$$cjm = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\psi_{jm}^r(t)}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{l=1}^{M}\psi_{jl}^r(t)} \tag{6.8}$$

6.1.5 Modélisation du langage

Le modèle du langage réduit efficacement les hypothèses du modèle acoustique. Une probabilité de caractéristiques acoustiques d'une transcription de mots donnés $P(a|w)$ estimée par un modèle acoustique est combinée avec la probabilité d'une transcription de mots $P(w)$ estimée par un modèle du langage pour un domaine donné dans le but de calculer la probabilité à posteriori de la transcription $P(w|a) = \frac{P(a|w) \times P(w)}{P(a)}$.

Le modèle du langage statistique attribue à une séquence donnée de mots une probabilité donnée par l'équation 6.9. Les plus utilisés sont les modèles du langage n-gram. Ils calculent la probabilité de k séquences de mots w (voir l'équation 6.9) [143]. L'hypothèse de Markov se rapproche de la probabilité en supposant que seuls les plus récents $n-1$ mots sont pertinents pour prédire le mot suivant.

$$P(W) = P(w_k, w_{k-1}, w_{k-2}, ..., w_1) \approx \Pi_{i=1}^{k} P(w_i|w_{i-n+1}^{i-1}) \tag{6.9}$$

Les probabilités $P(w_i|w_{i-n+1}^{i-1})$ pour chaque mot w_i sont estimées à l'aide des fréquences relatives des n-grams, (n-1)-grams, (n-2)-grams, ... sur l'apprentissage des données. Le modèle du langage estime la probabilité en comptant les fréquences relatives sur le corpus de textes qui est généralement choisi en fonction du domaine du RAP ciblé.

6.1.6 Décodage de la parole

Les décodeurs MMC de parole ont pour objectif de trouver les séquences de mots les plus probables en recherchant les séquences de phones qui correspondent aux mots. Les phones sont généralement représentés comme des triphones dans le modèle acoustique. L'utilisation de la combinaison des probabilités des modèles acoustiques et du langage, comme décrit dans l'équation 6.2, ne produit pas plus de précision dans les transcriptions de parole. Généralement un poids du modèle du langage w_{ml} est utilisé pour améliorer la précision de la reconnaissance de la parole. Il est réglé sur tout l'ensemble des données d'apprentissage et équilibre l'impact des deux modèles. En utilisant la meilleure séquence de mots, w_{ml} est trouvé selon l'équation 6.10.

$$w* = argmax_w P(a|w) \times P((w)^{w_{ml}} \tag{6.10}$$

La reconnaissance est effectuée en construisant les modèles de mots à partir des règles de composition des modèles de phones décrites dans un lexique [27]. Les phrases sont par la

27. Le lexique est l'ensemble des règles permettant de décomposer un mot en sous-unités (phonèmes,

suite formées par concaténation des modèles de mots. Les modèles de mots sont connectés via un modèle MMC. Pour une reconnaissance de mots isolés, le modèle MMC est évalué pour chaque possibilité de mots. Le modèle MMC intègre les unités décomposées chacune en 3 états. En général, un modèle de phonème est composé de 3 états émetteurs servant à décrire la forme (*pattern*) du phonème à reconnaître [144]. Pour la reconnaissance de la parole, on considère généralement que les phonèmes (contextuels ou non contextuels) possèdent trois parties, chacune correspondant à un état :

- un contexte gauche correspondant à la transition entre le phonème précédent et le phonème courant ;
- une partie centrale considérée comme stationnaire ;
- un contexte droit correspondant à la transition entre le phonème courant et le suivant.

Certains systèmes proposent une liste des N-meilleures hypothèses voire même un treillis. À partir des probabilités générées par le modèle acoustique et des probabilités des séquences de mots obtenues avec le modèle de langage et le lexique, l'algorithme de décodage permet de produire la meilleure hypothèse de phrase. L'objectif de l'algorithme est de trouver le chemin optimal permettant de maximiser la probabilité de la suite d'observations étant donné le modèle utilisé sans pour autant générer la totalité du treillis.

6.2 La boîte à outils Kaldi

Kaldi est une boîte à outils de reconnaissance de la parole constituée d'une bibliothèque, de programmes et des scripts en ligne de commande pour la modélisation acoustique. Il utilise l'apprentissage Viterbi pour estimer les modèles acoustiques et implémente plusieurs décodeurs pour leur évaluation.

L'architecture de Kaldi se compose des librairies Kaldi et des scripts d'apprentissage. Les scripts accèdent aux fonctionnalités offertes par les librairies à travers des programmes en ligne de commande. La bibliothèque C++ de Kaldi est basée sur la librairie *OpenFST* [145] et utilise les librairies optimisées pour l'algèbre linéaire telles que *BLAS* et *LAPACK*.

Kaldi utilise les transducteurs à états finis (FST), par son implémentation *OpenFST*, pour représenter partiellement les modèles acoustiques et du langage, le lexique et aussi la transformation entre le texte, la prononciation et les triphones. La librairie *FST* fournit des opérations de graphe qui peuvent être utilisées efficacement pour la modélisation

phonèmes en contexte, syllabes).

acoustique. En utilisant le *FST*, la tâche de décodage de la parole s'exprime comme étant un problème de recherche heuristique dans un graphe. Les graphes *FST* utilisés pour l'apprentissage du modèle acoustique et du décodage de la parole sont construits sous forme de séquence d'opérations normalisées de l'*OpenFST* [146]. Le décodage est effectué sur ce qu'on appelle *graphe de décodage HCLG* qui est construit à partir de graphes simples *FST* comme illustré dans l'équation 6.11 où le symbole *o* représente une opération binaire associative de la composition sur les *FSTs*.

$$HCLG = H \ o \ C \ o \ L \ o \ G \tag{6.11}$$

Les fonctionnalités du transducteur à partir de l'équation 6.11 sont :

1. G est destiné à coder le modèle du langage ou la grammaire.
2. L représente le lexique. Ses symboles d'entrée sont des phones et les symboles de sortie sont des mots.
3. C représente la relation entre les phones contextuels de l'entrée et de la sortie.
4. H contient les définitions MMC qui prennent comme id d'entrée le nombre de fonctions de densité de probabilités et retournent les phones contextuels.

L'équation 6.12 illustre comment Kaldi crée le graphe de décodage à partir des opérations *FST* [146].

$$HCLG = asl(min(rds(det(H'omin(det(Comin(det(LoG))))))))) \tag{6.12}$$

Kaldi utilise des exécutables qui chargent leur entrée à partir des fichiers et stockent les résultats dans des fichiers adéquats. Parmi les exécutables, il y a généralement des alternatives pour toutes les tâches de reconnaissance de parole comme le montre la liste suivante :

1. la paramétrisation de la parole
   - apply-mfcc
   - compute-mfcc-feats
   - compute-plp-feats
   - ...
2. la transformation de caractéristiques

- apply-cmvn
- compute-cmvn-stats
- acc-lda
- fmpe-apply-transform
- ...

3. les décodeurs
   - gmm-latgen-faster
   - gmm-latgen-faster-parallel
   - gmm-latgen-biglm-faster
   - ...
4. l'évaluation et les utilitaires
   - compute-wer
   - show-alignments
   - ...

Kaldi fournit, en plus des exécutables (voir figure 6.1), des scripts standards très utiles qui ajoutent aux librairies de nouvelles fonctionnalités. Les scripts sont logés dans les répertoires *utils* et *steps* de Kaldi et sont utilisés dans de nombreuses recettes de scripts d'apprentissage pour différents corpora de données. Dans ce chapitre, nous avons créé une nouvelle recette d'apprentissage pour la parole en continue en Fongbe en utilisant l'infrastructure de Kaldi et les corpora audio et de textes collectés et décrits dans le chapitre 2. Par la suite, nous décriverons les différents scripts de modélisation acoustique qui composent la recette.

6.3 La boîte à outils SRILM

SRILM [28] est une boîte à outils destinée à la construction et à l'application de modèles du langage statistiques. Il est employé dans la reconnaissance de la parole, dans la segmentation et l'annotation statistique et la traduction automatique. SRILM est une collection de bibliothèques C++, de programmes exécutables et de scripts d'assistance conçus pour permettre à la fois la production et l'expérimentation de modèles de langue statistiques [147]. SRILM est conçu et mis en œuvre en trois couches [147] :

28. http ://www.speech.sri.com/projects/srilm/

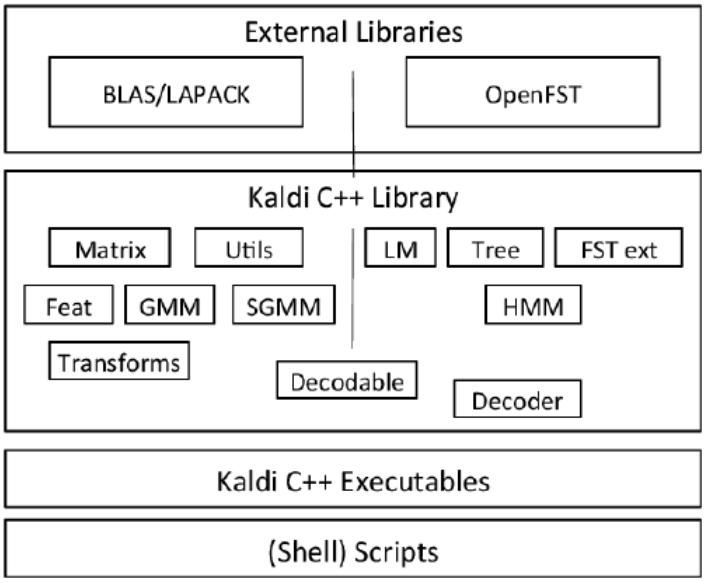


FIGURE 6.1 – Architecture de la boîte à outils Kaldi.

- la couche des bibliothèques contenant environ 50 classes C++ pour la modélisation du langage et des objets tels que les tables de symboles du vocabulaire, les N-meilleures listes et les graphes de mots ;
- la couche des programmes contenant 14 principaux outils exécutables, tels que *ngram-count*, *ngram*, et *taggers* écrits en C++ et positionnés au dessus de l'API fournit par les bibliothèques ;
- la couche des scripts implémentés pour la plus part en shell ;

SRILM utilise des formats de fichiers standards tels que le format ARPA [29] pour les modèles N-grammes. Il est disponible pour les utilisateurs non-commerciaux sous une licence communautaire Open Source.

6.4 Modélisation du langage pour le Fongbe

Pour comparer les performances du système de reconnaissance du Fongbe, nous avons construit deux différents modèles du langage basés sur le même corpus de textes. Le premier modèle du langage LM1 est construit avec les textes après une première normalisation et contenant les différents tons des voyelles. L'utilisation des diacritiques implique que le système doit traiter 26 voyelles (au lieu de 12) considérées comme étant différentes

29. Format utilisé pour coder un modèle du langage avec SRILM

à cause des tons. Le deuxième modèle du langage LM2 est construit avec les textes originaux modifiés par une seconde phase de normalisation sur les différentes voyelles tonales du corpus de textes. La normalisation est faite en supprimant les tons des voyelles et le remplacement des caractères accentués par des caractères simples faciles à manipuler par le système. Le résultat est que nous avons de nouvelles entrées avec leurs transcriptions dans notre dictionnaire de vocabulaire. Par exemple, le mot *axɔ́sú* qui signifie *Roi* devient dans le dictionnaire *ax ɔsu*. Le tableau 6.1 présente les différents changements effectués sur les voyelles pendant la seconde normalisation.

TABLE 6.1 – Normalisation des voyelles.

Tonal vowels	Normalisation
á	/a/
à	/a/
ã	/a/
ó	/o/
ò	/o/
õ	/o/
é	/e/
è	/e/
ẽ	/e/
ú	/u/
ù	/u/
ũ	/u/
́ı	/i/
̀ı	/i/
̃ı	/i/
ɛ́	/ ɛ/
ɛ̀	/ ɛ/
ɛ̃	/ ɛ/
ɔ́	/ ɔ/
ɔ̀	/ ɔ/
ɔ̃	/ ɔ/

Nous avons utilisé la boîte à outils SRILM pour entrainer deux différents modèles du langage LM1 et LM2 sur 995.339 mots (10.095 unigrams). L'apprentissage des deux modèles est effectué avec les données d'apprentissage (1.054.724 mots, 33.153 phrases) extraites du corpus de textes sans les énoncés (5.490 mots and 1.500 phrases) utilisés pour le corpus de parole.

LM1 est entraîné avec les textes originaux pendant que LM2 est entraîné avec les textes modifiés par la normalisation des voyelles. Pour représenter l'incertitude des deux modèles du langage, nous avons calculé les valeurs de perplexité de toutes les transcrip-

tions d'énoncés à l'exception de ceux qui composent les données de test. Ces valeurs de perplexité sont calculées pour évaluer la performance des deux modèles. La perplexité est l'exponentiel de la log-vraisemblance. Elle est basée sur les probabilités et est calculée directement à partir de l'entropie-croisée du modèle sur un jeu de phrases de test. Plus elle est petite, plus le modèle du langage est performant. La perplexité d'un modèle du langage pour un texte D est calculée à partir de l'équation 6.13.

$$2^{\frac{-\sum_{w_1^T \in D} log\hat{P}(w_1^T)}{N}} \tag{6.13}$$

Les probabilités sont estimées à partir du corpus d'apprentissage, puis sont comparées à celles observées sur le texte D. Le terme exponentiel dans l'équation de la perplexité peut être considéré comme la cross-entropie des deux distributions. Le tableau 6.2 présente les différentes valeurs de perplexité et le taux de mots hors-vocabulaire (MHV). Les mots hors-vocabulaire sont des mots rencontrés par le système mais qui ne figurent pas dans le vocabulaire construit pour l'apprentissage du système. Si un utilisateur prononce un mot que le système ne connaît pas, il est remplacé par un ou plusieurs mots connus qui s'en rapprochent acoustiquement. Les mots hors-vocabulaire sont calculés à postériori après la construction du modèle. La perplexité peut varier par la prise en compte ou non des mots hors-vocabulaire sur l'ensemble des données test.

TABLE 6.2 – La perplexité des modèles du langage.

LM	Vocab (mots)	MHV	PPL
LM1	10.130	9,1%	591
LM2	8.244	4,96%	138

On peut donc remarquer à travers le tableau 6.2 que la normalisation des voyelles après modification des textes originaux a un impact postif sur la qualité du modèle du langage en réduisant les mots hors-vocabulaire de $9,1\%$ à $4,96\%$. Ceci conduit à une amélioration significative de la valeur de perplexité du modèle LM2 comparé au modèle LM1. Finalement on obtient un système basé sur un lexique de 10.130 graphèmes. Le tableau 6.3 illustre un exemple du contenu du lexique obtenu après le pré-taitrement sur les textes.

TABLE 6.3 – Exemple du contenu du lexicon.

	Mots	Graphèmes
Texte original	axɔ́sú du du hãgbɛ́	a x ɔ́ s ú d u d u h ã g b ɛ́
Normalisation des voyelles	ax ɔs u du du h a agb ɛ	a x ɔ s u d u d u h a a g b ɛ

6.5 Apprentissage des modèles acoustiques pour le Fongbe

Dans cette section nous présentons les différents scripts réalisés pour la modélisation acoustique des données du Fongbe. Nous décrivons les méthodes utilisées pour l'apprentissage et le test des deux configurations (FSC1 et FSC2) du corpus de parole décrites dans le tableau 2.4. Les enregistrements effectués et leurs transcriptions sont utilisés pour cette modélisation.

Les modèles acoustiques ont été entraînés et testés sur les données acoustiques provenant des deux configurations par les scripts de modélisation acoustique Kaldi que nous avons adapté pour produire les scripts Kaldi du Fongbe. Nous n'avons pas seulement exploré les méthodes d'apprentissage des modèles acoustiques mais aussi expérimenté l'impact de la présence des tons dans la transcription des énoncés du corpus audio. L'impact a été étudié avec les deux modèles du langage dont LM1 (avec diacritiques) et LM2 (sans diacritiques) construits pour la comparaison. Ainsi, l'apprentissage de FSC1 et FSC2 a été réalisé non seulement avec les mêmes scripts réalisés mais aussi avec les deux dictionnaires de prononciation issus des deux modèles du langage. Dans notre contexte d'étude, les dictionnaires de prononciation sont basés sur des graphèmes avec 49 graphèmes pour LM1 et 28 pour LM2. Le choix du graphème comme unités de modélisation acoustique présente, dans le cas de la langue peu dotée Fongbe, un potentiel très intéressant car il rend plus simple la construction automatique du dictionnaire de prononciation. Ce choix est aussi motivé par le fait que nous n'avons pas de connaissances linguistiques approfondies sur la langue cible (le Fongbe).

Les modèles acoustiques sont entraînés avec 13 caractéristiques MFCCs dont les coefficients sont triplés avec $\triangle + \triangle\triangle$ par le calcul des dérivées premières et secondes à partir des coefficients obtenus. Nous avons aussi employé d'autres techniques de transformation de caractéristiques telles que LDA et MLLT qui ont gagné une amélioration substantielle sur la transformation $\triangle + \triangle\triangle$. Par la suite, une adaptation au locuteur a été appliquée avec la technique fMLLR.

La figure 6.2 et le tableau 6.4 présentent l'hiérarchie des méthodes d'apprentissage

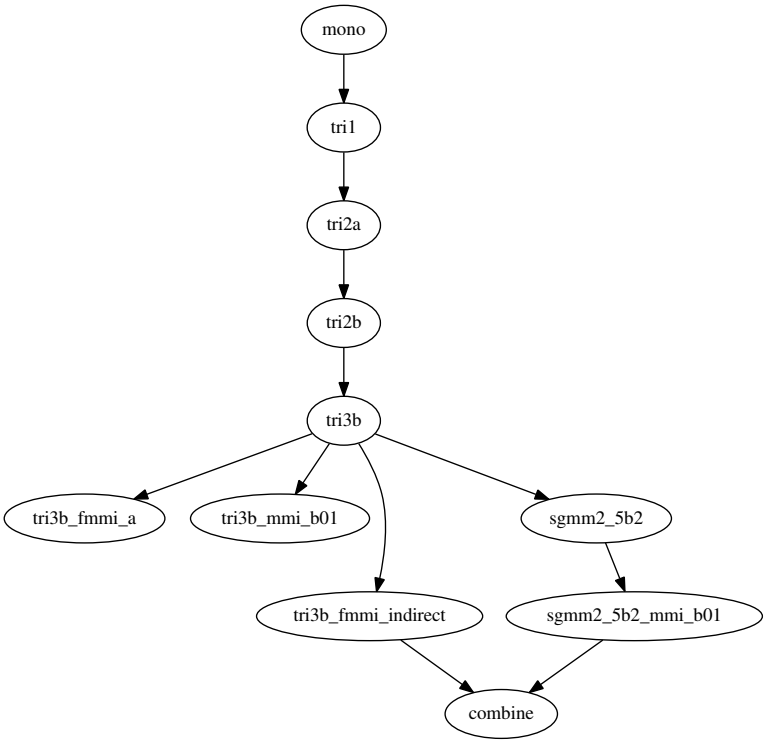


FIGURE 6.2 – Hiérarchie des modèles acoustiques entraînés.

utilisées pour entraîner les modèles acoustiques dans nos expériences. Dans cette hiérarchie, nous avons commencé par l'apprentissage d'un modèle monophone en utilisant les caractéristiques MFCC et fini par l'apprentissage du modèle SGMM en utilisant les caractéristiques transformées fMMI. Les modèles triphones intermédiaire ont aussi été entraînés comme illustré dans la figure 6.2. Comme exemple, dans un premier temps, le modèle monophone a été entraîné d'abord avec les caractéristiques MFCCs, ensuite avec les dérivées $\triangle + \triangle\triangle$. Les vecteurs caractéristiques subissent un alignement forcé aux états MMC à partir des transcriptions des énoncés. Dans un second temps, nous avons re-entraîné le modèle triphone (*tri1a*) qui devient *tri2a* après un apprentissage avec les caractéristiques MFCC $\triangle + \triangle\triangle$. Un peu plus loin, nous avons utilisé la transformation MLLT + LDA pour entraîner un autre modèle acoustique tri2b. La branche s'arrête au modèle tri3b à partir duquel quatre autres modèles sont entraînés de manière discriminative.

Pour le décodage, nous avons utilisés les différents modèles acoustiques entraînés avec, comme données test, les énoncés du corpus test. Pour chaque modèle acoustique entraîné, nous avons utilisé les mêmes méthodes de paramétrisation de la parole et de transformation des caractéristiques qui ont été utilisées pour le modèle acoustique en question au moment de l'apprentissage.

TABLE 6.4 – Modèles acoustiques.

Méthode d'apprentissage	Script
Monophone	mono
Triphone	tri1
$\Delta + \Delta\Delta$	tri2a
LDA + MLLT	tri2b
LDA + MLLT + SAT + FMLLR	tri3b
LDA + MLLT + SAT + FMLLR + fMMI	tri3b_fmmi_a
LDA + MLLT + SAT + FMLLR + MMI	tri3b_mmi_b0.1
LDA + MLLT + SAT + FMLLR + fMMI + MMI	tri3b_fmmi_indirect
LDA + MLLT + SGMM	sgmm2_5b2
LDA + MLLT + SGMM + MMI	sgmm2_5b2_mmi_b0.1
LDA + MLLT + SGMM + fMMI + MMI	combine*

6.6 Evaluation

Les expériences portent sur la comparaison de la qualité des hypothèses du système RAP évaluée par la mesure WER. Cette mesure est calculée sur les modèles acoustiques entraînés par les différentes méthodes de modélisation. Pour obtenir le meilleur chemin, nous avons suivi les procédures standards de Kaldi afin de reporter les meilleurs taux WER. Les expériences ont été d'abord effectuées sur LM1 en utilisant les deux configurations du corpus de parole. Ensuite, nous avons mené des expériences, basées sur les mêmes procédures, sur LM2 incluant des textes sans diacritiques. L'intérêt est de mesurer l'impact de l'utilisation des diacritiques dans la modélisation du langage. Nous montrons également dans cette section l'influence des configurations du corpus de parole sur la qualité des modèles acoustiques mesurée par le taux WER.

6.6.1 Evaluation du système RAP sans normalisation des voyelles

Dans cette section nous présentons les résultats obtenus avec les différentes méthodes d'apprentissage acoustiques selon la configuration du corpus de parole. Le tableau 6.5 présente les différents résultats des modèles acoustiques entraînés pour LM1. Nous pouvons remarquer que le modèle acoustique monophone a le mauvais taux de reconnaissance WER alors que les meilleures performances sont réalisées avec le modèle acoustique *sgmm2_5b2* pour la configuration FSC1 et avec le modèle *sgmm2_5b2_mmi* pour la configuration FSC2. Nous notons donc que le modèle acoustique monophone est généralement utilisé pour initialiser les modèles triphones d'où son mauvais taux de reconnaissance. La qualité de la reconnaissance de la parole varie en fonction de la méthode d'apprentissage discriminative utilisée. Il se présente dans notre évaluation que la transformation de caractéristiques

LDA+MLLT est plus efficace que l'utilisation des caractéristiques $\triangle + \triangle\triangle$. Il en ressort des différences subtiles dans la performance des méthodes acoustiques entraînées de manière discrimative.

TABLE 6.5 – WER du SRAP basé sur LM1 (sans diacritiques) pour les différentes méthodes d'apprentissage monophone et triphone.

Config du corpus de parole/ méthode	WER %
FSC1-config	
Monophone (a)	69.44
Triphone (b)	69.13
$\Delta + \Delta\Delta$ (c)	70.21
LDA + MLLT (d)	65.7
LDA + MLLT + SAT + FMLLR (e)	54.96
LDA + MLLT + SAT + FMLLR + fMMI (f)	55.36
LDA + MLLT + SAT + FMLLR + MMI (g)	51.11
LDA + MLLT + SAT + FMLLR + fMMI + MMI (h)	55.60
LDA + MLLT + SGMM (i)	**44.04**
LDA + MLLT + SGMM + MMI (j)	47.11
LDA + MLLT + SGMM + fMMI + MMI (k)	49.83
FSC2-config	
Monophone (a)	71.97
Triphone (b)	60.37
$\Delta + \Delta\Delta$ (c)	59.74
LDA + MLLT (d)	57.52
LDA + MLLT + SAT + FMLLR (e)	51.47
LDA + MLLT + SAT + FMLLR + fMMI (f)	53.06
LDA + MLLT + SAT + FMLLR + MMI (g)	52.75
LDA + MLLT + SAT + FMLLR + fMMI + MMI (h)	52.37
LDA + MLLT + SGMM (i)	49.85
LDA + MLLT + SGMM + MMI (j)	**44.09**
LDA + MLLT + SGMM + fMMI + MMI (k)	44.17

Le taux WER sur les deux configurations du corpus de parole, pour LM1 fixé, varie autour de 44%. Ceci s'explique par la complexité du Fongbe dans la modélisation des diacritiques qui influence la qualité du modèle du langage. La valeur de perplexité obtenue (PPL = 591) pour le modèle LM1, dans le tableau 6.13 vient confirmer cette complexité à modéliser les diacritiques. La variation dans les performances des méthodes d'apprentissage acoustiques est illustrée dans la figure 6.3.

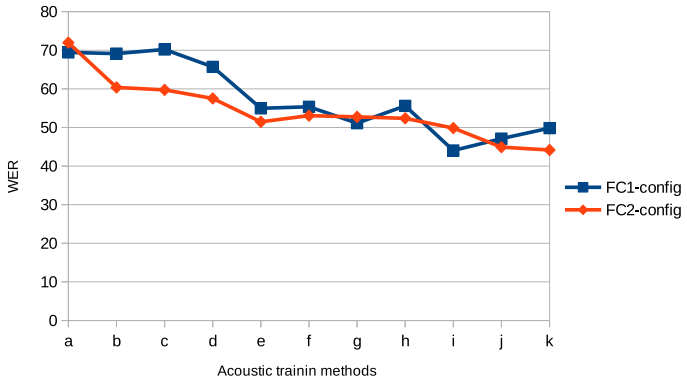


FIGURE 6.3 – Influence de la configuration du corpus de parole sur la qualité de la reconnaissance de la parole. LM1 est fixé et seulement les données et les modèles acoustiques varient. Les lettres en abscisse représentent les méthodes d'apprentissage étiquetées dans le tableau 6.5.

6.6.2 Evaluation du système RAP avec normalisation des voyelles

Le tableau 6.6 présente deux taux WER des méthodes d'apprentissage en fonction de la configuration du corpus de parole. Dans la seconde colonne (RAP basée sur LM2), nous avons inséré les résultats de la reconnaissance réalisée après la normalisation des voyelles (sans diacritiques). Il ressort de ces résultats que les modèles triphones améliorent de façon significative la performance des modèles monophones. Le modèle acoustique tri2b+SAT+fMLLR réduit de 6% le taux d'erreur de reconnaissance de mots pour les deux configurations du corpus audio. Le taux d'erreur WER de la reconnaissance utilisant FSC1-config est inférieur à 20% pour les méthodes discriminatives basées sur le modèle tri3b. Pour FSC2-config, ces méthodes d'apprentissage ont réduit le taux d'erreur de 20%.

Les meilleurs résultats sont obtenus avec la méthode SGMM, avec un taux d'erreur global de $14,83\%$ pour FSC1-config et de $28,93\%$ pour FSC2-config. Les données audio divisées par locuteurs ont permit d'obtenir un gain relatif de 14% avec un taux d'erreur final de $14,83\%$. Ceci conduit à choisir les méthodes d'apprentissage basées sur SGMM comme méthodes servant de comparaison de performance entre les deux configurations de corpus FSC1-config et FSC2-config. La figure 6.4 illustre l'évolution du taux d'erreur en fonction des modèles acoustiques et avec le modèle du langage LM2..

Il est à remarquer que le modèle du langage LM2 donne des résultats très satisfaisants de décodage de parole comparé au modèle LM1 (sans diacritiques). L'ajout des diacritiques dans le corpus de texte avant la modélisation du langage rends le système de reconnaissance de la parole moins efficace en augmentant le taux d'erreur à $44,04\%$ comparé au $15,23\%$ obtenu avec le modèle sans les diacritiques. Les informations addi-

TABLE 6.6 – WER de la reconnaissance basée sur LM2 (sans les diacritiques) et de la reconnaissance basée sur LM1' (suppression des diacritiques depuis les hypothèses et les références).

Configuration du corpus de parole/ méthode	RAP basée sur LM2	RAP basée sur LM1'
FSC1-config		
Monophone (a)	36.36	59.05
Triphone (b)	28.19	46.8
$\Delta + \Delta\Delta$ (c)	28.21	46.98
LDA + MLLT (d)	24.4	41.52
LDA + MLLT + SAT + FMLLR (e)	17.83	29.29
LDA + MLLT + SAT + FMLLR + fMMI (f)	19.72	31.34
LDA + MLLT + SAT + FMLLR + MMI (g)	18.93	35.59
LDA + MLLT + SAT + FMLLR + fMMI + MMI (h)	18.26	35.44
LDA + MLLT + SGMM (i)	15.23	**20.56**
LDA + MLLT + SGMM + MMI (j)	15.3	20.68
LDA + MLLT + SGMM + fMMI + MMI (k)	**14.83**	21.39
FSC2-config		
Monophone (a)	52.26	57.89
Triphone (b)	38.72	47.47
$\Delta + \Delta\Delta$ (c)	38.58	46.39
LDA + MLLT (d)	35.34	42.45
LDA + MLLT + SAT + FMLLR (e)	30.74	35.63
LDA + MLLT + SAT + FMLLR + fMMI (f)	35.36	37.46
LDA + MLLT + SAT + FMLLR + MMI (g)	32.38	36.19
LDA + MLLT + SAT + FMLLR + fMMI + MMI (h)	32.94	37.52
LDA + MLLT + SGMM (i)	31.64	**31.58**
LDA + MLLT + SGMM + MMI (j)	31.36	32.75
LDA + MLLT + SGMM + fMMI + MMI (k)	**28.93**	32.02

tionnelles apportées par les signes diacritiques augmentent, en plus du taux d'erreur, la perplexité et le taux de mots hors-vocabulaire du modèle du langage.

Par la suite, pour une comparaison efficace de performances, nous avons supprimé les signes diacritiques des références et des sorties (les hypothèses) du système de reconnaissance construit avec LM1. Ceci conduit à obtenir un système RAP basé sur le modèle LM1' (LM1 sans signes diacritiques). Les résultats obtenus pour cette évaluation sont inclus dans la troisième colonne du tableau 6.6. Ces résultats sont comparés à ceux obtenus avec le système basé sur le modèle LM2. Cette comparaison nous amène à affirmer que la suppression des diacritiques pour les différents modèles est plus efficace et offre un système RAP efficace.

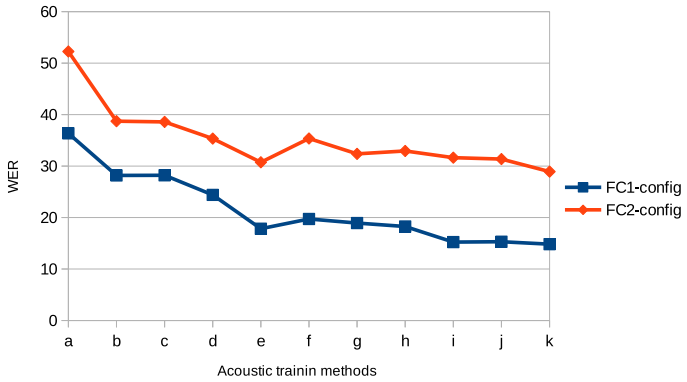


FIGURE 6.4 – Influence de la configuration du corpus de parole sur la qualité de la reconnaissance de la parole. LM1 est fixé et seulement les données et les modèles acoustiques varient. Les lettres en abscisse représentent les méthodes d'apprentissage étiquetées dans le tableau 6.6

Conclusion

Dans ce chapitre, nous avons introduit un premier système de reconnaissance automatique de la parole continue en Fongbe. Nous avons présenté le développement de modèles acoustiques avec l'outil Kaldi et de modèles du langage avec l'outil SRILM. Nous avons aussi démontré l'effet de l'usage des tons sur la qualité du système de reconnaissance automatique de la parole. Ainsi pour le système actuel construit, nous pouvons conclure que la modélisation du langage sans diacritiques améliore nettement les performances de reconnaissance en diminuant le taux d'erreur de reconnaissance de mots à $15,23\%$ pour la configuration du corpus de parole liée aux locuteurs et à $28,93\%$ pour la configuration liée aux catégories de textes.

En utilisant la recette de scripts Kaldi et les ressources linguistiques que nous offrons par cette thèse, un utilisateur lambda pourra désormais construire un système de reconnaissance automatique de la parole en Fongbe avec les mêmes taux WER obtenus dans ce chapitre.

Conclusion générale et perspectives

"*dosú gudunö klen afö bo*
j'ayi : e đö gbe kpo đö nukön"

Conclusion

Dans le contexte du traitement automatique de la parole en langue naturelle, nous avons étudié la faisabilité d'un système de reconnaissance automatique de la parole pour le Fongbe (une langue peu dotée). Nous concluons dans ce chapitre les différents travaux réalisés dans le cadre de cette thèse.

En adoptant la méthode développée par V. Berment, nous avons démontré que le Fongbe est une langue peu dotée de la classe des $langues - \pi$ dont l'indice ($\sigma = 6,02$) a révélé l'inexistence de ressources numériques nécessaires aux services du traitement de l'oral. Les premiers travaux se sont axés sur le recueil de ressources du language (corpus audio et textuel) et de ressources lexicales (vocabulaires et dictionnaires de prononciations). Pour le recueil des données textuelles, la méthodologie adoptée et décrite dans le présent manuscrit consiste à extraire, dans un premier temps, des pages web ayant des contenus en Fongbe et, dans un second temps, à traiter les pages web afin d'en extraire les textes. Malgré la faible qualité et quantité des sources de données textuelles, il a été recueillit dans nos travaux un corpus de 34.653 phrases dont 1.060.214 mots et un vocabulaire de 10.130 mots. Il a été ensuite créé, à base de l'approche graphémique, deux différents dictionnaires de prononciations.

Pour la collecte des signaux audio, il a été procédé à la construction de deux corpora pour l'étude acoustique de la parole en Fongbe. Le premier corpus, contenant uniquement les signaux de 32 unités phonémiques considérées, a servi à l'étude acoustique des sons isolés du Fongbe et à la réalisation d'un système de reconnaissance automatique des phonèmes basée sur la fusion de deux classifieurs différents. Il comprend 4929 sigaux (environ 4h de lecture) enregistrés et digitalisés (à 44100Hz) avec le logiciel Audacity. Cent soixante quattorze (174) locuteurs (natifs ou non) ont participé aux enregistrements.

La réalisation d'un système de reconnaissance de parole continue large vocabulaire est fortement liée à un corpus audio incluant des paroles continues bien segmentées suivant un alignement textuel bien précis. Dans ce sens, le second corpus audio collecté est un corpus de parole continue devant servir à l'apprentissage de modèles acoustiques pour la reconnaissance statistique et automatique de la parole continue. Ainsi nous avons collecté 10.302 signaux à l'aide du logiciel Lig-Aikuma en enregistrant, à 16 Khz, vingt-huit (28) locuteurs. Vingt (20) locuteurs ont été utilisés pour créer les données d'apprentissage et huit (8) locuteurs pour les données test.

Pour facliter la reconnaissance d'identités phonétiques des sons du Fongbe, nous nous sommes interressés à leurs configurations acoustiques en fournissant une description acoustique des phonèmes du Fongbe basée d'une part sur l'analyse formantique des voyelles et d'autre part, sur la durée moyenne, la fréquence fondamentale et l'intensité des consonnes. L'étude acoustique est effectuée, dans un contexte isolé, sur les données du corpus audio *FongbePhones-FLDataset*. Avec les mêmes données, nous avons proposé une approche de reconnaissance des phonèmes dans un contexte de multiclassification. L'insuffisance des données et le déséquilibre entre les signaux des consonnes et ceux des voyelles ont conduit à opter pour une fusion de décisions des classifieurs probabilistiques et neuronaux. L'objectif a été de résoudre l'incertitude dans les résultats des deux classifieurs en gérant les conflits dans les décisions. Pour une proposition effective de système de reconnaissance de phonèmes isolés, nous avons exploré la performance de trois méthodes de fusion de décisions par la combinaison intelligente et adaptative des classifieurs et des paramètres acoustiques. Il en résulte que le système de logique floue, exploitant le raisonnement humain, performe au mieux la stratégie de fusion développée sur les données audio du Fongbe.

En exploitant les données du corpus de la parole continue, nous avons proposé, dans cette thèse, deux algorithmes de segmentation automatique de la parole continue et développé des modèles du language et acoustiques pour le Fongbe. Le premier algorithme se base sur le calcul, en chaque point du signal de la parole, des exposants de singularité combinés à une mesure entropique que nous avons définie à partir de l'entropie de Rényi. Le deuxième algorithme exploite des connaissances expertes que fournissent les mécanismes d'un système de logique floue pour une segmentation automatique de la parole indépendante de la transcription textuelle. L'évaluation des deux algorithmes a révélé de meilleures performances sur les données du Fongbe. Pour finir, nous avons procédé à la construction de modèles acoustiques et de modèles du language pour un décodeur de parole continue en Fongbe. Les résultats expérimentaux ont été en adéquation avec l'intèrêt d'une modélisation acoustique graphémique vu que le Fongbe ne disposait pas d'un dic-

tionnaire phonétique. Dans ce travail, nous avons fournit une recette complète de scripts Kaldi pour la construction d'un système de reconnaissance de parole continue en Fongbe. Nous nous sommes aussi interressés à l'usage du ton dans les expressions orales en étudiant l'impact des diacritiques sur les différents modèles et sur la performance globale du système. Nous sommes donc arrivés à la conclusion que la suppression des diacritiques des transcriptions textuelles des énoncés en Fongbe rend plus efficace les modèles acoustiques et offre un système performant de reconnaissance automatique de la parole en Fongbe.

L'ensemble des travaux réalisés dans le cadre de cette thèse est fortement orienté d'une part dans le domaine temporel par l'usage des paramètres acoustiques temporels tels que l'énergie à court-terme, le taux de passage par zéro, les exposants de singularité et d'autre part, dans le domaine fréquentiel par l'usage des coefficients MFCC, PLP et Rasta-PLP. Les méthodes d'apprentissage basées sur les réseaux de neurones profonds, les réseaux de neurones de croyance profonds et les SVM ont été explorées dans cette thèse pour expérimenter leurs performances sur le Fongbe. Nous nous sommes fortement interressés aux connaissances expertes que fournissent l'approche de la logique floue pour proposer des méthodes efficaces de segmentation et de reconnaissance automatique de la parole.

Perspectives

Pour la suite des travaux, nous envisageons des perspectives propres à chaque thématique de recherche abordée dans cette thèse. En d'autres termes, une certaine amélioration et des extensions significatives seront appliquées sur les algorithmes et méthodes développés dans le cadre de cette thèse.

- Dans un premier temps, nous envisageons élargir les ressources linguistiques en augmentant les données des corpora d'entrainement audio et textuels. L'impact direct serait l'amélioration de la qualité des modèles acoustiques et du language pour un meilleur taux de reconnaissance.

- Nous avons réalisé la modélisation acoustique sans tenir compte du caractère tonal du Fongbe. Vu l'importance des tons de la langue et de leur influence sur l'évolution temporelle des noyaux vocaliques dans une syllabe ou un mot, nous envisageons améliorer la modélisation acoustique en allant vers un système de diacritisation du texte Fongbe.

- La modélisation graphémique a été utilisée dans notre contexte d'étude à cause de l'intèrêt qu'elle présente en cas d'absence d'un dictionnaire phonétique pour une

langue peu dotée. Bien que dans nos travaux, l'utilisation des graphèmes a révélé de meilleures performances, nous envisageons recueillir un dictionnaire phonétique de bonne qualité dans l'espoir qu'il améliore la qualité du système de reconnaissance.

- Concernant les algorithmes de segmentation proposés dans ce manuscrit, nous envisageons les étendre aux unités phonémiques pour une détection efficace de frontières de phonèmes. Ceci dans le but de proposer une boîte à outils intégrant des techniques d'extraction de caractéristiques, des algorithmes de segmentation et de reconnaissance de phonèmes typiquement pour le Fongbe.

- Pour finir, nous envisageons appliquer nos algorithmes sur d'autres langues peu dotées du Bénin afin d'étudier leurs performances et ressortir les proximités entre elles dans le domaine du traitement automatique et statistique de la parole.

Glossaire

ARPA : Advanced Research Projects Agency
ASCII : American Standard Code for Information Interchange
ASR : Automatic Speech Recognition
BLAS : Basic Linear Algebra Subprograms
DBN : Deep Beliefs Networks
DCT : Discrete Cosine Transform
DNN : Deep Neural Network
fMMI : feature-space Maximum Mutual Information
fMLLR : feature-space Maximum Likelihood Linear Regression
FFT : Fast Fourier Transform
FN : False Negative
FP : False Positive
FSC : Fongbe Speech Configuration
FSS : Frame Step Size
FST : Finite State Transducter
GETALP : Groupe d'Etude pour la Traduction Automatique et le traitement automatisé des Langues et de la Parole
HTK : Hidden Markov Model Toolkit
IPA : International Phonetic Alphabet
LAPACK : Linear Algebra PACKage
LM : Language Model
LPC : Linear Predictive Coding
LRN : Likehood Ratio Negative
LRP : Likehood Ratio Positive
LVQ : Learning Vector Quantization
MFCC : Mel-Frequency Cepstral Coefficients
RBM : Restrited Boltzmann Machines
RLAT : Rapid Language Adaptation Toolkit

SE : Singularity Exponents
SER : Sentence Error Rate
SGMM : Subspace Gaussian Mixture Models
SLAM : Simultaneous Loxalization And Mapping
SNR : Signal-Noise Rate
SPPAS : Segmentation Phonetisation Alignement Syllabation
STE : Short-Term Energy
SVM : Support Vector Machine
TIC : Technologie de l'Information et de la Communication
TN : True Negative
TP : True Positive
UTF-8 : Universal Character Set Transformation Format - 8 bits
VTLN : Vocal Tract Length Normalization)
WSJ : Wall Street Journal
ZCR : Zero Crossing Rate

Annexe A

Liste des publications

Conférence internationale avec acte et comité de lecture

- **Fréjus A. A. Laleye**, Laurent Besacier, Eugène C. Ezin et Cina Motamed. **(2016)** *First Automatic Fongbe Continuous Speech Recognition System : Development of Acoustic Models and Language Models.* Proc. Federated Conference on Computer Science and Information Systems, Vol. 8, pp. 483–488, Pologne, Septembre 2016. DOI : 10.15439/2016F153

- **Fréjus A. A. Laleye**, Eugène C. Ezin et Cina Motamed. **(2016)** *Automatic fongbe phoneme recognition from spoken speech signal.* Proc. The 13th International Conference on Informatics in Control, Automation and Robotics, vol 1, pages 102–109, Juillet 2016. DOI : 10.5220/0006004101020109

- **Fréjus A. A. Laleye**, Eugène C. Ezin et Cina Motamed. **(2015)** *An algorithm based on fuzzy logic for text-independent fongbe speech segmentation.* Proc. IEEE The 11th International Conference on Signal Image Technology & Internet Based Systems, vol 1, pages 1–6, Novembre 2015. DOI : 10.1109/SITIS.2015.72

- **Fréjus A. A. Laleye**, Eugène C. Ezin et Cina Motamed. **(2015)** *Adaptive Decision-Level Fusion For Fongbe Phoneme Classification Using Fuzzy Logic and Deep Belief Networks.* Proc. The 12th International Conference on Informatics in Control, Automation and Robotics, pages 15–24, Juillet 2015. DOI : 10.5220/0005536100150024

- **Fréjus A. A. Laleye**, Eugène C. Ezin et Cina Motamed. **(2014)** *Weighted combination of naive bayes and lvq classifier for fongbe phoneme classification.* Proc. IEEE The 10th International Conference on Signal Image Technology & Internet Based Systems, vol 1, pages 7–13, Novembre 2014. DOI : 10.1109/SITIS.2014.84

Conférence nationale avec acte et comité de lecture

- **Fréjus A. A. Laleye**, Eugène C. Ezin et Cina Motamed. **(2015)** *Traitement automatique statistique de la parole en langue Fongbe.* Acte. 5ième Colloque de l'Université d'Abomey-Calavi des Sciences, Cultures et Technologies, page 542, Octobre 2015.

Revue internationale

- **Fréjus A. A. Laleye**, Eugène C. Ezin et Cina Motamed. **(2016)** *Fuzzy-Based Algorithm For Fongbe Continuous Speech Segmentation.* Pattern Analysis and applications, Springer International Publishing, 2016. *Accepté pour publication* (IF. 1.104)

- **Fréjus A. A. Laleye**, Eugène C. Ezin et Cina Motamed. **(2016)** *Automatic Text-Independent Syllable Segmentation Using Singularity Exponents And Rényi Entropy.* Journal of Signal Processing Systems, Springer International Publishing, pages 1–13, 2016. DOI : 10.1007/s11265-016-1183-9 (IF. 0.508)

- **Fréjus A. A. Laleye**, Eugène C. Ezin et Cina Motamed. **(2016)** *Automatic Text-Independent Syllable Segmentation Using Singularity Exponents And Rényi Entropy.* Multimedia Tools and Applications, Springer International Publishing, pages 1–22, 2016. DOI : 10.1007/s11265-016-1183-9 (IF. 1.331)

- **Fréjus A. A. Laleye**, Eugène C. Ezin et Cina Motamed. **(2016)** *Speech Phoneme Classification by Intelligent Decision-Level Fusion.* Lecture Notes in Electrical Engineering, Springer International Publishing, vol 383, pages 63–78, 2016.

Bibliographie

[1] L.V. Bác. Reconnaissance automatique de la parole pour des langues peu dotées. *Thèse de doctorat de l'Université J. Fourier - Grenoble I, France*, page 368, Juin, 2006.

[2] F. A.A. Laleye, L. Besacier, E. C. Ezin, and C. Motamed. First automatic fongbe continuous speech recognition system : Development of acoustic models and language models. In *Proceedings of Federated Conference on Computer Science and Information Systems*, pages 483–488, 2016.

[3] Fréjus A. A. Laleye, Eugène C. Ezin, and Cina Motamed. Automatic boundary detection based on entropy measures for text-independent syllable segmentation. *Multimedia Tools and Applications*, pages 1–22, 2016.

[4] Fréjus A. A. Laleye, Eugène C. Ezin, and Cina Motamed. Fuzzy-based algorithm for fongbe continuous speech segmentation. *Pattern Analysis and applications*, 2016.

[5] F. A.A. Laleye and C. Motamed E. C. Ezin. Speech phoneme classification by intelligent decision-level fusion. *Lecture Notes in Electrical Engineering*, 383 :63–78, 2016.

[6] V. Berment. Méthodes pour informatiser des langues et des groupes de langues peu dotées. *Thèse de doctorat de l'Université J. Fourier - Grenoble I, France*, page 368, Mai, 2004.

[7] J. Ndamba, C. Nstadi, Véronique. Rey, and J. Véronis. Traitement informatique des langues africaines : problèmes et perspectives. pages 810–819.

[8] C. Lefebvre and A.M. Brousseau. A grammar of fonge, de gruyter mouton. page 608, 2001.

[9] J. Greenberg. Languages of africa. *La Haye Mouton*, page 117, 1996.

[10] A. B. AKOHA. Syntaxe et lexicologie du fon-gbe : Bénin. *Ed. L'harmattan*, page 368, 2010.

[11] S. Nefti. Segmentation automatique de parole en phones. correction d'étiquetage par l'introduction de mesures de confiance. *Thèse de doctorat de l'Université de Rennes 1, France*, page 223, Décembre, 2004.

[12] S. Jarifi. Segmentation automatique de corpus de parole continue dédiés à la synthèse vocale. *Thèse de doctorat de l'ENST Bretagne, France*, page 180, Janvier, 2007.

[13] K. Sj olander and J. Beskow. Wavesurfer - an open source speech tool. 2000.

[14] C.W. Wightman and D.T. Talkin. The aligner : text-to- speech alignment using markov models. *J.P.H. van Santen, R.W. Sproat, J.P. Olive and J.Hirschberg, Progress in speech synthesis*, 1996.

[15] J.P van Hemert. Automatic segmentation of speech. volume 39, pages 1008–1012, 1991.

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

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

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

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

[20] M. Basseville and A. Benveniste. Sequential detection of abrupt changes in spectral characteristics of digital signal. In *IEEE Transactions on Information Theory*, volume 29, pages 709–724, 1983.

[21] R. André-Obrecht. A new statistical approach for the automatic segmentation of continous speech signals. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 33, pages 29–40, 1988.

[22] T. Styger, B. Gabioud, and E. Keller. Méthodes informatiques pour l'analyse de paramètres primaires en parole pathologique. *C.A.L.A.P*, 12, 1993.

[23] L. X. Hung. Extraction des traits non-linguistiques pour l'indexation des documents audio-visuels. *Rapport technique, Groupe MRIM - CLIPS-IMAG*, 2003.

[24] J. Laroche, Y. Stylianou, and E. Moulines. Hnm : A simple, efficient harmonic plus noise model for speech. In *IEEE ASSP Workshop on Applications of signal processing to audio and acoustics*, pages 169–172, 1993.

[25] D. Talkin. A robust algorithm for pitch tracking (rapt). *Speech coding and synthesis, Elsevier*, pages 495–518, 1995.

[26] E. Wesfreid and M. V. Wickerhauser. Adapted local trigonometric transform and speech processing. In *IEEE Transactions on Signal Processing*, volume 41, pages 3596–3600, 1993.

[27] C. Wendt and A. P. Petropulu. Pich determination and speech segmentation using the discrete wavelet transform. In *IEEE International Symposium on Circuits and Systems*, volume 2, pages 45–48, 1996.

[28] M. Baudry. Étude du signal vocal dans sa représentation amplitude-temps. algorithmes de segmentation et de reconnaissance de la parole. *Thèse de doctorat de l'Université PARIS VI*, 1978.

[29] M. Ito and R. Donaldson. Zero-crossing measurements for analysis and recognition of speech sounds. In *IEEE Transactions on Audio and Electroacoustics*, volume 19, pages 235–242, 1971.

[30] R. Scarr. Zero crossings as a means of obtaining spectr al information in speech analysis. In *IEEE Transactions on Audio and Electroacoustics*, volume 16, pages 247–255, 1968.

[31] P. Ravier. Détection de transitoires par ondelettes adaptées - critères d'adaptation fondés sur les statistiques d'ordre supérieur. *Thèse de doctorat de l'Institut National Polytechnique de Grenoble*, 1998.

[32] F. Brugnara, D. Falavigna, and M. Omologo. Automatic segmentation and labeling of speech based on hidden markov models. *Speech Communication, Elsevier*, 12 :370–375, 1993.

[33] B. Petek, O. Andersen, and P. Dalsgaard. On the robust automatic segmentation of spontaneous speech. In *ICSLP*, volume 2, pages 913–916, 1996.

[34] J. M. Makhoul. Linear prediction : a tutorial review. In *Proceedings of the IEEE*, volume 63, pages 561–580, 1975.

[35] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *IEEE*, 77 :257–286, 1989.

[36] L. R. Rabiner and B. H. Juang. Fundamentals of speech recognition. *Prentice Hall, Englewood Cliffs, New Jersey*, 1993.

[37] A. Vorstermans, J.-P. Martens, and B. Van Coile. Automatic segmentation and labelling of multi-lingual speech data. *Speech Communication, Elsevier*, 19 :271–293, 1996.

[38] D.S. Shete, S.B. Patil, and S.B. Patil. Zero crossing rate and energy of the speech signal of devanagari script. *Journal of VLSI and Signal Processing (IOSR-JVSP)*, 4(1) :01–05, 2014.

[39] F. Lahouti, A.R. Fazel, A.H. Safavi-Naeini, and A.K. Khandani. Single and double frame coding of speech lpc parameters using a lattice-based quantization scheme. *IEEE Transaction on Audio, Speech and Language Processing*, 14(5) :1624–1632, 2006.

[40] Ovidiu Buza1, Gavril Toderean1, Alina Nica1, and Alexandru Caruntu1r. Voice signal processing for speech synthesis. In *IEEE International Conference on Automation, Quality and Testing Robotics*, volume 2, pages 360–364, 2006.

[41] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *Acoustical Society of America Journal*, 87 :1738–1752, 1990.

[42] Zarader J.-L. et Chavy C. Gas B. A new approach to speech co- ding : the neural predictive coding. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 4(1) :120–127, 2000.

[43] Paliwal K.-K. et Atal B.-S. Efficient vector quantization of lpc parameters at 24 bits/frame. *IEEE Transactions on Speech and Audio Processing*, 1(1) :3–14, 1993.

[44] A. Acero et H-W. Hon. X. Huang. Spoken language processing – a guide to theory, algorithm, and system development. *Prentice Hall, Englewood Cliffs, New Jersey*, 2001.

[45] H-F. Silverman et D-P. Morgan. The application of dynamic programming to connected speech recognition. *IEEE ASSP magazine*, 7 :6–25, 1990.

[46] Bouallègue Mohamed. L'analyse factorielle pour la modélisation acoustique des systèmes de reconnaissance de la parole. *Thèse de doctorat de l'Université d'Avignon et des Pays de Vaucluse*, page 121, Décembre, 2013.

[47] S. F. Chen et J. Goodman. An empirical study of smoo- thing techniques for language modeling. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 310–318, 1996.

[48] Alexandre A. et Jean-Luc G. Construction automatique du vocabulaire d'un système de transcription. In *Proceedings of JEP 2004, Journées d'Etude de la Parole, Fès Maroc*, 2004.

[49] Frédéric Béchet. Un système complet de phonétisation de textes. *Traitement Automatique des Langues - TAL*, 42(1) :47–67, 2001.

[50] D. Vaufreydaz. Modélisation statistique du langage à partir d'internet pour la reconnaissance automatique de la parole continue. *Thèse de doctorat de l'Université J. Fourier - Grenoble I, France*, page 226, Janvier, 2002.

[51] K-P. Scannell. Automatic thesaurus generation for minority languages : an irish example. In *Atelier TALN'03, Batz-sur-Mer, France*, volume 2, pages 203–212, 2003.

[52] R. Jones et D. Mladenic. R. Ghani. Building minority language corpora by learning to generate web search queries. *Knowledge and Information Systems*, 7(1) :56–83, Janvier, 2005.

[53] A. W. Black et T. Schultz. Rapid language adaptation tools and technologies for multilingual speech processing. In *Automatic Speech Recognition & Understanding, IEEE Workshop,*, page 51, 2009.

[54] I. Daubechies. Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math*, 41(1) :909–996, 1988.

[55] D.L. Donoho. De-noising by soft thresholding. *IEEE transactions on Information Theory*, pages 613–627, 1995.

[56] D. Blachon, E. Gauthier, L. Besacier, G.-N. Kouarata, M. Adda-Decker, and Rialland. Parallel speech collection for under-resourced language studies using the lig-aikuma mobile device app. In *Proceedings of Spoken Language Technologies for Under-Resourced Languages, Yogyakarta, Indonesia*, 2016.

[57] F. A.A. Laleye, E. C. Ezin, and C. Motamed. Automatic fongbe phoneme recognition from spoken speech signal. In *Proceedings of the 13th International Conference on Informatics in Control, Automation and Robotics*, volume 1, pages 102–109, 01/2016.

[58] A. B. AKOHA. Ecrire et lire la langue fon. *CAAREC Editions Collection Education*, page 118, 2010.

[59] Christine Meunier. Phonétique acoustique. In Auzou P., editor, *Les dysarthries*, pages 164–173. Solal, 2007.

[60] F. A.A. Laleye, E. C. Ezin, and C. Motamed. Adaptive decision-level fusion for fongbe phoneme classification using fuzzy logic and deep belief networks. In *Proceedings of the 12th International Conference on Informatics in Control, Automation and Robotics, Colmar, Alsace, France*, volume 1, pages 15–24, 07/2015.

[61] Jensen Wong Jing Lung, Md. Sah Hj. Salam, Mohd Shafry Mohd Rahim Amjad Rehman, and Tanzila Saba. *Fuzzy Phoneme Classification Using Multi-speaker Vocal Tract Length Normalization.* IETE Technical Review, London, 2nd edition, 2014.

[62] Matthew Ager, Zoran Cvetkovic, and Peter Sollich. *Phoneme Classification in High-Dimensional Linear Feature Domains.* Computing Research Repository, 2013.

[63] M. Genussov, Y. Lavner, and I Cohen. Classification of unvoiced fricative phonemes using geometric methods. In *Proceedings of the 12th International Workshop on Acoustic Echo and Noise Control*, Tel-Aviv, Israel,, 2010.

[64] Dimitri Palaz, Ronan Collobert, and Mathew Magimai.-Doss. End-to-end phoneme sequence recognition using convolutional neural networks. *Idiap-RR*, 2013.

[65] E.C. Ezin. Neural networks and neural fuzzy systems for speech applications. *Thèse de doctorat de l'Université d'Abomey-Calavi (UAC) - Institut de Mathématiques et de Sciences Physiques (IMSP), Bénin*, page 121, March, 2001.

[66] A. Esposito, E.C. Ezin, and M. Ceccarelli. Preprocessing and neural classification of english stop consonants [b, d, g, p, t, k]. In *The 4th International Conference on Spoken Language Processing*, pages 1249–1252, Philadelphia, 1996.

[67] A. Esposito, E.C. Ezin, and M. Ceccarelli. Phoneme classification using a rasta-plp preprocessing algorithm and a time delay neural network : Performance studies. In *Proceedings of the 10th Italian Workshop on Neural Nets*, pages 207–217, Salerno,, 1998.

[68] Jibran Yousafzai, Zoran Cvetkovic, and Peter Sollich. Tuning support vector machines for robust phoneme classification with acoustic waveforms. In *10th Annual conference of the International Speech communication association*, pages 2359 – 2362, England, 2009. ISCA-INST SPEECH COMMUNICATION ASSOC.

[69] Viet-Bac Le and Besacier L. Automatic speech recognition for under-resourced languages : Application to vietnamese language. In *IEEE Transactions on Audio, Speech, and Language Processing*, pages 1471–1482. IEEE, 2009.

[70] Thomas Niesler and Philippa H. Louw. Comparative phonetic analysis and phoneme recognition for afrikaans, english and xhosa using the african speech technology telephone speech database. In *South African Computer Journal*, pages 3–12, 2004.

[71] Tim Schlippe, Edy Guevara Komgang Djomgang, Ngoc Thang Vu, Sebastian Ochs, and Tanja Schultz. Hausa large vocabulary continuous speech recognition. In *The third International Workshop on Spoken Languages Technologies for Under-resourced Languages*, Cape-Town, 2012.

[72] Emily M Mugler, James L Patton, Robert D Flint, Zachary A Wright, Stephan U Schuele, Joshua Rosenow, Jerry J Shih, Dean J Krusienski, and Marc W Slutzky. Direct classification of all american english phonemes using signals from functional speech motor cortex. *J. Neural Eng*, 2014.

[73] G. Rogova. Combining the results of several neural networks classifiers. *Neural Networks*, pages 777–781, 1994.

[74] S-B. Cho and J.H. Kim. Combining multiple neural networks by fuzzy integral and robust classification. *IEEE Transactions on Systems, Man, and Cybernetics*, pages 380–384, 1995.

[75] J. Kittler, M. Hatef, R.P.W Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Patterns Analysis and Machine Intelligence*, pages 226–239, 1998.

[76] R.A. Jacobs. Methods for combining experts's probability assessments. *Neural Computation*, pages 867–888, 1995.

[77] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, and G.E. Hinton. Adaptive mixture of local experts. *Neural Computation*, pages 79–87, 1991.

[78] A. Metallinou, S. Lee, and S. Narayanan. Decision level combination of multiple modalities for recognition and analysis of emotional expression. In *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 2462–24665, 2010.

[79] T. W. Lewis and D. M.W. Powers. Improved speech recognition using adaptive audio-visual fusion via a stochastic secondary classifier. *International Symposium on Intelligent Multimedia, Video and Speech Processing*, 1 :551–554, 2001.

[80] G. Iyengar, H.J. Nock, and C. Neti. Audio-visual synchrony for detection of monologue in video archives. In *IEEE International Conference on Multimedia and Expo*, volume 1, pages 329–332, 2003.

[81] C. Neti, B. Maison, A. Senior, G. Iyengar, P. Decuetos, S. Basu, and A. Verma. Joint processing of audio and visual information for multimedia indexing and human-computer interaction. In *Sixth International Conference RIAO. Paris, France*, pages 294–301, 2000.

[82] A. Corradini, M. Mehta, N. Bernsen, J. Martin, and S. Abrilian. Multimodal input fusion in human–computer interaction. In *NATO-ASI Conference on Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management*, 2003.

[83] N. Pfleger. Context based multimodal fusion. In *ACM International Conference on Multimodal Interfaces*, pages 265–272, 2004.

[84] V. Pitsikalis, A. Katsamanis, G.Papandreou, and P. Maragos. Adaptive multimodal fusion by uncertainty compensation. In *Ninth International Conference on Spoken Language Processing. Pittsburgh*, volume 7, pages 423–435, 2006.

[85] G.F. Meyer, J.B. Mulligan, and S.M. Wuerger. Continuous audio-visual digit recognition using n-best decision fusion. *Information Fusion*, 5 :91–101, 2004.

[86] H. Xu and T.S. Chua. Fusion of av features and external information sources for event detection in team sports video. *ACM Trans. Multimed. Comput. Commun. Appl.*, 2 :44–67, 2006.

[87] S. Foucher, F. Laliberte, G. Boulianne, and L. Gagnon. A dempster-shafer based fusion approach for audio-visual speech recognition with application to large vocabulary french speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, 2006.

[88] P. O'Connor, D. Neil, Liu SC, T. Delbruck, and M. Pfeiffer. Real-time classification and sensor fusion with a spiking deep belief network. *Front. Neurosci*, 2013.

[89] M. Malcangi, K. Ouazzane, and K. Patel. Audio-visual fuzzy fusion for robust speech recognition. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pages 1 – 8, Dallas, 2013. IEEE.

[90] H. Zhang. Exploring conditions for the optimality of naïve bayes. *IJPRAI*, 19 :183–198, 2005.

[91] Richard J. Ye, J. Povinelli, and Michael T. Johnson. Phoneme classification using naive bayes classifier in reconstructed phase space. In *Proceedings of IEEE 10th Digital Signal Processing Workshop*, pages 37–40, 2002.

[92] L. Tóth, A. Kocsor, and J. Csirik. On naive bayes in speech recognition. *International Journal of Applied Mathematics and Computer Science*, 15 :287–294, 2005.

[93] T. Kohonen. An introduction to neural computing. *Neural Networks*, 1 :3–16, 1988.

[94] P. Borne, M. Benrejeb, and J. Haggege. Les réseaux de neurones, présentation et applications. *TECHNIP Editions*, page 90, 2007.

[95] J. Mantysalo, K. Torkkolay, and T. Kohonen. Lvq-based speech recognition with high-dimensional context vectors. In *Proceedings of International Conference on Spoken Language Processing*, pages 539–542, 1992.

[96] S. Katagiri and E. McDermott. Speaker-independent large vocabulary word recognition using an lvq/hmm hybrid algorithm. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 553–556, 1991.

[97] N. Dave. Feature extraction methods lpc, plp and mfcc in speech recognition. *IJARET*, 1, 2013.

[98] H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *IJARET*, 87 :1738–1752, 1990.

[99] F. A.A. Laleye, E. C. Ezin, and C. Motamed. Weighted combination of naive bayes and lvq classifier for fongbe phoneme classification. In *Proceedings of IEEE 10th International Conference on Signal Image Technology & Internet Based Systems*, pages 7 – 13. IEEE, 2014.

[100] Y. Bengio, Lamblin P., Popovici D., and Larochelle H. Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems*, 2006.

[101] G. Hinton, Osindero S., and Y.W Teh. A fast learning algorithm for deep belief nets. *Neural Comput*, 18 :1527–1554, 2006.

[102] S. Wang and X. Yao. Diversity analysis on imbalanced data sets by using ensemble models. *IEEE Symp.Comput. Intell. Data Mining*, pages 324–331, 2009.

[103] A. Origlia, F. Cutugno, and V. Galatà. Continuous emotion recognition with phonetic syllables. *Speech Communication*, 57 :155 – 169, 2014.

[104] P. Mermelstein. Automatic segmentation of speech into syllabic units. *Journal of the Acoustical Society of America*, 58 :880–883, 1957.

[105] X. Zhao and D. O'Shqughnessy. A new hybrid approach for automatic speech signal segmentation using silence signal detection, energy convex hull, and spectral variation. In *Canadian Conference on Electrical and Computer Engineering*, pages 145–148. IEEE, 2008.

[106] H.R. Pfitzinger, S. Burger, and S. Heid. Syllable detection in read and spontaneous speech. In *Proceedings of the Fourth International Conference on Spoken Language (ICSLP)*, volume 2, pages 1261–1264. IEEE, 1996.

[107] N. Jittiwarangkul, S. Jitapunkul, S. Luksaneeyanavin, V. Ahkuputra, and C. Wutiwiwatchai. Thai syllable segmentation for connected speech based on energy. In *The Asia-Pacific Conference on Circuits and Systems*, pages 169–172. IEEE, 1998.

[108] L. Wu, M. Shire, S. Greenberg, and N. Morgan. Integrating syllable boundary information into speech recognition. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 987–990. IEEE, 1997.

[109] M. Petrillo and F. Cutugno. A syllable segmentation algorithm for english and italian. In *Proceedings of 8th European Conference on Speech Communication and Technology, EUROSPEECH, Geneva*, pages 2913–2916, 2003.

[110] G. Sheikhi and A. Farshad. Segmentation of speech into syllable units using fuzzy smoothed short term energy contour. In *Proceedings of International Conference on Acoustics, Speech and Signal Processing*, pages 195–198. IEEE, 2011.

[111] F. Pan and N. Ding. Speech denoising and syllable segmentation based on fractal dimension. In *International Conference on Measuring Technology and Mechatronics Automation*, pages 433–436. IEEE, 2010.

[112] N. Obin, F. Lamare, and A. Roebel. Syll-o-matic : An adaptive time-frequency representation for the automatic segmentation of speech into syllables. In *International Conference on Acoustics, Speech and Signal Processing*, pages 6699–6703. IEEE, 2013.

[113] R. Villing, J. Timoney, T. Ward, and J.Costello. Automatic blind syllable segmentation for continuous speech. In *In Proceedings of the Irish Signals and Systems Conference*, pages 41–46. Belfast, UK, 2004.

[114] C-H. Chou, P-H. Liu, and B. Cai. On the studies of syllable segmentation and improving mfccs for automatic birdsong recognition. In *Asia-Pacific Services Computing Conference*, pages 745–750. IEEE, 2008.

[115] L. Shastri, S. Chang, and S. Greenberg. Syllable detection and segmentation using temporal flow neural networks. In *Proceedings of the Fourteenth International Congress of Phonetic Sciences*, pages 1721–1724, 1999.

[116] H. Ching-Tang, S. Mu-Chun, L. Eugene, and H. Chin. A segmentation method for continuous speech utilizing hybrid neuro-fuzzy network. *Journal of Information Science and Engineering*, 15(4) :615–628, 1999.

[117] T. Demeechai and K. M akel ainen. Recognition of syllables in a tone language. *Speech Communication, Elsevier*, 33(3) :241 – 254, 2001.

[118] V. Khanagha, K. Daoudi, O. Pont, and H. Yahia. Improving text-independent phonetic segmentation based on the microcanonical multiscale formalism. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4484 – 4487. IEEE, 2011.

[119] T.A. Hall. Encyclopedia of language and linguistics. *Elsevier*, 12, 2006.

[120] O. Pont, A. Turiel, and H. Yahia. An optimized algorithm for the evaluation of local singularity exponents in digital signals. In *Combinatorial Image Analysis*, pages 346–357. Springer Berlin Heidelberg, 2011.

[121] V. Khanagha. Nouvelles méthodes multi-échelles pour l'analyse non-linéaire de la parole. *Thèse de doctorat de l'Université de Bordeaux 1, France*, page 158, Janvier, 2013.

[122] A. Turiel and N. Parga. The multi-fractal structure of contrast changes in natural images : from sharp edges to textures. *Neural Computation*, 12 :763 – 793, 2000.

[123] A. Turiel, C.J. Pérez-Vicente, and J. Grazzini. Numerical methods for the estimation of multifractal singularity spectra on sampled data : A comparative study. *Journal of Computational Physics*, 216(1) :362 – 390, 2006.

[124] C.E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27 :379 – 423, 1948.

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

[126] C. Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 52(1-2) :479 – 487, 1998.

[127] R.G Baraniuk, P. Flandrin, A.J.E.M Janssen, and O.J.J Michel. Measuring time-frequency information content using the renyi entropies. In *IEEE Transactions on Information Theory*, volume 47, pages 1391–1409. IEEE, 2001.

[128] B. Boashash. Time frequency signal analysis and processing : A comprehensive reference. In *Elsevier, Oxford*. Elsevier, 2003.

[129] F. A.A. Laleye, E. C. Ezin, and C. Motamed. An algorithm based on fuzzy logic for text-independent fongbe speech segmentation. In *Proceedings of IEEE 11th International Conference on Signal Image Technology & Internet Based Systems*, volume 1, pages 1–6, 11/2014.

[130] O.J. Rasanen, U.K. Laine, and T. Altosaar. An improved speech segmentation quality measure : the r-value. In *Proceedings of INTERSPEECH*, pages 1851 – 1854, 2009.

[131] A. Howitt. Vowel landmark detection. *Journal of the Acoustical Society of America*, 112(5) :2279, 2002.

[132] H. Y. Lo and H. M. Wang. Phonetic boundary refinement using support vector machine. In *International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, Honolulu, HI*, pages 933–936. IEEE, 2007.

[133] I. Mporas, T. Ganchev, and N. Fakotakis. Speech segmentation using regression fusion of boundary predictions. *Computer Speech & Language*, 24(2) :273–288, 2010.

[134] E. Gauthier, L. Besacier, S. Voisin, M. Melese, and U. P. Elingui. Collecting resources in sub-saharan african languages for automatic speech recognition : a case study of wolof. In *10th edition of the Language Resources and Evaluation Conference, Slovenia*, pages 23–28, 2016.

[135] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, IEEE Signal Processing Society*, 2011.

[136] Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 28.

[137] Mark JF Gales. Semi-tied covariance matrices for hidden markov models. In *IEEE Transactions on Speech and Audio Processing*, volume 7.

[138] Ramesh A Gopinath. Maximum likelihood modeling with gaussian distributions for classification. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, volume 2.

[139] Daniel Povey, G. Zweig, and A. Acero. The exponential transform as a generic substitute for vtln. In *IEEE ASRU*.

[140] Sirko Molau, Florian Hilger, and Hermann Ney. Feature space normalization in adverse acoustic conditions, acoustics, speech, and signal processing. In *IEEE ICASSP'03*, volume 1.

[141] Xuedong Huang, Alejandro Acero, and Hsiao-Wuen Hon. Spoken language processing. In *Prentice Hall PTR New Jersey*.

[142] Luis Javier Rodríguez and Inés Torres. Luis javier rodríguez and inés torres. *Comparative study of the Baum-Welch and Viterbi training algorithms applied to read and spontaneous speech recognition*, pages 847–857, 2003.

[143] Thorsten Brants, Ashok C Popat, Peng Xu, Franz J Och, and Jeffrey Dean. Large language models in machine translation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Citeseer*.

[144] L. Barrault. Diagnostic pour la combinaison de systèmes de reconnaissance automatique de la parole. *Thèse de doctorat de l'Université d'Avignon et des Pays de Vaucluse, France*, page 184, Juillet, 2008.

[145] Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. *OpenFst : A General and Efficient Weighted Finite-State Transducer Library*, pages 11–23. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[146] Mehryar Mohri, Fernando Pereira, and Michael Riley. Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1) :69–88, 2002.

[147] A Stolcke. Srilm – an extensible language modeling toolkit. In *Proceedings of International conference on on Spoken Language Processing, Denver*.

Résumé

L'une des difficultés d'une langue peu dotée est l'inexistence des services liés aux technologies du traitement de l'écrit et de l'oral. Dans cette thèse, nous avons affronté la problématique de l'étude acoustique de la parole isolée et de la parole continue en Fongbe dans le cadre de la reconnaissance automatique de la parole.

La complexité tonale de l'oral et la récente convention de l'écriture du Fongbe nous ont conduit à étudier le Fongbe sur toute la chaîne de la reconnaissance automatique de la parole. En plus des ressources linguistiques collectées (vocabulaires, grands corpus de texte, grands corpus de parole, dictionnaires de prononciation) pour permettre la construction des algorithmes, nous avons proposé une recette complète d'algorithmes (incluant des algorithmes de classification et de reconnaissance de phonèmes isolés et de segmentation de la parole continue en syllabe), basés sur une étude acoustique des différents sons, pour le traitement automatique du Fongbe. Dans ce manuscrit, nous avons aussi présenté une méthodologie de développement de modèles acoustiques et de modèles du langage pour faciliter la reconnaissance automatique de la parole en Fongbe. Dans cette étude, il a été proposé et évalué une modélisation acoustique à base de graphèmes (vu que le Fongbe ne dispose pas encore de dictionnaire phonétique) et aussi l'impact de la prononciation tonale sur la performance d'un système RAP en Fongbe.

Enfin, les ressources écrites et orales collectées pour le Fongbe ainsi que les résultats expérimentaux obtenus pour chaque aspect de la chaîne de RAP en Fongbe valident le potentiel des méthodes et algorithmes que nous avons proposés.

Mots-clés: Fongbe, reconnaissance automatique de la parole, segmentation automatique de la parole, entropie de Rényi, modélisation acoustique graphémique, modélisation du langage, fusion de décisions, multi-classification, DBN, logique floue.

Abstract

One of the difficulties of an unresourced language is the lack of technology services in the speech and text processing. In this thesis, we faced the problematic of an acoustical study of the isolated and continuous speech in Fongbe as part of the speech recognition.

Tonal complexity of the oral and the recent agreement of writing the Fongbe led us to study the Fongbe throughout the chain of an automatic speech recognition. In

addition to the collected linguistic resources (vocabularies, large text and speech corpus, pronunciation dictionaries) for building the algorithms, we proposed a complete recipe of algorithms (including algorithms of classification and recognition of isolated phonemes and segmentation of continuous speech into syllable), based on an acoustic study of the different sounds, for Fongbe automatic processing. In this manuscript, we also presented a methodology for developing acoustic models and language models to facilitate speech recognition in Fongbe. In this study, it was proposed and evaluated an acoustic modeling based on grapheme (since the Fongbe don't have phonetic dictionary) and also the impact of tonal pronunciation on the performance of a Fongbe ASR system.

Finally, the written and oral resources collected for Fongbe and experimental results obtained for each aspect of an ASR chain in Fongbe validate the potential of the methods and algorithms that we proposed.

Keywords: Fongbe, automatic speech recognition, automatic speech segmentation, Rényi Entropy, grapheme-based acoustical modeling, language modeling, fusion of decisions, multiclass classification, DBN, fuzzy logic.